Generalizing the Results from Social Experiments: Theory and Evidence [*]

Michael Gechter[†]

October 13, 2014
Preliminary and incomplete

Abstract

To what extent are causal effects estimated in one region or time period informative about another region or time? This paper provides a formal answer by developing methods to quantify the assumptions on heterogeneity in individual-specific causal effects that are required for causal effects estimated in one population to allow researchers to reject hypotheses about causal effects in a population of interest. For example, the method delivers the assumptions required to reject a zero causal effect or an average cost per unit of improvement deemed excessive by policymakers. Hypotheses that can be rejected under a wide range of assumptions constitute more robust inferences about the causal effects in the population of interest. I empirically investigate what assumptions are required for experimental results on the return to cash transfers to male microentrepreneurs in one Mexican city in 2006 to speak to the returns among male microentrepreneurs in urban Mexico in 2012. The experimental results yield narrow bounds on the average causal effect for male microentrepreneurs in urban Mexico in 2012 under a wide variety of assumptions on heterogeneity. 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 standard methods are unsuccessful.

[*] I am grateful to my advisors, 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, Isabella Dobrescu, Iván Fernández-Val, Claudio Ferraz, Patrik Guggenberger, Rema Hanna, Stefan Hoderlein, Asim Khwaja, Horacio Larreguy, David Lam, Leigh Linden, Shanthi Nataraj, Rohini Pande, Justin Sandefur, Johannes Schmieder, Duncan Thomas, Nate Young and seminar participants at Boston University and the The Western Economic Association International Graduate Student Workshop for useful conversations and feedback. I especially thank David McKenzie for discussions and sharing the data from McKenzie and Woodruff (2008).

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

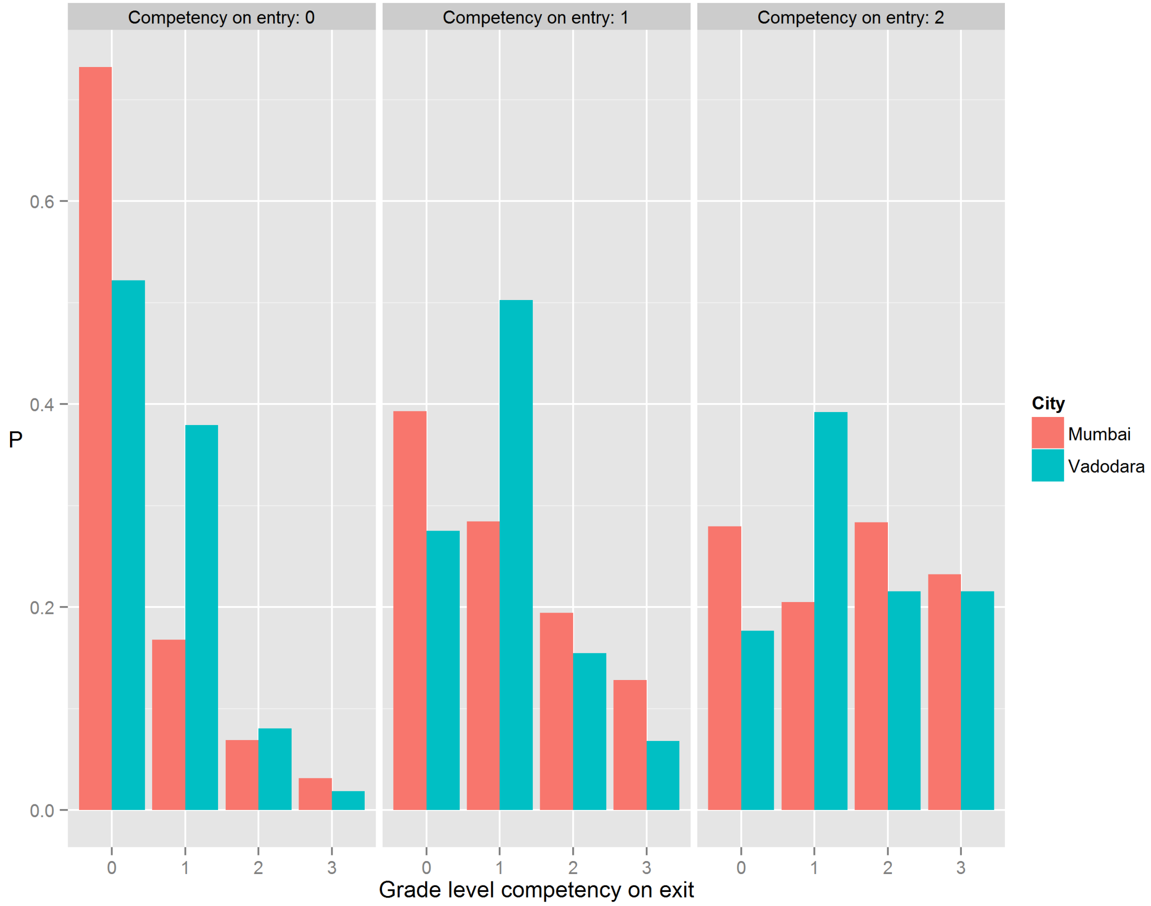
1 Introduction

The “credibility revolution in empirical economics” (Angrist and Pischke (2010)) has focused on finding populations where the economic variable or “treatment” for which we would like to estimate a causal effect can be randomly assigned or has been assigned in a manner as good as random. While this focus on research design has led to an increase in the credibility of estimated causal effects in studies of populations where treatment assignment is random or quasi-random (“internal validity”), a number of recent papers have questioned whether causal effects estimated in this way apply to populations of interest outside of the original study, the question of “external validity” (e.g. Deaton (2010)). The question of external validity is particularly pressing in the field of development economics. With lower costs, development economists have been able to implement a large number of randomized field experiments in recent years (Duflo et al. (2008)). However, development economists offer policy advice across an arguably wider variety of contexts than any other field and, as a result, there has been an active debate in the field over the role that randomized experiments should play in influencing policy decisions in populations other than the ones where they were originally carried out.

Consider the following example from Banerjee et al. (2007). A randomized evaluation of a remedial education program is carried out in Mumbai, India. The program is found to raise the average grade level competency of third-grade students by a significant margin. Should education policy makers in another Indian city, Vadodara, implement this same policy? To this point, economists and policy makers have lacked effective tools to assess whether the population on which the experiment was conducted (Mumbai) is sufficiently similar to the population of interest (Vadodara) for the experimental results to be useful (Banerjee and Duflo (2009); Allcott (2014)). This had led some authors to protest against what they see as the implicit extrapolation of unadjusted results from a small number of experiments to a wide variety of disimilar contexts in policy recommendations (Pritchett and Sandefur (2013)). This paper aims to fill the methodological gap by investigating the assumptions on heterogeneity in individual-specific causal effects required for the results from an experiment to reject hypotheses about the average causal effect in the population of interest.

The standard response to heterogeneity in causal effects would be to estimate average causal effects conditional on covariates in the Mumbai sample and to weight these conditional average casual effects according to the distribution of covariates in a sample from Vadodara (see Allcott (2014)). The approach based on weighting conditional average causal effects has been unsuccessful at recovering causal effects when using the results from one or more randomized experiments to predict the results of other experiments examining the same policy

Figure 1: controls - grade level competency on exiting 3rd grade conditional on grade level competency on entering 3rd grade



(Allcott (2014)). The reweighting approach also ignores information from the distributions of outcomes of individuals with the same treatment status (treated or control) and covariates, which often differ between the experimental sample and the population of interest. To fix ideas, I will assume in what follows that everyone in the population of interest is control, as occurs when the experiment evaluates a pilot like the remedial education program[1]. Figure 1 provides an example of the information ignored by standard extrapolation methods: conditional on grade level competency when entering third grade, the distribution of grade level competency on exiting third grade (the outcome of interest) differs substantially between controls (no remedial education) in Mumbai and Vadodara.

In this paper, I make use of the information in Figure 1, the distributions of control outcomes for individuals with the same covariates in the population of interest (Vadodara)

[1] 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 et al. (2014) for an example of such an experiment).

and the experimental sample (Mumbai), to derive sharp bounds on the average causal effect in the population of interest. We can only bound the average causal effect in the population of interest because we do not know how causal effects were distributed among individuals with the same covariates in the experiment. If we assume that large causal effects accrue to individuals with control outcomes that are more common in the population of interest than in the experiment, then the treatment will look like it would have a big effect in the population of interest. If, in contrast, small causal effects accrue to individuals with control outcomes that are more common in the population of interest than in the experimental sample, then the treatment will look like it would have a small effect in the population of interest[2]. The width of the bounds will depend primarily on the extent of difference between the distributions of control outcomes in the experimental sample and the population of interest[3] and secondarily on the amount of mass in the tails of the outcome distributions and the heterogeneity in the quantile treatment effects.

Not every distribution of causal effects is equally plausible. In particular, distributions characterized by causal effects of very large magnitude but different signs may be implausible. I therefore index possible distributions of causal effects by a parameter measuring the extent of heterogeneity in causal effects among individuals with the same covariates: the correlation between an individual's rank in her observed outcome distribution and her counterfactual outcome distribution. I derive computationally tractable bounds on the average causal effect in the population of interest subject to the restriction that the causal effect heterogeneity for individuals with same covariates must be less than a specified value of the parameter. These bounds allow us to investigate the robustness of hypotheses about the average causal effect in the population of interest to the degree of causal effect heterogeneity allowed for individuals with the same covariates. For example, we can ask how much heterogeneity we can introduce and still reject an average causal effect of zero or a critical lower bound on

[2] It is worth mentioning that, while outside the scope of this paper, structural models do not provide an immediate solution to the issues raised here. Direct extrapolation to alternative populations using a structural model requires that the distributions of any parameters that affect policy response and cannot be identified using data on only control individuals be the same across populations. For example, the marginal utility of a schooling subsidy can be identified only using treatment group data in Attanasio et al. (2012). If the marginal utility of the subsidy varies across populations, we would expect an extrapolation to predict the effect of introducing the same school subsidy in another context using a structural model but keeping the same marginal utility to fail (as such an extrapolation does in Attanasio et al. (2003)). The implausibility of transporting all such parameters across contexts motivates attempting a sensitivity analysis similar to the one in this paper, this time with respect to differences in the distributions of structural parameters identified using the treatment group data. The width of bounds obtained in this way and their performance in predicting average treatment effects across contexts are a matter for future research.

[3] In this sense, the approach can be thought of as determining the extent to which the experimental sample is representative (as done informally in McKenzie and Woodruff (2008) and Heckman et al. (2010)) of the population of interest and translating that representativeness into bounds on the average causal effect.

cost-per-impact (for example average expenditure per grade level comptency increase) in the population of interest. When heterogeneity is the minimum possible, the bounds on the average causal effect reduce to a point and is an estimator from Athey and Imbens (2006) (henceforth AI).

I empirically investigate the extent of causal effect heterogeneity required to undermine the conclusion of a non-zero average causal effect in extrapolating the results from a small experiment conducted in Leon, Mexico and documented in McKenzie and Woodruff (2008) (henceforth MW). The experiment is part of a series (including experiments in Sri Lanka described in de Mel et al. (2008) and Ghana described in Fafchamps et al. (2014)) examining the returns to cash transfers to microentrepreneurs. The Leon experiment, like others in the series, finds large returns to the transfers in terms of microenterprise profits, in this case an increase in monthly profits of about 33% of the transfer. I investigate to what extent this notable finding generalizes across space and time within Mexico. To represent the population of interest, I use data from a subsample of the 2012 nationally-representative Mexican microenterprise survey with the same covariates as the Leon experiment, facilitated by the unique fact that the questionnaire in the Leon trial was intended to be compatible with the contemporaneous microenterprise survey. Unsurprisingly, given the small size of the Leon trial (just over 200 entrepreneurs), we cannot reject a zero average causal effect in the full Mexican sample at any level of heterogeneity in causal effects. However, the conditional distributions of profits in the 2006 Leon control group and the 2012 microenterprise survey are sufficiently similar that the bounds on the average causal effect remain quite narrow, even when allowing a substantial amount of heterogeneity in causal effects. Furthermore, the results separate uncertainty about the average causal effect among male entrepreneurs in urban Mexico in 2012 into uncertainty due to differences in the control outcome distributions and due to the sample size of the experiment where existing methods account only for the sample size of the experiment.

To check the results of the methods advocated here against measured causal effects, I use data from randomized evaluations of a remedial education program implemented in two Indian cities and described in Banerjee et al. (2007) (henceforth BCDL). Here, I can compare the performance of standard methods based on reweighting with methods based on the distributions of control outcomes for individuals with the same covariates. To do this, I treat one city's data as the experimental sample, e, and the other as the population of interest or alternative population, a. I hold out city a's treated group data and compare it against predicted average causal effects for city a obtained using city e's treated and control groups and city a's control group. As in previous work, I find that the predicted average causal effects of methods based on reweighting are rejected. In contrast, predictions using information in the distributions of control outcomes in cities e and a are able to recover the average causal effect in city a. The level of heterogeneity required for the predictive bounds to contain the average causal effect in city a are within the range required to reject a zero average causal effect.

The rest of the paper is organized as follows. The following subsection reviews the foundational theoretical literature. Section 2 sets up the problem and notation. Section 3 provides a more detailed review of approaches based on reweighting of average causal effects or treated outcomes conditional on covariates. In section 4, I then show that an estimator from AI allows us to use the distributions of control outcomes for individuals with the same covariates to point-identify the average causal effect in the population of interest at the cost of assuming minimal heterogeneity in causal effects for individuals with the same covariates. I argue that minimal heterogeneity is often too restrictive an assumption and derive bounds on the average causal effect for the population of interest when we allow a specified level of heterogeneity in causal effects for individuals with the same covariates. These bounds allow us to investigate the maximal level of heterogeneity under which a hypothesis about the average causal effect in the population of interest can be rejected. Section 5 presents the empirical results for generalizing from the 2006 experiment providing cash transfers to microenterpreneurs in Leon, Mexico to urban locations in Mexico in 2012. Section 6 investigates using the results from each of the two remedial education experiments to try to predict the results in the other experiment. Section 7 concludes by offering advice to researchers carrying out randomized experiments and concerned about their generalizability to other contexts.

1.1 Foundational literature

In focusing on the distributions of control outcomes, I extend the main strand of theoretical literature on extrapolation of results from randomized experiments to new environments which begins with Hotz et al. (2005) (henceforth HIM). HIM and the papers following (Cole and Stuart (2010), Stuart et al. (2011), Flores and Mitnik (2013)) assume that the joint distribution of the potential outcomes (treated and control) is independent of the population conditional on covariates. This approach generates the testable implication that the distribution of control outcomes should be independent of the population conditional on covariates. If a test for equality of the conditional control group outcome distributions rejects, we conclude experiment provides no useful information about causal effects in the population of interest. If the test fails to reject, we weight the conditional mean treated outcomes from the experiment by the distribution of covariates in the population of interest and subtract the mean control outcome in the population of interest to obtain the average causal effect. In practice, testing is often abandoned due to lack of power in small experiments (Flores and Mitnik (2013)). If samples were large, in contrast, we might reject even when equality of distributions holds approximately and believe that we have learned nothing from the experiment. My approach, instead, translates differences in the conditional distributions of control outcomes into a set of assumptions required to reject hypotheses about the average causal effect in the population of interest.

In moving from a testing framework to an approach based on evaluating the breakdown point in terms of assumptions required to reject a hypothesis regarding the causal effect, my paper is related Altonji et al. (2005) and Altonji et al. (2013) who move from testing whether observed covariates related to an outcome are also related to a candidate instrument to a framework which bounds the treatment effect on the basis of the magnitude of the relationship between the covariates and the instrument. Oster (2014) takes a similar approach in translating changes in coefficients of interest when covariates are included in linear regressions to bounds on the true coefficient. While these papers operate within the parametric context of a linear regression model, my approach is non-parametric and is similar to Kline and Santos (2013) who explore the sensitivity of conclusions about conditional distributions of outcomes to deviations from the assumption that missing outcomes are missing at random. Kline and Santos (2013) measure deviations non-parametrically by the Kolmogorov-Smirnov distance between the conditional outcome distributions for individuals with missing and non-missing outcomes.

I also draw on the literature on distributions of individual-specific causal effects consistent with control and treated group outcome distributions, which begins in economics with Heckman et al. (1997) and continues with Djebbari and Smith (2008), Fan and Park (2010) and Kim (2014). Like Kim (2014), I approach the distribution of individual-specific treatment effects as an optimal transportation problem (c.f. Villani (2009)), but with a different objective function and constraints.

2 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 (for example if her class receives a remedial education teacher) and $Y_0 \in \mathcal{Y}_0 \subseteq \mathcal{Y}$ is her outcome if she does not (no remedial education teacher is assigned). Only one of these two outcomes is ever observed, the other is hypothetical. If a student's class receives a remedial education teacher, we observe Y_1 and her outcome in the event that her class had not received a remedial education teacher (Y_0) 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 $\Delta \subseteq \mathbb{R}$, the effect for her of having a remedial education teacher assigned to her class:

$$\Delta = Y_1 - Y_0$$

We have data on 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 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, allowing us to identify the average individual-specific treatment effect Δ in population e[4]:

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

We are, however, interested in the average treatment effect in the population of interest, $E^a[\Delta]$, of which we can identify only one component:

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

[4] Putting perfect compliance with treatment assignment another way, the estimand of interest is the intention-to-treat (ITT) effect, often thought to be the object of policy interest since compliance can rarely be mandated in policy settings.

If the treatment effect were constant for all individuals and equal to $\overline{\Delta}$, $E^a[\Delta]$ would simply be equal to $E^e[\Delta]$. We rarely, however, believe that this is the case (and can often reject it empirically). To investigate heterogeneity in individual-specific treatment effects, I now introduce some additional notation. Suppose observe a vector of observable 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: 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. Treatment status and covariates combine to produce the outcome through a function common across populations, $g : \{0,1\} \times \mathcal{X} \times \mathcal{U} \to \mathbb{R}$. We can rewrite the potential outcomes as:

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

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:

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

noting 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:

$$\begin{aligned} ATE^a =& \int_{\mathcal{X}} \left[\int_{\mathcal{U}} g(1, x, u) - g(0, x, u) dF^a_{U|X}(u|x) \right] dF^a_X(x) && (1) \\ =& \int_{\mathcal{X}} \left[\int_{\mathbb{R}^2} y_1 - y_0 dF^a_{Y_0,Y_1|X}(y_0, y_1|x) \right] dF^a_X(x) && (2) \\ & \int_{\mathcal{X}} \left[\int_{\mathbb{R}} \delta dF^a_{\Delta|X}(\delta|x) \right] dF^a_X(x) && (3) \end{aligned}$$

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 show that the invariance 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. Different methods of extrapolation will make different assumptions about the relationships of the conditional distributions $F^a_{U|X}(u|x)$ and $F^e_{U|X}(u|x)$ and their equivalent counterparts.

2.1 Example: remedial education in India

To make the above discussion concrete, I describe a simple parametric model using the example of remedial education India. Suppose Mumbai is the experimental population, e, and Vadodara as the alternative population, a, where we would like to predict the average treatment effect. We will leave the observable covariates X as a vector, but break the vector U into the two components discussed above, latent skill S and parental input I. Similar form for $g(\cdot)$ 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.

Note that ATE_e 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. We will return to this parametric model to build intuition for key points throughout.

We now turn to a description of previous methods that have been used to identify ATE^a.

3 Previous methods

We now review the most common methods used for extrapolation of causal effects identified in an experimental population to a target population of interest.

3.1 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 (2014)):

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

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

where $\perp\!\!\!\perp$denotes statistical independence[5]. 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 control outcomes, $F_{Y_0}^a(y_0|x)$ and $F_{Y_0}^e(y_0|x)$ such as the difference between the distributions of grade level competency at the end of third grade without remedial education teachers conditional on grade level competency at the beginning of third grade between Mumbai and Vadodara shown in figure 1. 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 the outcome is binary and all individuals in the population of interest already have outcome 1. Predictions will also depend on the scaling of Y, for example, whether they are in levels or logs.

Even more substantively, differences in the conditional distributions of control outcomes are indicative of some unobservable differences between the experimental population and the population of interest. To see this, note that:

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

Then

$$F_{Y_0|X}^a(y_0|x) \overset{d}{\neq} F_{Y_0|X}^e(y_0|x) \implies F_{U|X}^a(u|x) \overset{d}{\neq} F_{U|X}^e(u|x)$$

where $\overset{d}{=}$ indicates equality in distribution. If the elements of U whose difference in condi-

[5] 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.

tional distribution produces the difference in the conditional distribution of control outcomes also influence the individual-specific treatment effect, 6 will not hold.

3.2 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 HIM and papers following them have suggested testing equality of the distributions or their moments. As mentioned briefly in section 1.1, 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, an issue also raised in Flores and Mitnik (2013). Second, if we do reject the null hypothesis, we must conclude that the experiment

Table 1: 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	-

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'.

3.2.1 Example: remedial education in India

Table 1 replicates the information in figure 1 in tabular form: the distributions of grade level competency in math on leaving third grade among the control groups in both of the 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 the distributions $F^e_{Y_0|X}(y_0|x)$ and $F^a_{Y_0|X}(y_0|x)$ for each x representing a grade level competency on entering third grade. The test rejects at the 5% level for all values of x.

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.

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

4.1 Identification

In investigating the role of the conditional control 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 control 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 control 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-control-outcome-conditional distributions $F_{Y_1|Y_0,X}(y_1|y_0,x)$ associated with the covariate-conditioned marginal control 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 $C_1(u,v|x) = \frac{\partial C(u,v|x)}{\partial u}$. 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 control and treated outcomes in the experimental population for individuals with the same covariates. A copula function combined with the marginal distributions of control ($F^e_{Y_0,X}(y_0|x)$) 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. We will impose the following assumption.

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

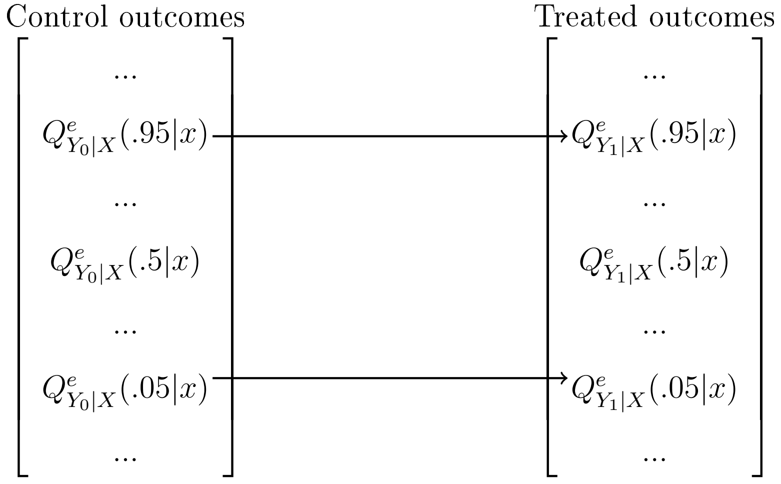
$$F^a_{Y_1|Y_0,X}(y_1|y_0,x) = C_1(F^e_{Y_0|X}(y_0|x), F^e_{Y_1|X}(y_1|x)|x)$$

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

Assumption 3 states that we must be able to express the control-outcome conditional distribution of the treated outcome as one of the conditional distributions consistent with the distributions of control and treated outcomes in the experiment, all for individuals with the same covariate values.

To make assumption 3 more concrete, we 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. Figure 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 the treatment would still be the highest-scoring student with the treatment. The crossing arrows in figure

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



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.

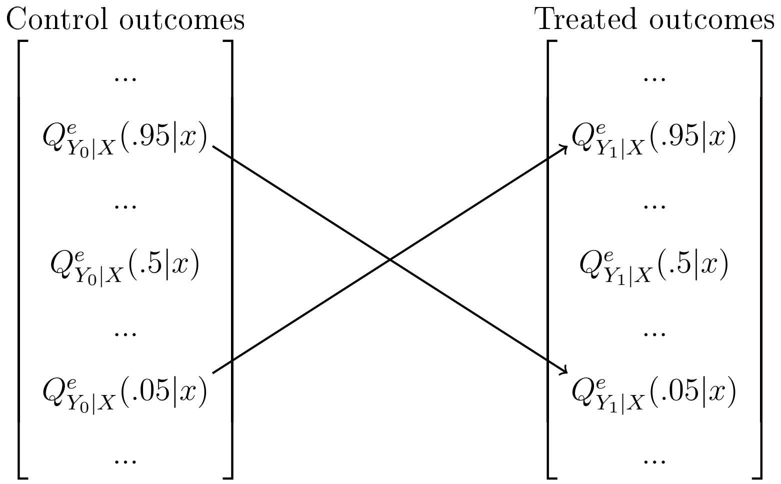
A joint distribution $F^e_{Y_0,Y_1|X}(y_0, y_1|x)$ consistent with the experimental marginal distributions of control and treated outcomes also determines the extent of heterogeneity in treatment effects for individuals with covariates x. When the treatment perfectly preserves individuals' ranks in the outcome distributions, treatment effect heterogeneity due to unobservables is minimized. 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 Δ have the largest possible magnitude.

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 popuilation and the population of interest, the conditional treated outcomes have the same distribution as well. Formally:

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

A sufficient 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:

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



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

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 et al. (2011))? In terms of the underlying unobservables, a sufficient condition for 9, in turn, is:

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

We will look at relaxing assumption 3 in an extension.

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

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

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

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.

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

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 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 their 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. Normalized Spearman's ρ:

$$\rho(Y_0, Y_1|x) = \frac{Cor_C(R(Y_0|x), R(Y_1|x))}{Cor_M(R(Y_0|x), R(Y_1|x))}$$

where $R(Y_t|x) = F_{Y_t|X}(Y_t|x)$ when Y_t is continuously distributed and $R(Y_t|x) = \frac{F_{Y_t|X}(Y_t|x) + F_{Y_t|X}(Y_t-|x)}{2}$ when Y_t takes a finite number of values. The notation $F_Z(z-)$ denotes $P(Z < z)$. The correlation is evaluated under the joint distribution generated by copula C in the numerator and the joint distribution generated under comonotonicity (M) in the denominator.

Under comonotonicity

$$F_{Y_0,Y_1|X}(y_0,y_1|x) = \min\left\{F^e_{Y_0|X}(y_0|x), F^e_{Y_1|X}(y_0|x)\right\}.$$

The corresponding unique copula acheiving $Cor_M(R(Y_0|x), R(Y_1|x))$ is

$$C(u,v) = \min\{u,v\}.$$

The definition of Normalized Spearman's ρ is chosen to coincide with the standard calculation of Spearman's ρ in the numerator. In the denominator, when Y_0 and Y_1 are continuously distributed, the correlation between $R(Y_0)$ and $R(Y_1)$ under M is 1 so that the calculation is completely standard. The only difference 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 4. *C is an element of $\mathcal{C}(\rho^L)$, the set of copula functions such that $\rho(Y_0, Y_1|x) \geq \rho^L$ where $\rho^L \in [0,1]$.*

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

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

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 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 latent 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 et al. (2005)).

AI 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 AI). $U \perp\!\!\!\perp T$ 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, AI point out that assumption 3 is implied by monotonicity in $\underline{u}$ 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 pick up the parametric example introduced in section 2.1. Assume the parental input I is excluded from the production function so unobservables are one-dimensional[6] and the potential outcomes can be written as

$$\begin{aligned} Y_0 &= \beta_0 + \beta_{0X} X + \beta_{0S} S \\ Y_1 &= \beta_1 + \beta_{1X} X + \beta_{1S} S \end{aligned}$$

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

[6] This is not the only way to generate 1-dimensional unobservables in the linear production function described in section 2.1. We could make use of a single index specification for the unobservables where

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

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:

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

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:

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

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:

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

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.

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

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.

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

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

$\mathcal{C}(\rho^L)$ is a set of potentially infinite-dimensional objects making the search over them for the minimizing and maximizing copulas a computational task beyond the scope of this paper. We therefore confine attention to the case where outcomes and covariates are discrete or discretized, illustrating both in the empirical work.

Assumption 5. Finite support of the potential outcomes and covariates: $\mathcal{Y}_0 = \{y_{0,1}, \ldots, y_{0,j}, \ldots, y_{0J}\}$, $\mathcal{Y}_1 = \{y_{1,1}, \ldots, y_{1,k}, \ldots, y_{1K}\}$ and $\mathcal{X}$ is finite-dimensional.

In the discrete outcome and covariate case, the challenging search over $\mathcal{C}(\rho^L)$ for the copula functions that yield the minimal and maximal values for the average treatment effect in the population of interest becomes the solution to a linear programming problem which can be solved quickly using software provided by the author. The solution to the problem also illustrates the way in which differences in the conditional distributions of control outcomes $F^e_{Y_0|X}(y_0|x)$ and $F^a_{Y_0|X}(y_0|x)$ affect the bounds on ATE^a, which I take up at the end of the section.

We leave conditioning on x implicit to economize on notation. Given ρ^L, bounds on ATE^a can be computed by solving a discrete optimal transportation problem with a non-standard cost function and additional linear constraint. The upper bound is obtained by solving the following linear programming problem (the lower bound is obtained by replacing the max operator with min).

$$
\begin{aligned}
\max_{\{P^e(y_{0j},y_{1k})\}} & \sum_{j=1}^{J}\sum_{k=1}^{K} y_{1k}\frac{P^a(y_{0j})}{P^e(y_{0j})} \times P^e(y_{0j},y_{1k}) \\
& - \sum_{j=1}^{J} y_{0j}P^a(y_{0j}) \qquad (12)
\end{aligned}
$$

subject to

$$\sum_{k=1}^{K} P^e(y_{0j},y_{1k}) = P^e(y_{0j}) \ \forall j \in \{1,...,J\} \qquad (13)$$

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

$$
\begin{aligned}
& \sum_{j=1}^{J}\sum_{k=1}^{K}\left(R(y_{0j})-\frac{1}{2}\right)\left(R(y_{1k})-\frac{1}{2}\right)P^e(y_{0j},y_{1k}) \\
& \geq \rho^L\left[\max_{\{P^e(y_{0j},y_{1k}\}}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(R(y_{0j})-\frac{1}{2}\right)\left(R(y_{1k})-\frac{1}{2}\right)P^e(y_{0j},y_{1k})\right] \qquad (15) \\
& P^e(y_{0j},y_{1k}) \geq 0 \ \forall j \in \{1,...,J\}, k \in \{1,...,K\}
\end{aligned}
$$

Maximization is with respect to the elements of the matrix defining the joint distribution of Y_0 and Y_1 in population e,$\{P^e(y_{0j},y_{1k})\}$. Line 12 is simply a normalization so that the value of the objective function of the problem can be interpreted as ATE^a. Constraints 13 and 14 require that the minimizing/maximizing joint distribution be consistent with the marginal outcome distributions in e.

Table 2 shows an example of the choice variables and constraints 13 and 14 in the context of the remedial education in India example where Mumbai is treated as e and we condition on a competency level of zero on entering third grade, $x = 0$. The choice variables are highlighted in blue, while 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 control outcomes) equal the probability in the column labelled "All" and that the sums down the columns (for the treated outcomes) equal the probability in the row labeled "All."

The coefficients on the elements of $\{P^e(y_{0j},y_{1k})\}$ are $\left\{y_1\frac{P^a(y_{0j})}{P^e(y_{0j})}\right\}$. Together with constraint 14, this shows the role of the distributions of control outcomes $\{P^a(y_{0j})\}$ and $\{P^e(y_{0j})\}$ in determining the bounds. If $P^a(y_0) \approx P^e(y_0)$, $\frac{P^a(y_0)}{P^e(y_0)} \approx 1$ and constraint 14 implies that the

Table 2: Choice variables - $P^e(y_{0j}, y_{1k}|\text{pre-competency} = 0)$, e =Mumbai

		y_1: post-competency (treatment)				
		0	1	2	3	All
	0	$P^e(0,0)$	$P^e(0,1)$	$P^e(0,2)$	$P^e(0,3)$	0.73
y_0: post-competency (control)	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	0.66	0.20	0.10	0.04	1

counterfactual $E^a[Y_1] = E^e[Y_1]$ [7]. All 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. For example, table 3 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 15 on the dependence between Y_0 and Y_1 in e 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 4 shows the joint distributions 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.

The linear programming problem can be estimated using a sample analog.

[7] 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 third line follows from substituting in constraint 14.

Table 3: Contribution of choice variables to the objective -$P^e(y_{0j}, y_{1k}|\text{pre-competency} = 0)$, e =Mumbai

		y_1: post-competency			
		(treatment)			
		0	1	2	3
	0	0	0.71	2×0.71	3×0.71
y_0:	1	0	2.26	2×2.26	3×2.26
post-competency	2	0	1.16	2×1.16	3×1.16
(control)	3	0	0.60	2×0.60	3×0.60

Table 4: $P^e(y_{0j}, y_{1k}|\text{pre-competency} = 0)$, $\rho^L = 1$ e =Mumbai

		y_1: post-competency				
		(treatment)				
		0	1	2	3	All
	0	0.66	0.07	0	0	0.73
y_0:	1	0	0.13	0.04	0	0.17
post-competency	2	0	0	0.06	0.01	0.07
(control)	3	0	0	0	0.03	0.03
	All	0.66	0.20	0.10	0.04	1

4.3 Inference

Confidence intervals with a fixed asymptotic coverage probability of containing the true value of ATE^a conditional on ρ^L can be computed using the method of Imbens and Manski (2004) (henceforth IM). IM provide a method for computing the upper and lower bounds of the confidence interval given standard errors for the upper and lower bounds on ATE^a under the high-level assumption that the asymptotic distribution of the bounds is Gaussian. The asymptotic distribution of the bounds is not available in closed form, so I compute standard errors for the bounds using the bootstrap under the high-level assumption that they are normally distributed (more detail to be added).

4.4 Extensions

4.4.1 2-dimensional sensitivity analysis: relaxing assumption 3

4.4.2 Multiple experimental populations

4.4.3 Failure of support assumptions 1 and 2

5 Transfers to Mexican microenterprises: results

The experiment carried out in 2006 (baseline Oct. 2005) in Leon, Mexico. The treatment was a 1,500 (≈ \$140) peso transfer (50% “in-kind”). Y is monthly profits. $ATE^e \approx 600$ pesos. Uniquely, the questionnaire used in the experiment was based on the national microenterprise survey: Encuesta Nacional de Micronegocios (ENAMIN). The sample in the experiment was the following:

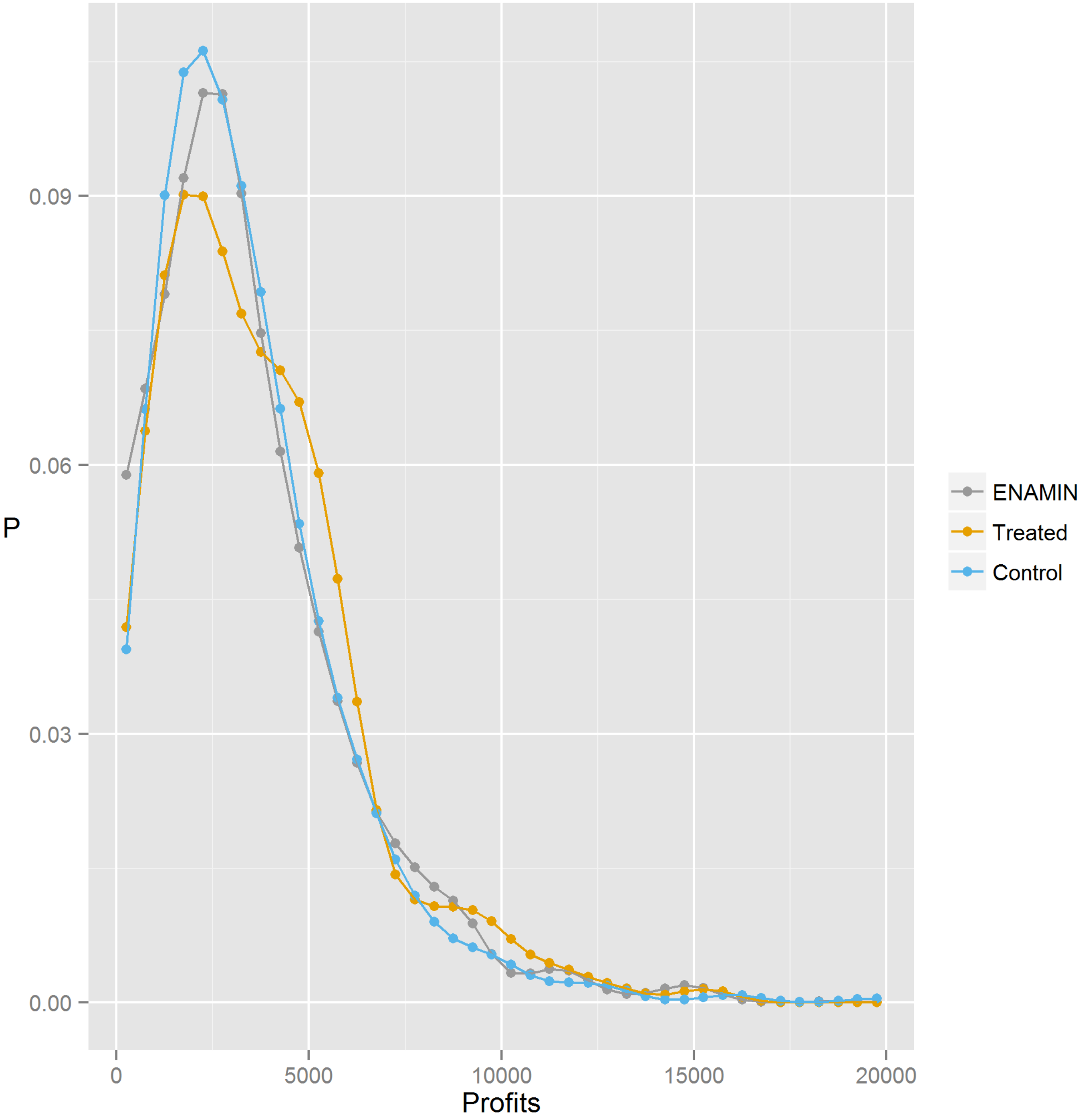
- 22-55 year old male entrepreneurs
- Working in retail
- Capital stock $\leq$ 10,000 pesos
- No paid employees
- Working 35+ hours per week in microenterprise

I select a sample using the same (inflation-adjusted) criteria from the 2012 ENAMIN with the additional restriction that the entrepreneurs be working in urban areas. I trim profit reports of more than 15000 pesos. Since sample selection chooses a restricted set of individuals, I will not condition on any X. In addition, while the ENAMIN sample has 903 microentrepreneurs fitting the criteria, the experiment had only 207 unique microentrepreneurs. Unsurprisingly, confidence sets are very wide. However, we will see that the bounds are fairly narrow over a wide range of assumptions on heterogeneity.

4 shows the outcome distributions in ENAMIN as well as the treated and control groups in the experiment. Since heaping is an issue in reported profits, I have smoothed them using a kernel density estimator before discretizing to 500 peso bins. The reason for the narrow bounds is clear from figure 4, which shows that the experimental control group and the ENAMIN sample have very similar outcome distributions.

The combination of small sample size and similar control outcome distributions yields figure 5, which shows bounds (in dark blue) on the average treatment effect of providing cash transfers to male microentrepreneurs in urban Mexico in 2012 as a function of the level

Figure 4: Outcome distributions



of treatment effect heterogeneity allowed, ρ^L. Imbens and Manski (2004) 95% confidence regions (translucent blue) are computed using 100 bootstrap replications for each ρ^L, clustering at the firm level for the experiment. The plot shows two aspects of the procedure: 1) the similar control outcome distributions yield narrow bounds on the average treatment effect for male microentrepreneurs in urban Mexico in 2012 for a wide range of possible levels of treatment effect heterogeneity and 2) the experimental sample size is sufficiently small that we cannot reject an average treatment effect of zero at any level of heterogeneity. 2) is actually a feature of the procedure. Under previous methodologies, we would test the equality of the control distributions in figure 4. Having been unable to reject due to the small size of the experimental sample, we would predict the average treatment effect for male microentrepreneurs in urban Mexico in 2012 to be equal to the average treatment effect in the experiment, with an identical confidence interval. Since MW were able to reject a zero average treatment effect in the original experiment, we would do the same in extrapolating to male microentrepreneurs in urban Mexico in 2012, despite 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.

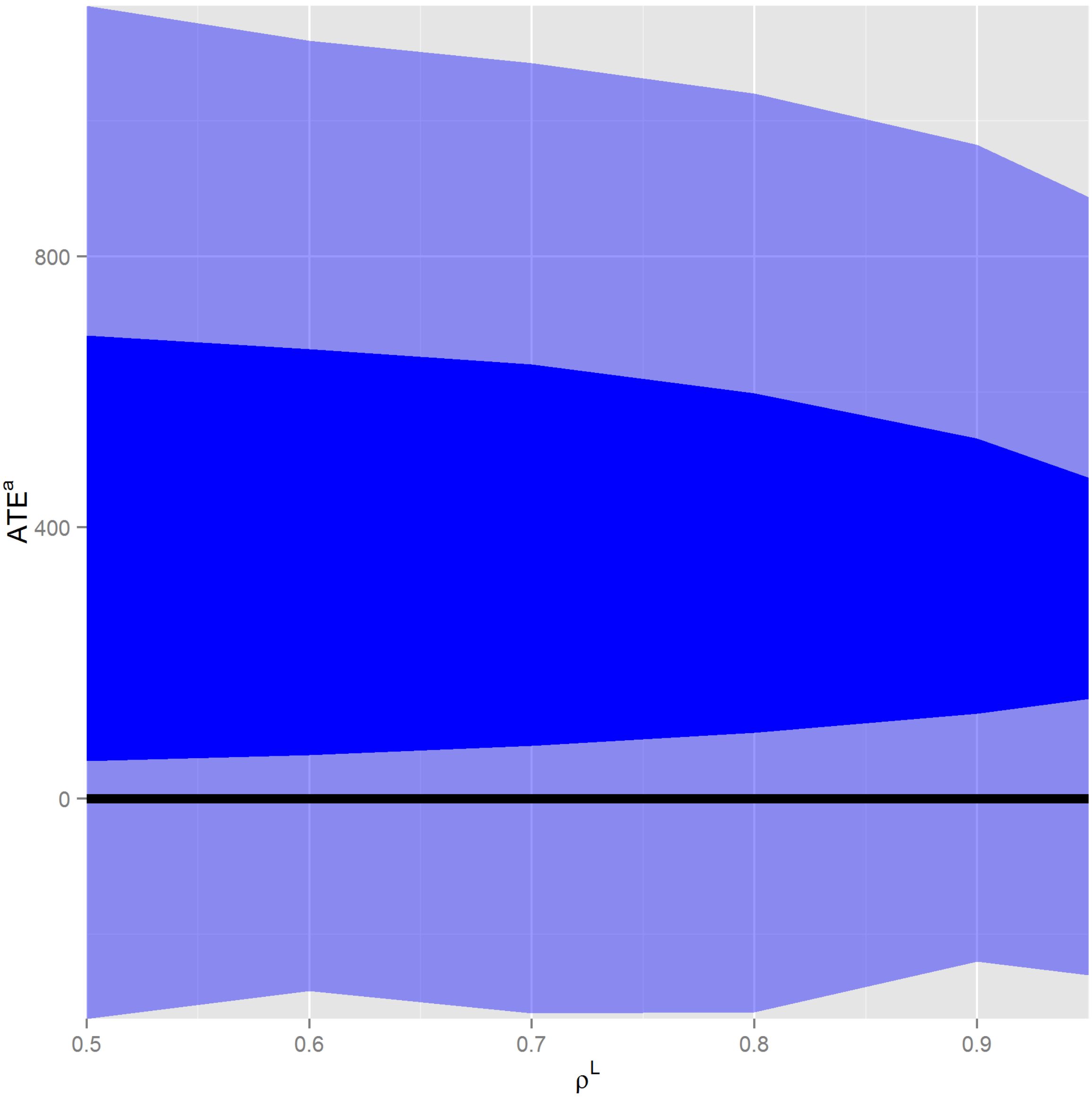
6 Remedial education in India: results

The remedial education program was implemented by the same NGO, Pratham, in both cities. 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.

6.1 Data

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 average treatment effects, 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. To focus on the issue of individual heterogeneity in treatment response, rather than idiosyncratic issues affecting all members of each population, I consider the Mumbai population as made up of

Figure 5: Bounds on ATE^a



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. Idiosyncratic issues affecting all members of a population are an important barrier to the generalization of experimental results (see, for example, Bold et al. (2013)), but lie outside the scope of this paper.

A challenge in applying the methods discussed above in this data set is that the researchers administered different math tests in the two samples. 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. We foreshadowed the solution in previous sections. The tests also recorded the students' grade level competency: that is, whether the student successfully answered questions showing mastery of the subjects taught in each grade. This measure of achievement should be relatively comparable across populations.

With the exception of the pre-test score, relatively little data on students are available consistently across the two samples. Tables 5 and 6 show summary statistics for the pre-test scores in the two samples 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. There is no evidence of treatment effect heterogeneity on either of these characteristics, so we ignore them and focus on the pre-test score, as we did in the theoretical exposition.

Table 5: Vadodara

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

Table 6: Mumbai

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

Table 7 shows the difference in the unconditional average treatment effects. The first line shows the average treatment effect in Vadodara. In Vadodara, the treatment raised students' maximum grade level competency in math by .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 treatment effect in Mumbai is estimated at .059 grade levels, .103 less than the Vadodara ATE.

Table 7: Unconditional ATEs

	Post-test: maximum 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.

6.2 Using Mumbai to predict Vadodara

We now move to investigating the level of treatment effect heterogeneity needed for 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).

Turning to the method developed in this paper, figure 6 plots bounds on the predicted values of the average treatment effect in Vadodara as a function of the degree of treatment effect heterogeneity allowed for individuals with the same grade level competency on entering third grade. The bounds are plotted in dark blue, while the translucent light blue region represents a 95% Imbens and Manski (2004) confidence interval, based on 100 bootstrap replications[8]. The bounds become a point when we impose minimum treatment effect heterogeneity for individuals with the same competency level on entering third grade. 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 table 1. A zero average treatment effect in Vadodara can only be rejected using the Mumbai results

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

if $\rho^L > .925$.

The red line plots the estimated of ATE in the Vadodara sample, while the translucent red region shows the 95% confidence interval. In terms of the unconditional prediction of ATE^a, we see that though the point estimate with maximum rank correlation (minimum unobserved treatment effect heterogeneity) 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.

6.3 Using Vadodara to predict Mumbai

Figure 7 shows the results of using Vadodara to predict Mumbai. The results show the difficulty that arises when assumption 1 support fails. As shown in table 1, 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 7 assign these students the lower bound of the support 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 possible magnitudes of treatment effect heterogeneity (¡ .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. Using theory that will be developed in section 4.4.3, I will explore 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.

7 Conclusions

I conclude with a few suggestions for applied researchers concerned about external validity. First, it is important to specify the population of interest other than the one where the study was conducted. The robustness of inferences on average causal effects in the alternative population of interest will depend in large part on the extent of difference in the control outcome distributions between the study population and the alternative population of interest. In the empirical results, we saw that the distributions of student acheivement without remedial education were sufficiently different between Mumbai and Vadodara that a

Figure 6: Using Mumbai to predict the Vadodara ATE (blue) and Vadodara estimated ATE (red)

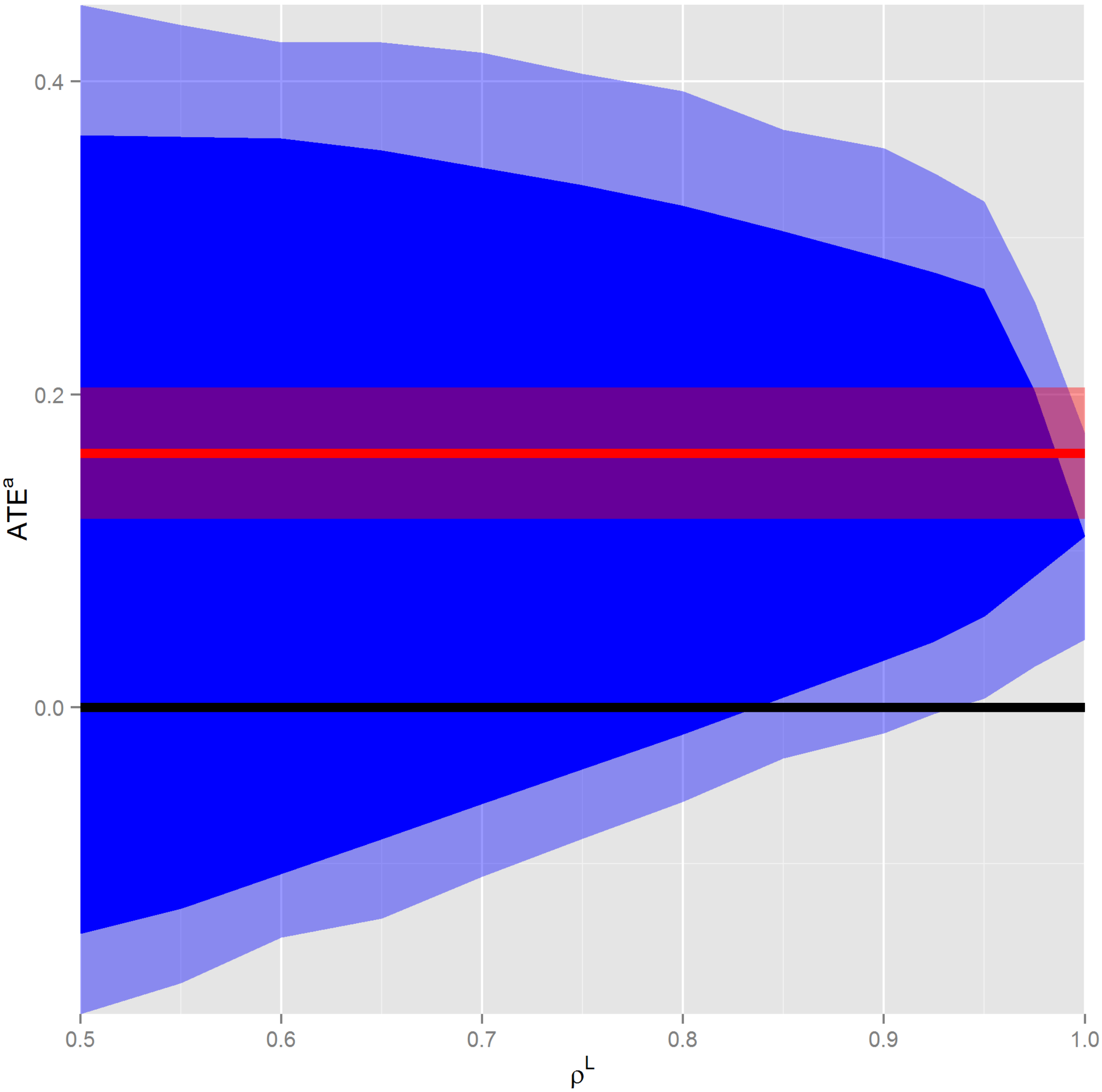
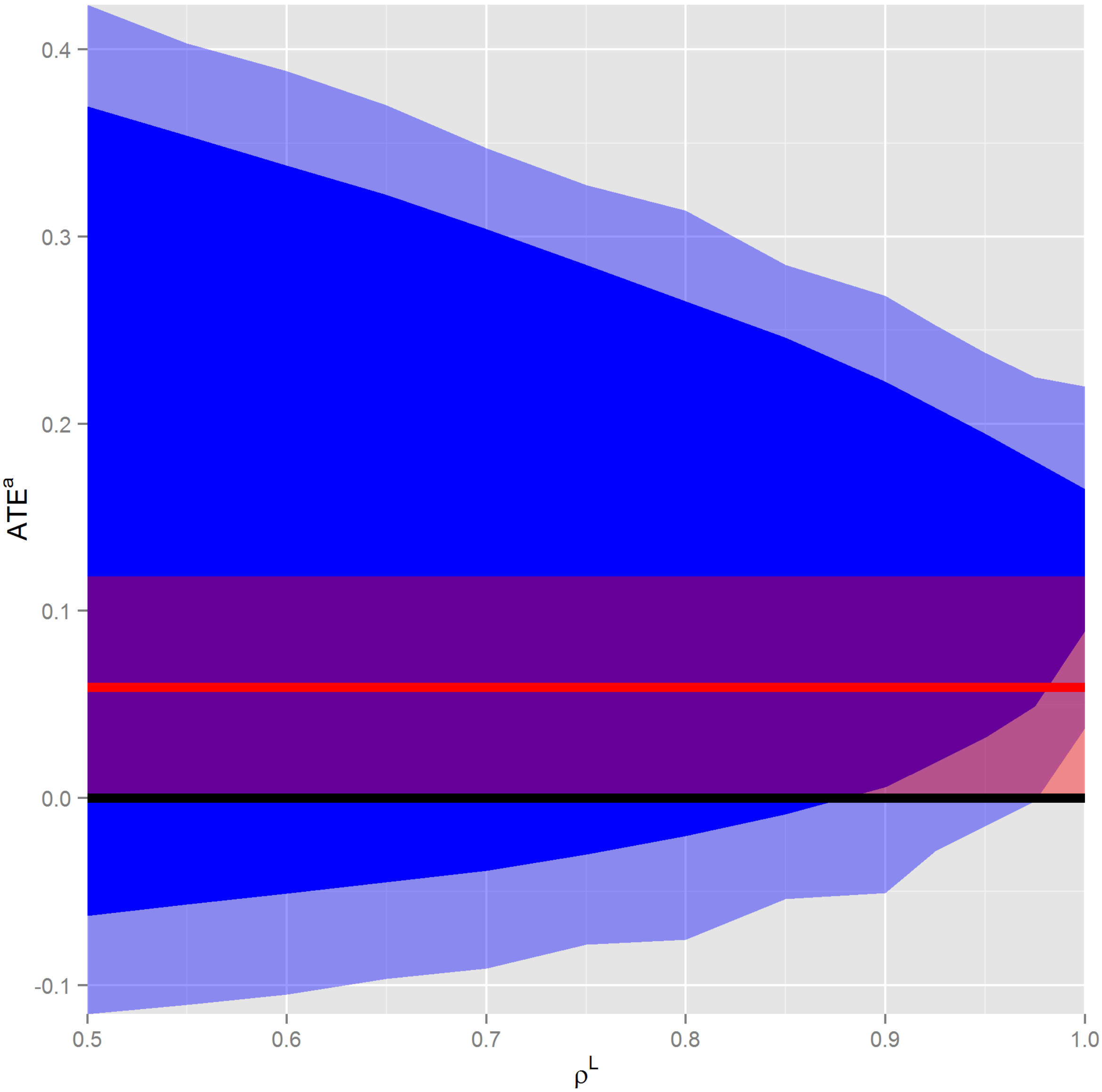


Figure 7: Using Mumbai to predict the Vadodara ATE (blue) and Vadodara estimated ATE (red)



zero causal effect for Vadodara using the Mumbai experiment could be rejected only under a small set of assumptions and vice versa. Second, researchers should report the assumptions under which hypotheses involving the average causal effect in the alternative population can be rejected to let readers determine the credibility of the inferences.

References

Allcott, H. (2014). Site Selection Bias in Program Evaluation.

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).

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

Angrist, J. and J. Pischke (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives 24*(2), 3–30.

Athey, S. and G. W. 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.

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.

Banerjee, A. V. and E. Duflo (2009). The experimental approach to development economics. *Annual Review of Economics* (1), 151–78.

Bitler, M., T. Domina, and H. Hoynes (2014). Experimental evidence on distributional effects of head start.

Bold, T., M. Kimenyi, G. Mwabu, A. Ng'ang'a, and J. Sandefur (2013). Scaling-up What Works: Experimental Evidence on External Validity in Kenyan Education.

Cole, S. R. and E. a. Stuart (2010, July). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American journal of epidemiology 172*(1), 107–15.

de Mel, S., D. McKenzie, and C. Woodruff (2008). Returns to Capital in Microenterprises: Evidence from a Field Experiment*. *Quarterly Journal of Economics*.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature 48*(2), 424–455.

Djebbari, H. and J. Smith (2008, July). Heterogeneous impacts in PROGRESA. *Journal of Econometrics 145*(1-2), 64–80.

Duflo, E., R. Glennerster, and M. Kremer (2008). Using randomization in development economics research: A toolkit. *Handbook of development economics 4*(07).

Fafchamps, M., D. McKenzie, S. Quinn, and C. Woodruff (2014, January). Microenterprise growth and the flypaper effect: Evidence from a randomized experiment in Ghana. *Journal of Development Economics 106*, 211–226.

Fan, Y. and S. Park (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 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* (4451).

Fortin, N., T. Lemieux, and S. Firpo (2011). Decomposition methods in economics. *Handbook of Labor Economics 4*(11), 1–102.

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. W. Imbens, and J. H. Mortimer (2005, March). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics 125*(1-2), 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.

Kline, P. and A. Santos (2013). Sensitivity to missing data assumptions: theory and an evaluation of the US wage structure. *Quantitative Economics 4*(2013), 231–267.

McKenzie, D. and C. Woodruff (2008). Experimental evidence on returns to capital and access to finance in Mexico. *The World Bank Economic Review*, 457–482.

McNeil, A., R. Frey, and P. Embrechts (2005). *Quantitative risk management: concepts, techniques, and tools.* Princeton University Press.

Oster, E. (2014). Unobservable Selection and Coefficient Stability: Theory and Validation.

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

Stuart, E. a., S. R. Cole, C. P. Bradshaw, and P. J. Leaf (2011, April). 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 $C : [0,1]^2 \rightarrow [0,1]$ 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:

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