GENERALIZING THE RESULTS FROM SOCIAL EXPERIMENTS: THEORY AND EVIDENCE FROM MEXICO AND INDIA

Michael Gechter[1]

How informative are treatment effects estimated in one region or time period for another region or time? In this paper, I derive bounds on the average treatment effect in a context of interest using experimental evidence from another context. The bounds provide robustness against the possibility of unobserved treatment effect moderators whose distribution differs across contexts. Empirically, I use results from an experiment on returns to cash transfers given to microentrepreneurs in Leon, Mexico to predict average returns among microentrepreneurs in other Mexican cities. I show that the benchmark extrapolation method from the literature yields implausibly precise predictions for other cities considering the very small experimental sample. Using data from a pair of remedial education experiments carried out in India, I show the bounds are able to recover average treatment effects in one location using results from the other while the benchmark method cannot.

Keywords: external validity, partial identification, sensitivity analysis.

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 va-

Department of Economics, The Pennsylvania State University, 609 Kern Building, University Park, PA 16802. Email: mdg5396@psu.edu.

[1] First version: April 15, 2014. This version: December 22, 2016. This paper is a revised version of the first chapter of my Ph.D. dissertation. I am grateful to my advisors Dilip Mookherjee, Hiroaki Kaido and Kevin Lang for guidance, support and encouragement. I also thank Kehinde Ajayi, Manuel Arellano, Sam Bazzi, Xavier d'Haultfoeuille, Rajeev Dehejia, Isabella Dobrescu, Iván Fernández-Val, Claudio Ferraz, Andy Foster, Xin Geng, 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 numerous seminar and conference participants for useful conversations and feedback. I especially thank David McKenzie for discussions and sharing the data from McKenzie and Woodruff [2008]. Part of this work was completed while I was at the Markets, Trade, and Institutions Division of IFPRI. I thank IFPRI and Maximo Torero in particular for supporting this project. Financial support from the Boston University Department of Economics is gratefully acknowledged. Gloria Sarmiento-Becerra provided excellent research assistance.

riety of dissimilar contexts (Deaton [2010], Deaton and Cartwright [2016], 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]). Relevant unobserved differences between contexts 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 treatment 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 in 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 for any pair of contexts using the software accompanying this paper. They formalize the idea that the conclusions we can draw about the average treatment 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 focus on settings where a randomized evaluation of a pilot program has been run and we 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. Data are available on characteristics and outcomes of individuals participating in the experiment. There are also data available 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, the distri-

[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 where 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). It is also possible to consider the case where individuals in the context of interest choose their treatment status. In this case, the experimental data can be used to bound the counterfactual expected untreated

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

For each set of characteristics, the bounds I derive on the average treatment effect in the context of interest 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. That is, the conditional distribution in the context of interest can be generated from one of the possible joint distributions of the potential outcomes in the experiment.[4] This is a relatively weak restriction on the average treatment effect because the experiment does not rule out any level of dependence between treated and untreated outcomes.[5] Except in extreme cases we expect positive dependence between treated and untreated outcome levels for any individual, to varying degrees depending on the nature of the program.

I therefore develop tighter bounds, indexed by the minimum level of dependence between an individual's treated and untreated outcomes the researcher is willing to consider. When treated and untreated outcomes are perfectly dependent, there is only a single joint distribution of untreated and treated outcomes consistent with the experimental results.[6] As we move away from perfect dependence, different associations between treated and untreated outcomes become possible. These different associations produce uncertainty about the average treatment 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 treatment effect. They also allow researchers to assess the assumptions on dependence necessary to draw specific conclusions about the effect of the program in the context of interest,

outcome among individuals choosing to take the treatment and the expected treated outcome among individuals who choose not to take the treatment. Alternative assumptions available in the treatment choice setting, such as first stage monotonicity, may lead to tighter bounds (see, e.g., Kowalski [2016] who uses first stage monotonicity along with other assumptions). I focus the exposition on settings where the evaluated policy is a pilot program because it is common in the development literature and because of its relative simplicity.

[4] The joint distribution of treated and untreated outcomes is not point-identified because of the fundamental problem of causal inference: we cannot observe an individual in both the treated and untreated state at the same time. However the experiment allows us to identify the marginal distributions of the potential outcomes, which place restrictions on their joint distribution.

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

[6] In the continuous outcome case, each untreated outcome is linked to a single treated outcome.

such as its ability to exceed a cost-effectiveness threshold.

Computation of the bounds is challenging because it requires solving an infinite-dimensional optimization problem over the space of possible joint distributions of treated and untreated outcomes. I address this difficulty when outcomes and characteristics are discrete by deriving an alternative representation of the problem based on optimal transportation theory (c.f. Galichon [2016], Villani [2009]). I show how this representation, which I use in estimation and inference, can be solved quickly using linear programming techniques.

I empirically evaluate the results of my bounding procedure compared to the current benchmark method for extrapolating treatment effects: Hotz, Imbens, and Mortimer [2005] (henceforth HIM). HIM assume that the joint distribution of untreated and treated outcomes for individuals with the same observed characteristics is independent of context. They also suggest using untreated outcomes for individuals with the same characteristics to assess generalizability, but within a testing framework. If we reject independence across contexts of the untreated outcome distributions conditional on a set of observed characteristics, we conclude that the experiment teaches us nothing about causal effects in the context of interest. Otherwise, the HIM framework leads us to conclude that the experiment point-identifies the treatment effect of interest.

I first examine the generalizability of a small experiment investigating the returns to loosening credit constraints by providing cash transfers to very small-scale entrepreneurs in Leon, Mexico in 2006 (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 microentrepreneurs with the same characteristics in urban Mexico in 2012 of participating in the transfer program, holding program scale constant. The distributions of untreated outcomes are fairly similar in the Leon and 2012 urban Mexico samples[7] so the estimated bounds are narrow for a wide range of assumptions on dependence between profits with and without the transfer. Accounting for the unobserved differences between the contexts along with sampling variation in the small experiment and the national survey 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, given the test's lack of power. Using the HIM method, we would compute a narrower confidence interval on

[7] The 2012 data are obtained from that year's microenterprise survey.

the predicted treatment effect for urban Mexico in 2012 than on the treatment effect in the original experiment.

Second, to check the predictions against measured average treatment 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]. 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 treatment effects - such as a zero effect - in one city using the results from the other. The observed causal effects are consistent with predictions based on strong dependence between the treated and untreated outcomes.

This paper extends the literature on generalizing treatment effects to new contexts based on invariance assumptions on distributions of untreated and treated outcomes or treatment effects for individuals with the same observed characteristics (HIM, Attanasio et al. [2003], Angrist and Fernández-Val [2013], Angrist and Rokkanen [2015], Cole and Stuart [2010], Stuart, Cole, Bradshaw, and Leaf [2011], Pearl and Bareinboim [2014], Flores and Mitnik [2013], Dehejia, Pop-Eleches, and Samii [2015]).[8] The literature on external validity reviewed in Muller [2014] emphasizes the failure of generalization when there are unobserved variables producing treatment effect heterogeneity whose distribution differs across contexts. The methodology developed here allows for treatment effect heterogeneity due to unobserved variables that play a role in determining untreated outcomes. In related work, Athey and Imbens [2006] use outcome distributions for different groups of individuals in the same time period to capture unobserved differences across groups when generalizing the linear difference-in-differences estimator. One of their estimators coincides with mine, with time playing the role of treatment, under perfect dependence between treated and untreated outcomes.[9]

[8] Meta-analysis is another methodology discussed in the context of external validity (e.g. Vivalt [2016], Meager [2016]). With a sizable set of studies of the same treatment, meta-analysis can be used to evaluate the ability of one or more studies' results to extrapolate to others in the set. When used to extrapolate to contexts outside the set of studies, meta-analysis generates a point estimate that assumes individuals with the same characteristics will have the same treatment response.

[9] My linear-programming-based estimator can also be used to examine sensitivity of estimates using this estimator to Athey and Imbens [2006]'s assumption of perfect dependence in 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 treatment effect whose width depends on the magnitude of the relationship between the covariates and the instrument.[10] While Altonji et al. [2005] and Altonji et al. [2013] work within the context of a linear model, my identification analysis is non-parametric and thus similar in spirit to Kline and Santos [2013]'s analysis of the sensitivity of conclusions about conditional distributions to missing data. Gerard, Rokkanen, and Rothe [2015] similarly move away from the McCrary [2008] test for a manipulated running variable in regression discontinuity designs to provide bounds on treatment effects.

The rest of the paper is organized as follows. Section 2 describes the intuition behind the proposed methods by means of a simple example. Section 3 sets up the problem and notation and describes HIM's approach formally. In Section 4, I present the derivation of the bounds. Section 6 presents the empirical results for generalizing from the 2006 Leon microenterprise experiment to microentrepreneurs in urban locations in Mexico in 2012. Section 7 investigates using the results from one of the two remedial education experiments to try to predict the results in the other experiment. Section 8 concludes. All proofs are collected in appendices.

2. INTUITION FOR THE METHODOLOGY: A SIMPLE EXAMPLE

To illustrate the intuition behind the methodological contribution, 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 covariates.

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

across time.

[10] Alternative forms of sensitivity analysis (of unconfoundedness assumptions) are explored in Imbens [2003] and Rosenbaum [2002].

without the CCT in location a. How will the difference in the no-CCT enrollment rates moderate the average treatment 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[11], as follows.

$$\begin{aligned}
&ATE^a \\
&= P^a(\text{enroll w/o CCT})P^a(\text{enroll w/CCT} \mid \text{enroll w/o CCT}) \\
&\quad + P^a(\text{out of school w/o CCT})P^a(\text{enroll w/CCT} \mid \text{out of school w/o CCT}) \\
&\quad - P^a(\text{enroll w/o CCT}) \\
&= \frac{1}{2}P^a(\text{enroll w/ CCT} \mid \text{enroll w/o CCT}) \\
&\quad + \frac{1}{2}P^a(\text{enroll w/ CCT} \mid \text{out of school w/o CCT}) - \frac{1}{2}
\end{aligned}$$

The average treatment effect in a depends on two unknowkn probabilities: (1) the probability that a child who *does not* enroll without the CCT would instead enroll with the CCT and (2) the probability that a child who enrolls in school without the CCT would also enroll with the CCT. Note that enrollment with and without the CCT are contemporaneous potential outcomes and thus can never be jointly observed for any child.

The rationale behind (1), children who do not enroll without the CCT but do enroll 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), children 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. Some households may be more sensitive to child wages than they are to enrollment subsidies and would respond to the 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 the CCT in place as wage effects, although in principle there may be

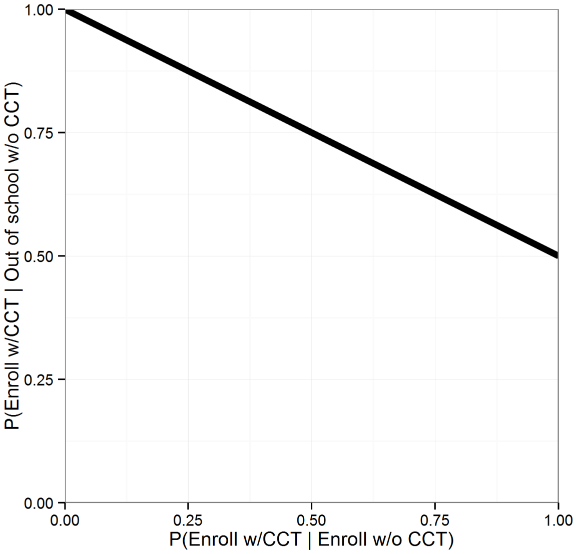
[11] Throughout the paper, I will use context-specific superscripts to index quantities conditioned on context. For example, letting $D \in \{a, e\}$ denote context,

$$ATE^a = E[Y_1 - Y_0 | D = a]$$

where Y_1 and Y_0 denote outcomes with and without the CCT program.

other ways for the CCT to cause children who would otherwise enroll to not enroll.

FIGURE 1.— Permissible values for P^a(enrolled with CCT | enrollment without CCT)



The paper's key identifying assumption is that the conditional probabilities

$$\{P^a(\text{enroll w/ CCT} \mid \text{enroll w/o CCT}),\\ P^a(\text{enroll w/ CCT} \mid \text{out of school w/o CCT})\}$$

in a are consistent with the experimental results from e. That is, pairs of conditional probabilities which cannot be ruled out by the results in e. There are many possible pairs of conditional enrollment probabilities that are consistent with the experimental results. The possible pairs are given by the black line in Figure 1. Despite the fact that a continuum of pairs is possible under consistency with the experimental results, the assumption has substantial identifying power. Without it, any pair of conditional probabilities is possible.

To see why a continuum of pairs is possible, note that

$$\begin{aligned} &P(\text{enroll w/ CCT} \mid \text{enroll w/o CCT}) \\ &= \frac{P(\text{enroll w/ CCT \& enroll w/o CCT})}{P(\text{enroll w/o CCT})}. \end{aligned} \tag{1}$$

Equation 1 makes explicit the fact that P(enroll with CCT | enroll without CCT) relies on knowledge a child's enrollment status with and without the CCT at the same time, knowledge that is unavailable to us. 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 I 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 available from the experiment. The rows of Table I must sum to the control group results and the columns to the treatment group results.

TABLE I

THE JOINT DISTRIBUTION OF ENROLLMENT WITH AND WITHOUT 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	$\frac{1}{3}$	$\frac{2}{3}$	

Additional assumptions about the way wage effects of the CCT impact children who do and don't enroll without the CCT generate different predictions for the causal effect of the CCT program in location a. To see this, first consider assuming that 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. The assumption allows us to fill in all the entries of Table I, as shown in Table II.[12] The probability of enrolling with the CCT if a child is out of school without the CCT is $\frac{1}{2}$. This is the right endpoint of the line in Figure 1. Under this "no wage effect" assumption, the increase in the fraction enrolled in location a is $\frac{1}{4}$.

Now consider another assumption about 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

[12] In this binary outcome case the "no wage effect" assumption is analogous to the "no defiers" assumption, imposed on treatment choice, in the Imbens and Angrist [1994] Local Average Treatment Effect framework.

TABLE II

Case 1: 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	$\frac{1}{3}$	$\frac{2}{3}$	

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 I, as shown in Table III. This case represents the left endpoint of the line in Figure 1. In this rather unbelievable case, we predict no change in the fraction enrolled in location a. The case is unbelievable because it seems unlikely that preferences for schooling with and without the CCT would be negatively correlated.

TABLE III

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 I can be filled in as shown in Table IV. The predicted increase in the fraction employed is $\frac{1}{6}$.

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

TABLE IV

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	$\frac{1}{3}$	$\frac{2}{3}$	

et al. [1997b] and measure dependence using the rank correlation[13] between treated and untreated outcomes for any individual. The "no wage effects" assumption generates the maximum possible rank correlation between a child's enrollment decision with and without the CCT. The second assumption, that the CCT makes all enrollees drop out and all dropouts enroll, generates the minimum possible rank correlation. The third assumption, that wage effects are independent of enrollment status without the CCT, generates a rank correlation of zero. As I have shown, 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.[14] Then, the bounds on the enrollment gain in location a are $\frac{1}{6}$ and $\frac{1}{4}$. A second option, which I will emphasize, 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 of the CCT is only possible when the rank correlation between enrollment with and without the CCT is the minimum possible. Since this only occurs in the implausible case where children who enroll without the CCT all drop out with the CCT, we would feel confident in our conclusion that the CCT will have a positive effect on enrollment location a.

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

[14] A related approach, pursued in the NBER working paper draft of Heckman et al. [1997b], is to specify a prior over possible values of the rank correlation.

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 produce bounds on the increase in the enrollment rate of 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 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. Observed characteristics can be conditioned on. Non-binary outcomes present a particular challenge because the lower and upper bound on ATE^a are no longer guaranteed to occur at the endpoints of the multi-dimensional equivalent of Figure 1. To face this difficulty in estimation, I derive the aforementioned characterization of the bounds in terms of an optimal transportation problem. Taking sampling into account, I produce confidence intervals for ATE^a robust to the parameter's being only bounded when using experimental data from e and data from untreated individuals in a. Readers primarily interested in applications of the methodology may, however, wish to skip to the empirical results beginning in Section 6.

3. ECONOMETRIC SETUP

In general, 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: $Y_1 \in \mathcal{Y}$ is her outcome if she receives treatment and $Y_0 \in \mathcal{Y}$ is her outcome if she does not. Only one of these two outcomes is ever observed, the other is hypothetical. The observed outcome Y can be written as:

$$Y = TY_1 + (1-T)Y_0. \tag{2}$$

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

$$\Delta = Y_1 - Y_0.$$

Data come from two contexts, indexed by $D \in \{e,a\}$. e is the context in which an experimental evaluation of T was conducted and a is the alternative context of interest. d-superscripts will index context-specific distributions and their attributes.

In context e I assume an evaluator assigns T independently of all other random variables with perfect compliance.[15] The probability of assignment to treatment in context e is bounded away from zero and one.

ASSUMPTION 1 ***Random assignment in context e.*** $0 < P^e(T = 1) < 1$. *$T|D = e$ is independent of all other random variables.*

Under Assumption 1, we can identify the marginal distribution of untreated outcomes in e, $F^e_{Y_0}(y_0)$:

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

where $F^e_{Y|T}(y|T = 0)$ denotes the marginal distribution of Y conditional on the treatment indicator being equal to zero. By the same argument, we can also identify the marginal distribution of treated outcomes, $F^e_{Y_1}(y_1)$. We can similarly identify the average treatment effect in e:

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

As in previous sections, I maintain that all members of the alternative population are untreated.

ASSUMPTION 2 ***Treatment assignment in context a.*** *$T = 0|D = a$.*

In context a, we can identify the distribution of untreated outcomes:

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

In this paper, the object of interest is the average treatment effect in the alternative context, $E^a[\Delta]$, which depends on our ability to identify the expectation of the

[15] Incorporating dependence between T and covariates (conditional random assignment) introduces no special complications. Additionally, while the analysis can be extended to settings where the object of interest is the effect of a treatment with imperfect compliance, for the purpose of exposition the estimands of interest will be intention-to-treat (ITT) effects including any participation decisions.

counterfactual treated outcome $E^a[Y_1]$:

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

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. [1997b], 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.[16]

To demonstrate the role of heterogeneity in observed and unobserved characteristics on the average treatment effect in a, I 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 $U \in \mathcal{U} \subseteq \mathbb{R}^{d_U}$ affecting the outcome. An equivalent representation for the potential outcomes is that treatment status and covariates combine to produce the outcome through a deterministic function common across contexts, $g : \{0,1\} \times \mathcal{X} \times \mathcal{U} \to \mathbb{R}$. In this representation, the potential outcomes are:

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

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:

$$ATE^a = E^a[Y_1 - Y_0] = \int_{\mathcal{X}} \left[\int_{\mathcal{U}} g(1,x,u) - g(0,x,u) dF^a_{U|X}(u|x) \right] dF^a_X(x) \tag{3}$$

$$= \int_{\mathcal{X}} \left[\int_{\mathbb{R}^2} (y_1 - y_0)\, dF^a_{Y_0,Y_1|X}(y_0, y_1|x) \right] dF^a_X(x) \tag{4}$$

where $F^a_{U|X}(u|x)$ denotes the joint distribution of observed and unobserved covariates in population a. $F^a_{U|X}(u|x)$ in general differs from $F^e_{U|X}(u|x)$.

[16] See e.g. Bitler, Gelbach, and Hoynes [2006].

3.1. *The HIM approach*

Within this general setup, I briefly describe HIM's approach to identifying ATE^a.[17] HIM assume that the joint distribution of potential outcomes is independent of the population conditional on the observed covariates:

$$(Y_0, Y_1) \perp\!\!\!\perp D|X \tag{5}$$

or equivalently, that all unobserved covariates determining the outcome are independent of the context indicator: $U \perp\!\!\!\perp D|X$. It is straightforward to show that (5) implies $E^a[Y_1|x] = E^e[Y_1|x]$. As long as $\mathcal{X}^a \subseteq \mathcal{X}^e$ we can identify the average treatment effect in the population of interest by reweighting the conditional expectation of the treated outcome from e by the distribution of covariates in a and subtracting the expectation of the untreated outcome in a:

$$ATE^a = \int_{\mathcal{X}} E^e[Y_1|x] dF^a_X(x) - E^a[Y_0]. \tag{6}$$

For (5) to hold, the conditional distributions of untreated 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. 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 context of interest. First, considering the small sample sizes of many social experiments, we will often be underpowered to reject equality of the conditional outcome distributions (a point raised also 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. Suppose we have two alternative contexts of interest, a and a', and we can 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 treatment effect in both a and a'. In the following section, I depart from the testing framework and derive bounds on the average treatment effect in the population of interest as a function of the differences in the conditional distributions of untreated outcomes

[17] The treatment in this section is general. Readers interested in an example should consult appendix (A.1), which uses a parametric linear model for illustration.

between the population of interest and the experimental population.

4. IDENTIFICATION OF BOUNDS ON ATE^A

In this section, I derive bounds on the average treatment effect in context a that incorporate information about unobservables provided by the conditional untreated outcome distributions. Formalizing and generalizing the intuition from Section 2, I begin by deriving bounds imposing only the assumption that $F^a_{Y_1|X}(y_1|x)$ is consistent with the experimental results. I then derive bounds on ATE^a that impose additional restrictions on the dependence between treated and untreated outcomes for any individual.

4.1. *Bounds under consistency with experimental results*

Recall that we can already identify $E^a[Y_0]$ (simply the expected outcome in the population of interest). Therefore, what we need to identify $E^a[Y_1] - E^a[Y_0]$ is the counterfactual $E^a[Y_1]$. By the law of iterated expectations, the expected value of the treated outcome in the context of interest can be written as follows:

$$E^a[Y_1] = \int_{\mathcal{X}} \left(\int_{\mathbb{R}} \left[\int_{\mathbb{R}} y_1 dF^a_{Y_1|Y_0,X}(y_1|y_0,x) \right] \underbrace{dF^a_{Y_0|X}(y_0|x)}_{identified} \right) \underbrace{dF^a_X(x)}_{identified}. \tag{7}$$

We are missing information on the distribution of treated outcomes that individuals with a particular untreated outcome would experience in the context of interest. Since no one is treated in the context of interest, we must turn to the experiment to learn about this object. I begin by deriving bounds on $E^a[Y_1|x]$ which I will use to bound $E^a[Y_1 - Y_0]$ by integrating over the distribution of X in a and subtracting $E^a[Y_0]$.

4.1.1. *Bounds on $E^a[Y_1|x]$*

I use the experimental data to provide information about $F^a_{Y_1|Y_0,X}(y_1|y_0,x)$ by imposing a generalization of the consistency with experimental results assumption described in Section 2. I assume $F^a_{Y_1|Y_0,X}(y_1|y_0,x)$ is a distribution not ruled out by the marginal distributions identified in the experiment, $F^e_{Y_0|X}(y_1|x)$ and $F^e_{Y_1|X}(y_1|x)$. $F_{Y_1|Y_0,X}(y_1|y_0,x)$ 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 and only 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 $C : [0,1]^2 \rightarrow [0,1]$ is a copula function (see appendix C.1 for the definition) and $C_1(v,w|x) = \frac{\partial C(v,w|x)}{\partial v}$ (Nelsen [2006]). Informally, a copula function is a bivariate CDF where both arguments are defined on the unit interval and 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 express the assumption that $F^a_{Y_1|Y_0,X}(y_1|y_0,x)$ is not ruled out by the experimental marginals formally as follows.

Assumption 3 ***Consistency with experimental results.*** *The conditional distribution of treated outcomes in the population of interest is consistent 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 $C \in \mathcal{C}$.

A distribution $F^a_{Y_1|Y_0,X}(y_1|y_0,x)$ obtained from Assumption 3 is defined only for y_0 on the support of $F^e_{Y_0|X}(y_0|x)$. Therefore, I will also assume that the support of $Y_0|x$ in a is a subset of the support in the experiment.

Assumption 4 ***Support of*** $Y_0|X = x$***.*** *The support of $Y_0|X = x$ in the context of interest is a subset of the support in the experiment for all values of x in the support of X in the experiment: $Supp^a(Y_0|X = x) \subseteq Supp^e(Y_0|X = x)$ for any $x \in \mathcal{X}^e$.*

Note that it is still possible to obtain identification without Assumption 3 if Y is bounded, as I discuss in more detail in Section 5.2. I state the following identification result in terms of observable quantities.

Lemma 1 *Suppose Assumptions 1, 2, 3, 4 and 5 hold. If $x \in \mathcal{X}^e$,*

$$
\begin{aligned}
&E^a[Y_1|x] \in \\
&\left[\min_{C \in \mathcal{C}} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} y_1 dC_1(F^e_{Y|T,X}(y|T=0,x), F^e_{Y|T,X}(y|T=1,x)|x) \right) dF^a_{Y|X}(y|x), \right.\\
&\left. \max_{C \in \mathcal{C}} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} y_1 dC_1(F^e_{Y|T,X}(y|T=0,x), F^e_{Y|T,X}(y|T=1,x)|x) \right) dF^a_{Y|X}(y|x) \right]
\end{aligned}
$$

and the bounds are sharp.

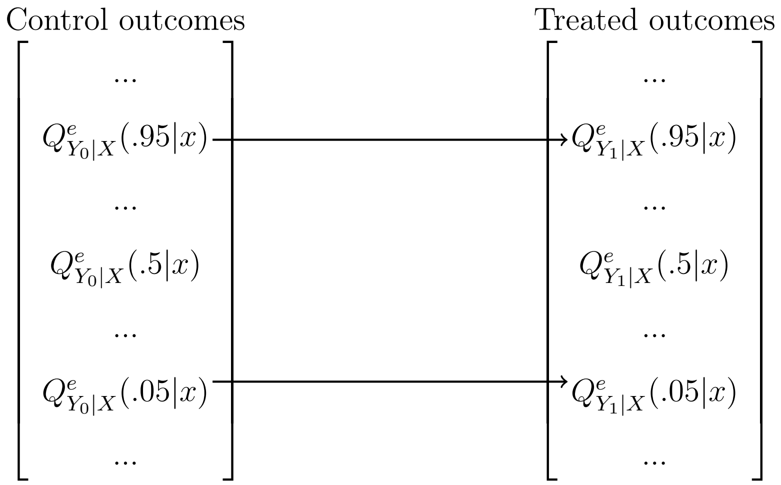
Proof: See Appendix D.1. *Q.E.D.*

4.1.2. *Discussion of Assumption 3*

To make Assumption 3 more concrete, I illustrate two examples of copula functions and show how they define a joint distribution of potential outcomes $F_{Y_0,Y_1|X}(y_0,y_1|x)$, which can in turn be used to define a conditional distribution, and discuss their relationship to individual-level treatment effects. Let $Q^e_{Y_0|X}(\alpha|x)$ denote the α-quantile of $Y_0|X$ in the experiment and $Q^e_{Y_1|X}(\alpha|x)$ the α-quantile of $Y_1|X$. 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.

A joint distribution $F_{Y_0,Y_1|X}(y_0,y_1|x)$ consistent with the experimental marginal distributions of untreated 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'

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



ranks in the outcome distributions, the treatment effects have the largest possible magnitude.

A sufficient condition for Assumption 3 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. Formally:

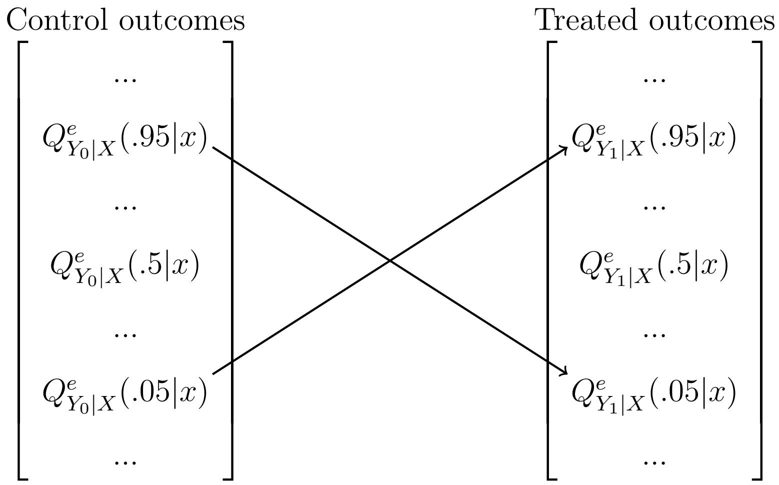
$$Y_1 \perp\!\!\!\perp D|Y_0, X \text{ or, equivalently, } U \perp\!\!\!\perp D|Y_0, X, \tag{8}$$

recalling that in conditioning on $Y_0 = y_0$ and $X = x$ we are conditioning on a function of U and X, $g(0, u, x)$, and x itself. Athey and Imbens [2006] also use this assumption to point identify average treatment effects in their generalized difference-in-differences model with discrete outcomes.

The condition in (8) is stronger than necessary. While the condition in Equation (8) states that $F^a_{Y_1|Y_0,X}(y_1|y_0, x) = F^e_{Y_1|Y_0,X}(y_1|y_0, x)$, Assumption 3 does not require equality. In terms of the example in Section (2),

$$\begin{aligned}&\{P^e(\text{enrolled w/ CCT} \mid \text{enrolled w/o CCT}),\\ &\ P^e(\text{enrolled w/ CCT} \mid \text{out of school w/o CCT})\}\end{aligned}$$

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



is known to be a point on the line in Figure (1). Assumption 3 requires that

$$\begin{aligned}&\{P^a(\text{enrolled w/ CCT} \mid \text{enrolled without CCT}),\\ &\ P^a(\text{enrolled w/ CCT} \mid \text{out of school w/o CCT})\}\end{aligned}$$

also be a point on the line in Figure (1). Equation (8) requires that the two points be the same. In practice, (8) may be easier to evaluate economically than the more general Assumption 3.

4.1.3. *Bounds on ATE^a*

Turning now to identification of ATE^a, note that the bounds from Lemma 1 are only defined for values x on the support of X in the experiment. Therefore, to produce potentially informative bounds on the unconditional expectation $E^a[Y_1]$ I assume that all values x on the support of X in the context of interest are contained in the support of X in the experiment. Formally, I impose the following.

Assumption 5 ***Support of X.*** *The support of X in the population of interest is a subset of the support in the experimental population: $\mathcal{X}^a \subseteq \mathcal{X}^e$.*

As in the case of Assumption 4, identification is still possible if Assumption 5 fails but Y is bounded. Finally, to ensure that the bounds on ATE^a are well-defined, I require existence of the expectation of Y_0 in a.

ASSUMPTION 6 ***Expectation of*** Y_0**.** *Y_0 has finite expectation: $E[|Y_0|] < \infty$.*

Under Assumptions 1, 2, 3, 4, 5, and 6, bounds on the unconditional average treatment can be recovered by weighting the minimal and maximal conditional average treatment effects by the distribution of covariates in the context of interest as stated in the following proposition.

PROPOSITION 1 *Suppose 1, 2, 3, 4, 5 and 6 hold. Then*

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

and the bounds are sharp.

PROOF: See Appendix D.2. *Q.E.D.*

4.2. *Bounds with restricted dependence*

By considering the full set of possible copulas, we consider copulas that may not be credible. 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.

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

Definition 1 ***Normalized Spearman's ρ.*** *For any two random variables U and V, Normalized Spearman's ρ is given by:*

$$\rho(U,V) = \begin{cases} \frac{Cov_C(R(U),R(V))}{Cov_M(R(U),R(V))} & \text{if } Cov_C(R(U),R(V)) \geq 0 \\ -\frac{Cov_C(R(U),R(V))}{Cov_W(R(U),R(V))} & \text{if } Cov_C(R(U),R(V)) < 0 \end{cases}$$

where $R(u) = F_U(u)$ when U is continuously distributed and $R(u) = \frac{F_U(u)+F_U(u-)}{2}$ when U takes a finite number of values, and equivalently for V. The notation $F_U(u-)$ denotes $P(U < u)$, and equivalently for V. $Cov_C(R(U),R(V))$ refers to the covariance between $R(U)$ and $R(V)$ under copula C:

$$\int \left(R(u) - \frac{1}{2}\right)\left(R(v) - \frac{1}{2}\right) dC(F_U(u), F_V(v)).$$

$Cov_M(R(U),R(V))$ is the maximum covariance possible between $R(U)$ and $R(V)$ (comonotonicity: see Appendix C.2):

$$\int \left(R(u) - \frac{1}{2}\right)\left(R(v) - \frac{1}{2}\right) d\left(\min\left\{F_U(u), F_V(v)\right\}\right).$$

$Cov_W(R(U),R(V))$ is the minimum possible covariance (countermonotonicity: see Appendix C.3):

$$\int \left(R(u) - \frac{1}{2}\right)\left(R(v) - \frac{1}{2}\right) d\left(\max\left\{F_U(u) + F_V(v) - 1, 0\right\}\right).$$

Nešlehová [2007] shows that $12Cov_C(R(U),R(V))$ is equivalent to the standard definition of Spearman's ρ found in, e.g., Nelsen [2006]:

$$\rho^{\text{standard}}(U,V) = 3\left(P\left[(U-U')(V-V') > 0\right] - P\left[(U-U')(V-V') < 0\right]\right),$$

where U' and V' are distributed independently with the same marginals as U and V. That is, $(U',V') \sim F_U(u)F_V(v)$. When U and V are continuously distributed, $Cov_M(R(U),R(V)) = 1/12$ and $Cov_W(R(U),R(V)) = -1/12$ so that the calculation is completely standard. However, when U and V take a finite number of values, $|Cov_M(R(U),R(V))|$ or $|Cov_W(R(U),R(V))|$ may not equal $1/12$ and (Genest and Nešlehová [2007]). The only difference with the standard calculation is the normalization in the discrete case.[18]

I produce bounds on $E^a[Y_1|x]$ subject to the restriction that we only consider copula functions generating dependence greater than a specified level. These are represented in the following assumption and proposition.

ASSUMPTION 7 ***Restricted dependence between*** Y_0 ***and*** Y_1***.*** *C is an element of $\mathcal{C}(\rho^L)$, the set of copula functions such that $\rho^e(Y_0, Y_1|X = x) \geq \rho^L$ where $\rho^L \in [-1, 1]$.*

PROPOSITION 2 *Suppose Assumptions 1, 2, 3, 4, 5, 6 and 7 hold. Then*

$$\begin{aligned}&E^a[Y_1|x] \in \\ &\Bigg[\min_{C\in\mathcal{C}(\rho^L)} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} y_1 dC_1(F^e_{Y|T,X}(y|T=0,x), F^e_{Y|T,X}(y|T=1,x)|x)\right) dF^a_{Y|X}(y|x), \\ &\quad \max_{C\in\mathcal{C}(\rho^L)} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} y_1 dC_1(F^e_{Y|T,X}(y|T=0,x), F^e_{Y|T,X}(y|T=1,x)|x)\right) dF^a_{Y|X}(y|x)\Bigg]\end{aligned}$$

and

$$\begin{aligned}&ATE^a \in \\ &\left[\int_{\mathcal{X}} (\min\ E^a[Y_1|x])\, dF^a_X(x) - E^a[Y], \int_{\mathcal{X}} (\max\ E^a[Y_1|x])\, dF^a_X(x) - E^a[Y]\right]\end{aligned}$$

PROOF: See Appendix D.3. *Q.E.D.*

Note that imposing $\rho^L = -1$ recovers the bounds from Proposition 1. At the opposite extreme, $\mathcal{C}(1)$ is a singleton and the bounds shrink to a point.[19] For interested readers, appendices A.2 and A.3 provide a discussion of the structure underlying the choice of ρ^L within the context of a parametric model.

5. ESTIMATION AND INFERENCE

While the identification results in Section 4 are general, for the purposes of estimation and inference 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, I will show that the optimization over the restricted space

[18] An equivalent normalization for Kendall's τ is described in Genest and Nešlehová [2007].

[19] The single $E^a[Y_1]$ identified when $\rho^L = 1$ is identical to Athey and Imbens [2006]'s counterfactual for the treated group in a generalized difference-in-difference setting with T indexing time rather than treatment. (8) must also hold when outcomes are discrete.

of copulas $\mathcal{C}(\rho^L)$ can be represented as the solution to a linear programming problem. In particular, the bounds on the average treatment 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. Galichon [2016] provides a review of applications in economics). 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 accompanying this paper. The linear programming representation is crucial for practical estimation and inference because the space of copulas is infinite-dimensional and, as mentioned in Section 2, the minimizing and maximizing arguments do not lie at the comonotonic/countermonotonic boundary points of the space.[20]

A similar representation as a continuous optimal transportation problem exists with continuous outcomes but there is no analogous tractable method to compute the solution. 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 the sample size $N \to \infty$. Exploring this possibility is left to future research. For the purposes of the present paper, I assume the following.

Assumption 8 **Finite support of the potential outcomes and covariates.** (i) Finite support of Y: Y takes finite values in $\mathcal{Y} = \{y_1, \ldots, y_j, \ldots, y_J\}$. (ii) Finite support of X: X takes finite values in a finite set $\mathcal{X} = \{x_1, \ldots, x_l, \ldots, x_L\}$..

5.1. *Linear programming representation*

To illustrate the linear programming representation, I begin by considering the case where there are no covariates X. For clarity, I refer to the supports of the potential outcomes as $\mathcal{Y}_0 = \{y_{01}, \ldots, y_{0j}, \ldots, y_{0J}\}$ and $\mathcal{Y}_1 = \{y_{11}, \ldots, y_{1j}, \ldots, y_{1J}\}$, but it should be understood that $y_{0j} = y_{1j} = y_j$ for any $j \in \{1, \ldots, J\}$. For $\rho^L \in [0, 1]$ (positive dependence, which I have argued is often most plausible),[21] 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).

[20] In contrast to the minimizing and maximizing arguments for other problems, such as that of computing the variance of Δ as in Cambanis et al. [1976].

[21] The representation for $\rho^L \in [-1, 1]$ is given in Appendix E.2.

PROPOSITION 3 *Suppose Assumptions 1, 2, 3, 4, 5, 7 and 8 hold and $\rho^L \in [0,1]$. Then the upper bound in Proposition 2, $\tau^U(\rho^L)$, is equivalent to the solution to the following linear programming problem, expressed in terms of observable quantities.*

$$\max_{\pi\in[0,1]^{J^2}} \sum_{j=1}^{J}\sum_{k=1}^{J} \frac{y_{1k}P^a(y_j)}{P^e(y_j|T=0)}\pi_{jk} - \sum_{j=1}^{J} y_{0j}P^a(y_j) \tag{9}$$

subject to

$$\sum_{k=1}^{J}\pi_{jk} = P^e(y_j|T=0)\ \forall j \in \{1,...,J\} \tag{10}$$

$$\sum_{j=1}^{J}\pi_{jk} = P^e(y_k|T=1)\ \forall k \in \{1,...,J\} \tag{11}$$

$$\begin{aligned}\rho^L + 4\sum_{j=1}^{J}\sum_{k=1}^{J}\pi_{jk}&\left(\frac{P^e(Y\le y_j|T=0)+P^e(Y\le y_{j-1}|T=0)-1}{2}\right)\\ &\times\left(\frac{P^e(Y\le y_j|T=1)+P^e(Y\le y_{j-1}|T=1)-1}{2}\right)\\ \ge \rho^L\sum_{j=1}^{J}\sum_{k=1}^{J}\Bigg[&P^e(y_j|T=0)P^e(y_k|T=1)\\ &\times\Big(\min\{P^e(Y\le y_j|T=0),P^e(Y\le y_j|T=1)\}\\ &+\min\{P^e(Y< y_j|T=0),P^e(Y< y_j|T=1)\}\\ &+\min\{P^e(Y< y_j|T=0),P^e(Y\le y_j|T=1)\}\\ &+\min\{P^e(Y\le y_j|T=0),P^e(Y< y_j|T=1)\}\Big)\Bigg]\end{aligned} \tag{12}$$

The lower bound in Proposition 2 can be obtained by replacing the max *operator with the* min *operator in the statement of the problem above.*

PROOF: See Appendix E.1. *Q.E.D.*

The choice variables of the linear programming representation are the elements of the matrix defining a possible joint distribution of Y_0 and Y_1 in context e, $\pi = \{P(y_{0j}, y_{1k})\}_{j=1,...,J}^{k=1,...,J}$. The second term in the objective function (9) 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 (10) and (11) require that the minimizing/maximizing joint distribution be consistent with the marginal outcome distributions in e. Constraint (12) enforces that Normalized Spearman's ρ (see Definition 1) applied to the potential outcomes Y_0 and Y_1 in e may not be below ρ^L. Together, constraints (10), (11) and (12) make maximizing over the elements of the joint distribution of Y_0 and Y_1 equivalent to maximizing over the restricted space of copulas, $\mathcal{C}(\rho^L)$.

The coefficients on the elements of π are $\left\{\frac{y_{1k}P^a(y_j)}{P^e(y_j|T=0)}\right\}_{j=1,\dots,J}^{k=1,\dots,J}$. Together with constraint (10), this shows the role of the distributions of untreated outcomes $\{P^a(y_j)\}_{j=1,\dots,J}$ and $\{P^e(y_j|T=0)\}_{j=1,\dots,J}$ in determining the bounds. If $P^a(y_j) = P^e(y_j|T=0)$, $\frac{P^a(y_j)}{P^e(y_j|T=0)} = 1$. If this equality applies for all j, constraint (11) implies that the counterfactual $E^a[Y_1] = E^e[Y_1]$ since

$$\sum_{j=1}^{J}\sum_{k=1}^{J} y_{1k}P(y_{0j}, y_{1k}) = \sum_{k=1}^{J} y_{1k}\sum_{j=1}^{J} P(y_{0j}, y_{1k}) = \sum_{k=1}^{J} y_{1k}P^e(y_{1k}) = E^e[Y_1].$$

The second equality follows from substituting in constraint (11).

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 (j large) when $\frac{P^a(y_j)}{P^e(y_j|T=0)}$ is large and to low values on the support of Y (j is small) when $\frac{P^a(y_j)}{P^e(y_j|T=0)}$ is small. Constraint (12) limits our ability to do so to a degree we find a priori implausible. For a concrete example, readers are encouraged to consult Appendix B, which describes one of the linear programs used in the empirical results.

5.2. *Estimation and inference with covariates*

Our object of interest is the unconditional average treatment effect in site a shown in equation 4. For the purposes of estimating ATE^a, I produce a single, unconditional linear program which results from stacking the constraints of the covariate-specific linear programs from Proposition 3 and using a covariate-weighted objective function. π is now a vector containing possible elements of the joint distribution of Y_0 and Y_1 in e conditional on each x_l (i.e., $[P^e(y_{01}, y_{11}|X = x_1), \dots P^e(y_{0j}, y_{1k}|X =$

$x_l), \ldots P^e(y_{0J}, y_{1J}|X = x_L)]')$.[22] The objective function is given by

$$\max_{\pi \in [0,1]^{J^2 \times L}} \sum_{l=1}^{L} \sum_{j=1}^{J} \sum_{k=1}^{J} y_{1k} \frac{P^a(y_j|X = x_l)P^a(x_l)}{P^e(y_j|X = x_l, T = 0)} \pi_{jkl} - \sum_{j=1}^{J} y_{0j} P^a(y_j). \tag{13}$$

Maximization is subject to constraints (10), (11), and (12) for each x_l. Let $\tau^U(\rho^L)$ now denote the solution to the unconditional problem.

Denote the population joint probability mass function as

$$p = \begin{bmatrix} p_{110a} \\ \vdots \\ p_{jltd} \\ \vdots \\ p_{JL1e} \end{bmatrix} = \begin{bmatrix} P(Y = y_1, X = x_1, T = 0, D = a) \\ \vdots \\ P(Y = y_j, X = x_l, T = t, D = d) \\ \vdots \\ P(Y = y_J, X = x_L, T = 1, D = e) \end{bmatrix}.$$

Note that $\tau^U(\rho^L)$ depends on the data through p only. Now consider an arbitrary vector $\tilde{p} \in \mathbb{R}^{J \times L \times 3}$ satisfying $\tilde{p} \geq 0$ and $\sum_{j=1}^{J} \sum_{l=1}^{L} \sum_{td \in \{0a,0e,1e\}} \tilde{p}_{jltd} = 1$. Abusing notation somewhat, I denote conditional and marginal probabilities derived from $\tilde{p}$ by indexing the probability measure P with a tilde, i.e.:

$$\tilde{P}^a(x_l) = \sum_{j=1}^{J} \tilde{p}_{jl0a} \text{ and } \tilde{P}^e(y_j|X = x_l, T = 0) = \frac{\tilde{p}_{jl0e}}{\sum_{j=1}^{J} \tilde{p}_{jl0e}}.$$

[22] Note that ρ^L could in principle be set by the researcher as a function of x. For simplicity, in the rest of the paper I consider a fixed value for ρ^L.

Let $\phi^U(\tilde{p};\rho^L)$ be the unconditional linear program applied to some $\tilde{p}$.

$$\begin{aligned}
&\phi^U(\tilde{p};\rho^L) = \\
&\max_{\left\{\begin{array}{c}\pi \in [0,1]^{J^2\times L}, \\ \overline{y} \in \{y_1, y_J\}\end{array}\right\}} \sum_{l=1}^{L}\sum_{j=1}^{J}\sum_{k=1}^{J} \psi(y_{1k}, \bar{y}, \tilde{P}^e(y_j|X=x_l, T=0))\tilde{P}^a(y_j, x_l)\pi_{jkl} \\
&\qquad\qquad\qquad\qquad - \sum_{j=1}^{J} y_{0j}\tilde{P}^a(y_j)
\end{aligned} \tag{14}$$

subject to

$$\sum_{j=1}^{J}\sum_{k=1}^{J}\pi_{jkl} = 1 \ \forall l \in \mathcal{X}^e$$

$$\tilde{P}^e(T=0, x_l)\sum_{k=1}^{J}\pi_{jkl} = \tilde{P}^e(y_j, x_l, T=0) \ \forall j \in \{1,...,J\}, x^l \in \mathcal{X}^e$$

$$\tilde{P}^e(T=1, x_l)\sum_{j=1}^{J}\pi_{jkl} = \tilde{P}^e(y_j, x_l, T=1) \ \forall k \in \{1,...,J\}, x^l \in \mathcal{X}^e$$

$$\begin{aligned}
&\rho^L + 4\sum_{j=1}^{J}\sum_{k=1}^{J}\pi_{jkl}\left(\frac{\tilde{P}^e(Y\le y_j|T=0,x_l) + \tilde{P}^e(Y \le y_{j-1}|T=0,x_l) - 1}{2}\right) \\
&\qquad\qquad \times \left(\frac{\tilde{P}^e(Y\le y_j|T=1,x_l) + \tilde{P}^e(Y \le y_{j-1}|T=1,x_l) - 1}{2}\right) \\
&\ge \rho^L \sum_{j=1}^{J}\sum_{k=1}^{J}\Big[\tilde{P}^e(y_j|T=0,x_l)\tilde{P}^e(y_k|T=1,x_l) \\
&\qquad \times \Big(\min\Big\{\tilde{P}^e(Y\le y_j|T=0,x_l), \tilde{P}^e(Y \le y_j|T=1,x_l)\Big\} \\
&\qquad\quad + \min\Big\{P^e(Y< y_j|T=0,x_l), \tilde{P}^e(Y < y_j|T=1,x_l)\Big\} \\
&\qquad\quad + \min\Big\{\tilde{P}^e(Y< y_j|T=0,x_l), \tilde{P}^e(Y \le y_j|T=1,x_l)\Big\} \\
&\qquad\quad + \min\{P^e(Y\le y_j|T=0,x_l), P^e(Y < y_j|T=1,x_l)\}\Big)\Big] \\
&\qquad\qquad \forall x^l \in \mathcal{X}^e
\end{aligned}$$

where

$$\sum_{k=1}^{J} \psi(y_{1k}, \bar{y}, \tilde{P}^e(y_j|X = x_l, T = 0)) = \begin{cases} \frac{y_{1k}}{\tilde{P}^e(y_j|X=x_l,T=0)} & \text{if } \tilde{P}^e(y_j|X = x_l, T = 0) \neq 0 \\ \bar{y} & \text{otherwise} \end{cases}.$$

$\psi(\cdot)$ imposes the bounds on the support of Y (y_1 and y_J) as the bounds on $\tilde{E}^a[Y_1|Y_0 = y_j, X = x_l]$ for x_l off the support of $\tilde{P}^e(x|T = 0)$ and y_j off the support of $\tilde{P}^e(y|T = 0, X = x_l)$. These are cases where Assumption 5 or 4 do not hold for $\tilde{p}$. [23] Use of the $\psi(\cdot)$ function makes these assumptions unnecessary.

Let $\phi : \mathbb{R}^{J^2\times L\times 3} \to \mathbb{R}^2$ denote the vector valued function

$$\phi(p; \rho^L) = \left[\phi^L(p; \rho^L), \phi^U(p; \rho^L)\right]'.$$

I estimate p using the consistent "frequency estimator" $\hat{p}$ (e.g. Li and Racine [2007]). Each element of $\hat{p}$ is given by

$$\hat{p}_{jltd} = \frac{1}{N}\sum_{i=1}^{N} 1\{Y_i = y_j, X_i = x_l, T_i = t, D_i = d\}.$$

For the purposes of consistent estimation and inference for the bounds, I will assume that $\phi(\cdot; \rho^L)$ is differentiable at p.[24]

Assumption 9 ***Estimation and inference.*** *(i) Sampling.* $Z_i = (Y_{0i}, Y_{1i}, X_i, D_i)$ *for* $i = 1, \ldots, N$ *are i.i.d. (ii) Differentiability of the linear programming representations.* $\phi(\cdot; \rho^L)$ *is differentiable at* p. *(iii) Finite bounds.* $\tau^U(\rho^L) - \tau^L(\rho^L) <$

[23] Using the bounds of the support follows Manski [1990]. More informative bounds could be obtained by imposing additional assumptions such as, for example, Monotone Treatment Response (Manski [1997]).

[24] Differentiability of $\phi(p; \rho^L)$ at p is required to use the bootstrap to perform inference for τ (Fang and Santos [2015]), which is crucial since the asymptotic distribution of the bounds cannot be characterized in closed form. Fang and Santos [2015] show that it is possible to use the bootstrap under the weaker condition that $\phi(p; \rho^L)$ is directionally differentiable in p. Inference could be performed if the directional derivative can be consistently estimated. Hong and Li [2016] (in progress) provides a promising avenue for consistent estimation of the directional derivative by numerical methods.

As an alternative, in appendix E.4, I provide a simple simulation-based approach to carrying out Bayesian inference for τ that does not require differentiability of $\phi(p; \rho^L)$ in p.

∞. *(iv) Positive, finite asymptotic variance of* $\sqrt{N}(\phi(\hat{p};\rho^L)-\phi(p;\rho^L))$. $[\underline{\sigma}^2,\underline{\sigma}^2]' \leq \nabla\phi(p;\rho^L)\Sigma\nabla\phi(p;\rho^L)' \leq [\bar{\sigma}^2,\bar{\sigma}^2]$ *where* $\underline{\sigma}^2$ *and* $\bar{\sigma}^2$ *are positive and finite and*

$$\Sigma = \begin{bmatrix} p_{110a}\left(1-p_{110a}\right) & -p_{110a}p_{210a} & \cdots \\ -p_{110a}p_{210a} & p_{210a}\left(1-p_{210a}\right) & \\ \vdots & & \ddots \end{bmatrix}.$$

Under Assumption 9, consistent estimation of the bounds by $\phi(p;\rho^L)$ follows immediately from the consistency of the frequency estimator and the continuous mapping theorem. Performing inference is more challenging, since $\phi^L(\cdot;\rho^L)$ and $\phi^U(\cdot;\rho^L)$ are not available in closed form. I propose the following bootstrap procedure to generate $(1-\alpha)$-percent confidence intervals for $\tau(\rho^L)$.

1. Generate $\{Z_1^*,\ldots,Z_N^*\}$ from $\hat{p}$.
2. Compute $\hat{p}^*$ by applying the frequency estimator to $\{Z_1^*,\ldots,Z_N^*\}$
3. Compute $[\tau^L(\rho^L)^*,\tau^U(\rho^L)^*]' = \phi(\hat{p}^*;\rho^L)$.
4. Repeat steps 1-3 B times. Compute $\hat{\sigma}^L = \sqrt{N}\times SD(\tau^L(\rho^L)^*)$, $\hat{\sigma}^U = \sqrt{N}\times SD(\tau^U(\rho^L)^*)$ and $\hat{\varrho} = Cor(\tau^L(\rho^L)^*,\tau^U(\rho^L)^*)$.
5. Form the $(1-\alpha)$-percent confidence interval for $\tau(\rho^L)$ as

$$CI_\alpha(\rho^L) = \left[\phi^L(\hat{p};\rho^L) - \frac{\hat{\sigma}^L c^L}{\sqrt{N}}, \phi^U(\hat{p};\rho^L) - \frac{\hat{\sigma}^U c^U}{\sqrt{N}}\right]$$

where $[c^L,c^U]$ solve

$$\begin{aligned}
&\min_{c^L,c^U} \hat{\sigma}^L c^L + \hat{\sigma}^U c^U \\
&\textit{subject to} \\
&P\left(-c^L \leq M_1, \hat{\varrho}M_1 \leq c^U + \frac{\sqrt{N}(\phi^U(\hat{p};\rho^L)-\phi^L(\hat{p};\rho^L))}{\hat{\sigma}^U} + M_2\sqrt{1-\hat{\varrho}^2}\right) \\
&\qquad \geq 1-\alpha \\
&P\left(-c^L - \frac{\sqrt{N}(\phi^U(\hat{p};\rho^L)-\phi^L(\hat{p};\rho^L))}{\hat{\sigma}^L} - M_2\sqrt{1-\hat{\varrho}^2} \leq \hat{\varrho}M_1, M_1 \leq c^U\right) \\
&\qquad \geq 1-\alpha
\end{aligned}$$

and M_1 and M_2 are independent standard normal random variables (Stoye [2009][25]).

PROPOSITION 4 *Suppose Assumptions 1, 2, 3, 7, 8 hold. Let $\mathcal{P}$ be the set of distributions for which Assumption 9 holds. Then,* $\lim_{N\to\infty} \inf_{P\in\mathcal{P},\tau(\rho^L)\in[\phi^L(\rho^L),\phi^U(\rho^L)]} P(\tau(\rho^L) \in CI_\alpha(\rho^L))= 1-\alpha$.

PROOF: See Appendix E.3. *Q.E.D.*

In the next sections I move on to apply the theoretical results derived so far in two empirical examples, contrasting my approach with HIM's.

6. TRANSFERS TO MEXICAN MICROENTERPRISES

My first application considers the problem of generalizing from a small scale experiment. I argue that HIM's approach is unreasonably confident in the generalizability of small experiments while the bounds approach is appropriately cautious. The setting is McKenzie and Woodruff [2008]'s (henceforth MW) experiment, carried out in 2006 (baseline Oct. 2005) in Leon, Mexico. The experiment was designed 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 receive a transfer, randomly assigned a wave in which to receive it. The transfers were valued at 1,500 pesos (about $140). Half the transfers were randomly determined to be in-kind, meaning that a member of the survey team accompanied the entrepreneur to purchase equipment or inputs of his choice. MW found that there was no difference in treatment effects by type of transfer, so I ignore the distinction here.

To ensure that the transfers be large relative to each firm's scale of operation, MW 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

[25] Stoye [2009] appendix B provides a method, which I implement, for solving this problem without simulation.

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: the 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 [1997a], 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 areas. Since sample selection already chooses a restricted set of individuals, I do not condition on any covariates in the analysis.[26]

It is important to note that the thought experiment here is not implementing a transfer program in the whole country. There would surely be large equilibrium effects from such a program. Instead, the thought experiment I consider concerns obtaining the weighted average of treatment effects for each entrepreneur in the ENAMIN sample of participating in a cash transfer program of similar scale to MW's. Each entrepreneur's weight in the average is his inverse sampling probability, provided by ENAMIN.

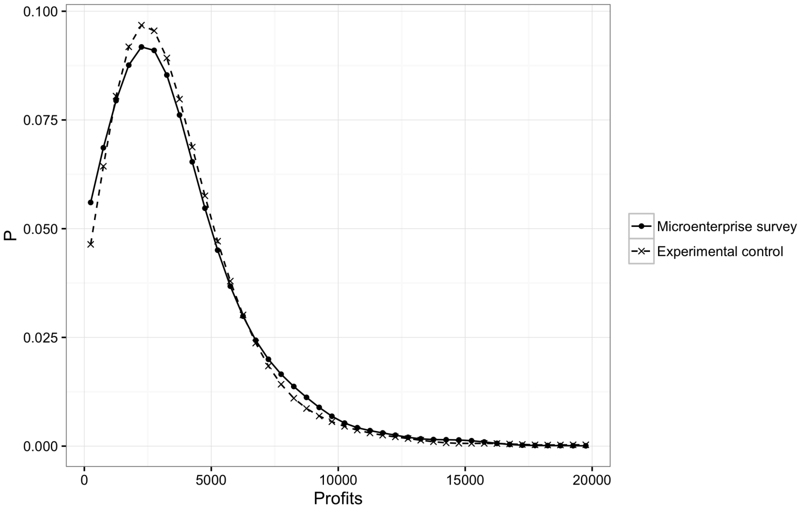
I trim profit reports of more than 20,000 pesos in both samples. This trimming keeps slightly more observations than MW who exploit the panel structure and base their trimming procedure on percentile changes in reported profits. Since ENAMIN is a cross-section, I cannot implement the MW procedure. Results 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 shows the outcome distributions in ENAMIN and the control group from the experiment, which provide one key input to computation of the bounds. Since heaping is a substantial issue in reported profits, particularly in ENAMIN, I first

[26] The conclusions I draw in this section are not sensitive to the choice of not conditioning on covariates.

smooth profits using a kernel density estimator with a Gaussian kernel and a bandwidth of 1200 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.

FIGURE 4.— Distribution of profits: McKenzie and Woodruff [2008] control group and 2012 ENAMIN



Notes: Author's calculation based on data from 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 1200 pesos before discretizing to 500 peso bins.

I now explore implications of the differences in the distributions of untreated profits for what we can learn about the average return to cash transfers for entrepreneurs in urban Mexico in 2012 using 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. Stoye [2009] 90% confidence regions (in translucent gray) are computed using 1500 bootstrap replications for each ρ^L, clustering at the firm level for the experiment[27]. The information in the plot is repeated in Table V.

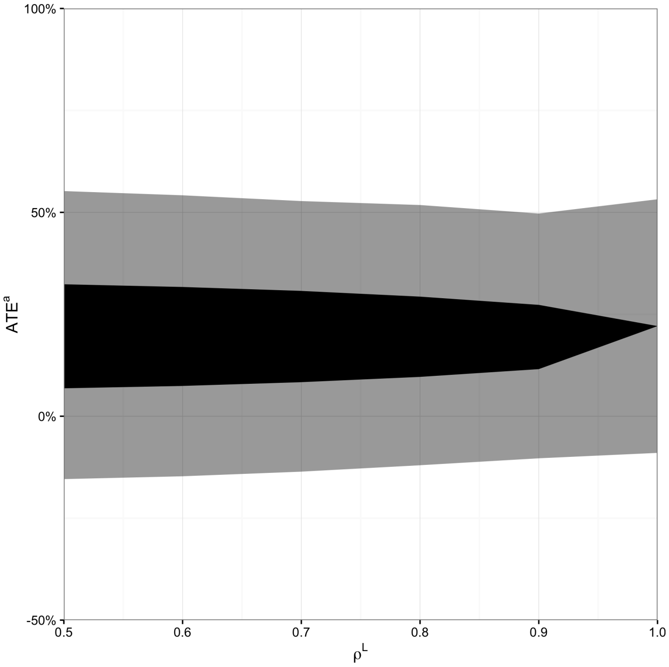
[27] This requires replacing the individual-level indicator i with a cluster-level indicator g in Assumption 9.

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 90% 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.

Recall that HIM suggest 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 test is 0.865). 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 treated 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. In sum, 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 sampling variation, which is particularly important given the small size of the Leon experiment.[28] Considering that the 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.

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

Figure 5.— Bounds on the average return to cash transfers in urban Mexico in 2012 using experimental data from McKenzie and Woodruff [2008]



Notes: For each lower bound on the dependence between profits with and without cash transfers, ρ^L, the solid black region shows the bounds on ATE^a for microentrepreneurs in urban Mexico in 2012 selected according to the criteria in McKenzie and Woodruff [2008]. The translucent gray region shows a a Stoye [2009] 90% confidence interval for ATE^a, based on 1500 block bootstrap replications, clustered at the firm level.

7. REMEDIAL EDUCATION IN INDIA

I now consider a setting where I can compare predicted average treatment effects derived using my bounds approach to experimentally estimated average treatment effects. I take advantage of Banerjee et al. [2007]'s (henceforth BCDL) evaluation of 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. I will use the results from the Vadodara experiment, combined with the control group data from Mumbai to predict the average treatment effect in Mumbai.[29]

[29] It is also possible to use Mumbai to predict Vadodara. However, this prediction will be associated with wider confidence intervals because the Mumbai sample is roughly have the size of the Vadodara

TABLE V

Bounds on the average return to cash transfers among microentreprenuers 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	0.069	0.074	0.084	0.097	0.116	0.221
ATE^a upper bound	0.324	0.317	0.307	0.293	0.273	0.221
90% Stoye [2009] confidence interval lower bound	-0.154	-0.147	-0.136	-0.120	-0.103	-0.090
90% Stoye [2009] confidence interval upper bound	0.552	0.542	0.528	0.518	0.497	0.532

Notes: based on 1500 bootstrap replications block bootstrap replications, clustering at the firm level. Author's calculations using data from McKenzie and Woodruff [2008].

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. To abstract from the problems with compliance, I will work with the sample of third graders surveyed during the first year of the experiment in Mumbai. In Vadodara, in contrast, both grade levels were represented in each of the first two years. I implicitly condition on grade level and thus do not consider the Vadodara fourth graders.

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 HIM's method 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

sample.

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 and focus on math scores.

With the exception of competency at baseline, relatively little data on students are available consistently across the two samples. Table VI shows 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 conditioning on the maximum competency level at baseline.

TABLE VI

Summary statistics for Mumbai and Vadodara samples

	(1) Vadodara	(2) Mumbai
Pre-test: maximum competency	0.29	0.54
	(0.57)	(0.79)
Male	0.29	0.47
	(0.45)	(0.50)
Number of students in class	63.94	89.51
	(27.81)	(40.23)
Observations	10049	4429

Notes: Student-level sample means and standard deviations (in parentheses) for student and classroom characteristics. Students are third graders from years 1 and 2 of the Banerjee et al. [2007] experiment in Vadodara and year 1 in Mumbai. Author's calculations based on data from Banerjee et al. [2007].

I now move to using the results from Vadodara and the Mumbai control group to predict the average treatment effect in Mumbai, comparing HIM's procedure and the bounds developed in this paper. Testing equality of the conditional distributions of maximum grade level competency in math for the two control groups is the first step in applying HIM to this example. Table VII 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 panel B shows the p-value associated with a χ^2 test of equality for each conditional distribution, treating each classroom-year pair as a cluster. The tests reject equality of the conditional distributions at the 5% level for all values of grade level competency on entering third grade. Following the HIM methodology, we would conclude that we cannot learn anything about the treatment effect in Mumbai from the Vadodara experiment: the students in the two cities are too different.

TABLE VII

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

A) 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

B) Vadodara		Post-competency					
		0	1	2	3	N	P(M = V)
Pre-competency	0	0.64	0.29	0.05	0.03	3749	<1e-4
	1	0.31	0.48	0.11	0.10	1014	<1e-4
	2	0.21	0.39	0.16	0.24	70	0.035
	3	0.49	0.37	0.01	0.12	67	<1e-4

Notes: The final column of panel B reports p-values for χ^2 tests of the equality of the conditional distributions, accounting for classroom-year level clusters.

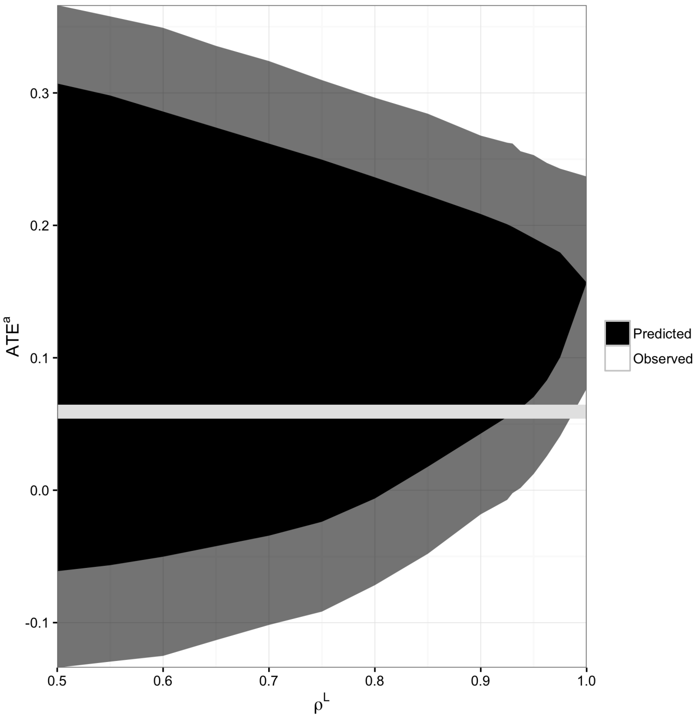
Turning to the bounds developed in this paper, Figure 6 plots bounds on the predicted values of the average effect of the remedial education program on maximum math grade level competencies in Vadodara as a function of the minimum rank correlation, ρ^L, between outcomes with and without the remedial 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 90% Stoye [2009] confidence interval, based on 1500 block bootstrap replications clustering at the classroom-year level. Table VIII replicate the key results from Figure 6 in tabular form.

A notable feature of the bounds in this example is that, in contrast with the previous example, 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 VII. A zero average treatment effect in Vadodara can only be rejected using the Mumbai results if ρ^L is greater than or equal to .937. This reflects a strong, but plausible level of dependence. BCDL's result that the ATT is about one half of the test score gain a control group child realizes from completing a year of school suggests the remedial education program is unlikely to have the effect of raising participating students' grade level competency by more than one grade. Furthermore, BCDL did not find evidence of significant effects on non-participants. Importantly, in this example I can investigate the assumption of strong positive dependence by comparing the predictions obtained from the bounds to the observed results in Mumbai. The light gray line plots the measured average effect of remedial education on maximum grade level competency in math in Mumbai. We see that the point estimate with maximum rank correlation over-predicts the average treatment effect in Mumbai. Some amount of shuffling in the potential outcome distributions induced by the remedial education treatment is needed to recover the Mumbai point estimate. This is, again, plausible given that not all students work with the remedial education teacher.

8. 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 method for considering generalizability as more than 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

Figure 6.— Bounds on the change in average grade level competency in Mumbai using experimental results from Vadodara and untreated outcomes from Mumbai



Notes: 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 treatment effect in Mumbai. The translucent gray region shows a Stoye [2009] 90% confidence interval for ATE^a, based on 1500 bootstrap replications block bootstrap replications clustering at the classroom-year level. The light gray line shows the point estimate of the average treatment effect in Mumbai, using the experimental results.

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

TABLE VIII

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

Rank correlation	0.5	0.7	0.9	0.9375	0.95	1
ATE^a lower bound	-0.061	-0.034	0.043	0.062	0.070	0.157
ATE^a upper bound	0.307	0.262	0.209	0.196	0.190	0.157
90% Stoye [2009] confidence interval lower bound	-0.134	-0.102	-0.018	0.002	0.012	0.077
90% Stoye [2009] confidence interval upper bound	0.366	0.324	0.268	0.256	0.253	0.237

Notes: based on 1500 bootstrap replications block bootstrap replications, clustering at the classroom-year level. Author's calculations using data from Banerjee et al. [2007].

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 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.

REFERENCES

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

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

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

J. Angrist and I. Fernández-Val. 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, 2013.

J. Angrist and M. Rokkanen. Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff. *Journal of the American Statistical Association*, 110(512): 1331–1344, 2015.

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

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

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

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

M. Bitler, J. Gelbach, and H. Hoynes. What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments. *American Economic Review*, 96(4):988–1012, 2006.

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

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

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

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

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

A. Deaton and N. Cartwright. Understanding and Misunderstanding Randomized Controlled Trials. *National Bureau of Economic Research Working Paper Series*, No. 22595, 2016.

R. Dehejia, C. Pop-Eleches, and C. Samii. From Local to Global: External Validity in a Fertility Natural Experiment. *NBER Working Paper*, 21459, 2015.

J. Diaz and S. Handa. 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, 2006.

F. J. DiTraglia and C. Garcia-Jimeno. A Framework for Eliciting, Incorporating, and Disciplining

Identification Beliefs in Linear Models. *Mimeo*, 2015.

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

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

Z. Fang and A. Santos. Inference on Directionally Differentiable Functions. *Mimeo*, 2015.

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

A. Galichon. *Optimal Transport Methods in Economics.* Princeton University Press, 2016.

A. Gelman, J. B. Carlin, H. S. Stern, D. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis.* 2014.

C. Genest and J. Nešlehová. A Primer on Copulas for Count Data. *ASTIN Bulletin*, 37(2):475–515, dec 2007.

F. Gerard, M. Rokkanen, and C. Rothe. Identification and Inference in Regression Discontinuity Designs with a Manipulated Running Variable. *Mimeo*, 2015.

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

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

J. J. Heckman, J. Smith, and N. Clements. Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts. *The Review of Economic Studies*, 64(4):487–535, 1997b.

H. Hong and J. Li. The Numerical Delta Method and Bootstrap. *Mimeo*, 2016.

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

G. Imbens and J. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

G. W. Imbens. Sensitivity to Exogeneity Assumptions in Program Evaluation. *American Economic Review, Papers and Proceedings*, 93(2):126–132, 2003.

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

B. Kline and E. Tamer. Bayesian inference in a class of partially identified models. *Quantitative Economics*, 7(2):1–53, 2016.

P. Kline and A. Santos. Sensitivity to Missing Data Assumptions: Theory and an Evaluation of the US Wage Structure. *Quantitative Economics*, 4(2013):231–267, 2013.

A. E. Kowalski. Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Experiments. *Working Paper*, 2016.

Q. Li and J. S. Racine. *Nonparametric Econometrics: Theory and Practice.* Princeton University Press, 2007.

C. Manski. Monotone Treatment Response. *Econometrica*, 65(6):1311–1334, 1997.

C. F. Manski. Nonparametric Bounds on Treatment Effects. *American Economic Review, Papers and Proceedings*, 80(2):829–823, 1990.

J. McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714, 2008. .

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

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

R. Meager. Understanding the Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of 7 Randomised Experiments. 2016.

M. Mesfioui and A. Tajar. On the properties of some nonparametric concordance measures in the discrete case. *Journal of Nonparametric Statistics*, 17(5):541–554, 2005.

H. R. Moon and F. Schorfheide. Bayesian and Frequentist Inference in Partially Identified Models. *Econometrica*, 80(2):755–782, 2012. .

S. Muller. Randomized Trials for Policy: A Review of the External Validity of Treatment Effects. *Mimeo*, 2014.

R. Nelsen. *An Introduction to Copulas*. Springer, second edition, 2006.

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

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

L. Pritchett and J. Sandefur. Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix. *Journal of Globalization and Development*, 4(2):161–197, 2013.

P. Rosenbaum. *Observational Studies*. Springer, New York, 2002.

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

J. Stoye. New Perspectives on Statistical Decisions Under Ambiguity. *Annual Rev. Econ. forthcoming*, 4:257–282, 2012.

E. Stuart, S. Cole, C. Bradshaw, and P. Leaf. The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386, 2011.

A. W. van der Vaart. *Asymptotic Statistics*, volume 3. 1998.

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

E. Vivalt. How Much Can We Generalize From Impact Evaluations? *Mimeo*, 2016.

APPENDIX: FOR ONLINE PUBLICATION ONLY

APPENDIX A: A PARAMETRIC EXAMPLE

In this appendix, I describe the concepts covered in Section 4 using the example of remedial education in India and a simple parametric model. Suppose students from the city of Vadodara represent context e, and students from the city of Mumbai context a. Let Y be a continuous measure of academic performance.[30] I leave the observed covariates X as a vector, but break the vector U into two components, latent skill S and parental input I. $g(\cdot)$ is a linear production function with different parameters depending on treatment status.

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

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 a, the average deviation from the common effect due to latent skill in 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.

[30] In practice, in Section 7, the measure is discrete.

A.1. *HIM*

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 the expected values of site characteristics along which treatment effects are heterogeneous. 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, (5), to hold.

A.2. $\rho^L = 1$

It is straightforward to show that assuming $\rho^L = 1$ is equivalent to assuming an individual's control and treated potential outcomes, Y_0 and Y_1, are both generated by a single unobserved characteristic of the individual (see, for example, the proof of proposition 5.16 in McNeil, Frey, and Embrechts [2005]). In the terminology of Equation (3) with a continuous outcome, U is one-dimensional and the structural functions $g(0,x,u)$ and $g(1,x,u)$ are each weakly increasing in u (Athey and Imbens [2006]).

To gain some intuition for this result, assume the parental input I is excluded from the production function so unobservables are one-dimensional[31] and the potential outcomes can be written as

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

and

$$Y_1 = \beta_1 + \beta_{1X}X + \beta_{1S}S.$$

With a one-dimensional unobservable, the way in which the distributions of observables $F^e_{X,Y}(x,y)$

[31] This is not the only way to generate 1-dimensional unobservables in the linear production function described in Section A.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:

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

in the experimental population change with treatment 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 expectation from each quantile to remove the common and x-specific structural effects,

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

and

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

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

$$\begin{aligned} \frac{Q^e_{Y_1|X}(\alpha|x) - E^e[Y_1|x]}{Q^e_{Y_0|X}(\alpha|x) - E^e[Y_0|x]} &= \frac{\beta_{1S}\left(Q^e_{S|X}(\alpha|x) - E^e[S|x]\right)}{\beta_{0S}\left(Q^e_{S|X}(\alpha|x) - E^e[S|x]\right)} \\ &= \frac{\beta_{1S}}{\beta_{0S}} \qquad (15) \end{aligned}$$

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 (15) 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.

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

A.3. $\rho^L < 1$

When we introduce multidimensional heterogeneity, we can no longer directly link differences in $F^e_{X,Y_0}(x, y_0)$ and $F^a_{X,Y_0}(x, y_0)$ to changes in the structural function $g(t, x, u)$ in response to treatment.

This is easy to see when we reintroduce independent variation in I. Consider the treatment-to-control ratio of α-quantile deviations from the x-specific subgroup expectations in e:

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

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 a specific change in the structural function.

In the empirical sections (6 and 7), I show that small deviations from 1-dimensional unobserved heterogeneity ($\rho^L = 1$) can generate non-trivial bounds, depending on the extent of difference in the conditional distributions of untreated outcomes between a and e. 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.

APPENDIX B: EXAMPLE LINEAR PROGRAM

This appendix provides an example of a linear program derived in 3 that is used in estimation in Section 7. The unconditional bounds on ATE^a are derived using $x-$conditional linear programs, averaged over the distribution of x in context a, as shown in (13). Table IX shows the choice variables and constraints (10) and (11) in the context of the remedial education in India example where Mumbai is treated as a and Vadodara as e. As is discussed in more detail in Section 7, I measure student achievement by the discrete grade level competency of third graders on completion of the grade. In Table IX, I condition on a competency level of zero in math when beginning the 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.”

Example Table X shows the coefficient on each choice variable $P(y_{0j}, y_{1k})$. 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 $Y_0 = 2$ and the lowest treatment effects to

TABLE IX

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

			Remedial ed				
			Competency on exiting grade 3				
			0	1	2	3	All control
No remedial ed	Competency	0	$P(0,0)$	$P(0,1)$	$P(0,2)$	$P(0,3)$	0.64
		1	$P(1,0)$	$P(1,1)$	$P(1,2)$	$P(1,3)$	0.29
		2	$P(2,0)$	$P(2,1)$	$P(2,2)$	$P(2,3)$	0.05
		3	$P(3,0)$	$P(3,1)$	$P(3,2)$	$P(3,3)$	0.03
		All treated	0.57	0.31	0.07	0.05	

Notes: Choice variables for the linear program described in Section 5.1. Treatment and control group probability mass functions computed using third graders entering grade 3 with grade level competency 0 in math in years 1 and 2 of the Banerjee et al. [2007] experiment in Vadodara and year 1 of the Mumbai experiment. Probability mass functions do not sum to 1 due to rounding.

individuals with $Y_0 = 1$.

Constraint (12) on the dependence between Y_0 and Y_1 in Vadodara renders us unable to do so arbitrarily.To gain some intuition for the joint distributions implied by different values of ρ^L, Table XI shows the joint distribution implied by assuming $\rho^L = 1$. In this case, the majority of the mass in the joint distribution lies on the principal diagonal. Most individuals (85%) have a treatment effect of zero, with a few individuals experiencing a positive treatment effect of at most 1 competency level.

APPENDIX C: OMITTED DEFINITIONS

C.1. *Copula*

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

1. Boundary conditions:
   (a) $C(0,u) = C(v,0) = 0 \;\forall\; u,v \in [0,1]$
   (b) $C(u,1) = u$ and $C(1,v) = v \;\forall\; u,v \in [0,1]$
2. Monotonicity condition:
   (a) $C(u,v) + C(u',v') - C(u,v') - C(u',v) \geq 0 \;\forall\; u,v,u',v'$ s.t. $u \leq u', v \leq v'$

TABLE X

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

			Remedial education			
			Competency on exiting grade 3			
			0	1	2	3
No remedial ed	Competency	0	0	1.14	2×1.14	3×1.14
		1	0	0.59	2×0.59	3×0.59
		2	0	1.48	2×1.48	3×1.48
		3	0	1.2	2×1.2	3×1.2

Notes: Objective function for the linear program described in Section 5.1. Author's calculations based on 3 with grade level competency 0 in math in years 1 and 2 of the Banerjee et al. [2007] experiment in Vadodara.

C.2. *Comonotonicity*

When two random variables U and V are comonotonic

$$F_{U,V}(u,v) = \min\left\{F_U(u), F_V(v)\right\}.$$

C.3. *Countermonotonicity*

When two random variables U and V are countermonotonic

$$F_{U,V}(u,v) = \max\left\{F_U(u) + F_V(v) - 1, 0\right\}.$$

APPENDIX D: OMITTED PROOFS FROM SECTION 4

D.1. *Proof of Lemma 1*

Under Assumption 1, $F^e_{Y|T,X}(y|T=0,x) = F^e_{Y_0|T,X}(y_0|T=0,x) = F^e_{Y_0}(y_0|x)$ where the last equality follows from T's being independent of Y_0 and X conditional on $D = e$. $F^e_{Y|T,X}(y|T=1,x) = F^e_{Y_1}(y_1|x)$ by the same argument. $F^a_{Y|X}(y|x) = F^a_{Y_0|X}(y|x)$ by Assumption 2. Under Assumption 4, the conclusion follows from Assumption 3. The bounds are sharp by construction, since each element of $\mathcal{C}$ defines a valid possible conditional distribution $F^a_{Y_1|Y_0,X}(y_1|y_0,x)$.

TABLE XI

$P^e(y_{0j}, y_{1k}|$COMPETENCY ON ENTERING THIRD GRADE $= 0)$, $\rho^L = 1$ $e =$VADODARA

			Remedial education				
			Competency on exiting grade 3				
			0	1	2	3	All Control
No remedial ed	Competency	0	0.57	0.07	0	0	0.64
		1	0	0.24	0.05	0	0.29
		2	0	0	0.02	0.03	0.05
		3	0	0	0	0.03	0.03
		All Treatment	0.57	0.31	0.07	0.05	

Notes: Joint distribution generated by the linear program described in Section 5.1 when $\rho^L = 1$. Treatment and control group probability mass functions computed using third graders entering grade 3 with grade level competency 0 in math in years 1 and 2 of the Banerjee et al. [2007] experiment in Vadodara. Probability mass functions may not sum to 1 due to rounding.

D.2. *Proof of Proposition 1*

Using Assumption 5 and Lemma 1, we obtain sharp bounds on $E^a[Y_1|x]$ for all values of x on the support of $F^a_X(x)$ in terms of observable quantities. min $E^a[Y_1|x]$ refers to the lower bound and max $E^a[Y_1|x]$ to the upper bound. Integrating the lower bounds over $F^a_X(x)$ provides a lower bound for $E^a[Y_1]$. Doing the same for the upper bounds provides the upper bound on $E^a[Y_1]$. Sharpness of the unconditional bounds follows from sharpness of the conditional bounds in Lemma 1. Under Assumption 2, $E^a[Y] = E^a[Y_0]$. $ATE^a = E^a[Y_1 - Y_0] = E^a[Y_1] - E^a[Y]$. The conclusion follows by substituting the lower and upper bounds on $E^a[Y_1]$ for $E^a[Y_1]$.

D.3. *Proof of Proposition 2*

The proof of (i) is identical to the proof of Lemma 1 with the additional restriction in Assumption 7 imposed. The proof of (ii) is identical to the proof of Proposition 1, with the bounds in (i) substituting for the bounds under Lemma 1.

APPENDIX E: OMITTED PROOFS AND RESULTS FROM SECTION 5

E.1. *Proof of Proposition 3*

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 (a) $F_{Y_0,Y_1}(y_0, \infty) = F^e_{Y_0}(y_0)$ and (b) $F_{Y_0,Y_1}(\infty, y_1) = F^e_{Y_1}(y_1)$. $Y = TY_0 + (1 - T)Y_1$ and independence of T from the potential outcomes gives us $F^e_{Y_0}(y_0) = F^e_{Y|T}(y|T = 0)$ and $F^e_{Y_1}(y_1) =$

$F^e_{Y|T}(y|T=1)$. The CDF corresponding to any joint probability mass function π satisfying constraints (10) and (11) can thus be represented as $C(F_{Y_0}(y_0), F_{Y_1}(y_1))$ for some $C \in \mathcal{C}$. Conversely, the PMF corresponding to $F_{Y_0,Y_1}(y_0,y_1) = C(F_{Y_0}(y_0), F_{Y_1}(y_1))$ for any $C \in \mathcal{C}$ can be represented as some $\pi \in [0,1]^{J^2}$ satisfying constraints (10) and (11).

Mesfioui and Tajar [2005] show that $\rho^{standard,e}(Y_0, Y_1)$ can also be written as

$$\begin{aligned}
&-3+3\sum_{j=1}^{J}\sum_{k=1}^{J}[P^e(y_{0j})P^e(y_{1k}) \\
&\qquad\qquad \times \Big(P^e(Y_0 \le y_{0j}, Y_1 \le y_{1k}) \\
&\qquad\qquad\quad + P^e(Y_0 < y_{0j}, Y_1 \le y_{1k}) \\
&\qquad\qquad\quad + P^e(Y_0 < y_{0j}, Y_1 \le y_{1k}) \\
&\qquad\qquad\quad + P^e(Y_0 \le y_{0j}, Y_1 < y_{1k})\Big)\Big]
\end{aligned}$$

By definition C.2, Spearman's ρ under comonotonicity is therefore given by

$$\begin{aligned}
&\rho^{standard,e}_M(Y_0, Y_1) = \\
&-3+3\sum_{j=1}^{J}\sum_{k=1}^{J}\Bigg[P^e(y_j|T=0)P^e(y_k|T=1) \\
&\qquad\qquad \times \Big(\min\{P^e(Y \le y_j|T=0), P^e(Y \le y_j|T=1)\} \\
&\qquad\qquad\quad + \min\{P^e(Y < y_j|T=0), P^e(Y < y_j|T=1)\} \\
&\qquad\qquad\quad + \min\{P^e(Y < y_j|T=0), P^e(Y \le y_j|T=1)\} \\
&\qquad\qquad\quad + \min\{P^e(Y \le y_j|T=0), P^e(Y < y_j|T=1)\}\Big)\Bigg].
\end{aligned}$$

For discrete outcomes

$$R^e(y_j|T=t) = \frac{P^e(Y \le y_j|T=t) + P^e(Y \le y_{j-1}|T=t)}{2}$$

so

$$\rho^{standard,e}_\pi(Y_0, Y_1) = 12Cov_C(R^e(Y|T=0), R^e(Y|T=1)),$$

using independence of the potential outcomes and treatment assignment in e to equate $R^e(Y_t)$ and $R^e(Y|T=t)$. Thus, the inequality constraint (12) restricts $\mathcal{C}$ to the set $\mathcal{C}(\rho^L)$. The objective function is $\int_{\mathbb{R}} \left(\int_{\mathbb{R}} y_1 dC_1(F^e_{Y_0}(y_0), F^e_{Y_1}(y_1))\right) dF^a_{Y_0}(y_0) - E^a[Y_0]$, which is the object to be maximized/minimized in Proposition 2.

E.2. *Linear programming representation when $\rho^L \in [-1,1]$*

Proposition 5 *Suppose Assumptions 1, 2, 3, 4, 5, 7 and 8 hold. Then the upper bound in Proposition 2, $\tau^U(\rho^L)$, is equivalent to the solution to the following linear programming problem, expressed in terms of observable quantities.*

$$\max_{\pi \in [0,1]^{J^2}} \sum_{j=1}^{J}\sum_{k=1}^{J} y_{1k} \frac{P^a(y_j)}{P^e(y_j|T=0)} \times \pi_{jk} - \sum_{j=1}^{J} y_{0j} P^a(y_j)$$

subject to

$$\sum_{j=1}^{J}\sum_{k=1}^{K} \pi_{jk} = 1$$

$$\sum_{k=1}^{J} \pi_{jk} = P^e(y_j|T=0) \ \forall j \in \{1,...,J-1\}$$

$$\sum_{j=1}^{J} \pi_{jk} = P^e(y_k|T=1) \ \forall k \in \{1,...,J-1\}$$

$$\frac{G}{sign(G)(3H-3)} \geq \rho^L$$

where

$$G = \sum_{j=1}^{J}\sum_{k=1}^{J} \pi_{jk} \left(\frac{P^e(Y \leq y_j|T=0) + P^e(Y \leq y_{j-1}|T=0) - 1}{2} \right)$$

$$\left(\frac{P^e(Y \leq y_j|T=1) + P^e(Y \leq y_{j-1}|T=1) - 1}{2} \right)$$

and

$$H = \begin{cases} \sum_{j=1}^{J}\sum_{k=1}^{J}\Bigg[P^e(y_j|T=0)P^e(y_k|T=1) \\ \quad \times \Big(\min\{P^e(Y\le y_j|T=0), P^e(Y\le y_j|T=1)\} \\ \quad + \min\{P^e(Y< y_j|T=0), P^e(Y< y_j|T=1)\} \\ \quad + \min\{P^e(Y< y_j|T=0), P^e(Y\le y_j|T=1,x_l)\} \\ \quad + \min\{P^e(Y\le y_j|T=0), P^e(Y< y_j|T=1,x_l)\}\Big)\Bigg] \\ if\ G\ge 0 \\ \sum_{j=1}^{J}\sum_{k=1}^{J}\Bigg[P^e(y_j|T=0)P^e(y_k|T=1) \\ \quad \times \Big(\max\{P^e(Y\le y_j|T=0)+P^e(Y\le y_j|T=1)-1,0\} \\ \quad + \max\{P^e(Y< y_j|T=0)+P^e(Y< y_j|T=1)-1,0\} \\ \quad + \max\{P^e(Y< y_j|T=0)+P^e(Y\le y_j|T=1)-1,0\} \\ \quad + \max\{P^e(Y\le y_j|T=0)+P^e(Y< y_j|T=1)-1,0\}\Big)\Bigg] \\ otherwise \end{cases}$$

The lower bound in Proposition 2 can be obtained by replacing the max *operator with the* min *operator in the statement of the problem above.*

Proof: The proof is identical to that of 3, allowing for both cases in Definition 1 instead of only the first. The first case is all that is needed if dependence is restricted to be positive. *Q.E.D.*

E.3. *Proof of Proposition 4*

The frequency estimator is asymptotically normal: $\sqrt{N}(\hat{p}-p)\to\mathcal{N}(0,\Sigma)$. Theorem 23.4 in van der Vaart [1998] shows that the bootstrap provides a consistent estimate of Σ. Under Assumption 9, $\sqrt{N}(\phi(\hat{p})-\phi(p))\to\mathcal{N}(p,\nabla\phi(p)\Sigma\nabla\phi(p)')$ by the delta method. Theorem 23.5 in van der Vaart [1998] shows that the bootstrap consistently estimates this distribution and thus its functionals, σ^L, σ^U and ϱ. Thus, Assumption 1(i) of Stoye [2009] is satisfied. Assumption 1(ii) of Stoye [2009] holds by Assumption (9) (iii) and (iv). Recall that $\phi^L(p;\rho^L)$ is obtained by switching the max operator to min in equation (14), subject to the same set of constraints. $P(\phi^L(\cdot;\rho^L)\le\phi^U(\cdot;\rho^L))=1$ is thus satisfied by construction for any argument, so Assumption 3 of Stoye [2009] is satisfied. The conclusion follows from Proposition 2 of Stoye [2009].

E.4. *Bayesian inference for ATE^a*

I follow Moon and Schorfheide [2012]'s recommended approach to Bayesian inference in partially identified models.[32] I first perform inference for the point-identified, reduced form parameter vector

[32] See DiTraglia and Garcia-Jimeno [2015] for another recent application.

p. The sample vector

$$C = \begin{bmatrix} \sum_{i=1}^{N} 1\{Y_i = y_1, X_i = x_1, T_i = 0, D_i = a\} \\ \vdots \\ \sum_{i=1}^{N} 1\{Y_i = y_j, X_i = x_l, T_i = t, D_i = d\} \\ \vdots \\ \sum_{i=1}^{N} 1\{Y_i = y_J, X_i = x_L, T_i = 1, D_i = e\} \end{bmatrix}$$

constitutes a draw from a multinomial distribution with parameters N and p. The likelihood is characterized by the relationship

$$P(c|p) \propto \prod_{j=1}^{J} \prod_{l=1}^{L} \prod_{td \in \{0a, 0e, 1e\}} p_{jltd}^{c_{jltd}},$$

and an uninformative conjugate prior[33] for p by a Dirichlet distribution with the parameter vector given by a vector of ones of dimension $J \times L \times 4$ (see Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin [2014]). The posterior distribution is Dirichlet with parameter vector C.

We can use the linear program in (14) to produce bounds for any draw p^b from the posterior. To produce a posterior distribution for ATE^a, I specify that ATE^a is distributed uniformly on $[\tau^L(p^b; \rho^L), \tau^U(p^b; \rho^L)]$ (Moon and Schorfheide [2012]).[34] The distribution of ATE^a can be characterized by simulation, according to the following steps.

1. Draw p^b from a Dirichlet(C) distribution B times.
2. For each b, compute $\tau^L(p^b; \rho^L)$ and $\tau^U(p^b; \rho^L)$ using the linear program described in (13).[35]
3. Draw τ^{bs} from a Uniform$(\tau^L(p^b; \rho^L), \tau^U(p^b; \rho^L))$ distribution S times.

The posterior distribution function is approximated by

$$F_{BS}(ATE^a|c) = \frac{1}{BS} \sum_{b=1}^{B} \sum_{s=1}^{S} 1\{\tau^{bs} \leq ATE^a\},$$

which converges to the posterior, $F(ATE^a|c)$, uniformly almost surely as $BS \to \infty$.

[33] Kline and Tamer [2016] show that specifying a prior is not necessary for conducting inference on the identified set itself. However, the object of policy interest is not the identified set, but ATE^a itself.

[34] Other priors may be more relevant depending on the loss function employed. See Stoye [2012].

[35] Note that Assumption 9 (ii) is not required here.