

Preliminary draft – comments welcome

Causal Inference when Assignment May Have Been Random:

Peer Effects in North Carolina Elementary Schools

Jacob Vigdor and Thomas Nechyba*

Duke University and NBER

December, 2008

Abstract

Administrative datasets frequently lack information on the protocol used to assign subjects to treatment conditions. In this paper, we show that under certain conditions, consistent causal estimates can be recovered from data where the actual assignment protocol is uncertain, but the probability of random assignment is known. Predictably, estimators are less consistent when the probability of random assignment is unknown and must itself be inferred. We introduce a method of inferring the probability of random assignment ex post based on a scalar measure of the degree of balance in observed characteristics across treatment conditions. We then apply this method to the study of peer effects in education production. Results indicate that a considerable portion of OLS peer effects estimates reflect selection – intuitively, the results tend to be stronger in schools that depart obviously from random assignment. One relatively robust result, however, is that average achievement is higher in 5th grade classrooms with heterogeneous, rather than homogeneous, student ability levels. The benefit to low-ability students from exposure to high-ability peers exceeds the cost to high-ability students from exposure to low-ability peers.

* Vigdor: Sanford Institute of Public Policy, Box 90312, Durham NC 27708, jacob.vigdor@duke.edu.

Nechyba: Department of Economics, Box 90097, Durham NC 27708, nechyba@econ.duke.edu

We are grateful to Erika Martinez, Kata Mihaly, Jane Cooley, Natalie Goodpaster and Dan Hungerman for outstanding research assistance, to Jeffrey Zabel and other participants at the 2004 APPAM meetings and participants in the Harvard University public economics seminar for comments on an earlier draft of this paper, and to participants in the 2004 AEA meetings and the CESifo-PEPG conference on equity and efficiency in education for comments on our earlier work.

1. Introduction

Can schools improve the performance of individual children simply by placing them in a room with higher-ability peers? Would the higher-ability peers themselves suffer as a result of such a placement? Analyses of these questions, or ones very similar to them, have consumed a great deal of effort in recent economic literature

Over the past decade, there have been two unmistakable trends in empirical microeconomics. The first involves the use of randomized experiments to identify causal relationships of economic interest. The second is the use of large administrative databases, often not originally collected for research use, to study the impact of variation in local conditions and policies. There are limitations to both practices. The results of randomized experiments avoid concerns related to selection bias, but raise concerns regarding generalizability and scalability. The history of social science is replete with examples of policy interventions that generated large effects when implemented on small experimental samples, but failed to generate similar results when taken to a larger scale. Administrative data provide an opportunity to assess heterogeneity in effects across subpopulations, but only rarely provide opportunities for researchers to exploit truly exogenous variation in a policy variable of interest.

Recent literature in the economics of education provides numerous examples of both research methodologies. Researchers have conducted and analyzed experiments randomly assigning students to class sizes, peer groups, and teachers (Krueger 1999; Graham 2007), or randomly assigning students or teachers to incentive programs (Bettinger 2008). Other researchers have analyzed a broadly similar set of research questions using administrative data. Efforts to isolate the impact of class size using nonexperimental data have exploited strict rules governing class size (Angrist and Lavy 1999). Studies of the impact of peer characteristics on

achievement have studied presumably idiosyncratic variation in peer group composition across cohorts within schools (Hoxby 2000). Studies of the impact of incentive programs have utilized traditional difference-in-difference methods or regression discontinuity designs exploiting strict eligibility criteria (Clotfelter, Glennie, Ladd, and Vigdor 2008).

Administrative data may provide opportunities to exploit random assignment to environmental characteristics, but these opportunities are often difficult to discern. For example, anecdotal evidence suggests that some schools use a random algorithm to assign students to classrooms. Administrative data generally do not provide information on the type of assignment algorithm used, let alone whether the algorithm was followed consistently. Collecting information on the assignment protocol used *ex post*, particularly the degree of adherence to any random assignment protocol used, may be either costly or impossible, depending on the quality of record-keeping by individuals in charge of the process and the degree of turnover among administrative staff.

In this paper, we present a strategy for deriving causal impacts of environmental variables when assignment may have been random in at least some subgroups of a large sample. Using a Monte Carlo simulation, we demonstrate that consistent estimates can be derived from data where some individuals are assigned to an environmental factor randomly, and others assigned in a manner that yields a correlation between the environmental factor and unobserved determinants of the outcome of interest, even when the actual assignment algorithm is unknown. These consistent estimates can be derived so long as the probability that an individual was part of a randomly assigned subgroup is known. Even if the probability of random assignment cannot be reliably estimated, this method provides information about the direction of bias in uncorrected

estimates.

We then present a method for computing the probability of random assignment on the basis of observed variables. Intuitively, this method can be traced to a common statistical check on the reliability of random assignment algorithms, the comparison of pre-treatment characteristics for individuals of different assignment status. When the distribution of pre-treatment characteristics is comparable for individuals with different assignment status, the posterior probability of random assignment is high. When there are obvious differences, the posterior probability of random assignment is low. The criterion for comparability is a goodness of fit characteristic, rather than a measure of statistical significance, since the latter are sensitive to sample size and will tend to overpredict the probability of random assignment in small groups.

We apply this method to the study of peer group composition effects in elementary school classrooms. The question of whether students perform better when assigned to more-able peers, or a more homogeneous group of peers, has inspired a considerable amount of recent research, with no clear consensus in existing literature (Arcidiacono and Nicholson, 1999; Betts and Zau, 2002; Boozer and Cacciola 2001; Caldas and Bankston 1997; Gaviria and Raphael 2001; Hanushek et al. 2001; Hoxby 2000; Link and Mulligan 1991; Robertson and Symons 1996; Sacerdote 2001; Slavin 1987; Slavin 1990; Zimmer and Toma 1999). Definitive answers to this question would be useful to policymakers interested in maximizing measures of student achievement, and to theorists interested in better calibrating general equilibrium models of parental choices and student outcomes (Epple and Romano 1998, Nechyba 2000, Ferreyra 2002). Unfortunately, existing empirical literature has pointed to no clear consensus on the existence or magnitude of peer effects in educational settings (Nechyba et al. 1999).¹ This lack of consensus

¹ The absence of consensus on the existence and magnitude of peer effects contrasts with, for example, the

may in part reflect serious methodological issues involved in estimating causal relationships between peer characteristics and student outcomes (Duncan et al. 1997; Manski 1993; Moffitt 1998).

Recent efforts in the peer effects literature show increasing econometric sophistication. Graham (2007) uses random variation in the variability in peer group characteristics, associated with random variation in class size in the Tennessee STAR experiment, and concludes that more-able peers improve individual performance. It is not clear from this exercise, however, whether peer effects are nonlinear – a critical question for matters of policy design. Cooley (2007) finds evidence of important nonlinearities in peer effects using nonexperimental data; Hoxby and Weingarth (2006) exploit variation in peer group composition generated by changes in school attendance zone boundaries and show evidence of complex nonlinear effects.

Our results show that several patterns apparent even in sophisticated OLS models, restricting effect estimates to within-school variation across classrooms, reflect selection rather than true effects of peer characteristics on outcomes. In OLS models, student achievement appears significantly related to the mean and standard deviation of peer prior achievement, classroom racial and gender composition; our preferred method of distinguishing selection from treatment effects suggests that many of these effects suffer from selection bias. Estimates indicate that the source of bias is the selection of students with better unobserved determinants of achievement into classrooms with peers who are higher achieving, more likely to be white, and more likely to be female. Selection actually obscures the true magnitude of another important effect: Hispanic students, who in North Carolina are very likely to be first or second generation

estimated impact of teacher experience of student achievement. Multiple recent studies have reached the conclusion that students assigned to highly experienced teachers can expect to outperform equivalent students assigned to novice teachers by roughly one-tenth of a standard deviation on standardized tests (Clotfelter, Ladd and Vigdor 2004).

immigrants who speak Spanish at home, perform significantly better when they share classrooms with a higher proportion of Hispanic classmates. This effect is obscured by the fact that such classrooms tend to be populated by Hispanic students with poor unobserved determinants of achievement. Peer ability is consistently associated with higher test scores, and we also find evidence that students perform better when assigned to classrooms with heterogeneous ability levels. These results are broadly consistent with reliable estimates in the existing literature.

Section 2 reviews the basic concerns with selection bias in inference from non-experimental data, and reviews the results of a Monte Carlo simulation indicating how information on the likelihood of random assignment can be used in cases where assignment may have been random. Section 3 then discusses a method for inferring the probability of random assignment on the basis of the distribution of observed characteristics across treatment groups. Section 4 introduces our application to the study of peer effects, introducing the basic empirical specification, data, and describing the specific method for deriving the posterior probability of random assignment. Section 5 presents results, and section 6 concludes.

2. Causal inference under random and nonrandom assignment

In a typical setting, an econometrician wishes to infer the impact of a contextual variable Z on an outcome Y , while controlling for a vector of individual characteristics X . The outcome Y is also influenced by individual characteristics W , which are unobserved to the econometrician.

The process determining outcome Y can thus be described as:

$$(1) Y = \alpha + \beta_1 X + \beta_2 Z + \beta_3 W + \varepsilon$$

where the term ε refers to idiosyncratic factors that influence the outcome but cannot be related

to any individual or contextual characteristic. The equation estimated by the econometrician, by contrast, takes the form

$$(2) \quad Y = \tilde{\alpha} + \tilde{\beta}_1 X + \tilde{\beta}_2 Z + \tilde{\epsilon}$$

where the tilde marks indicate statistical estimates. It is well established that $\tilde{\beta}_2$ is a consistent estimator of β_2 only in the case where W , the unobserved determinant of Y , is uncorrelated with Z . In the estimation of causal treatment impacts, there are several strategies for ensuring that this condition holds. One solution is to randomly assign research subjects to different values of Z . When random assignment is conducted appropriately, this ensures that Z will be uncorrelated with W .² Alternative solutions include identifying situations where at least some factors uncorrelated with W have a role in determining Z . When neither of these options are available, the econometrician's estimate $\tilde{\beta}_2$ is given by equation (3):

$$(3) \quad \tilde{\beta}_2 = \beta_2 + \beta_3 \frac{\text{cov}(Z, W)}{\text{var}(Z)} .$$

It is clear from equation (3), the standard formula for omitted variable bias, that $\tilde{\beta}_2$ will generally be inconsistent unless β_3 is zero or there is no correlation between Z and W .

Suppose that an econometrician observes individuals i in groups j , and from within these groups they are assigned to subgroups k with common values of Z . Suppose further that some groups assign individuals to subgroups using a random algorithm, ensuring that there is no correlation between Z and W . Other groups assign individuals to subgroups in a manner that generates a correlation between Z and W . Finally, suppose the econometrician does not observe

²Random assignment studies are often complicated, however, by noncompliance with assignment, and with potentially nonrandom attrition from the study sample. The former problem can be addressed using the initially assigned value of Z as an instrument for the actual value. The latter problem can be addressed by modeling the process of sample attrition, particularly in scenarios where there exists some factor that influences attrition rates without having any correlation on the outcome of interest.

the actual algorithm used to assign individuals to subgroups, but does observe for each group the probability that random assignment was used. In its simplest form, one could imagine that groups are randomly assigned to use a random algorithm or a nonrandom algorithm for subgroup assignment, with some groups exogenously assigned a higher likelihood of using random assignment. Call this probability of using a random algorithm p . In such a scenario, it is possible to recover an unbiased estimate of β_2 . The estimated coefficient $\tilde{\beta}_2$ can be expressed as:

$$(4) \quad \tilde{\beta}_2 = p\beta_2 + (1-p)\left(\beta_2 + \beta_3 \frac{\text{cov}(Z, W)}{\text{var}(Z)}\right) = \beta_2 + (1-p)\beta_3 \frac{\text{cov}(Z, W)}{\text{var}(Z)} .$$

This implies that estimating a modified form of equation (2), adding an interaction term between the contextual variable Z and the probability of non-random assignment $(1-p)$ can effectively partition the coefficient $\tilde{\beta}_2$ into two components, one reflecting the true causal impact of Z and the other selection bias:

$$(5) \quad Y = \hat{\alpha} + \hat{\beta}_1 X + \hat{\beta}_2 Z + \gamma(1-p)Z + \hat{\epsilon} .$$

The selection bias present in $\tilde{\beta}_2$ loads onto the \square term in equation (5), leaving $\hat{\beta}_2$ an unbiased estimator of β_2 .

To document this, and to study the relative efficiency and bias associated with different methods of utilizing information on the probability of random assignment, we conducted a Monte Carlo simulation. In each simulation, one million individuals are assigned random draws of X and W from independent standard normal distributions. They are randomly assigned to one thousand different groups, each having one thousand members. Within each group, individuals are assigned to one of four subgroups using either a random or nonrandom algorithm. Reflecting

the application to the study of peer effects below, the contextual variable Z for individual i in subgroup k of group j is equal to the average value of X for all subgroup members except i . The outcome measure Y is defined as the sum of X , W , and Z , plus an error term drawn from an independent standard normal distribution.

In each group, the probability of using a random assignment is drawn from a uniform distribution on the interval $[0,1]$. In random assignment groups, individuals are sorted on the basis of a random number drawn from a uniform distribution on the interval $[0,1]$. In nonrandom assignment groups, individuals are sorted according to the sum of a random number drawn from a uniform distribution on the interval $[0,1]$ and their rank in the distribution of ability, where ranks are normalized to fall within the interval $[0,1]$. The random number is multiplied by 0.9 and the ability rank by 0.1. The relationship between X , W , and the outcome Y was then estimated using a within-group fixed effect estimator, first using equation (2) and then using equation (5). We also estimated two additional specifications where the probability of random assignment was used as a weighting variable, rather than as an interacted regressor. The first specification used a discrete form of weighting, including only observations with a probability of random assignment above 0.5. The second is a continuous form where the probability of random assignment is used as a weight for each observation.

Table 1 shows the results of the Monte Carlo. Table entries reflect the results of 200 iterations of the Monte Carlo simulation. Reported here are the coefficients at the median, 97.5th percentile, and 2.5th percentile from the rank-ordered distributions. The first row describes the range of coefficients obtained from a standard OLS-fixed effects estimator. Recalling that the true coefficient on the contextual variable Z is one, these estimates clearly show the impact of

omitted variable bias generated by the correlation between the unobserved variable W and Z in those schools not using random assignment. Fixed effects estimates range between 1.43 and 1.49, with a median estimate of 1.46.

The second row reports the results of specifications mirroring equation (5), with the addition of group-specific intercepts. The estimated coefficients on Z show only slight signs of bias, with a median value of 1.03, and a range between 0.905 and 1.11. The interval is not exactly symmetric around one, but relative to the OLS-fixed effect model, the amount of bias present in the median estimate has been reduced by more than 90 percent. As expected, most of the bias present in the OLS estimate has been loaded onto the interaction between Z and the probability of non-random assignment. This reduction in bias does not come without cost, however; OLS coefficients fell in a narrow range of width 0.06; the interval containing 95% of the coefficients from the interacted specification is more than three times wider.

The third and fourth row report the results of using weighting schemes rather than the interacted specification in equation (5). Restricting the analysis sample to groups with a probability of random assignment greater than 0.5 yields coefficients that are less biased than those obtained in the standard OLS-fixed effects model, but these estimates continue to display significant bias, with a median estimate of 1.23, and a range between 1.16 and 1.30. The median estimate from the restricted sample thus shows about half the bias in the median OLS estimate. The interval containing 95% of all estimated coefficients is narrower than the interval from the interacted specification, but wider than the OLS interval.

Using the probability of random assignment as a weight yields the smallest relative reduction in bias – the magnitude of the median coefficient here is more than two-thirds the

magnitude of the OLS median coefficient. Ninety-five percent of the coefficients fell within a range between 1.27 and 1.37, making this the most efficient of the three alternate estimates.

The primary obstacle to implementing a model such as equation (5) to purge estimates of selection bias is obtaining an estimate of the probability of random assignment. In section 3 below, we outline a method for deriving posterior estimates of the probability of random assignment. A question remains, however, regarding how well the estimator performs when the probability of random assignment is not accurately estimated. To assess this question, we conducted an additional simulation where a noisy estimate of the probability of random assignment was substituted for the actual probability. The noisy estimate was obtained by multiplying the true probability by 0.5 and adding a random number drawn from a uniform distribution on the interval $[0,0.5]$.

Table 2 reports estimates of the alternative Monte Carlo with noisy estimates of the probability of random assignment. In this case, implementation of equation (5) results in biased estimates, though the bias is small relative to most other estimators considered to this point. The median estimated coefficient is 1.08, with a range between 0.98 and 1.19. Whereas use of the true probability of random assignment as an interaction term is associated with a reduction in bias of over 90%, use of the noisy measure produces a reduction closer to 80%. Even with a noisy estimate of the probability of random assignment, equation (5) continues to load most of the bias onto the interaction term and not the main effect. The use of a noisy probability estimate does not have much impact on the precision of the estimates; 95% of the coefficients obtained in the Monte Carlo procedure fell within an interval of width 0.21, nearly identical to the width of the interval found using the actual probabilities.

Table 2 also shows results obtained from using the noisy estimates of the probability of random assignment as weights rather than as interacted regressors. Restricting the sample to observations with an observed probability of random assignment above 50% resulted in a 50% reduction in bias for the median coefficient, relative to the median OLS coefficient. Using the noisy probability measure eliminates a sizable fraction of this bias reduction. The median coefficient obtained in this case is 1.32, which represents a 30% reduction in bias relative to the median OLS coefficient.

Using noisy estimates of the probability of random assignment as weights similarly degrades the bias reduction of that estimator relative to one that uses the true probability. The median estimate obtained using the true probability reflected a 30% reduction in bias relative to the median OLS estimate; the median estimate using a noisy probability measure reflects a more modest 13% reduction in bias.

Overall, then, the model in equation (5) has the advantage of achieving the greatest reduction in bias and showing the least sensitivity to inaccuracy in the estimated probability of random assignment. It is a relatively noisy estimator, however, with the widest interval of estimates of any alternative considered here.

3. Inferring the likelihood of random assignment

True random assignment of individuals to a contextual variable Z results in a lack of correlation between Z and any unobserved variables W . Random assignment also results in a lack of correlation between Z and observed individual characteristics X . A common test of the validity of a random assignment algorithm involves testing the hypothesis of equality in the mean values of X for individuals assigned to different values of Z . This basic logic – that random

assignment is more likely to be valid when there are no significant differences in pretreatment characteristics between subgroups – is the basis for the method of inferring random assignment described here.

In principle, there are statistical tests designed to test the null hypothesis of random assignment across categories by comparing observed characteristics, the χ^2 and F tests. A researcher might thus be tempted to use these tests as the basis of assessing the hypothesis that an assignment algorithm was effectively random.³ In practice, these tests are more useful for researchers intending to prove that assignment was not random, and less useful as tests of randomness, since random assignment is set up as the null hypothesis. As a consequence, there will be a tendency to erroneously conclude that a random algorithm was used, particularly in smaller samples. Any attempt to use a statistical significance standard to infer whether a random assignment algorithm was used will suffer from this problem.

Goodness-of-fit statistics, which are generally invariant to changes in sample size when the underlying relationships in the data are unaltered, provide a reasonable substitute for χ^2 or F -statistic based tests of randomness. This section outlines a procedure for inferring whether a random assignment algorithm was used on the basis of the strength of the correlation between observable characteristics and assignment status.

Suppose that the contextual variable of interest Z is discrete and takes on one of a finite number of values. This applies to any traditional experimental evaluation of treatment effects, or situations where individuals are assigned to discrete subgroups with different values of Z , as in the example in the preceding section. A summary measure of the ability of observed characteristics X to predict assignment to the categories of Z can be obtained by estimating a

³ For an example of exactly such a procedure, see Clotfelter, Ladd and Vigdor (2006).

multinomial logit equation, and referring to the goodness-of-fit statistic commonly known as the pseudo- R^2 , defined as follows:

$$(6) \quad \text{pseudo } R^2 = 1 - \frac{\ln L_F}{\ln L_C}$$

where L_F refers to the likelihood of the estimated model and L_C refers to the likelihood of a model estimated using only a constant term.⁴ Better-fitting models will have likelihoods closer to 1, implying log-likelihoods that are closer to zero. In the special case of a binary treatment, the method of estimation is a simple logit rather than multinomial logit.

If the fit of a model predicting assignment to categories of Z on the basis of observed characteristics is poor, the likelihood that a random assignment algorithm was successfully implemented is relatively high. If the fit of such a model is relatively good, the likelihood that an effective random assignment algorithm was used is relatively small. This basic intuition can be formalized in the Bayesian expression:

$$(7) \quad P(RA_j^{post} | G = G_j) = \frac{P(G = G_j | RA_j) P(RA_j)}{P(G = G_j | RA_j) P(RA_j) + P(G = G_j | RA_j^C) P(RA_j^C)}$$

where the event RA_j is the successful implementation of a random assignment algorithm in sample group j , G is a goodness of fit statistic which takes on a value G_j in sample group j , and RA^C corresponds to any other possible assignment algorithm. The superscript *post* denotes a posterior probability, $P(RA_j)$ represents the econometrician's prior beliefs regarding the use of random assignment in group j .

To transform a goodness-of-fit statistic into a posterior probability of random assignment, it is necessary to determine the likelihood of obtaining such a statistic under random assignment,

⁴ This formulation of the pseudo- R^2 is often referred to as McFadden's model.

the likelihood of obtaining that statistic under any other assignment protocol, and to specify a prior belief regarding the use of random assignment in group j . The first of these three quantities can be obtained using a Monte Carlo simulation, repeatedly assigning the members of j to subgroups using a truly random protocol and tracking the distribution of obtained goodness-of-fit statistics.

The second of the three quantities is more difficult to obtain. The set of non-random assignment protocols is infinite, ranging from protocols that explicitly segregate individuals on the basis of some observed or unobserved characteristic to protocols that attempt to enforce random assignment but allow some degree of non-compliance. The event RA^C in equation (7) thus refers to an uncountably large number of possible alternative protocols.

As a feasible alternative, the probability of finding a goodness-of-fit statistic G_j under non-random assignment can be obtained by selecting a single protocol suitable for manipulation in a Monte Carlo simulation, and using such a procedure to determine the probability distribution over various potential values of G_j . A simple non-random protocol would be analogous to the one described in section 2 above, combining a random component with a deterministic component based on observed characteristics X .

It should be clear at this point that this entire method rests on the presumption that selective assignment, if it takes place, is not purely selective on unobservable characteristics – as was the selection process simulated in section 2 above. While this may seem like a restrictive assumption, it is effectively invoked by every researcher supporting a claim of effective random assignment by pointing to balanced observable characteristics across treatment categories.

The final component of the procedure for inferring the probability of random assignment

is selecting a value for $P(RA)$, the prior probability of random assignment. This prior belief could be formed by surveying a sample of group administrators, or chosen arbitrarily. Prior beliefs may also be uniform across groups j or identical. Arbitrary changes in identical prior beliefs will scale the posterior probabilities in potentially nonlinear ways, but will not influence their rank-order.

Ultimately, recovery of accurate posterior probabilities of random assignment is an infeasible task, primarily because of their sensitivity to prior beliefs and the multiplicity of alternative assignment algorithms. The goal of this process, then, is not to recover exact probabilities, but to transform the information present in the distribution of observed characteristics across subgroups within group j into an ordinal measure of the likelihood that a random assignment protocol was used. As established in the preceding section, the technique of using noisy estimates of the probability of random assignment as interacted regressors yields at least some useful information about the direction of selection bias, and the resulting implications for estimates of true causal effects.

4. An Application to the Study of Peer Effects in North Carolina Elementary Schools

To illustrate this method, we now present an analysis of the impact of classroom composition on standardized test scores in reading and math for a sample of 5th grade students in North Carolina public schools.

4.1 Data

Our estimation employs a dataset recording information on every public school student in

the state of North Carolina between the 1994/95 and 2002/03 school years. Standardized end-of-grade tests in reading and math are administered annually to students in North Carolina beginning in 3rd grade and continuing until 8th grade. Each student test score record contains information on the school attended and the identity of the teacher administering the test. It is this teacher information that permits us to match students who share a classroom within a school. This method of sorting students into classrooms is only effective in elementary school, where students spend nearly all their time with the same group of peers and receive most of their instruction from a single teacher. For this reason, we focus our peer group analysis on the characteristics of a student's classmates and schoolmates in the 5th grade. Virtually all North Carolina school districts assign 5th grade students to elementary schools.

We form a set of basic classmate characteristic variables for each student in the dataset, including racial composition and measures of peer ability. Our primary peer ability variables use data on 4th grade standardized test performance for those classmates that appear in administrative datasets in consecutive years.⁵ Across the entire sample, we are able to link about 95% of fifth grade students to their fourth grade test scores.⁶ As in existing literature, we use the mean of this lagged achievement variable as a basic ability measure.⁷ We also introduce a measure of dispersion in peer ability, the standard deviation of classmate 4th grade test scores, to analyze potential nonlinearities in peer effects. In a specification controlling for mean peer achievement, the coefficient on the standard deviation indicates the impact of a mean-preserving spread in the

⁵ Test scores of students who are repeating the 5th grade are not used to compute peer characteristic measures. These students are in some cases included as observations in the analysis, however.

⁶ We have estimated specifications where peer ability measures are computed using 4th grade test scores, restricting the sample to those students with 3rd grade test scores available. The results of these specifications suggests that there is little if any bias associated with selective attrition in models that use 3rd grade test scores as measures of peer ability.

⁷ We omit a student's own lagged test score from the computation of mean classroom or grade level peer ability.

classroom ability level.

4.2 Basic empirical model

The analysis of education production functions, and of the impact of peer or classmate characteristics in particular, is hampered by a number of empirical difficulties. First among them, and the difficulty most directly addressed by the method outlined in sections 2 and 3 above, is the concern that students are in many cases not randomly assigned to classrooms. This is not the only concern, however; a basic outline of the empirical model will serve to illustrate the others.

Equation (8) is a conceptual relationship between the math or reading ability of student i , who attends classroom k in school j , at the end of school year t , Y_{ijkt} , to individual-level factors X_{it} , classroom factors X_{jkt} , peer characteristics including ability P_{jkt} , and an unobserved mean-zero component ϵ_{ijkt} .

$$(8) \quad Y_{ijkt} = \alpha + \delta Y_{ijkt-1} + \beta_1 X_{it} + \beta_2 X_{jkt} + \beta_3 P_{jkt} + \epsilon_{ijkt}$$

Notably, ability at the end of school year t is presumed to be influenced by ability at the beginning of the school year, Y_{ijkt-1} . The coefficient δ on the lagged dependent variable is expected to be less than one, reflecting the possibility of decay in ability over time, as well as accounting for measurement error. This is a so-called “value-added” specification of educational production. While the inclusion of a lagged dependent variable does introduce econometric difficulties, it also addresses concerns regarding serial correlation in the determinants of the year t increment to ability (Boardman and Murnane, 1979).

Non-random assignment introduces the possibility that peer characteristics and other

classroom factors correlate with unobserved determinants of ability at the end of a school year, subsumed into the error term ϵ_{ijkt} . Such a pattern might reflect the widely documented segregation of students into schools, and into classrooms within schools, so long as that segregation occurs along both observable and unobservable dimensions (see, for example, Clotfelter, Ladd and Vigdor 2003).

Beyond the issue of non-random assignment, there is the additional concern that ability is measured noisily with standardized test scores. Measurement error in the dependent variable, by itself, reduces the efficiency of econometric estimates but does not introduce any bias or inconsistency. Measurement error in the lagged dependent variable, if classical, is another rationale for expecting a coefficient less than one in this specification. If not classical, and particularly if there is serial correlation in measurement error, more severe problems may occur in this specification. Classmate test scores are commonly used as a peer characteristic of interest in education production functions; measurement error in these test scores produces attenuation bias (Arcidiacono et al. 2004). Thus, the estimates presented below which attempt to separate out the impact of selection from causal effects may understate the true impact of some peer characteristics when those characteristics are measured noisily.

Peer effect estimates may be upwardly biased, even under random assignment, when students are exposed to common factors that influence their performance on ability tests relative to their true ability. For example, students may share exposure to suboptimal testing conditions, or may suffer similarly when teachers omit instruction on a subject that ends up featuring prominently on the test. This is primarily a concern in studies that attempt to relate one student's performance at the end of year t to classmates' performance in year t . As noted by Manski

(1993), linear estimates of the impact of simultaneous peer achievement on individual achievement are not identified.

As in many existing studies of classmate characteristics, we avoid these concerns by using lagged measures of peer achievement instead of contemporaneous measures. In Manski's terminology, the effects we estimate will blend endogenous and exogenous social effects: that is, our estimates will combine the effects of peer background characteristics on ability with the direct effects of peer achievement on individual ability. It would be inappropriate to use our peer effect estimates to compute so-called "social multiplier" effects.

4.3 Computing the posterior probability of random assignment

Students sort nonrandomly into schools, and within schools may be further selected into classrooms. To address the concern of sorting across schools, we employ school fixed effects in our analysis. To address sorting within schools, we employ the method outlined in section 3 above to derive a posterior probability of random assignment for every elementary school/year observation in our dataset.

For each school and year, we fit a multinomial logit model for classroom assignments, using a number of observed student characteristics as predictors. The predictors include student race, gender, and participation in the Federal free and reduced price lunch subsidy program, a categorical indicator of parental education, an indicator for whether the student attended the same school in the previous year, and standardized test scores from the prior year. The pseudo- R^2 measure, defined in equation (6) above, was then used as a summary measure of the models' goodness of fit.

To translate this goodness-of-fit measure into a posterior probability of random assignment, we conducted two sets of Monte Carlo simulations to determine the distribution of pseudo- R^2 measures under random assignment and under a specific form of nonrandom assignment. In the random assignment simulation, we re-assigned students in each school to classrooms on the basis of random numbers drawn from a uniform distribution on the interval $[0,1]$. We arbitrarily ordered the classrooms in a school and assigned students to seats in classrooms on the basis of their rank in the randomized distribution. In the non-random assignment simulation, we followed a similar algorithm, with the exception that the random numbers drawn for a subset of students – those with college-educated parents – were multiplied by 0.5 prior to rank-ordering.

Figure 1 is a histogram showing the distribution of pseudo- R^2 measures obtained in the random assignment simulation. Figure 2 shows the distribution obtained in the non-random assignment simulation. The histograms are drawn using the same number and width of bins. The probability of obtaining any given pseudo- R^2 conditional on random assignment or non-random assignment is given by the relative height of bins in figures 1 and 2, respectively.

Figure 3 shows the distribution of pseudo- R^2 measures obtained through actual estimation of multinomial logit models for each school/year observation in the North Carolina data. Relative to the hypothetical distributions in Figures 1 and 2, the realized distribution represents an intermediate case. Observable student characteristics explain too much of the variation in student assignment to classrooms to accept the hypothesis that all schools employ randomization algorithms. But it is also clear that student assignment practices, on the whole, are not as extreme as the non-random algorithm employed in our second simulation.

Referring to equation (7) above, we require three pieces of information for each school/year observation in order to obtain a posterior probability of random assignment. The conditional probability of obtaining the school/year specific pseudo- R^2 measure under random and nonrandom assignment are taken from the distributions in figures 1 and 2, respectively. The prior probability of random assignment is set equal to 0.50 for each school. As noted above, when prior beliefs are uniform across schools, changing the common prior probability of random assignment will not influence the rank order of posterior probabilities.

Table 3 presents summary statistics for the students in our 5th grade sample, as well as statistics for the subset of students who attend schools in which the posterior probability of random assignment is greater than 50%. Several important patterns appear in these statistics. First, the set of schools with a higher posterior likelihood of random assignment are by several measures better off than the population as a whole. Standardized test scores are slightly higher, though equally well dispersed. The proportion of black students is lower, and parental education levels are slightly higher.

Second, these statistics provide straightforward evidence that students are not randomly assigned to classrooms. Under random assignment statewide, we would expect the standard deviation of test scores within each classroom to approximate the statewide level. Instead, the classroom level standard deviations are roughly 12% lower than the statewide statistic. As would be expected, the classroom-level standard deviation is higher in schools with a greater posterior likelihood of random assignment, but even among those schools with a posterior probability of 50% or greater the classroom-level standard deviation is 10% below the statewide level. Some of this discrepancy could reflect the fact that any individual school might serve a

less dispersed population than the state as a whole. In the estimation below, school fixed effects will account for any broad differences in the student population across schools.

Further evidence of sorting into classrooms can be gleaned from the standard deviation of peer mean test scores. Under random assignment, we would expect the standard deviation of the mean of classmate test scores to be approximately one divided by the square root of class size, or just over 0.2. Instead, the standard deviation is quite a bit higher – around 0.45 in the entire sample and 0.4 in the set of schools with higher posterior likelihoods of random assignment. Average classroom ability varies too much across classrooms to be consistent with random assignment. Again, some of this discrepancy could reflect across-school differences, which will be addressed with school fixed effects.

5. Results

Table 4 presents a set of baseline estimates of the relationship between peer characteristics and 5th grade achievement for all elementary schools in North Carolina. The estimates include all students who took 5th grade end-of-grade standardized tests in 1999, 2000 or 2001 except those with missing information for one or more covariates. Peer characteristics include racial and gender composition and two moments in the distribution of lagged test scores, the mean and standard deviation. All estimates presented in the table are derived from specifications that include school-by-year fixed effects, implying that the principal source of variation used to identify peer characteristic effects is that occurring across classrooms within schools.⁸ The specifications in columns 1 through 3 examine math test scores, while those in columns 4

⁸ In an earlier paper (Vigdor and Nechyba, 2004), we present estimates using no fixed effects and school fixed effects only. Results in that paper are generally comparable to those found here. We also find that the relationship between school-level peer characteristics and individual outcomes is negligible after controlling for classroom-level peer characteristics.

through 6 analyze reading test scores. In each set of three regressions, the first specification restricts covariates to a parsimonious set of peer characteristics, the second specification adds a classroom input measure widely believed to correlate with student achievement, teacher experience (Clotfelter, Ladd and Vigdor 2004), and the last incorporates additional student-level covariates and interactions between certain student and peer group characteristics.

Across specifications, estimates of the relationship between current and lagged ability are quite precisely estimated with coefficients on the order of 0.8. As discussed in section 2 above, the significant gap between these coefficients and unity might reflect either the impact of measurement error in achievement test scores or depreciation of ability as a form of capital.

For both math and reading test scores, mean lagged peer test scores correlate significantly with individual achievement in 5th grade, holding other factors constant. In both cases, the magnitude of the estimated effect is reduced slightly by the inclusion of teacher experience controls, and more substantially by the inclusion of student characteristics, including parental education estimates. This diminution of the estimate effect is consistent with the tendency for higher achieving students to sort into particular classrooms within schools, notably those served by more experienced teachers (Clotfelter, Ladd and Vigdor 2004). Even in complete models, however, estimates indicate that increasing classmates' prior test scores by an average of 0.4 standard deviations (the standard deviation of mean peer achievement) leads to an increase in own achievement of between 3 and 4 percent of a standard deviation.

Several other peer characteristics associate with achievement, particularly in the specifications that control for the broadest set of individual characteristics and teacher experience. Both math and reading test scores tend to be higher in classrooms where the

distribution of prior achievement levels is more dispersed, which suggests that the highest mean test scores are achieved when students are integrated rather than segregated by ability level. Overall achievement levels also tend to be lower in classrooms that serve a higher fraction of black or Hispanic students, relative to other classrooms in the same school. The specifications in columns (3) and (6) indicate that the apparent negative effects of black or Hispanic classmates are attenuated for members of the racial or ethnic group. Hispanic students' performance on reading tests actually appears higher when their group forms a higher proportion of the class. This pattern might reflect economies of scale in providing reading instruction to non-native English speakers.⁹ Finally, test scores tend to be higher in classrooms with a higher share of female students; this tendency is most pronounced (and most significant in a statistical sense) in models that control for individual student characteristics.

To what extent do these within-school across-classroom estimates reflect true causal effects, and to what extent are they spurious, driven by a tendency for students with better unobserved characteristics to sort into specific types of classroom? Before discussing results, it is worth reviewing basic hypothesized mechanisms linking classroom composition to student achievement (for a fuller discussion, see Hoxby and Weingarth 200x). Educational production may vary with classroom composition because peers themselves are direct inputs into the education production function, because teacher time allocation varies according to classroom composition, or because educational inputs including teachers themselves are distributed in part according to student characteristics. Highly able students may participate in the instruction of peers; conversely, less able peers may serve as a distraction (Jacob and Vigdor 2005; Lazear

⁹ North Carolina's Hispanic population consists almost entirely of first generation immigrants and their children. Between 1990 and 2000 the state's Hispanic population quintupled.

2001). Teachers may allocate their time unevenly in classrooms with heterogeneous ability levels, or may alter their instructional style to best serve particular types of students. Finally, substantial amounts of research have established that experienced teachers gravitate toward schools serving more advantaged students, implying that peer characteristics may proxy for unobserved aspects of teacher quality (or quality of other classroom resources). This last causal mechanism is less likely to be important in a within-school estimation, unless elementary schools can commit to offer some teachers consistently superior classrooms' worth of students. Such commitments are least likely to be entered into in schools practicing random assignment protocols. To the extent that the estimates presented here can be referred to as causal, they should be interpreted as a reflection of the partial-equilibrium impact of changing classroom assignment patterns within schools. They do not necessarily fully reflect the general-equilibrium impacts of reallocating students across schools, which could in turn lead to resorting in the teacher labor market.

With this caveat in mind, Table 5 presents the results of our basic strategy for partitioning the estimated relationship between peer characteristics and student achievement into selection and treatment effects. The table repeats regression results from the most complete models in Table 4 for purposes of comparison. To review, these specifications introduce (one minus) the posterior probability of random assignment as a regressor, included as a main effect and interacted with the contextual variables of interest. Under assumptions outlined above, coefficients on the un-interacted contextual variables can be interpreted as treatment effects; coefficients on the interactions between the contextual variables and one minus the probability of random assignment can be interpreted as selection effects.

In ten out of 14 cases, the main effects of peer characteristics are either closer to zero in the interacted model, or of opposite sign. As foreshadowed by the Monte Carlo simulation, standard errors are also substantially larger in the interacted models. Whereas the basic fixed effect models indicated that classmate race and gender were significant predictors of student achievement, the interacted models ascribe statistical significance only to the mean and standard deviation of peer lagged test scores. The effect of classmate mean lagged test scores is reduced by almost half in magnitude; the effect of the classmate standard deviation is attenuated only slightly in the math specification but by a more substantial amount in the reading model.

In the analysis of math test scores, the effects of peer race and gender composition are uniformly closer to zero or of opposite sign in the interacted model, indicating that a substantial portion of the apparent effects in Table 4 reflect selection on unobservables into classrooms with a higher proportion of female or white students. While the interaction terms themselves fail to attain statistical significance, their signs are uniformly consistent with this interpretation.

In reading, the interacted models also suggest that classroom gender composition effects are at least to some extent spurious. The effects of racial composition are more complicated. The point estimate on classmate percent black actually becomes more negative, suggesting that students selecting into classrooms with more black students have poorer unobserved math skills but better reading skills. A more striking pattern appears with regard to percent Hispanic. The effects of Hispanic student density on non-Hispanic students appear more strongly negative in the interacted model, but the effects are reversed for Hispanic students themselves. Altogether, the results indicate that the reading performance of Hispanic 5th grade students is substantially improved when they share a classroom with same-race peers, but the tendency for students with

lower unobserved ability to congregate in such classrooms masks the effect.

The Monte Carlo simulation in section 2 above indicates that estimators using the probability of random assignment as a discrete or continuous weight tend to be more biased than the preferred interaction-based estimator, but can still provide information on the direction of selection bias while imposing a lesser efficiency cost. Table 6 presents implementations of weighted estimators, using either a discrete cutoff in the posterior probability of random assignment or the probability itself as a weight. These results generally corroborate the patterns in Table 5. The relationship between classmates' mean lagged test score and achievement is attenuated relative to OLS. The degree of attenuation is smaller when examining the standard deviation of peer lagged achievement, suggesting that students with better unobserved determinants of achievement select into classrooms with higher means but not necessarily greater or lower dispersion. The estimates of the effect of classmate racial composition are close to the OLS estimates, consistent with the fact that the preferred estimator exhibits less bias than these weighted estimators. The effect of classmate gender composition, as in the interacted models, is attenuated toward zero. These estimates thus support most of the conclusions derived from Table 5.

The question of whether elementary students are better off in ability-integrated or stratified classrooms is of significant policy interest. Estimates presented to this point suggest consistently that mean-preserving spreads in the prior achievement distribution lead to improvements in average achievement. The last two tables probe this pattern further, to determine whether it is robust to changes in the parameterization of ability dispersion and whether the effect is heterogeneous across students.

Table 7 shows the results of replacing the peer standard deviation with dispersion measures based on quantiles in the peer lagged test score distribution. The first and third columns show estimates that control for the mean, 90th percentile, and 10th percentile in the lagged test score distributions. These estimates need to be interpreted cautiously; the coefficient on the 90th percentile, for example, illustrates the effect of increasing the 90th percentile while holding both the mean and 10th percentile constant. Implicitly, then, any increase in the 90th percentile must be accompanied by an offsetting shift in the test score distribution that negates the impact on the mean, while leaving the 10th percentile unchanged.¹⁰

With this caveat in mind, the results show a pattern consistent with those in earlier tables. The coefficients on peer 90th percentile are larger than those on the 10th percentile, implying that a mean-preserving spread in the peer ability test score distribution increases average achievement. The effect is stronger in math than in reading, consistent with the results in Table 5 above. The results also show very little evidence of selection into classrooms on the basis of peer 10th or 90th percentile, after permitting selection on the mean. Note that the effect of mean peer ability is estimated to be negative in both math and reading specifications. Technically, this coefficient needs to be interpreted as the effect of an increase in the mean while holding the 90th and 10th percentiles constant. This is not an intuitive interpretation.

The second and fourth columns impose a parameter restriction on the original models, forcing the effect of an decrease in the 10th percentile to be equal to the effect of an increase in the 90th percentile. Students are projected to perform better on math tests in classrooms where the 90-10 differential is wider. The effect is opposite-signed in reading test scores; evidence of

¹⁰ Similar concerns can be attributed to Hoxby and Weingarth's peer effect estimates, which control for the proportion of classmates in ten test score deciles as well as the mean. Theoretically, it should not be possible to increase the share of classmates in a decile while holding the share in nine other deciles and the mean constant.

dispersion effects in reading has been weaker in most specifications shown to this point. Both math and reading specifications indicate that increasing mean peer achievement, holding dispersion constant, improves individual performance.

The clearest theoretical argument for positive dispersion effects is that highly able students can improve the performance of their peers, perhaps by serving as a type of teacher's assistant. The specifications in Table 8 test for such a pattern, which would imply that the benefits of dispersion accrue primarily to students at the lower end of the test score distribution. By including interaction terms with one minus the posterior probability of random assignment, these specifications also allow for differential selection into classrooms with varying mean achievement or dispersion in achievement.

The main effects in these specifications indicate the impact of increasing peer lagged achievement, or the dispersion in lagged achievement, for a student with test scores equal to the statewide mean. The main effects indicate that such a student benefits slightly, and in the case of math insignificantly, from an increase in mean peer achievement. These students fare substantially better, however, when assigned to classrooms with more dispersed ability distributions. For such a student, switching from a classroom with ability dispersion equal to 0.9 to one where the dispersion equals the statewide level (1.0) is forecast to improve math performance by 1.6 percent of a standard deviation, and reading performance by 0.8 percent of a standard deviation. Interactions with one minus the posterior probability of random assignment indicate that students with average lagged test scores tend to be positively selected into classrooms with higher mean peer ability, and negatively selected into more dispersed classrooms. In other words, above average students who post test scores at the state average are

more likely to be placed in classrooms with higher performing peers, and with lower degrees of lagged test score dispersion.

The effects of mean peer ability are inconsistently heterogeneous across students. Estimates indicate that the math performance of more able students is more sensitive to peer ability, but reading performance is not. These across-subject differences may reflect the differential technology of instruction, with reading more likely to be ability-stratified in all classrooms.

As predicted, the effects of peer ability dispersion are more positive for students at the lower end of the achievement distribution. While dispersion has a strong positive impact on the achievement of the mean student, the effect is close to zero or negative for students one standard deviation or more above the statewide average.

Three-way interaction terms indicate that the phenomenon of positive selection on unobservables into classrooms with high mean peer ability attenuates as observed ability increases. This pattern might arise if, for example, certain classes were reserved for high ability students, and among those with average or low observed ability only those with strong unobservables were admitted. The pattern of negative selection into classrooms with high degrees of dispersion increases as own ability increases. Intuitively, among those students with very high ability, those found in low-dispersion (and hence uniformly high ability) classrooms tend to have better unobserved determinants of achievement than those who sort into more representative classrooms. At the other end of the spectrum, among low-observed-ability students, those with better unobservables sort into representative classrooms and those with poor unobservables are more likely to be in homogeneous low-ability classrooms. These selection patterns are entirely plausible.

6. Conclusions

The methods presented in this paper provide a potential means to disentangle selection and treatment effects in non-experimental data, so long as subjects in at least some cases are assigned to treatment groups in a method that approximates random assignment. The empirical application to peer effects in North Carolina public elementary schools shows evidence of some significant and policy-relevant causal effects, along with pervasive selection patterns that in some cases amplify and in others obscure those effects.

The results indicate that the gains to less-able students from exposure to higher-ability peers outweigh any losses from ability mixing at the high end, implying that overall mean test scores are maximized when classrooms serve heterogeneous groups of students. While the mean test score is not necessarily the most important summary measure from a social welfare perspective, the results do indicate that any instructional benefits associated with delivering ability-level-appropriate content are outweighed by the benefits of exposure to heterogeneity at the low end of the distribution. In turn, the results suggest that increasing judicial and public opposition to school busing will impose costs on society, even without considering the general equilibrium impacts flowing through the teacher labor market.

Results further indicate that most of the observed correlations between classmate gender and race composition and individual achievement are tainted by selection bias. The one prominent exception regards the reading performance of Hispanic students in classrooms with a high proportion of Hispanics. In spite of the fact that Hispanic students in such classrooms are negatively selected, their performance benefits. This may occur because there are economies of

scale in teaching reading to non-native English speakers, or because Hispanic students benefit from being able to communicate with one another in an environment where their native language is not the accepted mode of discourse.

In principle, the methods introduced here could be used in many alternative situations. Additional examples using administrative education data, where there is potential variation in assignment protocols across schools or districts, are numerous. Using administrative data on health outcomes, one could exploit potential variation in assignment protocols to physicians or treatments across hospitals or other facilities. Using administrative data on court records, one could exploit variation in assignment protocols to correctional facilities, probation officers, or other treatments. Advocating the use of random assignment more consistently for purposes of evaluation (and in some cases fairness) might improve econometricians' ability to estimate causal effects in the future, but these methods promise to make useful causal attributions using past and present data.

References

- Angrist, J. and V. Lavy (1999) "Using Maimonides' Rule To Estimate The Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* v.114 pp.533-575.
- Arcidiacono, P. and S. Nicholson (2001) "Peer Effects in Medical School." Forthcoming, *Journal of Public Economics*.
- Bettinger, E. (2008)
- Betts, J.R. and A. Zau (2002) "Peer Groups and Academic Achievement: Panel Evidence from Administrative Data." Unpublished manuscript.
- Black, S. (1999) "Do Better Schools Matter? Parental Valuation of Elementary Education." *Quarterly Journal of Economics* v.114 pp.577-600.
- Boardman, A.E. and R.J.Murnane (1979) "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement." *Sociology of Education* v.52 pp.113-121.
- Boozer, M.A. and S.E. Cacciola (2001) "Inside the 'Black Box' of Project STAR: Estimation of Peer Effects Using Experimental Data." Unpublished manuscript.
- Caldas, S. J. and C. Bankston (1997). "Effect of School Population Socioeconomic Status on Individual Academic Achievement." *Journal of Educational Research* v.90 pp.269-277.
- Clotfelter, C.T., E. Glennie, H.F. Ladd and J.L. Vigdor (2008) "Would Higher Salaries Keep Teachers in High-Poverty Schools? Evidence from a Policy Intervention in North Carolina." *Journal of Public Economics* v.92 pp.1352-1370.
- Clotfelter, C.T., H.F. Ladd and J.L. Vigdor (2004) "Teacher Sorting, Teacher Shopping, and the Assessment of Teacher Effectiveness." Duke University manuscript.
- Clotfelter, C.T., H.F. Ladd and J.L. Vigdor (2003) "Segregation and Resegregation in North Carolina's Public School Classrooms." *North Carolina Law Review* v.81 pp.1463-1511.
- Duncan, G., J. Connell, and P. Klebanov (1997) "Conceptual and Methodological Issues in Estimating Causal Effects of Neighborhoods and Family Conditions on Individual Development." in J. Brooks-Gunn, G. Duncan, and J. Aber, eds., *Neighborhood Poverty: Volume 1*. New York: Russell Sage Foundation.
- Epple, D., E. Newlon and R. Romano (2002) "Ability Tracking, School Competition, and the Distribution of Educational Benefits." *Journal of Public Economics* v.83 pp.1-48.

- Epple, D. and R. Romano (1998). .Competition Between Private and Public Schools, Vouchers, and Peer Group Effects.. *American Economic Review* 88(1), 33-62.
- Ferreira, Maria. .Estimating a General Equilibrium Model with Multiple Jurisdictions and Private Schools.. Carnegie Mellon University working paper, 2002.
- Gaviria, A. and S. Raphael (2001). "School-Based Peer Effects and Juvenile Behavior." *Review of Economics and Statistics*, v.83 pp.257-268.
- Hanushek, E., J. Kain, J. Markman and S. Rivkin. "Does Peer Ability Affect Student Achievement?" NBER Working Paper #8502, 2001.
- Hoxby, C. "Peer Effects in the Classroom: Learning from Gender and Race Variation." NBER Working Paper #7867, 2000.
- Hoxby, C., and G. Weingarth (2006). "Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects." Unpublished manuscript.
- Kane, T.J. and D.O. Staiger (2002) "Volatility in School Test Scores: Implications for Test-Based Accountability Systems." *Brookings Papers on Education Policy* pp.235-283.
- Katz, L., J. Kling and J. Liebman (2001) "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment." *Quarterly Journal of Economics* v. 116 pp.607-654
- Lazear, E.P. (2001) "Educational Production." *Quarterly Journal of Economics* v.116 pp.777-803.
- Link, C. R. and J. G. Mulligan (1991). "Classmates' Effects on Black Student Achievement in Public School Classrooms." *Economics of Education Review* v.10 pp.297-310.
- Ludwig, J., G. Duncan and P. Hirschfield (2001) "Urban Poverty and Juvenile Crime: Evidence from a Randomized Housing-Mobility Experiment." *Quarterly Journal of Economics* v.116 pp.655-680.
- Ludwig, J., H. Ladd and G. Duncan (2001) "The Effects of Urban Poverty on Educational Outcomes: Evidence from a Randomized Experiment." Unpublished working paper.
- Ludwig, J., G. Duncan and Pinkston (2004) "Neighborhood Effects on Economic Self-Sufficiency: Evidence from a Randomized Housing-Mobility Experiment." *Journal of Public Economics* v. pp.
- Manski, C.F. (1993) "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies*, v.60 pp.531-542.

- Moffitt, R. A. (1998). "Policy Interventions, Low-Level Equilibria and Social Interactions." in S. Durlauf and H.P. Young, eds., *Social Dynamics*. Cambridge: MIT Press.
- Nechyba, T. (2000). "Mobility, Targeting and Private School Vouchers." *American Economic Review*, v.90 pp.130-46.
- Nechyba, T., D. Older-Aguilar and P. McEwan (1999) "The Effect of Family and Community Resources on Education Outcomes." Unpublished manuscript.
- Robertson, D. and J. Symons (1996) "Do Peer Groups Matter? Peer Group versus Schooling Effects on Academic Attainment." unpublished manuscript, London School of Economics Centre for Economic Performance.
- Sacerdote, B. (2001) "Peer Effects with Random Assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics*, v.116 pp.681-704.
- Slavin, R. E. (1987). "Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis." *Review of Educational Research* v.57 pp.293-336.
- Slavin, R. E. (1990). "Achievement Effects of Ability Grouping in Secondary Schools: A Best-Evidence Synthesis." *Review of Educational Research* v.60 pp.471-499.
- Vigdor, J.L. and T.S. Nechyba (2004) "Peer Effects in North Carolina Public Schools." Kennedy School of Government Program on Education Policy and Governance Research Paper #04-20.
- Zimmer, R. W. and E. F. Toma (1999). "Peer Effects in Public Schools Across Countries." *Journal of Policy Analysis and Management*, v.19 pp.75-92.

Table 1: Monte Carlo results

Estimator	2.5 th percentile	Median	97.5 th percentile
OLS-Fixed Effects	1.426	1.462	1.493
P(Random Assignment) used as interaction term	0.905	1.029	1.112
P(Random Assignment) used to restrict sample	1.1581	1.231	1.304
P(Random Assignment) used as a weight	1.271	1.328	1.372

Note: Results taken from 200 replications of a Monte Carlo simulation described in the text.

Table 2: Using Noisy Estimates of P(Random Assignment)

Estimator	2.5 th percentile	Median	97.5 th percentile
P(Random Assignment) used as interaction term	0.978	1.079	1.188
P(Random Assignment) used to restrict sample	1.254	1.323	1.395
P(Random Assignment) used as a weight	1.357	1.401	1.445

Note: Results taken from 200 replications of a Monte Carlo simulation described in the text.

Table 3: Summary statistics for 5th grade students in North Carolina public schools

Variable	Entire sample		Posterior probability above 0.50	
	Mean	Std. Dev.	Mean	Std. Dev.
Math standardized test score	0.024	0.989	0.061	0.987
Reading standardized test score	0.024	0.988	0.059	0.985
Male	0.503	---	0.504	---
Black	0.310	---	0.263	---
Hispanic	0.024	---	0.022	---
Parent's education (as reported by previous year's teacher):				
high school	0.443	---	0.432	---
trade/business school	0.054	---	0.054	---
community/tech. college	0.132	---	0.134	---
4-year college	0.191	---	0.201	---
graduate school	0.074	---	0.075	---
Teacher experience:				
1-2 years	0.116	---	0.114	---
3-5 years	0.133	---	0.131	---
6-12 years	0.197	---	0.201	---
13-20 years	0.180	---	0.181	---
21-27 years	0.206	---	0.209	---
27+ years	0.096	---	0.093	---
Average 4 th grade score of classmates				
Math	0.028	0.465	0.069	0.410
Reading	0.028	0.447	0.066	0.390
Std. dev. in 4 th grade scores of classmates:				
Math	0.876	0.200	0.896	0.193
Reading	0.885	0.203	0.906	0.197
% black among classmates	0.293	0.253	0.251	0.232
% Hispanic among classmates	0.039	0.063	0.035	0.057
% female among classmates	0.494	0.095	0.493	0.088
Class size	23.528	3.858	24.114	3.373
N	624,898		352,901	

Table 4: OLS-fixed effect estimates using the full sample of North Carolina elementary schools

Independent variable	Dependent variable					
	Math test scores (5 th grade)			Reading test scores (5 th grade)		
4 th grade test score	0.833*** [0.004]	0.832*** [0.004]	0.784*** [0.005]	0.811*** [0.005]	0.811*** [0.005]	0.760*** [0.005]
Mean 4 th grade test score of 5 th grade classmates	0.106*** [0.009]	0.102*** [0.009]	0.084*** [0.008]	0.102*** [0.010]	0.099*** [0.010]	0.078*** [0.009]
Std. Dev. Of 4 th grade test score of 5 th grade classmates	0.167*** [0.026]	0.162*** [0.026]	0.149*** [0.026]	0.139*** [0.031]	0.136*** [0.032]	0.117*** [0.030]
% black among classmates	-0.046* [0.026]	-0.046* [0.026]	-0.117*** [0.026]	0.032 [0.024]	0.031 [0.024]	-0.042* [0.024]
% Hispanic among classmates	-0.102** [0.044]	-0.095** [0.044]	-0.088** [0.044]	-0.074** [0.037]	-0.071* [0.037]	-0.065* [0.036]
% Female among classmates	0.024 [0.020]	0.030 [0.020]	0.044** [0.021]	0.040** [0.017]	0.044*** [0.017]	0.067*** [0.017]
Class size	---	---	0.0004 [0.001]	---	---	0.001 [0.001]
Male	---	---	-0.014*** [0.002]	---	---	-0.028*** [0.002]
Black	---	---	-0.112*** [0.006]	---	---	-0.126*** [0.006]
%black*black	---	---	0.033** [0.013]	---	---	-0.0003 [0.013]
Hispanic	---	---	0.005 [0.009]	---	---	0.030*** [0.009]
% Hispanic*Hispanic	---	---	0.025 [0.089]	---	---	0.201** [0.098]
Teacher experience controls	No	Yes	Yes	No	Yes	Yes
Parent education controls	No	No	Yes	No	No	Yes
Observations	356,007	356,007	356,007	354,456	354,456	354,456
R-squared	0.746	0.746	0.754	0.705	0.705	0.714

Note: Standard errors, corrected for within-classroom clustering, in square brackets. All specifications include school-by-year fixed effects. Sample includes all 5th grade students with non-missing values for regression covariates.

*** denotes a coefficient significant at the 1% level; ** the 5% level; * the 10% level.

Table 5: Separating selection from treatment effects using the posterior probability of random assignment

Independent variable	Math		Reading	
	OLS/Fixed Effects	Posterior probability interactions	OLS/Fixed Effects	Posterior probability interactions
4 th grade test score	0.784*** [0.005]	0.785*** [0.005]	0.760*** [0.005]	0.760*** (0.005)
Mean 4 th grade test score of 5 th grade classmates	0.084*** [0.008]	0.045** [0.016]	0.078*** [0.009]	0.045*** [0.017]
Std. Dev. Of 4 th grade test score of 5 th grade classmates	0.149*** [0.026]	0.136*** [0.158]	0.117*** [0.030]	0.063* [0.038]
% black among classmates	-0.117*** [0.026]	-0.077 [0.048]	-0.042* [0.024]	-0.069* [0.039]
% black * black	0.033** [0.0130]	-0.065 [0.084]	-3.0*10 ⁻⁴ [0.013]	-0.017 [0.024]
% Hispanic among classmates	-0.088** [0.044]	-0.084 [0.086]	-0.065* [0.036]	-0.100 [0.267]
% hispanic* hispanic	0.025 [0.089]	0.010 [0.190]	0.201** [0.098]	0.496*** [0.175]
% Female among classmates	0.044** [0.021]	0.035 [0.038]	0.067*** [0.017]	0.040 [0.031]
Mean classmate test score*(1-P(RA))	---	0.062 [0.025]	---	0.059** [0.028]
Std. Dev. Of classmate test scores*(1-P(RA))	---	0.031 [0.093]	---	0.120 [0.103]
% black*(1-P(RA))	---	-0.065 [0.084]	---	0.060 [0.079]
% black*black*(1-P(RA))	---	0.040 [0.042]	---	0.022 [0.042]
% Hispanic*(1-P(RA))	---	-0.001 [0.142]	---	-0.054 [0.117]
% Hispanic*Hispanic*(1-P(RA))	---	0.062 [0.287]	---	-0.547 []
% Female*(1-P(RA))	---	0.015 [0.063]	---	0.053 [0.055]
Observations	356,007	349905	354,456	348373
R ²	0.754	0.754	0.714	0.714

Note: Specifications include all Table 4 covariates, including teacher experience and parental education controls, as well as school/year fixed effects.

*** denotes a coefficient significant at the 1% level; ** the 5% level; * the 10% level.

Table 6: Alternative uses of P(Random Assignment)

Independent variable	P(Random Assignment)>0.75 sample restriction		P(Random Assignment) as weight	
	Math	Reading	Math	Reading
Mean 4 th grade test score of 5 th grade classmates	0.053*** [0.015]	0.045** [0.015]	0.070*** [0.009]	0.066*** [0.009]
Std. Dev. Of 4 th grade test score of 5 th grade classmates	0.153*** [0.030]	0.097*** [0.031]	0.145*** [0.021]	0.093*** [0.023]
% black among classmates	-0.122*** [0.047]	-0.077** [0.036]	-0.100*** [0.028]	-0.049** [0.022]
% black*black	0.008 [0.023]	-0.017 [0.023]	0.026* [0.015]	-0.006 [0.015]
% Hispanic among classmates	-0.097 [0.085]	0.006 [0.064]	-0.090* [0.052]	-0.058 [0.041]
%Hispanic*Hispanic	-0.093 [0.191]	0.361** [0.162]	0.022 [0.114]	0.326*** [0.109]
% Female among classmates	0.0027 [0.038]	0.029 [0.030]	0.036 [0.024]	0.056*** [0.019]
Observations	143008	142522	349869	338337
R ²	0.75	0.706	0.75	0.71

Note: Specifications include all Table 4 covariates, including teacher experience and parental education controls, as well as school/year fixed effects.

*** denotes a coefficient significant at the 1% level; ** the 5% level; * the 10% level.

Table 7: Alternative parameterizations of dispersion in the test score distribution

	Math		Reading	
Mean 4 th grade test score of 5 th grade classmates	-0.052** [0.021]	0.032* [0.018]	-0.037 [0.022]	0.033* [0.017]
90 th percentile of peer test scores	0.104*** [0.011]	---	0.062*** [0.010]	---
10 th percentile of peer test scores	0.030*** [0.008]	---	0.046*** [0.008]	---
90-10 differential in peer test scores	---	0.016*** [0.006]	---	-0.013*** [0.005]
Peer mean*(1-P(RA))	0.044 [0.051]	0.058* [0.034]	-0.009 [0.058]	0.045 [0.039]
Peer 90 th percentile*(1-P(RA))	-0.017 [0.026]	---	0.057** [0.027]	---
Peer 10 th percentile*(1-P(RA))	0.016 [0.019]	---	0.008 [0.020]	---
90-10 differential*(1-P(RA))	---	-0.016 [0.011]	---	0.016* [0.009]
Observations	349,855	349,855	348,317	348,317
R ²	0.754	0.753	0.714	0.714

Note:

Table 8: Who Benefits from Higher Means or Greater Dispersion?

Independent variable	Math	Reading
Mean 4 th grade test score of 5 th grade classmates	0.023 [0.014]	0.027* [0.015]
Mean 4 th grade test score of 5 th grade classmates*own 4 th grade test score	0.032*** [0.008]	-0.021 [0.013]
Standard Deviation of 4 th grade test scores of 5 th grade classmates	0.159*** [0.020]	0.084*** [0.021]
Standard Deviation of 4 th grade test scores of 5 th grade classmates*own 4 th grade test score	-0.154*** [0.019]	-0.142*** [0.025]
Mean*(1-P(RA))	0.072*** [0.022]	0.025 [0.027]
Mean*Own*(1-P(RA))	-0.029* [0.017]	-0.005 [0.021]
SD*(1-P(RA))	-0.043*** [0.005]	-0.031*** [0.010]
SD*Own*(1-P(RA))	-0.184*** [0.049]	-0.176*** [0.049]
Observations	349,872	348,373
R ²	0.759	0.720

Note:

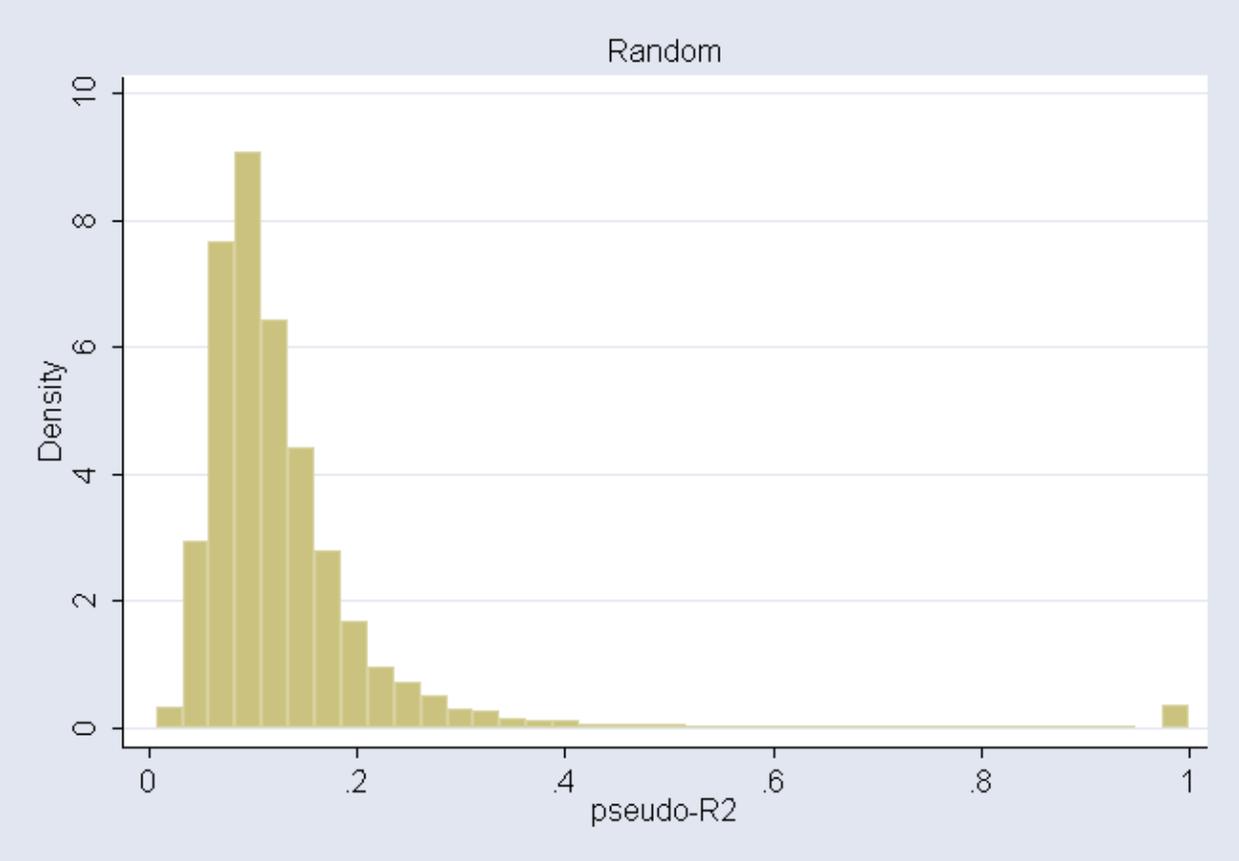


Figure 1: the distribution of pseudo-R-squared measures under random assignment.

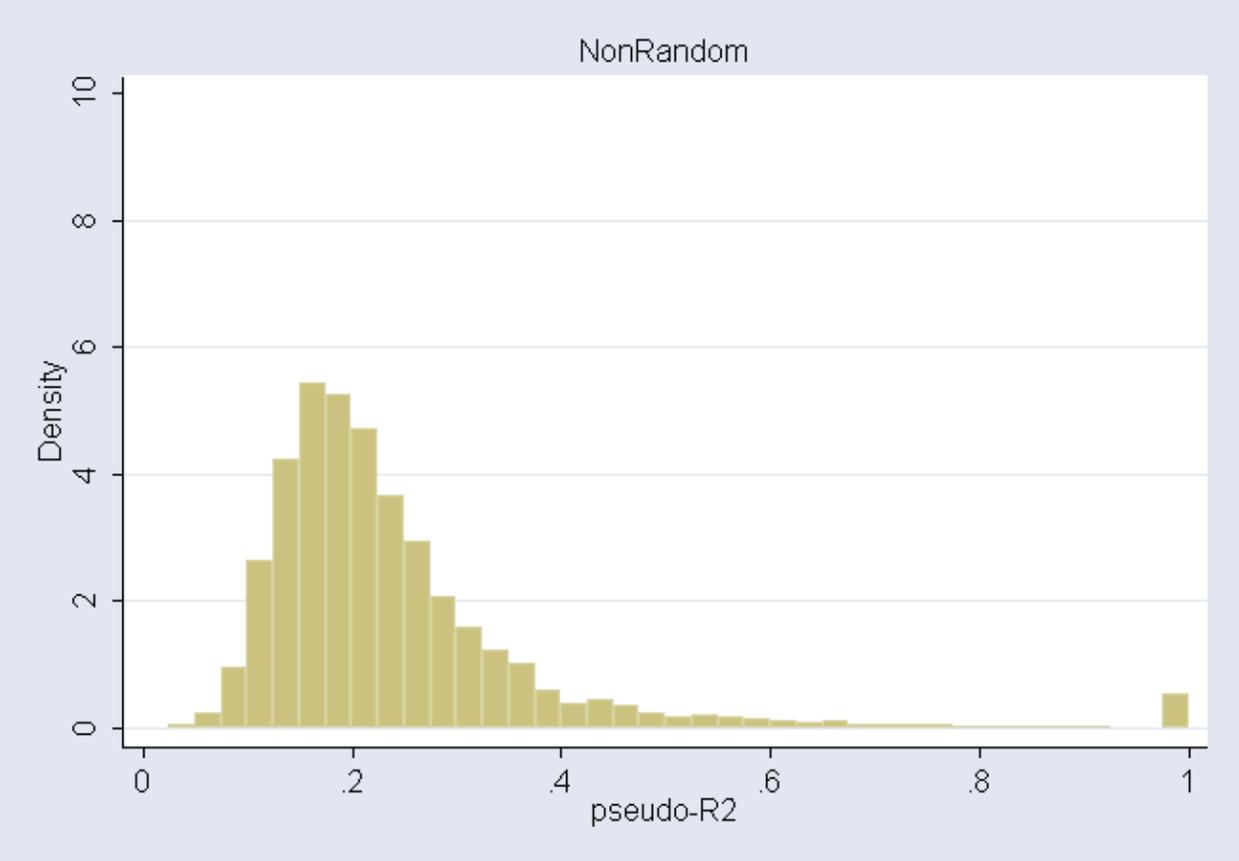


Figure 2: the distribution of pseudo-R-squared measures under a particular form of non-random assignment.

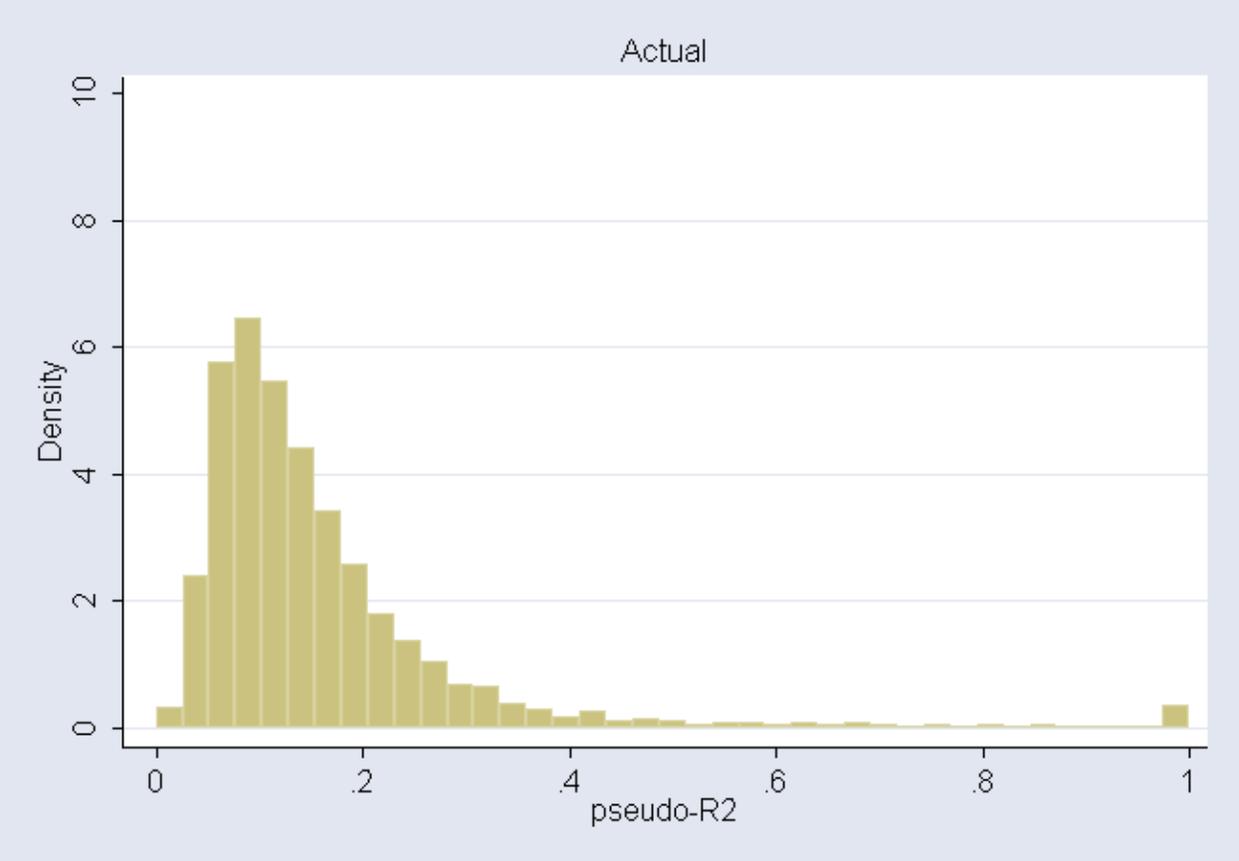


Figure 3: Observed distribution of pseudo-R-squared measures, North Carolina public elementary schools.