Identifying Distributional Effects of Teachers and Peers in Nonseparable Models

Sebastian Calonico†
University of Michigan
JOB MARKET PAPER
January 8, 2014

Abstract

This paper looks at the effects of teacher and peer characteristics on student achievement in the STAR Project conducted in Tennessee in the late 1980s. As in standard linear models, the proposed approach considers two types of unobservables: school-specific effects and idiosyncratic disturbances. It generalizes previous empirical research by allowing both effects to enter the structural function nonseparably. No functional form assumptions are needed for identification. Instead, it uses an exchangeability condition in the way that covariates affect the distribution of the school-specific effects. The model permits nonparametric distributional and counterfactual analysis of heterogeneous effects: it extends policy analysis beyond marginal or discrete changes to consider distributional effects originating from a counterfactual change in the distribution of characteristics of classrooms, peers and teachers. Also, these impacts can be analyzed on any feature of the distribution of student achievement, such as quantiles and inequality measures. The empirical analysis looks at the effects of class size, teacher experience and gender composition of the classroom on test scores. Findings suggest that nonseparable heterogeneity is an important source of individual-level variation in academic performance. The impact of class size is considerably larger using my approach: students in smaller classes benefit about 0.3 standard deviations, compared to a 0.16 effect obtained using a standard linear model. Also, teacher experience has a stronger, nonlinear impact. Still, the distributional analysis suggests that these gains are hard to achieve when facing resource constraints.

Keywords: Early Childhood, STAR Project, Unobserved Heterogeneity, Counterfactual Distributions, Nonseparable Models.

JEL Classification: C14, C23, C33, I21.

—I am especially thankful to Matias Cattaneo and Jeffrey Smith for their continuous support and advice. I am also indebted to Richard Hirth and Dan Ackerberg for thoughtful feedback and helpful discussions.

†Department of Economics. Email: calonico@umich.edu. Web: http://www.umich.edu/~calonico/
1 Introduction

The effects of educational inputs such as class size, teaching quality and school resources on student achievement have long been studied in the economic literature. In a highly influential work, Hanushek (1986) concludes that the literature does not provide strong evidence of a consistent relationship between school resources and student performance. A positive effect of school inputs, particularly teaching quality, has instead been highlighted in more recent work. For example, Card and Krueger (1992, 1996) find a positive relationship between school resources and student achievement, showing that both low pupil-teacher ratios and high quality school systems lead to higher future earnings for students. Mixed conclusions have been reached on the effect of class size on student performance: while some studies conclude that small classes do not improve student achievement (e.g., Hanushek (2003), Hoxby (2000)), others find evidence of a positive impact (e.g., Krueger (1999), Krueger and Whitmore (2001), Angrist and Lavy (1999)).

These contrasting results have usually been attributed to econometric problems that make it difficult to recover the causal effect of educational inputs on student performance, especially those related to omitted variable bias and reverse causality. Early studies have often relied on data in which the allocation of students to classes was not the result of an exogenous assignment. For example, schools might assign less able students to smaller classes, or better teachers to larger ones. In other cases, the allocation of students to classes is not exogenous due to parent decisions, for example parents more concerned about the education of their children may choose schools with a smaller class size or more experienced teachers.

With the aim to provide more reliable estimates, recent studies have relied on controlled randomized experiments or natural experiments. Most notably, a number of works have used data from the STAR Project, conducted in Tennessee from 1985-89. This was a large-scale, longitudinal experimental study of reduced class size, where students and teachers were randomly allocated to different class sizes. It motivated a large body of research on the effects of different classroom characteristics (not only class size, but also other factors such as teacher experience) on student performance both in the short and long-run. Most of the studies conclude that smaller classes increase student achievement, even after controlling for school fixed effects and teacher characteristics (e.g., Krueger (1999), Krueger and Whitmore (2001), Nye, Konstantopoulos, and Hedges (2004)).

Besides relying on an experimental setting, a common feature of all these studies is that they are based on linear models, where we can account for unobserved school heterogeneity by including school dummies, which can be handled with standard linear panel methods such as within-group transformations. I propose to extend this approach by considering a more general, nonseparable model that does not impose any functional form or parametric assumptions. In particular, no additivity or monotonicity assumptions are required for identification. Using the STAR Project data, I look at the influence of teacher experience, class size and gender composition of the classroom on student performance on standardized tests. The main motivation behind this paper is based on several important limitations of the standard model, which usually assumes a linear specification of the form:
where $Y_{ics}$ is a measure of achievement (e.g., kindergarten test scores) for student $i$ assigned to classroom $c$ in school $s$, $P_{cs}$ is a classroom characteristic (e.g., class size or teacher experience). Finally, $Z_{ics}'$ accounts for other student and teacher characteristics that affect student performance. The main interest is on the parameter $\beta$, e.g., the impact of class size or teacher experience on test scores. The model also includes two types of unobserved heterogeneity: a school fixed effect $\mu_s$ and an individual specific component, $\varepsilon_{ics}$.

The model presents several limitations, mostly derived from the linearity assumption. This is crucial for identification, since the school fixed effect $\mu_s$ is usually differenced out. Besides being subject to model misspecification, this imposes important limitations for the analysis of heterogeneity in terms of the relationship between the impact of the covariates and $\mu_s$. The marginal effect of $P_{cs}$ on $Y_{ics}$ (e.g., a marginal change in the gender composition of the classroom) is: $\partial Y_{ics}/\partial P_{cs} = \beta$, or $\beta (p_1 - p_0)$ for a discrete change (e.g., a reduction in class size). Given the additively separable assumption, the model fails to capture the heterogeneity that comes from $\mu_s$, such as unobserved school characteristics or other attributes that affect all members of the school but cannot be observed in the data. Additionally, it rules out the possibility of heterogeneous treatment effects, which is often an important feature of the data (e.g., Heckman, Smith, and Clements (1997) and Djebbari and Smith (2008)). For example, it does not account for the possibility that the effect of the same reduction in class size could be larger for schools with better reputations.

Additionally, the linearity assumption limits other aspects of the analysis. First, regarding what features of the distribution of student achievement are considered. Usually the analysis focuses on average outcomes; that is, the effects on the conditional expectation of $Y_{ics}$. Heterogeneity is most commonly accounted for by looking at subgroup impacts based on demographic characteristics. A small number of studies also look at quantile treatment effects (e.g., Jackson and Page (2013)), but the inclusion of fixed-effects is not straightforward as they can no longer be differenced out and additional assumptions are required. Second, it also limits the type of policy analysis that can be conducted. In a linear model, $\beta$ measures the impact of a marginal or discrete change in $P_{cs}$. This might not be very informative in terms of policy implementation. For example, if we care about reallocations of individuals across groups as opposed to infeasible increases in the population. These reallocations can be characterized by obeying a particular feasibility constraint that should be accounted for. For example, one might be interested on the distributional effects of a policy that reduces gender segregation in the classroom, while keeping the total number of students of both genders fixed.

I propose a general method trying to account for these limitations. First, I use a nonseparable model of the form:

$$ Y_{ics} = m(\mu_s, P_{cs}, Z_{ics}, \varepsilon_{ics}) $$

where the $m(\cdot)$ function is assumed unknown and left completely unspecified. Nonseparable mod-
els have been widely studied in the econometrics literature (e.g., Matzkin (2007, 2013)). In the model I employ, no additivity or monotonicity assumptions are required for identification of certain parameters related to the effect of teacher and peer characteristics on student achievement. In addition, both $\mu_s$ and $\varepsilon_{ics}$ can be of any dimension and interact with the covariates in general ways, in particular allowing for heterogeneous treatment effects. For example, for a marginal change:

$$\frac{\partial Y_{ics}}{\partial P_{cs}} = \frac{\partial m(p, z, \mu, \varepsilon)}{\partial p}.$$ 

Now the effect is allowed to be different even for students with the same observed characteristics. The same is true for a discrete change: $m(\mu_s, p_1, Z_{ics}, \varepsilon_{ics}) - m(\mu_s, p_0, Z_{ics}, \varepsilon_{ics})$.

The method I propose goes beyond the effect of marginal and discrete changes of the covariates. In particular, I extend the analysis to consider counterfactual changes in the marginal distribution of $P_{cs}$, and their effect on the unconditional distribution of the outcome. For example, my method allows me to study the effects of a policy that modifies the distribution of teacher experience by reducing the number of less experienced teachers. Additionally, the impact of these counterfactual policies can be identified on any feature of the distribution of $Y_{ics}$. This includes, for example, the mean, quantiles and other functionals such as inequality measures.

The identification strategy is based on a control function approach that disentangles the direct effect of $P_{cs}$ on $Y_{ics}$ by keeping the distribution of unobservables fixed. The main concern is the possible correlation between the school fixed effects $\mu_s$ and the policy variable $P_{cs}$. This is handled via an exchangeability assumption (Altonji and Matzkin (2005)) on the conditional distribution of $\mu_s$ given observable class characteristics, which imposes that they cannot be ordered in a particular way in each school.

Using data from the STAR Project, I look at the influence of class size, teacher experience and gender composition of the class on test scores. My findings suggest that nonseparable heterogeneity is an important source of individual-level variation in the academic performance of kindergarten students. Using the nonseparable model, the impact of class size is considerably larger: students in smaller classes benefit about 0.3 standard deviations, compared to a 0.16 effect obtained with a linear model. Also, teacher experience has a stronger, nonlinear impact: students assigned to more experienced teachers perform better in standardized test scores, and the gain increases with years of experience. Still, conducting a counterfactual distributional analysis I find that these gains in student performance are hard to achieve when facing resource constraints. For example, I find that a policy that reduces the size of some classes while keeping the number of students and teachers fixed generates a lower impact on test scores.

The remainder of the paper is organized as follows. Section 2 describes the STAR Project and discusses some of the related literature. Identification, estimation and implementation of the proposed model are discussed in Section 3. Section 4 presents my empirical findings, comparing them with previous approaches. Finally, I discuss the main conclusions in Section 5. Proofs to all theorems; tables and graphs; and a simulation study are included in the appendix.
2 STAR Project

The Student/Teacher Achievement Ratio (STAR) Project was conducted in Tennessee during 1985-89. It was a large-scale, 4-year, longitudinal, experimental study of reduced class size, where students and teachers were randomly assigned to classes of different sizes. It included 79 schools from inner-city, rural, urban, and suburban locations, and over 6,000 students per grade level (for students in kindergarten and grades 1 to 3).

A large body of research has looked at the relationship between class size and student performance in nonexperimental settings, but the STAR Project was the first large-scale experiment to address this issue. In the absence of an experiment, the effect of a policy may be confounded by other observed or unobserved factors that may be correlated with the policy. In this case, the experiment only manipulated class size and did not provide additional teacher training, new curriculum, or any other intervention.

In the original implementation of the experiment, students were to remain with the same randomly assigned class type from kindergarten through the end of the third grade. In practice, however, there were several deviations. Students who entered a participating school after the first year of the program were added to the experiment and randomly assigned to a class type. There was a substantial number of new entrants: 45 percent of eventual participants entered after kindergarten, due in part because, at the time, kindergarten was not required in Tennessee. A relatively large fraction of students exited the STAR Project schools (45 percent of overall participants) due to school moves, grade retention, or grade skipping. In addition, in response to parental concerns about fairness to students, all students in regular and regular-aide classes were randomized again in the first grade. Finally, a smaller number of students (about 10 percent of participants) were moved from one type of class to another in a nonrandom manner. Most of these moves reportedly were due to student misbehavior and not typically the result of parental requests to move their child to a small class. Still, if families felt that their child would be better served by attending smaller classes (or were upset that their child was randomly assigned to a regular class), this might yield a differential attrition rate or better attendance rate by class type. For these reasons, in this paper I focus only on the sample of students who entered the project in kindergarten.

Ideally one would check randomization with a pretest to ensure that there are no measurable differences in the dependent variable by class type before the program began. Unfortunately, no baseline survey was collected. Still, several authors (e.g., Krueger (1999)) investigated this issue by comparing student characteristics that are related to student achievement but cannot be manipulated in response to treatment, such as student race, gender and age, finding no systematic differences in observable characteristics across class type. Another drawback is that initial random assignment was not recorded, but instead initial enrollment was measured. This could be a concern if, for example, parents successfully lobbied for a class change in the days between class assignments and the beginning of school. Krueger (1999) presented evidence from a subset of the data suggesting that this was very unlikely. Finally, it is also important that teachers were randomly assigned. If the most effective teachers were disproportionately placed with small (or regular) classes, then the
class-size effect would pick up this effect as well. Based on the data available, Krueger (1999) finds no observed within-school differences across observed characteristics of teachers, such as race, gender, experience level, or highest level of education.

In terms of external validity, there are a few aspects of the sample that may limit the validity of generalizing the STAR Project findings to other settings. In order to be eligible to participate in the program, schools were required to have a minimum-size cohort of fifty-seven students, enough to sustain both a regular and a regular-aide classroom of twenty-two students and one small class of fifteen students. As a result, the schools that participated were about 25 percent larger, on average, than other Tennessee schools. Because of requirements imposed by the legislature for geographic diversity, schools in inner cities were overrepresented, and the students included were more economically disadvantaged and more likely to be African-American than those in the state overall. Even though the percentage of non-white participants closely mirrors the percentage in the United States overall (33 versus 31 percent), there were very few Hispanic and Asian students in Tennessee at the time compared to the rest of the nation. Finally, average school spending in Tennessee was about three-fourths of the nationwide average, and teachers were less likely to have a master’s degree. Krueger (1999) and Schanzenbach (2006) provide additional details on the implementation of the programs.

Numerous studies have used the STAR Project to show that class size, teacher quality, and peer characteristics have significant (both in a statistical sense and in magnitude) causal impacts on test scores (e.g., Schanzenbach (2006)). In addition, there is a large literature about their long-term impacts. Krueger (1999) finds that, on average, performance on standardized tests increases by four percentile points the first year students attend small classes, and this advantage expands by about one percentile point per year in subsequent years. The effects are larger for minority students and those on free/reduced lunch programs. Other studies have shown that students assigned to small classes are more likely to complete high school (Finn, Gerber, and Boyd-Zaharias (2005)), take the SAT or ACT college entrance exams and less likely to be arrested (Krueger and Whitmore (2001)). Chetty et al. (2011) analyze the long-term impacts of the STAR Project on college attendance, earnings, retirement savings, home ownership, and marriage by linking the original data to administrative data from tax returns. More recently, Dynarski, Hyman, and Schanzenbach (2011) also find that students in small classes are more likely to enroll and complete college. However, very few studies look at distributional impacts beyond subgroup analysis in the STAR Project. Jackson and Page (2013) find heterogeneity across achievement quantiles, with the largest test score gains being at the top of the achievement distribution.

I contribute to this literature by providing a nonparametric, distributional evaluation of the impact of teachers, peers, and other class attributes on student performance in standardized tests. By looking at the effect of some classmate characteristics, the approach also relates to the peer effects literature, in particular to “contextual” effects models as described in Manski (1993).\footnote{See, e.g., Durlauf (2004) and Sacerdote (2011) for reviews, and Bramoullé, Djebbari, and Fortin (2009), Boucher, Bramoullé, Djebbari, and Fortin (2012) for recent empirical applications.}
3 The Model

The performance on a standardized test for student $i$ in classroom $c$ from school $s$, $Y_{ics}$, is assumed to be generated through the nonseparable model:

$$Y_{ics} = m(\mu_s, P_{cs}, Z_{ics}, \varepsilon_{ics})$$ (2)

for $i = 1, \cdots, I_{cs}$, $c = 1, \cdots, C_s$ and $s = 1, \cdots, S$, where $(P_{cs}, Z_{ics})$ is a $d_x$-dimensional vector of covariates, with $P_{cs}$ a scalar that could be any classroom, teacher or peer characteristic whose effect on test scores we want to study. This flexible specification allows for general types of interaction between $\mu_s$ and $P_{cs}$, since no assumption is made on the functional form of $m(\cdot)$. This could be either a structural equation that describes the causal relationship between the variables, or a reduced form equation from a general structural system.

I consider several features of the relationship between test scores $Y_{ics}$ and the class characteristic $P_{cs}$. First, I look at two parameters that have a straightforward interpretation and can be compared to the $\beta$ coefficient from the linear model (1): a Weighted Average Derivative Function for continuous variables $P_{cs}$, and a Discrete Changes Function that evaluates $P_{cs}$ at different points, useful for discrete random variables such as class size or teacher experience. Finally, I introduce a Counterfactual Distribution Function that measures the effect of general changes in the distribution of $P_{cs}$ on the marginal distribution of $Y_{ics}$.

Definition 1 When $m(\mu, p, z, \varepsilon)$ is differentiable in $p$ and $p$ is continuously distributed, the Local Average Response is:

$$\delta_{ics}(p, z) = \int \frac{\partial m(p, z, \mu, \varepsilon)}{\partial p} dF_{\mu_s, \varepsilon_{ics}|P_{cs}, Z_{ics}}(\mu, \varepsilon|p, z)$$ (3)

where $F_{\mu_s, \varepsilon_{ics}|P_{cs}, Z_{ics}}(\mu, \varepsilon|p, z)$ is the distribution function of $(\mu_s, \varepsilon_{ics})$ conditional on $P_{cs} = p$ and $Z_{ics} = z$.

That is, $\delta_{ics}(p, z)$ is the partial effect of $P_{cs}$ on the expected value of $Y_{ics}$, evaluated at given values of $P_{cs}$ and $Z_{ics}$, averaged over the distribution of unobservables. For example, this could measure the average effect on test scores of a marginal change in the gender composition of the class, when the proportion of females is 0.5. Note that, without assumptions on the dependence relationship among students, classrooms and schools, the average derivative function is indexed by $(i, c, s)$. I will discuss this issue in more detail in Section 3.2.

One concern with the Local Average Response is that most common nonparametric estimators of (3) will exhibit low rates of convergence, especially when $Z_{ics}$ is high-dimensional. Besides, in some contexts the objective of the analysis is not to predict the entire derivative curve of a conditional expectation function at each data point. Instead, we might be interested in an average version of (3) over all values of $(P_{cs}, Z_{ics})$. Then, I also consider Weighted Average Derivatives.
Definition 2 The Weighted Average Derivative Function is

\[ \delta_{ics}^\omega = E \left[ \frac{\partial m(p, z, \mu, \varepsilon)}{\partial p} \omega(p, z) \right] \]  

(4)

where \( \omega(p, z) \) is some specified weight function.

The weighted average derivative function is a well known parameter, and its identification and estimation have been extensively studied in the nonparametric and semiparametric literature, in part because it is possible to construct nonparametric estimators of (4) that attain parametric convergence rates. Certain regularity conditions are usually required on the regression functions, the data and the weights \( \omega \), such as compact support on \( (Pcs, Zics) \), bounded higher moments of \( Yics \) and derivatives of the \( m(\cdot) \) function. See, e.g., Cattaneo, Crump and Janson (2010, 2013a, 2013b) and references therein for a more detailed discussion. Also, see Newey and Stoker (1993) for efficiency results for average derivative estimators. I discuss implementations issues of (4) in section 3.2.

For discrete variables such as class size or years of teacher experience, we are instead interested in finite changes rather than infinitesimal ones. In this case, we can use:

Definition 3 A Discrete Changes Function

\[ \Delta_{ics}(p'', p') = \int [m(p'', z, \mu, \varepsilon) - m(p', z, \mu, \varepsilon)] dF_{Zics,\mu,s,\varepsilon,ics | Pcs}(z, \mu, \varepsilon | p') \]  

(5)

is defined for a change between \( Pcs = p'' \) to \( Pcs = p' \).

Finally, I discuss a Counterfactual Distributions Function that measures the effect of a counterfactual change in the distribution of \( Pcs \) on the marginal distribution of \( Yics \). The parameter of interest in this case is:

Definition 4 The Counterfactual Distribution Function

\[ F_{Y*ics}(y) \equiv P[m(\mu_s, P*_{ics}, Zics, \varepsilon_{ics}) \leq y] \]  

(6)

is the marginal distribution of \( Y_{ics}^* \equiv m(\mu_s, P*_{ics}, Zics, \varepsilon_{ics}) \) obtained by evaluating the function \( m(\cdot) \) at values \( P*_{cs} \), where the distribution of \( Pcs \) changed from \( F_{Pcs} \) to \( F_{P*cs} \).

Now the research question is: how would the unconditional distribution of student performance \( Yics \) change if a policy maker could exogenously shift the values of \( Pcs \) to some \( P*_{cs} \), i.e., what is the difference between the distribution of \( Y_{ics} \) and that of the counterfactual random variable \( Y_{ics}^* \). This new distribution can be obtained in different ways. For example, it could come from a transformation of the original random variable (such as a policy that consists of reducing the number of less experienced teachers), or from a different population (e.g. the distribution of teacher experience from another state, different demographic groups or time periods, etc.). This type of counterfactual analysis have been extensively studied in other areas of economics (see, e.g., Fortin, Lemieux, and Firpo (2011)).
Note also that $P_{cs}^*$ can be dependent or independent of $P_{cs}$. Both cases can be considered in the same framework, with only different implications in terms of implementation, as I discuss in the next section. Finally, the proposed approach also works for the case in which $P_{cs}$ is different for each student, $P_{ics}$. For example, Dee (2004) looks at the effect of attending a class with teachers of similar characteristics as the students. Also, it does not have to consist only of classroom means (e.g. average age of peers). Glewwe (1997) points out the limitations associated with using the mean of peer characteristics without taking into account their overall distribution and how failure to do so can yield seriously misleading results. For example, $P_{ics} = (I^{-1} \sum_{i=1}^{I} Z_{ics}^{1-\zeta})^{1-\zeta}$ accounts for other characteristics of the distribution according to the parameter $\zeta$.

In all cases, the object of interest is the distribution $F_{Y_{ics}^*}$ and how it compares to $F_{Y_{ics}}$. The difference between them is called a distributional policy effect. In general, this approach can be used to conduct inference on $F_{Y_{ics}^*}$ as a whole, its moments and quantiles, or some functionals of it, such as inequality measures. The next section discusses the assumptions required for identification of all three parameters.

### 3.1 Identification

The main identification concern is the possible correlation between $P_{cs}$ and $\mu_s$. There are basically two ways in which $P_{cs}$ affects $Y_{ics}$: a direct effect through the function $m(\cdot)$, and an indirect effect through the distribution of $\mu_s$. In a linear approach, one could simply remove the effect of $\mu_s$ by differencing it out. This is no longer possible in a nonseparable model, so additional assumptions are required. In particular, I assume the existence of a vector $V_s$ including information at the school level, such that the school fixed effect is independent of $P_{cs}$ once we condition on $V_s$. Then, we can isolate the direct effect of $P_{cs}$ on $Y_{ics}$. For example, one approach would be to rely on a selection on observables type of assumption, and then construct the $V_s$ vector with a rich set of school characteristics. Instead, I employ an exchangeability assumption that fits well in the context of the STAR Project. I develop this idea in more detail in the next section. For identification purposes, it is only required that the vector $V_s$ satisfies:

**Assumption 1** $\mu_s \perp (Z_{ics}, P_{cs}) | V_s$

Note that $V_s$ has the role of a control function, and there could be many choices of $V_s$ satisfying this condition, each implying different restrictions on the model (see, e.g., Matzkin (2007, 2013) and references therein). I discuss a particular strategy to construct $V_s$ in Section 3.1.1. The main idea is that, by controlling for $V_s$, I can isolated the direct effec of $P_{cs}$ on $Y_{ics}$ without the influence of $\mu_s$.

The next assumption refers to the individual specific heterogeneity, $\epsilon_{ics}$. In the context of the STAR Project, the random allocation of teachers and students to classroom ensures that $\epsilon_{ics} \perp P_{cs}$. For example, let $\epsilon_{ics}$ represent family involvement in their children’s education. Given random assignment of students and teachers into classrooms, it is expected that this student-specific char-
acteristic is uncorrelated with the class size assigned to the student. More generally, I allow the independence of \( \varepsilon_{ics} \) and \( P_{cs} \) to be conditional:

**Assumption 2** \( \varepsilon_{ics} \perp P_{cs}|(\mu_s, Z_{ics}, V_s) \)

The first two assumptions are sufficient for identification of the average derivative and discrete changes functions, and have been previously proposed in a similar context by Altonji and Matzkin (2005). The result is given in Theorems 1 and 2.

**Theorem 1 (Identification of Weighted Average Derivatives)** Under (A.1)-(A.2):

\[
\delta_{ics}^\omega = \mathbb{E}\left[ \frac{\partial \mathbb{E}(Y_{ics}|P_{cs} = p, Z_{ics} = z, V_s = v)}{\partial P_{cs}} \omega(p, z) \right]
\]

which also requires \( \mathbb{E}\left[ | \partial \mathbb{E}(Y_{ics}|P_{cs} = p, Z_{ics} = z, V_s = v) / \partial P_{cs} | \right] < \infty. \)

The idea behind the identification of \( \delta_{ics}^\omega \) is straightforward. First, we calculate the partial effect of \( P_{cs} \) on \( Y_{ics} \) holding \( V_s \) constant. This holds the distribution of unobservables constant. Second, we compute the conditional distribution of \( V_s|P_{cs}, Z_{ics} \) and recover \( \delta_{ics}(p, z) \) by integrating out \( V_s \). From this result, identification of the density weighted average derivatives follows directly by integrating over the joint distribution of \( (P_{cs}, Z_{ics}) \). A similar identification strategy can be used for the discrete changes function:

**Theorem 2 (Identification of Discrete Changes)** Under (A.1)-(A.2):

\[
\Delta_{ics}(p'', p') = \left[ \int \mathbb{E}(Y_{ics}|P_{cs} = p'', Z_{ics} = z, V_s = v) \, dF_{Z_{ics}, V_s|P_{cs}}(z, v|p') \right] - \mathbb{E}(Y_{ics}|P_{cs} = p')
\]

The next two assumptions are specific to the counterfactual distribution analysis, and concern the type of distributions that can be considered for \( P_{cs}^* \). A general assumption regarding the relationship between the counterfactual random variable and the unobservables is:

**Assumption 3** \( \mu_s \perp (Z_{ics}, P_{cs}, P_{cs}^*)|V_s \) and \( \varepsilon_{ics} \perp (P_{cs}, P_{cs}^*)|(\mu_s, Z_{ics}, V_s) \)

This would be satisfied, for example, if \( P_{cs}^* \) is originated from a transformation of \( P_{cs} \), \( P_{cs}^* = \Gamma(P_{cs}) \). Finally, I also impose a common support condition:

**Assumption 4** \( sup(P_{cs}^*) \subseteq sup(P_{cs}) \)

This is required to achieve nonparametric identification due to the inability to extrapolate beyond the range observed in the data, restricting the policy experiments that can be considered to ones for which there is already some experience in the data. It could still be possible to give meaningful bounds on the counterfactual distribution when \( P_{cs}^* \) is allowed to take values outside of the support of \( P_{cs} \) with moderate probability. Identification follows:
Theorem 3 (Identification of Counterfactual Distributions) Under (A.1) – (A.4):

\[ F_{Y^*_ics}(y) = \mathbb{E}[F_{Y_{ics}|P_{cs},Z_{ics},V_s}(y,P^*_{cs},Z_{ics},V_s)] \]  

(9)

That is, we can identify the unobserved marginal distribution of \( Y^*_ics \) by first computing the conditional CDF of \( Y_{ics} \) given \( P_{cs},Z_{ics} \) and \( V_s \). As in the previous cases, holding \( V_s \) holds the distribution of the fixed effects constant. Finally, the unconditional distribution can be obtained by integrating over the distribution of \( (P^*_{cs},Z_{ics},V_s) \). Also note that, from these results, functionals such as quantiles and inequality measures are also identified.

Remark 1 In all cases, an implicit assumption is the nonparametric identification of the regression function \( \mathbb{E}(Y_{ics}|P_{cs},Z_{ics},V_s) \) for values of \( (P_{cs},Z_{ics},V_s) \) for which the conditional density of \( (V_s,Z_{ics}) \) given \( P_{cs} \) is positive. I discuss this issue in more detail in Section 3.1.1, after introducing the choice of \( V_s \).

The methodological contribution of this paper is to extend some previous results from the literature on distributional counterfactual effects and on nonseparable models, especially some recent contributions in a panel data context. Rothe (2010, 2012) proposes a nonparametric procedure to analyze counterfactual distributions using nonseparable models, but without accounting for group-invariant fixed effects. Recently, Chernozhukov, Fernández-Val, and Melly (2013) consider policy interventions that correspond to either changes in the distribution of covariates, or changes in the conditional distribution of the outcome given covariates, or both. This paper also contributes to research on nonseparable models, especially to some recent work for panel data. For a review of earlier contributions in a cross-sectional context, see Matzkin (2007, 2013). One important difference is that these models usually focus on the identification of local average structural derivatives (LASD), for which additional assumptions are required. For example, monotonicity on the unobservables (e.g., Altonji and Matzkin (2005), Evdokimov (2010)). Su, Hoderlein, and White (2010) discuss several limitations of this assumption. Alternatively, other papers restrict the analysis to a subpopulation for which the covariates do not change over time (e.g., Hoderlein and White (2012), Chernozhukov, Fernández-Val, Hahn, and Newey (2013)). None of these assumptions are employed in the identification results in Theorems 1 to 3.

3.1.1 Exchangeability

To find a vector \( V_s \) satisfying (A.1) I use the notion of exchangeability, first introduced in the nonseparable models literature by Altonji and Matzkin (2005). Graham, Imbens, and Ridder (2010) also use an exchangeability assumption but at the student level, and in a different model, to study segregation by gender in kindergarten. Without loss of generality, I assume that there are two classrooms for each school, \( C = 2 \). In the present context, exchangeability is defined as:

Definition 5 The conditional distribution of \( \mu_s \) given \( (X_{1s},X_{2s}) \) is exchangeable in \( (X_{1s},X_{2s}) \) if \( F_{\mu_s|X_{1s},X_{2s}}(\mu|x_1,x_2) = F_{\mu_s|X_{1s},X_{2s}}(\mu|x_2,x_1) \), where \( X_{cs} = (P_{cs},Z_{cs}) \) is a vector of classroom
characteristics.

This means that the conditional distribution $F_{\mu s | X_{1s}, X_{2s}}(\mu | x_1, x_2)$ is invariant to permutations of its arguments. That is, the subscript $c$ is uninformative, and the information that $(X_{1s}, X_{2s})$ provides is independent of the order in which the elements are collected. This does not imply that there are no classroom effects, but that classrooms cannot be ordered in a particular way for all schools. The order could be natural in other contexts, such as in panel data (if, for example, we were following classroom over time). In that context, it rules out any type of dynamic behavior. Here I assume that there are no a priori reasons for the first classroom to have a different effect than the second one on the distribution of $\mu s$. An important restriction is the implication that the same equality holds for any subset of the data. For $C = 2$ this implies $F_{\mu s | X_{1s}}(\mu | x_1) = F_{\mu s | X_{2s}}(\mu | x_2)$ which means that observable characteristics of each classroom provide the same information regarding the distribution of the school fixed effects. As opposed to a conditional independence assumption where we need to find a rich enough set of variables to include in $V_s$ such that Assumption 1 is satisfied, the exchangeability assumption can hold for any of the elements in $X_{cs}$.

**Example 1** Let $\mu s \in \{H, L\}$, so schools can be either low or high quality type. Also, $X_{cs} \in \{1, 2\}$ represents years of teacher experience. One possible scenario where the exchangeability assumption would not hold is when high quality schools always assign more experienced teachers to classroom 1. Then, it might be that $P(\mu s = H | X_{1s} = 1, X_{2s} = 2) = 0$ while $P(\mu s = H | X_{1s} = 2, X_{2s} = 1) > 0$.

**Example 2** Suppose classrooms are numbered by the extend of external distraction (e.g., nice views out the window, external noise, broken chairs, etc). Then, teacher assignment should be invariant to these choices.

**Example 3** We can also gain some intuition by looking at types of distributional assumptions that would lead to exchangeability. Let $P_{cs} = P_s + \bar{P}_{cs}$ and $\mu s = \theta P_s + \bar{\mu}_s$, where $P_s \sim N(0, 1)$, $\bar{P}_{cs} \sim N(0, 1)$, and $\bar{\mu}_s \sim N(0, 1)$, all i.i.d. Then, $F_{\mu s | P_{1s}, P_{2s}}(\mu | p_1, p_2) = F_{\mu s | V_s}(\mu | p_1 + p_2)$ by properties of the normal distribution.

To sum up, exchangeability is a reasonable assumption in the context of the STAR Project, where teachers and students were randomly assigned to each classroom, which, even though classrooms were of different size, this is observed and can be accounted for by including class size as one of the elements of $X_{cs}$. This assumption can be used to construct a vector $V_s$ satisfying (A.1). First, the Fundamental Theorem of Symmetric Polynomial Functions states that any symmetric polynomial can be written in terms of elementary symmetric functions. Together with the Weierstrass approximation theorem, this implies that if $F_{\mu s | X_{1s}, X_{2s}}(\mu | x_1, x_2)$ is exchangeable in $(X_{1s}, X_{2s})$, it can be approximated arbitrarily close by a function of the form:

$$F_{\mu s | X_{1s}, X_{2s}}(\mu | x_1, x_2) = F_{\mu s | V_s}(\mu | v)$$  \hspace{1cm} (10)

Note that i.i.d. $\Rightarrow$ exchangeability $\Rightarrow$ stationarity $\Rightarrow$ identically distributed.
where $V_s \equiv (V_{s1}^1, V_{s2}^2)$ are elementary symmetric polynomials of $(X_{1s}, X_{2s})$. For example, when $X_{cs}$ is a scalar, $V_s = (X_{1s} + X_{2s}, X_{1s}X_{2s})$. Finally, note that (10) implies (A.1): $\mu_s \perp X_{cs}\mid V_s$.

As mentioned before, an implicit assumption for the results in Theorems 1-3 is the nonparametric identification of $\mathbb{E}[Y_{ics}\mid P_{cs}, Z_{ics}, V_s]$ for values of $(P_{cs}, Z_{ics}, V_s)$ for which the conditional density of $(V_s, Z_{ics})$ given $P_{cs}$ is positive. This requires enough variability on $P_{cs}\mid Z_{ics}, V_s$. Altonji and Matzkin (2005) discuss several alternatives to guarantee this condition, but most of them require imposing additional restrictions to the model. Instead, I propose exploiting the additional variability arising from the inclusion of more elements in the vector of classroom characteristics $X_{cs}$ (which could also be at the student level). I use the results for elementary symmetric functions for vectors developed in Weyl (1939). For example, let $X_s = (X_{1s}, X_{2s})$, with $X_{cs} = (P_{cs}, Z_{cs})$. Then, $V_s = (P_{1s}Z_{1s} + P_{2s}Z_{2s}, P_{1s}Z_{1s}P_{2s}Z_{2s})$.

### 3.2 Implementation

The estimation of all parameters of interest can be based on the identification results in Theorems 1 to 3. First, I impose assumptions on the dependence across students, classrooms and schools.

**Assumption 5 (a)** The sequence $\{Y_s, X_s\}_{s=1}^S$ is i.i.d., where $(Y_s, X_s)$ is a vector including information for all classrooms and students in school $s$: $Y_s \equiv (Y_{1s}, \cdots, Y_{Cs})$ and $X_s \equiv ((P_{1s}, Z_{1s}), \cdots, (P_{Cs}, Z_{Cs}))$, where $Y_{cs} \equiv (Y_{1cs}, \cdots, Y_{Ics})$ and $Z_{cs} \equiv (Z_{1cs}, \cdots, Z_{Ics})$. **(b)** Additionally, I assume that observations are identically distributed across $i = 1, \cdots, I_{cs}$ and $c = 1, \cdots, C_s$.

Assumption 5 (b) arises naturally in a context of exchangeability of classrooms across schools, as stated in Definition 5. Then, we can omit the indexes $(i, c, s)$ from the left hand side of (4), (5) and (6). To estimate the Counterfactual Distribution Function Estimator I use:

$$
\hat{F}_{Y^*}(y) = \frac{1}{S} \sum_{s=1}^S \frac{C_s}{C_s} \left( \frac{1}{I_{cs}} \sum_{i=1}^{I_{cs}} \hat{F}_{Y_{ics}\mid P_{cs}, Z_{ics}, V_s}(y; P_{cs}^*, Z_{ics}, V_s) \right)
$$

(11)

where $\hat{F}_{Y_{ics}\mid P_{cs}, Z_{ics}, V_s}(y; p, z, v)$ is an estimator of the conditional distribution of $Y_{ics}$ given $(P_{cs} = p, Z_{ics} = z, V_s = v)$. This conditional distribution function can be estimated by either a semi-parametric approach (e.g., inverting a conditional quantile model), or by fully nonparametric methods (e.g., a kernel CDF estimator). I choose a semi-parametric approach with a prominent role in empirical work: a Distribution Regression Model. This approach was first developed in Foresi and Paracchi (1992) and recently extended by Chernozhukov, Fernández-Val, and Melly (2013). The estimator of the conditional CDF is:

$$
\hat{F}_{Y_{ics}\mid P_{cs}, Z_{ics}, V_s}(y; p, z, v) = \Lambda \left( \rho(p, z, v) \hat{\theta}(y) \right)
$$

(12)
that a linear combination can approximate the regression function 

\[ \hat{\theta}(y) = \arg \max_{b \in \mathbb{R}^{d_x + d_t}} \sum_{s=1}^{S} \sum_{c=1}^{C_s} \sum_{i=1}^{I_{cs}} (1 \{Y_{ics} \leq y\} \ln [\Lambda (\rho(P_{cs}, Z_{ics}, V_s)' b)] + 1 \{Y_{ics} > y\} \ln [1 - \Lambda (\rho(P_{cs}, Z_{ics}, V_s)' b)]) \]  

with \((P_{cs}, Z_{ics}) \in \mathbb{R}^{d_x}, V_s \in \mathbb{R}^{d_v}\) and \(\rho(\cdot)\) is a vector of transformations (polynomials or b-splines). The distribution regression model is flexible in the sense that, for any given link function \(\Lambda\), we can approximate the conditional distribution function arbitrarily well by using a rich enough \(\rho(\cdot)\). It generalizes location regression by allowing the slope coefficients \(\beta(y)\) to depend on the threshold index \(y\). As opposed to other semiparametric alternatives (such as a quantile regression model), it does not require smoothness of the conditional density, since the approximation is done pointwise in the threshold \(y\), and thus handles continuous, discrete, or mixed \(Y\) without any special adjustment (see Chernozhukov, Fernández-Val, and Melly (2013) for further details). In summary, the counterfactual distributions are estimated using the following algorithm:

**Algorithm 1 (Estimation of Counterfactual Distributions)**

(i) Apply the distribution regression model (12) to obtain estimates \(\hat{F}_{Y_{ics} | P_{cs}, Z_{ics}, V_s}\) using data on \((Y_{ics}, P_{cs}, Z_{ics}, V_s)\) for \(i = 1, \ldots, I_{cs}, c = 1, \ldots, C_s\) and \(s = 1, \ldots, S\). (ii) Compute the unconditional distribution \(\hat{F}_Y(y)\) in (11) by evaluating the estimator in (i) on \((y, P_{cs}^*, Z_{ics}, V_s)\) and taking the average over students, classrooms and schools.

Next, to estimate average derivatives I employ a simple unweighted version

\[ \delta = \mathbb{E} \left[ \frac{\partial m(p, z, \mu, \varepsilon)}{\partial p} \right] \]  

that can be compared with \(\beta\) in (1). An estimator of Average Derivatives is:

\[ \hat{\delta} = \frac{1}{S} \sum_{s=1}^{S} \sum_{c=1}^{C_s} \left( \frac{1}{I_{cs}} \sum_{i=1}^{I_{cs}} \frac{\partial \mathbb{E}(Y_{ics} | P_{cs}, Z_{ics}, V_s)}{\partial P_{cs}} \right) \]  

where \(\partial \mathbb{E}(Y_{ics} | P_{cs}, Z_{ics}, V_s) / \partial P_{cs}\) is a nonparametric series estimators of the first derivative of the regression function \(\mathbb{E}(Y_{ics} | P_{cs}, Z_{ics}, V_s)\) with respect to \(P_{cs}\). Let \(\bar{X}_{ics} = (P_{cs}, Z_{ics}, V_s)\) and \(g_0(x) = \mathbb{E}[Y_{ics} | \bar{X}_{ics} = x]\) denote the true conditional expectation. Series methods approximate the unknown \(g_0(x)\) with a flexible parametric function \(g_K(x, \vartheta)\) where \(\vartheta\) is an unknown coefficient vector. The integer \(K\) is the dimension of \(\vartheta\) and indexes the complexity of the approximation. Let \(\pi^K(x) = (\pi_{1K}(x), \ldots, \pi_{KK}(x))'\) be a vector of approximating (basis) functions having the property that a linear combination can approximate \(g_0(x)\), then a Linear Series Estimator of \(g_0(x)\) takes the form:

\[ \hat{g}(x) = \pi^K(x)' \hat{\vartheta} \]
with \( \hat{\vartheta} = (\Pi^0 \Pi)^{-1} \Pi^0 Y \), where \( Y \) is the vector containing all values of \( Y_{ics} \) and \( \Pi \) is a vector including \( \pi^K(x) \) for all values of \( X_{ics} \). From (16), we can construct a series estimator of the derivative of the regression function as:

\[
\frac{\partial g(x)}{\partial x} = \frac{\partial \pi^K(x)}{\partial x} \hat{\vartheta} \tag{17}
\]

Two popular choices for series estimators are power series and splines. Let \( r \) be the dimension of \( x \), and \( \lambda = (\lambda_1, \ldots, \lambda_r)^t \) a vector of nonnegative integers, i.e. a multi-index, with norm \( |\lambda| = \sum_{j=1}^r \lambda_j \), and let \( z^\lambda \equiv \prod_{j=1}^r (z_j)^{\lambda_j} \). For a sequence \( (\lambda(k))_{k=1}^\infty \) of distinct such vectors, a power series approximation has \( \pi_{k,K}(x) = x^{\lambda(k)} \). Regression splines are linear combinations of functions that are smooth piecewise polynomials of a given order with fixed knots (joint points). For additional references on series estimators, see, e.g., Newey (1997), Chen (2007), and Cattaneo and Farrell (2013). To sum up, the Average Derivative Estimator can be implemented using the following procedure:

**Algorithm 2 (Estimation of Density Weighted Average Derivatives)**

(i) Estimate the derivative of the regression function \( E(Y_{ics}|P_{cs}, Z_{ics}, V_s) \) using the series estimator (17) with data on \( (Y_{ics}, P_{cs}, Z_{ics}, V_s) \) for \( i = 1, \ldots, I_{ics}, c = 1, \ldots, C_s \) and \( s = 1, \ldots, S \). (ii) Compute (15) averaging over students, classrooms and schools.

Finally, for the Discrete Changes Estimator:

\[
\hat{\Delta}(p'', p') = \int \hat{E}(Y_{ics}|P_{cs} = p'', Z_{ics} = z, V_s = v) \hat{f}_{Z_{ics}, V_s|P_{cs}}(z, v|p') dz dv - \hat{E}(Y_{ics}|P_{cs} = p') \tag{18}
\]

where \( \hat{E}(Y_{ics}|P_{cs} = p'', Z_{ics} = z, V_s = v) \) and \( \hat{E}(Y_{ics}|P_{cs} = p') \) are nonparametric series estimators of the regression function, and \( \hat{f}_{Z_{ics}, V_s|P_{cs}}(z, v|p) \) is a nonparametric kernel estimator for the joint density of \( (V_s, Z_{ics}) \) conditional on \( P_{cs} = p \), given by:

\[
\hat{f}_{Z_{ics}, V_s|P_{cs}}(z, v|p) = \frac{S^{-1} \sum_{s=1}^S C_s^{-1} \sum_{c=1}^{C_s} I_{ics}^{-1} \sum_{k=1}^{I_{ics}} K_{h_0}(P_{cs} - p) K_{h_1}(Z_{ics} - z) K_{h_2}(V_s - v)}{S^{-1} \sum_{s=1}^S C_s^{-1} \sum_{c=1}^{C_s} I_{ics}^{-1} \sum_{k=1}^{I_{ics}} K_{h_0}(P_{cs} - p)} \tag{19}
\]

with \( K_{h}(u) = h^{-1} K(u/h) \) and \((h_0, h_1, h_2)\) the bandwidths associated with \((P_{cs}, Z_{ics}, V_s)\). The bandwidths can be obtained via cross-validation methods proposed in Fan and Yim (2004) and Hall, Racine, and Li (2004). The procedure can be summarized by:

**Algorithm 3 (Estimation of Discrete Changes)**

(i) Use the series estimator (16) to estimate the regression function \( E(Y_{ics}|P_{cs} = p'', Z_{ics} = z, V_s = v) \) using data on \( (Y_{ics}, P_{cs}, Z_{ics}, V_s) \) for \( i = 1, \ldots, I_{ics}, c = 1, \cdots, C_s \) and \( s = 1, \ldots, S \). (ii) Estimate the conditional density of \( (V_s, Z_{ics}) \) given \( P_{cs} \) using (19). (iii) Integrate the conditional expectation in (i) with respect to the density in (ii) to obtain the first term in (8). This can be done, for example, using Monte Carlo integration. (iv) Use the series estimator (16) to estimate the regression function \( E(Y_{ics}|P_{cs} = p) \) and substract it from (iii) to obtain the final estimator.
In all cases, I construct uniform confidence bands via nonparametric bootstrap with clusters at the school level.

4 Empirical Results

The primary Project STAR data consist of 11,601 students who participated for at least one year. It includes students demographic information, school and class identifiers, school and teacher information, experimental condition (class type) and achievement test scores. Achievement data continued to be collected through high school. This includes achievement test scores in grades 4 to 8, teachers’ ratings of student behavior in grades 4 and 8, students’ self-report of school engagement and peer effects in grade 8, mathematics, science, and foreign language courses taken in high school, SAT/ACT participation and scores and graduation/dropout information. The study also collected data on 1780 students in grades 1 to 3 in 21 comparison schools, matched with STAR schools but not participating in the experiment.

Table 1 presents the summary statistics of the final sample used for the empirical analysis. It consists of 5,781 students who started the project in kindergarten and have valid information on demographic characteristics and test scores. Females constitute 48 percent of the sample, average age at the beginning of 1985 is 4.7 years, 32 percent of the students are black, and 47 percent are eligible for the free/reduced lunch program. Mean years of teacher experience is 9.2, and classes have on average 19 students.

For all the analyses conducted below, the outcome $Y_{ics}$ consists of standardized (to have mean zero and standard deviation one) SAT scores averaged across subjects (math, reading, listening and word study skills), as is common in the literature. The policy variables $P_{ics}$ are class size, teacher experience and proportion of females in the classroom. In all the models, I include additional control variables $Z_{ics}$ accounting for student gender, race, age and free/reduced lunch status. I start by comparing the results obtained using a standard, linear fixed effects panel data model with the nonparametric estimates of density weighted average derivatives (4) and the discrete changes estimators (5). Tables 3 to 6 present these results for each policy variable, and for different implementations of the nonparametric series estimators.

The empirical analysis also includes a counterfactual study of the effect of different policies related to class size, teacher experience and proportion of females in the classroom, using the Counterfactual Distribution function (6). Results are presented in Figures 1 to 8. For each figure, the left panel (Panel (a)) displays the original distribution of the policy variable and the resulting counterfactual change. The right one (Panel (b)) reports Quantile Treatment Effects (QTE) for that policy, together with uniformly valid confidence intervals. The QTE estimator measures the impact of the counterfactual policy for quantile $q$ as the difference in outcomes between the $q$–th student in the countefactual (treatment) distribution and the $q$–th student in the original (control) one. For instance, we can compare the median test score for the students in the original distribution and subtract from it the median test score for the students under the counterfactual policy to estimate the QTE.
the effect at the median of the achievement distribution. Note that this estimator will not identify the impact of the policy on a particular student who would have been at the \( q \)-th percentile in the absence of the policy. This interpretation is only appropriate if the policy causes no re-ordering of achievement ranks within the distribution. As discussed in Heckman, Smith, and Clements (1997) and more recently in Djebbari and Smith (2008), quantile treatment effects are simply differences between the treatment and control distributions, and recovering quantiles of the treatment effect distribution requires specific assumptions about the joint distribution of outcomes in the treatment and control states (such as perfect positive or perfect negative dependence). Nevertheless, the QTE estimator provides substantial information about treatment effects heterogeneity.

4.1 Class Size

Empirical analyses in the STAR Project usually conclude that smaller classes increase student achievement, even after controlling for school fixed effects and teacher characteristics. Table 3 presents estimates of the effect of moving from a class size of 22 to 15 students (the median class sizes for regular and small classrooms in the STAR Project, respectively). In the linear model, this effect is simply \( \hat{\beta}(22 - 15) \), where \( \beta \) is the coefficient associated with class size in the linear model (1). Instead, the estimated effect using the discrete changes estimator (5) is:

\[
\hat{\Delta}(15, 22) = \int \hat{E}(Y_{ics} | P_{cs} = 15, Z_{ics} = z, V_s = v) \hat{f}_{Z_{ics}, V_s | P_{cs}}(z, v | 22) \, dz \, dv - \hat{E}(Y_{ics} | P_{cs} = 22)
\]

Using a fixed-effects linear panel data model (Column 1), I find that students benefit about 0.16 standard deviations from assignment to a small class. This is in line with previous findings. However, the nonparametric estimates are actually larger and statistically significant for all the specifications. For example, the effect of assignment to a small class is between 0.3 and 0.43 standard deviations using power series estimators of the regression function. This suggests that unobserved heterogeneity at the school level plays an important role on the impact of class size on student performance. In turn, it could help explain previous findings of different impact estimates for demographic groups, as in Schanzenbach (2006). Still, now the effect is more general since unobservable factor are also accounted for. For instance, it is possible that the positive effect of a smaller class size is larger in a school with a better management.

Next, I extend the analysis by looking at distributional effects of class size policies using the counterfactual distribution estimator (6). The goal here is to see what policies regarding class size would be able to generate the gains in students’ performance obtained in the previous analysis. I start with a policy that simply reduces class size in the largest classroom (Policy 1):

\[
P_{cs}^* = \begin{cases} 
  P_{cs} & \text{if } P_{cs} \leq 21 \\
  P_{cs} - 5 & \text{if } P_{cs} > 21
\end{cases}
\]

The QTE results are presented in Figure 1. The effect is positive throughout the achievement distribution, but heterogeneous, with the biggest impacts among children with scores near the top of
the distribution. For example, the test score of a student at the 90\textsuperscript{th} percentile in the counterfactual distribution is almost a third of a standard deviation higher than the test score of a 90\textsuperscript{th} percentile student in the original distribution, whereas the difference at the 10\textsuperscript{th} percentile of the distribution are less than a tenth of a standard deviation. These estimates are in line with, although lower in magnitude, the estimates comparing small versus large class sizes in Jackson and Page (2013). High achievers could benefit more from smaller classes if, for instance, teachers in small classes are better able to identify high achievers and use instructional approaches that work well for them.

One potential concern with the previous policy is that it does not take into account feasibility or resource constraints. For example, in order to reduce class size according to Policy 1, the school would need to hire additional teachers or to enroll some students in additional classrooms. For this reason, I also look at Policy 2 which keeps the number of students fixed by constructing the counterfactual variable as:

\[ P_{cs}^* = \begin{cases} 
P_{cs} + 5 & \text{if } P_{cs} \leq 21 \\
P_{cs} - 5 & \text{if } P_{cs} > 21 
\end{cases} \]

From the results in Figure 2 we can see that the QTE estimates are close to zero and statistically insignificant over the achievement distribution. This can be explained by the improvements in performance by the students in smaller classes being compensated by the worsening in performance by those in larger classes.

Overall, I conclude that the estimated impacts of class size are larger when using a nonseparable model, highlighting the relevance of accounting for heterogeneous treatment effects. The distributional analysis of counterfactual changes also suggests that the impacts are much smaller once we take into account feasibility constrains.

4.2 Teacher Experience

Teacher experience has traditionally been an important component of teacher policies in the U.S. school systems. Although recent debates have focused on the development and use of more direct measures of teacher performance (e.g., value-added models, standards-based evaluation), teacher experience continues to play a dominant role in most human resource policies. The underlying assumption is that experience promotes effectiveness and that experience gained over time enhances the knowledge, skills, and productivity of teachers.

Experience is among the most commonly studied teacher characteristic. Several studies find that the impact of experience is strongest during the first few years of teaching: on average, brand-new teachers are less effective than those with some experience (e.g., Rockoff (2004), Rivkin, Hanushek, and Kain (2005), Clotfelter, Ladd, and Vigdor (2007, 2010), Kane, Rockoff, and Staiger (2008), Harris and Sass (2011)), but the greatest productivity gains occur during their first few years on the job, after which their performance tends to level off. Empirical evidence suggests that, on average, students with teachers in their fifth year of teaching score between 5 and 15 percent of a standard deviation higher than students with teachers in their first year on the job (Atteberry, Loeb, and Wyckoff (2013)). There is also evidence that this effect is stronger than the effects of
other observable teacher characteristics such as advanced degrees, teacher licensure tests scores, and even class size (e.g., Clotfelter, Ladd, and Vigdor (2007), Rivkin, Hanushek, and Kain (2005)).

In the STAR Project, Krueger (1999) finds small but positive effects of teacher experience, with a peak at about twenty years: students in classes where the teacher has twenty years of experience tend to score about three percentile points higher than those in classes where the teacher has zero experience, all else being equal. As a whole, however, he concludes that measured teacher characteristics explain relatively little of student achievement on standardized tests. More recently, Chetty et al. (2011) finds that students randomly assigned to more experienced kindergarten teachers have higher test scores, with the effect being roughly linear. Schanzenbach (2006) analyze the indirect effect of teacher experience by comparing the performance of students in small versus regular class size with teachers of different experience, finding considerable heterogeneity of impacts: students with more experienced teachers show large, statistically significant gains from reduced class size. In contrast, students who have a teacher with fewer than five years of experience show smaller and often not statistically significant gains from small classes. Recently, Mueller (2013) finds that this pattern exists at all deciles of the achievement distribution, but is less pronounced at lower deciles.

In Tables 4 and 5, I compare the estimates from a linear panel data model (with a quadratic term for the experience variable) to the discrete changes estimator. The goal is to study nonlinear effects of teacher experience on student performance by comparing students with teachers of different years of experience. First, I look at a change from 5 to 10 years of experience (Table 4). Then, I consider a change from 10 to 15 years in Table 5. Using the nonseparable model (2), the effects are:

\[
\hat{\Delta} (10, 05) = \int \hat{E} (Y_{ics} | P_{cs} = 10, Z_{ics} = z, V_s = v) \hat{f}_{Z_{ics}, V_s | P_{cs}} (z, v | 05) \, dz \, dv - \hat{E} (Y_{ics} | P_{cs} = 05)
\]

\[
\hat{\Delta} (15, 10) = \int \hat{E} (Y_{ics} | P_{cs} = 15, Z_{ics} = z, V_s = v) \hat{f}_{Z_{ics}, V_s | P_{cs}} (z, v | 10) \, dz \, dv - \hat{E} (Y_{ics} | P_{cs} = 10)
\]

From Table 4, we can see that the estimate for \(\Delta (10, 05)\) is around 0.13. That is, students with a teacher with 10 years of experience perform 0.13 standard deviations higher than those with a teacher with only 5 years of experience. We can also see that the point estimates are precisely estimated. This effect is considerably larger than the one obtained using a quadratic model (0.027). Also, in Table 5, changing teacher experience from 10 to 15 years yields a larger impact (between 0.23 and 0.55), which is also larger than the one obtained from a quadratic model (0.063). That is, using the nonseparable model we find evidence of a strong and nonlinear effect of teacher experience. Again, this points to the importance of accounting for unobserved factors in the impact of teacher experience on student performance.

Finally, I look at distributional effects. Policy 1 consists of a general increase of five years in teacher experience,

\[P_{cs}^* = P_{cs} + 5\]

From Figure 3, we can see that the effect is positive for all percentiles, but is slightly larger for
those at the top quantiles. More importantly, the magnitude of the effect is considerably lower than what is obtained with the discrete changes estimator. To examine whether this could be due to a differential effect coming from teachers of different experience, the next two policies look at the differential effect of teacher experience for relatively new versus more experienced teachers. Policy 2 only affects classrooms with less experienced teachers:

$$P^*_{cs} = \begin{cases} 
P_{cs} & \text{if } P_{cs} > 5 \\
P_{cs} + 5 & \text{if } P_{cs} \leq 5
\end{cases}$$

while Policy 3 applies the same increase but for the more experienced teachers:

$$P^*_{cs} = \begin{cases} 
P_{cs} & \text{if } P_{cs} \leq 5 \\
P_{cs} + 5 & \text{if } P_{cs} > 5
\end{cases}$$

Results are presented in Figures 4 and 5. The effect of Policy 2 is roughly constant over the achievement distribution, and smaller in magnitude than the estimate obtained under Policy 3, which also shows more heterogeneity, with a larger impact for students at the higher percentiles. Overall, as with class size, the impacts of these policies are smaller than those obtained with the discrete changes estimators. This could be due to, for example, the impact of teacher experience coming from their interaction with other classrooms characteristics, such as class size (Mueller (2013)).

### 4.3 Proportion of Females

The idea that peers can affect student achievement is based on the assumption that students do not only learn from teachers but also from classmates. For example, students might teach one another by working in groups or having casual discussions, generating knowledge spillovers (see, e.g., Sacerdote (2011) for a review of this literature). One aspect of particular relevance in this context is the gender composition of the classroom. For example, the study of gender peer effects can shed light on the debate single-sex versus coeducational schools (Whitmore (2005)). Gender composition of the classroom could affect student performance in many ways. For example, a higher proportion of girls could improve classroom behavior, reduce classroom disruption and affect the level of violence, creating a better atmosphere for learning (Lavy and Schlosser (2011)). The presence of boys could intimidate girls from speaking up and influence student self-concepts or affect engagement with certain subjects. Finally, classroom composition could also affect the attitude and expectations of teachers towards the class, influencing the pace of teaching or their instructional methods (Cunningham and Andrews (1988)).

Several studies have examined the empirical role of gender composition of the classroom. Hoxby (2000) exploits gender variation in cohort composition in Texas elementary schools and finds that a higher share of girls raises student achievement in math and reading, both for boys and girls. Lavy and Schlosser (2011) find that, in Israeli middle-schools, a 10 percent point increase in the
proportion of female students increases girls’ math test scores by 3.7 percent of a standard deviation and boys’ scores by 2.2 percent.

Let $P_{cs}$ be the proportion of females in classroom $c$ in school $s$. Table 6 compares the fixed effect estimator of $\beta$ from model (1) to the density weighted average derivative estimator (3), for different choices of the series estimator of the derivative of the regression function. We can see that the impacts are considerably larger and statistically significant when we use the nonseparable model.

Next, I look at **distributional impacts**. First, *Policy 1* implies a 10 percent increase in the proportion of females for all classrooms,

$$P^{*}_{cs} = (1 + 0.1) \times P_{cs}$$

From the results in Figure 6, we can see that the effect of this policy is positive for all quantiles. There is some heterogeneity (with larger point estimates at the top of the distribution), but with wide confidence intervals. Overall, the impacts are smaller than those obtained in Table 6.

The next two policies try to disentangle the mechanism behind the positive effect of the proportion of females on student performance. For example, the effect could be coming from either having more girls in the classroom or more students of the same gender. Then, *Policy 2* increases the proportion of females in a classroom with majority of girls, and decreases the proportion in the classrooms with majority of boys:

$$P^{*}_{cs} = \begin{cases} (1 + 0.1) \times P_{cs} & \text{if } P_{cs} > 0.5 \\ (1 - 0.1) \times P_{cs} & \text{if } P_{cs} \leq 0.5 \end{cases}$$

From Figure 7 we can see that the impacts are now close to zero for all quantiles, suggesting that the effect is actually coming from a larger proportion of females, in line with previous findings in the literature. Finally, *Policy 3* takes into account feasibility constraints by changing the variance of the proportion of females without affecting the mean:

$$P_{cs}^{*} = P_{cs} + (1 - 0.1) \times (P_{cs} - \bar{P}_{cs})$$

We can see from Figure 8 that the effects are not statistically different from zero. As with class size, imposing feasibility constraints affects the magnitude of the impacts, suggesting that the implementation of policies regarding gender composition of the classrooms should take into account additional interactions and explore additional channels through which gender peer effects influence student performance.

### 5 Conclusion

In this paper, I look at the effects of teacher and peer characteristics on student achievement in the STAR Project conducted in Tennessee in the late 80s. As in standard linear models, I con-
sider two types of unobservables: school-specific effects and idiosyncratic disturbances. The model generalizes previous empirical research by allowing both effects to enter the structural function nonseparably. In particular, no functional form assumptions are needed for identification. Thus, the model permits nonparametric distributional and counterfactual analysis of heterogeneous effects. The main identification result uses an exchangeability assumption on the way that covariates affect the distribution of the school fixed effects. The model also extends policy analysis beyond marginal or discrete changes, to consider distributional effects originating from a counterfactual change in the distribution of characteristics of classrooms, peers and teachers. These impacts can also be analyzed on any feature of the distribution of student achievement, such as quantiles and inequality measures. In the empirical analysis, I look at the effects of class size, teacher experience and gender composition of the classroom on student test scores. Findings suggest that nonseparable heterogeneity is an important source of individual-level variation in the academic performance of kindergarten students in the STAR Project. Compared to previous results, the impact of class size is larger in magnitude and teacher experience has a stronger nonlinear impact. Still, conducting a counterfactual distributional analysis I find that these gains in student performance are hard to achieve when facing resource constraints.
References


25
## 6 Appendix I: Tables and graphs

### Table 1: Summary Statistics - Students

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>4.71</td>
<td>0.34</td>
<td>5,847</td>
</tr>
<tr>
<td>Race (Black)</td>
<td>0.32</td>
<td>0.46</td>
<td>5,847</td>
</tr>
<tr>
<td>Female</td>
<td>0.48</td>
<td>0.51</td>
<td>5,847</td>
</tr>
<tr>
<td>Free Lunch Eligible</td>
<td>0.48</td>
<td>0.52</td>
<td>5,847</td>
</tr>
<tr>
<td>Rural School</td>
<td>0.46</td>
<td>0.49</td>
<td>5,847</td>
</tr>
<tr>
<td>Total Math Score SAT</td>
<td>485</td>
<td>47.7</td>
<td>5,844</td>
</tr>
<tr>
<td>Total Reading Score SAT</td>
<td>436</td>
<td>31.7</td>
<td>5,763</td>
</tr>
</tbody>
</table>

### Table 2: Summary Statistics - Classrooms

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher Race (Black)</td>
<td>0.16</td>
<td>0.36</td>
<td>302</td>
</tr>
<tr>
<td>Teacher has Master Degree</td>
<td>0.36</td>
<td>0.48</td>
<td>302</td>
</tr>
<tr>
<td>Teacher Experience (years)</td>
<td>9.32</td>
<td>5.75</td>
<td>302</td>
</tr>
<tr>
<td>Class Size</td>
<td>19.4</td>
<td>4.14</td>
<td>302</td>
</tr>
</tbody>
</table>

Note: Original Sample Size: 6325, Sample with non-missing score information: 5907, Sample with non-missing values of class size, teacher experience and gender: 5886. The final sample excludes those with missing values in any of the covariates.
Table 3: Class Size

<table>
<thead>
<tr>
<th>OLS</th>
<th>Power Series</th>
<th>Regression Splines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effects</td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Coefficient</td>
<td>0.167</td>
<td>0.296</td>
</tr>
<tr>
<td>Std. Error</td>
<td>(0.028)</td>
<td>(0.043)</td>
</tr>
<tr>
<td>Parameter</td>
<td>-</td>
<td>$K = 2$</td>
</tr>
</tbody>
</table>

Notes: Regression Splines obtained using \textit{mgcv} R-package
### Table 4: Teacher Experience - 5 to 10 years

<table>
<thead>
<tr>
<th>OLS</th>
<th>Power Series</th>
<th>Regression Splines</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fixed Effects</td>
<td>(1)</td>
</tr>
<tr>
<td>Coefficient</td>
<td>0.027</td>
<td>0.125</td>
</tr>
<tr>
<td>Std. Error</td>
<td>(0.012)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>Parameter</td>
<td>-</td>
<td>$K = 2$</td>
</tr>
</tbody>
</table>

Notes: Regression Splines obtained using *mgcv* R-package
Table 5: Teacher Experience - 10 to 15 years

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>Power Series</th>
<th>Regression Splines</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fixed Effects</td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Coefficient</td>
<td>0.063</td>
<td>0.228</td>
<td>0.390</td>
</tr>
<tr>
<td>Std. Error</td>
<td>(0.013)</td>
<td>(0.025)</td>
<td>(0.041)</td>
</tr>
<tr>
<td>Parameter</td>
<td>-</td>
<td>$K = 2$</td>
<td>$K = 4$</td>
</tr>
</tbody>
</table>

Notes: Regression Splines obtained using \textit{mgcv} R-package.
Table 6: Proportion of Females

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>Power Series</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fixed Effects</td>
<td>(1)</td>
</tr>
<tr>
<td>Coefficient</td>
<td>0.376</td>
<td>0.580</td>
</tr>
<tr>
<td>Std. Error</td>
<td>(0.013)</td>
<td>(0.122)</td>
</tr>
<tr>
<td>Parameter</td>
<td>-</td>
<td>$K = 2$</td>
</tr>
</tbody>
</table>
Table 7: Simulation Results - Model 1

<table>
<thead>
<tr>
<th>γ</th>
<th>0</th>
<th>0.5</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
<th>4.5</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0.055)</td>
<td>(0.535)</td>
<td>(1.230)</td>
<td>(1.713)</td>
<td>(2.799)</td>
<td>(2.779)</td>
<td>(3.912)</td>
<td>(4.043)</td>
<td>(4.043)</td>
<td>(4.845)</td>
<td>(5.498)</td>
</tr>
<tr>
<td></td>
<td>(0.081)</td>
<td>(0.430)</td>
<td>(0.875)</td>
<td>(1.289)</td>
<td>(1.641)</td>
<td>(2.061)</td>
<td>(2.909)</td>
<td>(3.048)</td>
<td>(3.015)</td>
<td>(3.718)</td>
<td>(4.149)</td>
</tr>
<tr>
<td>( h_p )</td>
<td>0.002</td>
<td>0.109</td>
<td>0.045</td>
<td>0.101</td>
<td>0.168</td>
<td>0.076</td>
<td>0.107</td>
<td>0.065</td>
<td>0.012</td>
<td>0.078</td>
<td>0.096</td>
</tr>
<tr>
<td>( h_v )</td>
<td>0.038</td>
<td>0.170</td>
<td>0.037</td>
<td>0.132</td>
<td>0.186</td>
<td>0.030</td>
<td>0.198</td>
<td>0.156</td>
<td>0.035</td>
<td>0.082</td>
<td>0.175</td>
</tr>
</tbody>
</table>

Notes: 5,000 replications with \( N = 500 \) and \( C = 2 \).
Table 8: Simulation Results - Model 2

<table>
<thead>
<tr>
<th>( \gamma )</th>
<th>0</th>
<th>0.5</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
<th>4.5</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta )</td>
<td>1.000</td>
<td>2.336</td>
<td>3.672</td>
<td>5.008</td>
<td>6.343</td>
<td>7.679</td>
<td>9.015</td>
<td>10.351</td>
<td>11.687</td>
<td>13.023</td>
<td>14.358</td>
</tr>
<tr>
<td></td>
<td>(0.055)</td>
<td>(7.693)</td>
<td>(9.472)</td>
<td>(16.453)</td>
<td>(14.543)</td>
<td>(27.608)</td>
<td>(25.633)</td>
<td>(40.463)</td>
<td>(36.281)</td>
<td>(95.256)</td>
<td>(42.124)</td>
</tr>
<tr>
<td>( h_p )</td>
<td>0.199</td>
<td>0.074</td>
<td>0.180</td>
<td>0.056</td>
<td>0.199</td>
<td>0.080</td>
<td>0.166</td>
<td>0.030</td>
<td>0.038</td>
<td>0.105</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>0.185</td>
<td>0.031</td>
<td>0.094</td>
<td>0.034</td>
<td>0.177</td>
<td>0.047</td>
<td>0.056</td>
<td>0.027</td>
<td>0.147</td>
<td>0.193</td>
<td>0.165</td>
</tr>
</tbody>
</table>

Notes: 5,000 replications with \( N = 500 \) and \( C = 2 \).
Figure 1: Class Size - Policy I

Panel (a): Distributions

Panel (b): Quantile Policy Effect
Figure 2: Class Size - Policy II

Panel (b): Quantile Policy Effect

Panel (a): Distributions

Size

0 200 400 600 800 1000 1200

Quantiles

0.0 0.2 0.4 0.6 0.8 1.0

−0.02 0.00 0.02 0.04 0.06 0.08 0.10

Panel (a): Distributions

Size

0 10 20 30

0 200 400 600 800 1000 1200
Figure 3: Teacher Experience - Policy I

Panel (a): Distributions

Panel (b): Quantile Policy Effect
Figure 4: Teacher Experience - Policy II

Panel (a): Distributions
- Years: 0, 5, 10, 15, 20, 25
- Values: 0, 200, 400, 600, 800, 1000, 1200

Panel (b): Quantile Policy Effect
- Quantiles: 0.0, 0.2, 0.4, 0.6, 0.8, 1.0
- Values: -0.02, 0.00, 0.02, 0.04, 0.06, 0.08, 0.10
Figure 6: Proportion of Females - Policy I

Panel (a): Distributions

Panel (b): Quantile Policy Effect
Figure 7: Proportion of Females - Policy II

Panel (a): Distributions

Panel (b): Quantile Policy Effect
Figure 8: Proportion of Females - Policy III

Panel (a): Distributions

Panel (b): Quantile Policy Effect
Figure 9: Simulation Results - Model 1
Figure 10: Simulation Results - Model 2

\[ \gamma = 0 \]

\[ \gamma = 1 \]

\[ \gamma = 2 \]

\[ \gamma = 3 \]

\[ \gamma = 4 \]

\[ \gamma = 5 \]
7 Appendix II: Simulation Study

For all models, I assume $I = 1, C = 2, S = 500$. The $m(\cdot)$ function is given by:

$$m(\mu, p, \varepsilon) = \mu + \beta p + \gamma \Gamma(\mu, p) + \varepsilon$$

for different choices of $\Gamma(\cdot, \cdot)$. Note that $\gamma = 0$ implies that the linear model would be the correct specification, so a standard fixed effects panel data estimator would work in this case. I compare their performance to a nonparametric estimator for different values of $\gamma$. Also, the derivative of the function with respect to $p$ is given by:

$$m_p(\mu, p, \varepsilon) = \beta + \gamma \Gamma_p(\mu, p)$$

and so the Density-Weighted ADE parameter is given by:

$$\delta = \beta + \gamma \mathbb{E}[\Gamma_p(\mu, p)]$$

For the distributions, I assume that:

$$(\mu_s, P_{cs}) \sim N(0, \Sigma)$$

where

$$\Sigma = \begin{bmatrix}
\sigma_p^2 & \sigma_{pp} & \cdots & \sigma_{p\mu} \\
\sigma_{pp} & \sigma_p^2 & \cdots & \sigma_{p\mu} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{p\mu} & \sigma_{p\mu} & \cdots & \sigma_{\mu}^2
\end{bmatrix}$$

and $\sigma_p^2 = \sigma_{\mu}^2 = 1; \sigma_{pp} = 0.3$ and $\sigma_{p\mu} = 0.5$. Finally,

$$\varepsilon_{ics} \sim N(0, 1)$$

7.1 Model 1

In this specification, the marginal effect depends of both $p$ and $\mu$:

$$\Gamma(\mu, p) = \exp(p + \mu)$$

$$\implies \Gamma_p(\mu, p) = \exp(p + \mu)$$

Note that

$$\delta = \beta + \gamma \mathbb{E}[\exp(p + \mu)]$$

where, using known results for bivariate normal distributions:

$$\mu_s + P_{cs} \sim N(0, \sigma_p^2 + \sigma_{\mu}^2 + 2\sigma_{p\mu}\sigma_p\sigma_{\mu})$$
And then,
\[
\mathbb{E} [\exp(p + \mu)] = \exp \left( \frac{\sigma^2_p + \sigma^2_\mu + 2\sigma_{p\mu}\sigma_p\sigma_\mu}{2} \right) \\
= \exp(1 + \sigma_{p\mu})
\]

Finally,
\[
\delta = \beta + \gamma \exp(1 + \sigma_{p\mu})
\]

Table 4 presents results for Model 1, for different values of $\gamma$.

### 7.2 Model 2

In this specification, the marginal effect depends on both $p$ and $\mu$:
\[
\Gamma(\mu, p) = (p - 1)^2 \exp(p + \mu)
\]
\[
\implies \Gamma_p(\mu, p) = 2(p - 1) \exp(p + \mu) + (p - 1)^2 \exp(p + \mu)
\]
\[
= (p + 1)(p - 1) \exp(p + \mu)
\]

Table 5 presents results for Model 2, for different values of $\gamma$. 

44
8 Appendix III: Proofs

Proof Theorem I. Let

\[ \delta_{ics}(p, z) = \int \frac{\partial m(p, z, \mu, \varepsilon)}{\partial p} dF_{\mu, \varepsilon|ics|P_{ics}, Z_{ics}}(\mu|p, z) \]

Next, note that:

\[ \frac{\partial}{\partial P_{cs}} \left( \frac{\partial E(Y_{ics}|P_{ics}, Z_{ics}, V_{s})}{\partial P_{cs}} \right) = \int \frac{\partial}{\partial p} \left[ m(p, z, \mu, \varepsilon) dF_{\mu, \varepsilon|ics|P_{ics}, Z_{ics}, V_{s}}(\mu|p, z, v) \right] \]

Then,

\[ \int \frac{\partial E(Y_{ics}|P_{ics}, Z_{ics}, V_{s})}{\partial P_{cs}} dF_{V_{s}|P_{ics}, Z_{ics}}(v|p, z) = \int \frac{\partial m(p, z, \mu, \varepsilon)}{\partial p} dF_{\mu, \varepsilon|ics|P_{ics}, Z_{ics}, V_{s}}(\mu|p, z, v) dF_{V_{s}|P_{ics}, Z_{ics}}(v|p, z) \]

From this result, identification of the density weighted average derivatives follows directly:

\[ \delta_{ics} = \int \int \frac{\partial E(Y_{ics}|P_{ics}, Z_{ics}, V_{s})}{\partial P_{ics}} \omega(p, z) dF_{V_{s}|P_{ics}, Z_{ics}}(v|p, z) dF_{P_{ics}, Z_{ics}}(p, v) \]

\[ = \mathbb{E} \left[ \frac{\partial E(Y_{ics}|P_{ics}, Z_{ics}, V_{s})}{\partial P_{ics}} \omega(p, z) \right] \]
Proof Theorem II. Let

\[
\Delta (p'', p') = \int [m (p'', z, \mu, \varepsilon) - m (p', z, \mu, \varepsilon)] \, dF_{Z_{ics},\mu,\varepsilon_{ics}|P_{cs}} (z, \mu, \varepsilon|p')
\]

\[\equiv A - B\]

First, note that

\[
B = \int m (p', z, \mu, \varepsilon) \, dF_{Z_{ics},\mu,\varepsilon_{ics}|P_{cs}} (z, \mu, \varepsilon|p') = E (Y_{ics}|P_{cs} = p')
\]

and so is identified directly from the data. For the other term:

\[
A = \int m (p'', z, \mu, \varepsilon) \, dF_{Z_{ics},\mu,\varepsilon_{ics}|P_{cs}} (z, \mu, \varepsilon|p')
\]

\[= \int m (p'', z, \mu, \varepsilon) \, dF_{Z_{ics},\mu,\varepsilon_{ics}|P_{cs},V_{s}} (v, z, \mu, \varepsilon|p') \, dF_{V_{s}|P_{cs}} (v|p')
\]

\[= \int m (p'', z, \mu, \varepsilon) \, dF_{\mu_{cs},\varepsilon_{ics}|P_{cs},V_{s},Z_{ics}} (\mu, \varepsilon|p', v, z) \, dF_{Z_{ics},|P_{cs},V_{s}} (z|p', v) \, dF_{V_{s}|P_{cs}} (v|p') \tag{1}
\]

where

\[\tag{1} = \int m (p'', z, \mu, \varepsilon) \, dF_{\mu_{cs},\varepsilon_{ics}|P_{cs},V_{s},Z_{ics}} (\mu, \varepsilon|p', v, z)
\]

\[= \int m (p'', z, \mu, \varepsilon) \, dF_{\mu_{cs}|P_{cs},V_{s},Z_{ics}} (\mu, \varepsilon|p', v, z)
\]

\[= \int m (p'', z, \mu, \varepsilon) \, dF_{\mu_{cs}|P_{cs},V_{s},Z_{ics}} (\mu, \varepsilon|v, z)
\]

\[= \int m (p'', z, \mu, \varepsilon) \, dF_{\mu_{cs},\varepsilon_{ics}|P_{cs},V_{s},Z_{ics}} (\mu, \varepsilon|p''', v, z)
\]

\[= \int m (p'', z, \mu, \varepsilon) \, dF_{\mu_{cs},\varepsilon_{ics}|P_{cs},V_{s},Z_{ics}} (\mu, \varepsilon|p'', v, z)
\]

\[= E (Y_{ics}|P_{cs} = p'', Z_{ics} = z, V_{s} = v)
\]

so, finally

\[
A = \int E (Y_{ics}|P_{cs} = p'', Z_{ics} = z, V_{s} = v) \, dF_{Z_{ics},|P_{cs},V_{s}} (z|p', v) \, dF_{V_{s}|P_{cs}} (v|p')
\]

\[= \int E (Y_{ics}|P_{cs} = p'', Z_{ics} = z, V_{s} = v) \, dF_{Z_{ics},V_{s}|P_{cs}} (z, v|p')
\]

\[\blacksquare\]
Proof Theorem III. Let \( \phi(y) \) be any function of \( y \). Then,

\[
\mathbb{E} [\phi(Y_{ics}^*)] = \mathbb{E} \left[ \mathbb{E} \left[ \phi(m(P_{ics}^*, Z_{ics}, \mu_s, \varepsilon_{ics})) | P_{ics}^*, Z_{ics} \right] \right] = \int \int \phi(m(p, z, \mu, \varepsilon)) dF_{\mu_s, \varepsilon_{ics} | P_{ics}^*, Z_{ics}} (\mu, \varepsilon | p, z) dF_{P_{ics}^*, Z_{ics}} (p, z)
\]

where

\[
(1) = \int \int \phi(m(p, z, \mu, \varepsilon)) dF_{\mu_s, \varepsilon_{ics} | P_{ics}^*, Z_{ics}, V_s} (\mu, \varepsilon | p, z, v) dF_{V_s | P_{ics}^*, Z_{ics}} (v | p, z)
\]

and

\[
(2) = \int \phi(m(p, z, \mu, \varepsilon)) dF_{\varepsilon_{ics} | \mu_s, P_{ics}^*, Z_{ics}, V_s} (\varepsilon | \mu, p, z, v) dF_{P_{ics}^*, Z_{ics}, V_s} (\mu | p, z, v)
\]

Then, finally

\[
\mathbb{E} [\phi(Y_{ics}^*)] = \int \int \mathbb{E} [\phi(Y_{ics}) | P_{ics}, Z_{ics}, V_s] dF_{V_s | P_{ics}^*, Z_{ics}} (v | p, z) dF_{P_{ics}^*, Z_{ics}} (p, z)
\]

\[
= \mathbb{E} [\mathbb{E} [\phi(Y_{ics}) | P_{ics}, Z_{ics}, V_s]]
\]