Economic Analyses of Elementary Education in the United States: Three Essays

A Dissertation
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Brandon Trampe

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Judy Temple

June 2014

Acknowledgments

I am extremely appreciative of the guidance that my advisor, Judy Temple, has provided to me throughout the dissertation process. Members of my committee, Elizabeth Davis, Arthur Rolnick, and Aaron Sojourner, have graciously provided invaluable feedback along the way.

I am also grateful for the unwavering support of my wife and family throughout the graduate school journey. Finally, I would like to thank my good friend Jeff Moriarty for his assistance during earlier stages of my academic pursuits.

Dedication

To my parents

Abstract

This dissertation investigates three substantial issues in education policy. First, it examines whether or not unobserved student heterogeneity introduces bias into value-added models of teacher effectiveness. After developing a statistical test, the results indicate that the omission of typically unmeasured information on noncognitive skills generates statistically significant bias and results in the misclassification of teachers according to their value-added quintile.

Next, the dissertation explores the characteristics upon which students are assigned to teachers within schools. In contract to prior studies that analyzed data from limited geographic areas, this research employs a nationally-representative data set for the first time. While the analysis largely confirms earlier studies, I find that some of the matching documented in earlier studies, such as on prior test scores, might be an artifact of the chosen model, and that the sorting might instead take place on variables that are correlated with test scores.

Finally, the dissertation critically examines prior research on class size reduction in elementary grades and generates estimates of future labor market benefits based on several later outcomes. Though estimates of labor market benefits based on ACT scores are higher than the costs of class size reduction, estimates generated using other outcomes almost exclusively are not.

Table of Contents

List of Tables

Chapter 2 Tables

Chapter 3 Tables

Chapter 4 Tables

List of Figures

Chapter 2 Figures

Chapter 1

Introduction

Education has played a key role in economic philosophy from the very beginning. While Adam Smith, the father of modern capitalist economics, generally was against government intervention, he believed that the government should play a role in the promotion of education. At the same time, he demonstrated uncanny foresight in pinpointing some of the issues that might arise. Smith found the university salary system to be problematic, noting the disconnect between teachers' incentives and their compensation. Given that teachers would be paid the same amount regardless of their performance, self-interested teachers would maximize their utility by investing as little time as possible in teaching (Smith 1776).

Over two-hundred years later, the disconnect between pay and performance still plagues the United States' public education system. Teachers are generally paid according to their years of experience, license, and attained degrees, yet these factors seem to be only loosely related to student achievement gains (Corcoran and Goldhaber 2013). Viewed in the context of an education production function that relates family, school, and teacher inputs to academic achievement, such an arrangement could be viewed as economically inefficient (Hanushek 1986). One solution to the problem may be found in value added models (VAMs), which attempt to pinpoint exactly how much each teacher contributes to the education production process. If such measures are found to be reliable, then teachers could be paid according to their value. Unfortunately, the reliability of these models is still in question. The second chapter of this dissertation contributes to the existing literature on the topic both by providing a statistical test for bias in VAMs and by showing that such models are biased by the omission of students' typically unobserved noncognitive skills.

Related to the idea of VAMs is the question of upon which characteristics students are assigned to teachers within schools. The third chapter attempts to inves-

tigate the possibility of purposeful sorting of students to teachers versus the null hypothesis of random assignment. While VAMs are able to control for observed characteristics, any systematic matching of students to teachers based on unobserved characteristics can introduce bias into estimates, as demonstrated in the second chapter. Thus, an understanding of the sorting process can be an important aid to further research. The ramifications of teacher-student sorting are not limited to bias in VAM estimates, however, as they also provide insight into race dynamics within schools, for example. Further, the existence of sorting can cloud research on other important topics as well. If less-experienced teachers are assigned smaller class sizes, as this chapter finds, and if new teachers are less effective than more-experienced teachers, then the sorting based on experience might confound analyses of class size reduction.

Unobserved teacher differences are only one reason why studying class size reduction can be difficult. Other studies have noted a decrease in teacher quality corresponding to class size reduction initiatives, as the additional teachers schools must hire are typically less effective than the teachers they already employ (Jepsen and Rivkin 2009). Due in part to related problems, researchers have turned to the only modern randomized class size experiment, and have used the data to generally conclude that the labor market benefits of class size outweigh the cost. However, the fourth chapter of the dissertation suggests that such a conclusion may have been drawn in haste. Labor market benefit estimates only exceed the cost of the intervention when generated using test scores as the measured outcome, and generally under relatively liberal assumptions. When using other outcomes to generate benefit estimates, or when operating under stricter assumptions, the estimated benefits are exclusively lower than the costs.

Together, the three studies that comprise this dissertation address a number of

crucial questions in education research. Hopefully, these investigations will alleviate the concerns noted by Smith so long ago, and lead to an improved educational system which can better provide the learning experience that students deserve.

Chapter 2

Do Unobserved Noncognitive Skills Invalidate Value-Added Measures of Teacher Effectiveness?

Introduction

Value-added models (VAMs) have come to be looked at as a key analytical tool for improving educational outcomes. Staiger and Rockoff (2010), for example, propose that teachers should be assessed by value-added models for a few years and then retained or not based partially on this measure of their performance. Both Hanushek (2011) and Chetty, Friedman, and Rockoff (2013) suggest extremely large future labor market gains stemming from the replacement of the lowest-performing teachers. Given the high-stakes nature of VAM estimates, their accuracy is paramount. How could VAM estimates be utilized if they are in fact systematically biased by unobserved student information?

This paper adds to a growing empirical literature uncovering evidence that VAM estimates may be biased. Rothstein (2009) found that current test score gains are correlated with future teacher assignment, a relationship that cannot be causal and, in Rothstein's view, is indicative of bias. Subsequent research has tempered this conclusion, with Goldhaber and Chaplin (2011) noting that Rothstein's finding doesn't necessarily imply bias, and that bias only exists if the omitted variables are correlated with cognitive achievement and teacher assignment after conditioning on the observed variables.

While these earlier works suggest that bias might exist, no studies to this point have been able to show whether or not bias is actually present in VAM estimates, or which omitted variables serve as the culprits. The only relevant inquiry (Chetty, Friedman, and Rockoff 2013) found no substantial bias from typically-unobserved parental characteristics, but they tested just a small fraction of the variables which could be conditionally correlated with future achievement and only proxy for the student characteristics that are potentially problematic. Current knowledge about

the extent of bias resulting from unobserved characteristics is thus somewhat limited, and researchers therefore have an insufficient understanding of how accurate VAMs actually are.

Complicating the search for bias is the absence of an appropriate statistical method which would allow researchers to test a null hypothesis that VAM estimates (or any other coefficients) are not subject to omitted variable bias. Stated differently, there appears to be no suitable test for whether or not multiple coefficient estimates are sufficiently unaffected by the presence of a given variable. The lack of such a test may be especially surprising when one considers that a large proportion of observational studies argue that results are robust to unobserved variables based on a parameter stability heuristic. If coefficient estimates move by only a small amount when an additional control is added, researchers often take this as a sign, under the proportional selection assumption, that the coefficient estimates are likewise unaffected by the exclusion of other, unobserved covariates (Oster 2013). On the other hand, if coefficients demonstrate more than a small amount of movement as further variables are added to the model, this could be construed as indicating that results are not robust to unobserved heterogeneity. The problem with the use of this heuristic is that it is not clear what constitutes a small amount, and what differentiates it from being statistically significant.

This study considers the question of whether or not VAM estimates are robust to unobserved student heterogeneity. Using ECLS-K (Early Childhood Longitudinal Study Kindergarten Class of 1998-99) data, it contributes to the literature by demonstrating that the omission of noncognitive scores introduces substantial bias into VAM estimates. Additionally, a permutation test for omitted variable bias is developed and utilized to show that the bias is in fact statistically significant. The permutation test serves as a substantial methodological contribution, as it can be

used by practitioners to assess parameter stability and therefore robustness to unobserved information in a way that is much more scientific than the current use of heuristics. Finally, this analysis shows that a failure to account for typically-unobserved noncognitive scores leads to VAM estimates that are problematically inaccurate, incorrectly identifying teachers according to their value-added quintile at a rate of about 11.5%. Noting that noncognitive skills are only a portion of the student characteristics that might be sources of unobserved heterogeneity, and in light of the proportional selection assumption, the results raise serious concerns about the validity of VAM estimates.

This paper begins with a review of the literature before providing a conceptual framework of the conditions under which VAM estimates are biased. The following section empirically examines whether or not typically-unobserved noncognitive scores satisfy these conditions, finding that students' scores on an attribute described as "Approaches to Learning" is conditionally correlated with both teacher assignment and cognitive achievement. Then, a permutation test is developed and the identified bias is assessed with respect to its statistical significance. The practical implications of the bias are subsequently explored, with an examination of how much the bias affects the identification of the best and worst teachers, before concluding with a discussion of the results.

Previous Literature

The potential for bias due to unobserved student characteristics is greatly concerning when one considers the possible application of VAM estimates. Though researchers typically do not recommend making personnel decisions solely on a teacher's measured value-added, recent studies have pointed to the use of value-

added models in conjunction with other assessments to weed out low-performing teachers (Goldhaber and Hansen 2010, Gordon, Kane, and Staiger 2006, Staiger and Rockoff 2010). As mentioned earlier, estimates of labor market benefits in terms of higher future earnings of students resulting from the enaction of these policies suggest very large returns (Chetty, Friedman, and Rockoff 2013, Hanushek 2011).

If teacher retention policies based on VAM estimates are implemented, though, then value-added models must be sufficiently accurate, and it isn't clear whether this is the case. In fact, according to Corcoran and Goldhaber (2013), most researchers would agree that VAM estimates suffer from measurable bias. Their belief is driven primarily by a series of papers beginning with Rothstein (2009), who empirically demonstrated that future teacher assignments can predict current test score gains. Since a causal relationship between the quality of future teachers and current score gains is impossible, Rothstein's finding could indicate that students are assigned to subsequent teachers based on unobserved characteristics. Thus, Rothstein concludes that value-added estimates may be biased.

Later papers suggest that Rothstein's conclusion should perhaps have been a bit more nuanced. Koedel and Betts (2011) generally confirmed Rothstein's findings, but also noted that such bias can be greatly reduced by observing teachers over several waves of students. Indeed, after doing so, they find no evidence of bias. Kinsler (2012) finds that Rothstein's test performs poorly in small samples and that there is little evidence of tracking when accounting for sample size and allowing teacher effects to persist into later years. Goldhaber and Chaplin (2011) point out that Rothstein's falsification test in fact only detects tracking, which may not necessarily introduce bias unless there are omitted variables that are correlated with cognitive achievement and teacher assignment, conditional upon observed

covariates.

Perhaps the most relevant search for bias from unobservables is found in Chetty, Friedman, and Rockoff (2013). They include a set of typically-unobserved parental variables, such as mother's age at the child's birth and dummy variables for 401(k) contributions and home ownership, as predictors of test scores. The inclusion of the additional variables did not reveal any significant bias. As discussed previously, though, the additional variables they considered merely proxy for unobserved student information and make up only a small percent of typically-unobserved characteristics. Thus, while Rothstein's testing for the effect of future teachers on current gains suggests the potential for bias, there has been little subsequent evidence that the bias exists, or which unobserved variables might cause it.

One potential source of bias that has yet to be explored lies in noncognitive skills, such as self-discipline, persistence (Heckman 2000), attentiveness, temper control, and friendship formation. Several types of noncognitive skills have been demonstrated to be related to cognitive achievement. Swartz and Walker (1984) show that certain noncognitive abilities measured in kindergarten, such as interpersonal skills, are related to scores on the California Achievement Test. Similarly, Alexander, Entwisle, and Dauber (1993) find that a child's attentiveness and participation in classroom activities are predictors of future cognitive achievement gains. Other studies demonstrate that noncognitive skills are related to cognitive test scores even after controlling for prior achievement (McClelland, Morrison, and Holms 2000) or demographic characteristics (Fantuzzo et al. 2007, McClelland et al. 2007). Utilizing ECLS-K data, Li-Grining et al. (2010) show that the Approaches to Learning skill is an important factor in determining a student's academic trajectory.

Even if noncognitive skills do predict cognitive gains after conditioning on ob-

served student characteristic, Goldhaber and Chaplin explain that these skills would also have to be correlated with teacher assignment in order to bias VAM estimates. Unfortunately, there is suggestive evidence that this may be the case. While they didn't control for observed student characteristics, Kalogrides, Loeb, and Beteille (2013) demonstrate that students are matched to teachers based on prior behavior, which is often thought to be correlated with noncognitive ability. Additionally, Neidell and Waldfogel (2010) find evidence that the noncognitive scores in the ECLS-K data jointly predict a student's future teacher assignments. Thus, it is certainly plausible that noncognitive skills fulfill Goldhaber and Chaplin's requirements for bias.

Even in light of the potential for bias due to noncognitive scores or other unobserved student characteristics, Corcoran and Goldhaber (2013) argue that VAMs are still useful. Given the current body of research, they believe that the most important question at this point is whether or not VAMs are an improvement over other measures such as observational assessments by principals. Developing an answer to this question requires an examination of the sources and magnitude of bias in VAM estimates, precisely what this paper intends to do.

Data

Data from the ECLS-K will be used in this endeavor. Provided by the National Center for Education Statistics (NCES), the ECLS-K is a nationally representative sample of over 21,000 students who entered kindergarten in the fall of 1998 (Tourangeau et al. 2009). Approximately 1,000 schools were included in the sample, resulting in an average sample of around 21 students per school. About 8 students were sampled per classroom in kindergarten, though this figure drops to

just under 5 by the end of first grade. Weights are provided in the data to account for sampling variability, although researchers have noted that their estimates do not appear to be substantially affected by the use or choice of weight.[1] For the sake of simplicity with respect to the permutation exercises that follow, weights are not used in this analysis.

Information such as demographic characteristics, family background, subjective assessments, test scores, and school characteristics was gathered from teachers, parents, and administrators in the fall of kindergarten and the spring of kindergarten, 1st, 3rd, 5th, and 8th grade. However, the gaps between later grades limit this analysis to only the kindergarten and first grade year, as assessments from two or three years earlier do not serve as good baseline test scores in value added models. Further, in later grades the number of students per surveyed teacher is often only one or two, making it impossible to develop reliable VAM estimates for each teacher.

A particularly problematic aspect of the ECLS-K sample as it relates to this analysis is the prevalence of missing data. Only 15,526 students have reading and math scores recorded both for the spring of kindergarten and spring of first grade, and other variables used in the analysis contain missing values as well. One approach to dealing with this issue would be to replace missing values with school or sample means. However, since analyses will be conducted using school-level fixed effects, such a solution would attenuate coefficient estimates toward zero.

Making the situation even more troubling is the fact that the data are not missing at random. Students with missing test scores at the end of Kindergarten or first grade are substantially weaker, as measured by their fall kindergarten scores, than

[1] Fryer and Levitt (2004) also note that their estimates are not sensitive to the choice of weights or to weighting at all.

are students with valid scores in those waves. Therefore, if missing test scores were replaced with classroom means, then the replaced values would likely be artificially high. Linear predictions of later test scores based on the fall kindergarten score and other observed characteristics would also obscure test score gains between the spring of Kindergarden and spring of 1st grade, since they would be based off of the same underlying information. Thus, this study follows earlier research (Fryer and Levitt 2004, Claessens, Duncan, and Engel 2009) by simply dropping affected cases. The end result is an effective sample size of approximately 11,500 students.

Of special interest in this study are the well-validated measures of noncognitive skills offered in the ECLS-K data. Noncognitive scores are provided in five areas: Approaches to Learning, Externalizing Problem Behaviors, Internalizing Problem Behaviors, Interpersonal Skills, and Self-Control. While each score certainly measures the respective skill with error, they have been employed in several earlier articles (see Claessens, Duncan, and Engel 2009, Downey and Pribesh 2004, Finn and Pannozzo 2004, Hair et al. 2006, Ready et al. 2005). Neidell and Waldfogel (2010) find large and signficant behavioral peer effects when using the Externalizing Problem Behaviors score, for example. These scores will be utilized in this study as measures of the typically-unobserved noncognitive skills that might bias VAM estimates.

Theoretical Framework

Adapted from Hanushek and Rivkin (2010), Dieterle et al. (2012), and Goldhaber and Chaplin (2011), suppose the value added model takes the form of

$$A_{it} = \mu + S_{it} + \theta A_{i(t-1)} + X_{it}\gamma + U_{it}\delta + \sum \tau_{jit}\beta_j + \varepsilon_{it} \tag{1}$$

In the case above, A_{it} and $A_{i(t-1)}$ are measures of student achievement in year t and t-1, respectively, S_{it} is a school effect, and μ is both the intercept and the value added of the omitted teacher. X_{it} are observed characteristics, while U_{it} are typically-unobserved characteristics. θ, γ, and δ are unknown, and ε_{it} is a stochastic error term. τ_{jit} is a dummy variable which indicates whether or not student i was assigned to teacher j at time t, and β_j would be interpreted as a specific teacher's value added, relative to the omitted teacher.

Assuming that the typically-unobserved variables cannot be included in the analysis, the actual model used for estimation is

$$A_{it} = \mu + S_{it} + \theta A_{i(t-1)} + X_{it}\gamma + \sum \tau_{jit}\beta_j + \varepsilon_{it}$$
$$\varepsilon_{it} \equiv U_{it}\delta + u_{it}$$

If U_{it} is a linear function of X_{it} and $A_{i(t-1)}$, such that

$$U_{it}\delta = \alpha_1 A_{i(t-1)} + \alpha_2 X_{it} + v_{it}$$

where v_{it} is independently and identically distributed according to a normal distribution, then this is of little importance. The value-added model can be written as

$$A_{it} = \mu + S_{it} + (1 + \tfrac{\alpha_1}{\delta})\theta A_{i(t-1)} + (1 + \tfrac{\alpha_2}{\delta})X_{it}\gamma + \sum \tau_{jit}\beta_j + \epsilon_{it}$$

This, of course, would not affect regression estimates, since X_{it} and $A_{i(t-1)}$ are subject only to linear transformations. On the other hand, it may be the case that U_{it} is not a linear function of lagged achievement and observed characteristics, which could, as Goldhaber and Chaplin point out, be problematic. As they demonstrate, though, while the presence of U_{it} would bias some regression estimates, it wouldn't necessarily introduce bias into estimates of teacher value-added, β_j.

Adapting their work to the notation in this analysis, we can define the omitted variable bias as

$$Bias(\hat{\beta}_j) = E(\hat{\beta}_j) - \beta_j = \delta\pi_{je} \tag{2}$$

where π_{je} is the coefficient on τ_{jit} obtained from regressing U_{it} on all of the other independent variables in Equation 1 except ε_{it}. They then note that

$$\pi_{je} = Cov(U^*_{it}, \tau^*_{jit})/Var(\tau^*_{jit})^2 \tag{3}$$

where U^*_{it} and τ^*_{jit} "are the conditional values of [U_{it}] and [τ_{jit}] that exist after controlling for other variables in the model." (Goldhaber and Chaplin 2011, p.5-6)[3] In other words, the starred terms are the residuals found after regressing the respective unobserved variable on all of the other right-hand side variables in Equation 1.

Equation 2 demonstrates that bias exists only if δ and π_{je} are both nonzero. The δ coefficient,

$$\delta = Cov(A^*_{it}, U^*_{it})/Var(U^*_{it}) \tag{4}$$

is nonzero if U_{it} is not a linear function of X_{it} and A_{it}, and if U_{it} is correlated with A_{it} after conditioning upon all of the other observed variables. Stated differently, U_{it} must share a relationship with A_{it} that cannot be explained by the other covariates. π_{je} is nonzero only if U_{it} is correlated with τ_{jit} after conditioning upon the remaining control variables. In other words, if the unobserved characteristics are not a linear function of the observed characteristics and lagged achievement, if U_{it} explains A_{it} in a way not captured by the other variables, and if U_{it} is conditionally correlated with teacher assignment, then the unobservables will bias VAM

[2] See the footnote to page 5 in Goldhaber and Chaplin (2011).

[3] Goldhaber and Chaplin use ov_{ig} in place of U_{it} and $\tau_{t,i,g}$ in place of τ_{jit}.

estimates (Goldhaber and Chaplin 2011).

Of course, researchers have been mindful of the fact that unobserved characteristics could pose challenging for VAM estimation and have come up with methods to eliminate such problems. Perhaps the most popular of these is the use of student-level fixed effects, which has been employed in several recent papers including Harris and Sass (2006) and Koedel and Betts (2007, 2011). One problem that arises from this technique is that it generates a systematic relationship between the first-differenced error term and the lagged score gain (Koedel and Betts 2011). Researchers (Harris and Sass 2006, Koedel 2009, Koedel and Betts 2007, 2011) solve this problem by instrumenting for the lagged test score gains with the second lagged test score, as in Anderson and Hsiao (1981), via 2SLS (Koedel and Betts 2011).

While a theoretically-sound technique, it isn't perfect in its practical application. First, Kane and Staiger (2008) find that including student fixed effects results in VAM estimates that are attenuated toward zero. Perhaps even more important for an analysis of early grades, the requirement of a second-lagged test score makes it difficult or impossible to accurately assess kindergarten and first grade teachers, as each student might not have been assessed twice prior to teacher assignment. Additionally, districts might not have twice-lagged test scores for students moving into their district, increasing the likelihood that mobile students are left out of teacher assessments. Thus, it is important to understand the prospects for bias outside of the student fixed effects solution, too.

Do Typically-Unobserved Noncognitive Skills Bias Estimates of Teacher Effects?

The analysis begins by investigating whether or not noncognitive scores satisfy the conditions for bias laid out by Goldhaber and Chaplin. Doing so consists of estimating both δ and π_{je} so as to determine if their product is nonzero. The first step involves regressing composite test scores, constructed by averaging standardized reading and math scores, on lagged composite test scores, a school fixed effect, a vector of k lagged noncognitive scores, and dummy variables indicating gender, minority status, and free lunch recipiency. Only students who did not switch schools between kindergarten and first grade are considered, given the reduced amount of information schools would have about new students for the purpose of teacher assignment. Thus, the school fixed effect joins the female and minority indicators as constant across time.

$$A_{it} = \omega_0 + \omega_1 A_{i(t-1)} + \omega_2 Female_i + \omega_3 Minority_i + \omega_4 Lunch_{it} + \sum \omega_p S_i + \sum \tau_{jit}\beta_j + \sum NC_{ki(t-1)}\delta_k + e_{it} \tag{5}$$

In most administrative data sets, noncognitive scores are unobserved and so neither δ nor β_j (conditional upon noncognitive scores) could be observed. In this case, however, the typically-unobserved information is actually present in the selected data set. Because the noncognitive scores are observed, the δ's needed to estimate the bias from each individual noncognitive score are then simply the δ_k's in the model above. The model was estimated via Ordinary Least Squares in order to find the δ's of interest.

Results, with the estimated teacher effects omitted, are given in Table 2.1. Most of the coefficients on noncognitive scores are not statistically different from zero,

but the coefficient on the Approaches to Learning score is large and significant. The observed relationship is sensible given what the Approaches to Learning score measures. According to Tourangeau et al. (2009), it is comprised of teachers' ratings of students in six areas, including how eager a student is to learn new things and how well a student pays attention in class. The results demonstrate that noncognitive skills in this area predict achievement in a way not captured by the other control variables.

Finding π_{je} is similarly straightforward, as each teacher's π_{kje} is β_{jk}, the coefficient on τ_{jit} when an individual noncognitive score k is regressed on all of the other independent variables, including the other noncognitive scores.

$$\begin{aligned} NC_{ki(t-1)} = \zeta_0 + \zeta_1 A_{i(t-1)} + \zeta_2 Female_i + \zeta_3 Minority_i + \zeta_4 Lunch_{it} + \\ \sum \zeta_p S_i + \sum \tau_{jit}\beta_{jk} + \sum NC_{\sim ki(t-1)}\delta_{\sim k} + e_{it} \end{aligned} \tag{6}$$

After estimating π_{kje} for each teacher, the estimated bias from each noncognitive score is calculated as the product of δ_k and π_{kje}, and overall bias is constructed by summing the individual components.

$$Bias_{jk} = \delta_k \pi_{kje} \tag{7}$$

$$Bias_j = \sum_k \delta_k \pi_{kje} \tag{8}$$

Summary statistics for the observed bias from each noncognitive score and for the total bias are given in Table 2.2. In general, the bias is centered around zero, and the standard deviations are small, indicating that most of the noncognitive scores do not bias VAM estimates when they are left out of the model. On the other hand, there appears to be substantial bias generated by the omission of the Approaches

to Learning score, and the bias from this variable clearly drives estimates of the combined bias from all noncognitive scores.

This finding is compelling. The δ_k = 0.129 (0.011) for Approaches to Learning was large and positive. Students with higher scores perform better on standardized tests, even after controlling for other factors. Combined with π_{kje}, the presence of measurable bias shows that some teachers are assigned students with higher Approaches to Learning scores, and VAM estimates for these teachers are biased upwards. These teachers are matched to students who are more eager to learn and these teachers have their value-added overestimated when using the typical model. Other teachers are assigned students who are less eager to learn, and their value-added is biased down. This is perhaps the first confirmed example of a single student characteristic that generates bias in conventional value-added estimates.

Based on Rothstein's finding that including additional test scores from earlier years in the value-added model could eliminate much of the bias he observed, one potential solution to the bias problem identified here may also be found in adding extra lagged test scores. While the inclusion of an additional lag is often precluded by data limitations, and can make it impossible to evaluate kindergarten and first grade teachers in many schools, it is possible in this instance. Students in the ECLS-K sample were assessed in both the fall and spring of the kindergarten year, and so the fall score can be added to determine whether or not its inclusion eliminates all or part of the bias.

Since the bias results nearly entirely from the omission of the Approaches to Learning score, the process of identifying bias was carried out using just that score, both with and without an additional lag. That is, Equation 5 was modified and estimated separately as both Equations 9 and 10, where $L_{i(t-1)}$ is the lagged Ap-

proaches to Learning score.

$$A_{it} = \omega_0 + \omega_1 A_{i(t-1)} + \omega_2 Female_i + \omega_3 Minority_i + \omega_4 Lunch_{it} + \sum \omega_p S_i + \sum \tau_{jit} \beta_j + L_{i(t-1)} \delta_L + e_{it} \tag{9}$$

$$A_{it} = \omega_0 + \omega_1 A_{i(t-1)} + \omega_2 A_{i(t-2)} + \omega_3 Female_i + \omega_4 Minority_i + \omega_5 Lunch_{it} + \sum \omega_p S_i + \sum \tau_{jit} \beta_j + L_{i(t-1)} \delta_L + e_{it} \tag{10}$$

Table 2.3 offers regression results for Equation 9, while Table 2.4 provides results for Equation 10. In both cases, the coefficient on the Approaches to Learning score is statistically significant. The inclusion of the extra lag pushes the coefficient downward from .120 to .115, suggesting that some of the bias might be eliminated. It may also be the case that the teacher assignment based on the Approaches to Learning score may be somewhat mitigated when assignment on an additional lag is accounted for, and so further investigation of the bias is merited.

Table 2.5 provides summary statistics for the estimated bias both with and without the additional lag, while only considering bias stemming from the Approaches to Learning score. Though summary statistics such as the standard error, minimum, and maximum bias seem to indicate that the bias is reduced when the second lagged test score is added, the bias is far from eliminated. It is plausible that adding further lags would further reduce the bias, but the lack of a third lagged test score renders this option impossible in the ECLS-K data, and may be impractical in practice.

While the finding that VAM estimates are biased when the Approaches to Learning score is omitted is interesting, the magnitude of the bias is also important. Figure 1 is a histogram of the bias from the omission of the Approaches to Learning score for all teachers. The story it tells is both comforting and concerning at the

same time. For most teachers, the bias is relatively minor. About 47% of teachers show bias that is between -.05 and .05 standard deviations of student achievement. On the other hand, estimates for 15.7% of teachers exhibit bias greater than .1 in magnitude.

It may also be informative to explore how the bias varies with a teacher's estimated value-added. Figure 2 is a scatter plot with V_{conv}, the conventional value-added measure which does not include the Approaches to Learning score, on the X-axis, and the bias resulting from the omission of the Approaches to Learning score on the Y-axis. If V_{conv} was approximately unbiased, then the fitted line would be parallel to the X-axis. The upward slope of the line instead indicates that teachers evaluated favorably by the conventional measure have their value overestimated due to the bias. On the other hand, teachers evaluated unfavorably have their value underestimated. This provides further evidence that, while V_{conv} may measure teacher effectiveness in a manner that is close enough for many teachers, it may result in some teachers being evaluated with error.

Still, neither summary statistics nor graphs provide evidence of statistical significance. In this case, such a test would be useful to distinguish the identified bias from random noise. Unfortunately, several aspects of the bias rule out the use of typical tests. Most notably, this study is interested in bias across several teacher dummy variables, while existing tests only allow for the testing of one variable at a time[4]. Further, while the rankings of teachers differ based on inclusion of noncognitive scores, the distributions of VAM estimates generated both with and without these scores are essentially the same. The common mean and distribution rule out conventional t- and Kolmogorov-Smirnov tests, while the zero-sum nature of the bias renders useless such options as the paired t-, Mann-Whitney,

[4]See MacKinnon et al. (2002) for a summary for various methods.

and Wilcoxon signed-rank tests. Various rank statistics might be appropriate, but are not illuminating without a valid counterfactual. That is, computing a rank statistic is not helpful without knowing what the statistic would have been if noncognitive scores were not correlated with achievement and teacher assignment.

A Permutation Test for Omitted Variable Bias

The problem, then, set in the context of a hypothesis test, is that no known test is able to reject the null hypothesis:

$$H_0 : \prod_j \delta_k \pi_{k_{je}} = 0 \tag{11}$$

Given the generally-accepted requirement of statistical significance for causal conclusions, the lack of a usable test serves as a substantial obstacle.

In order to reject H_0 at some significance level α, a method of determining a p-value must be developed. To do so, it is helpful to think of the p-value as the likelihood that bias as extreme as that observed in the data would have been observed if H_0 was actually true. If noncognitive scores were randomly assigned to students after conditioning on observed student covariates, then

$$E[\prod_j \delta_k \pi_{k_{je}} | X] = 0. \tag{12}$$

The p-value can thus be considered to be the probability of observing such extreme bias under conditionally-random assignment.

Of course, conditionally-random assignment cannot be carried out in this case, as noncognitive scores simply cannot be assigned in this manner to students. Fortunately, earlier research points at permutation tests as a potential solution. Origi-

nating with Neyman (1923) and Fisher (1935), their theoretical basis is well established (Brown and Maritz 1982, Kempthorne 1952, Scheffé 1959). At that time, a lack of technology prevented their immediate application. With the advent of computers, though, researchers began investigating the use of permutation tests in linear regression (Gail, Tan and Piantadosi 1988, Levin and Robbins 1983, Manly 1991, Oja 1987, Welch 1990). To motivate the intuition behind a permutation test, a simple example from Kennedy (1995) will be used. Consider a linear regression that takes the following form:

$$Y = B_0 + B_1X + e$$

Suppose Y is income and X is gender, and suppose that one wishes to examine the relationship between gender and income among 20 individuals who are otherwise alike in every way. It is not possible to randomly assign the treatment of male or female to individuals, just as it is not possible to randomly assign noncognitive scores to students. However, it is possible to randomly shuffle the labels of male and female across cases under the null hypothesis that $B_1 = 0$, providing the error terms are homoscedastic and thus exchangeable. In carrying out this shuffling several times, a permutation test provides a multiple data sets as if gender was randomly assigned.

Unfortunately, permutation tests become more complicated in multiple regression. Again adapted from Kennedy, consider the case when education also varies across respondents, with Z serving as a dummy variable indicating whether or not the individual has a high school degree.

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3Z + e \tag{13}$$

Now, X_1 can no longer be permuted, as the collinearity between X_1, X_2 and Z

would no longer be maintained. Further, error terms would no longer be exchangeable, since it cannot be assumed that the variance of e is the same across highly educated and less educated individuals. Instead, values can only be permuted within groups, as in Kempthorne.

The case of a continuous regressor then poses a significant challenge, since there are no groups among which permuting is permissible. Several solutions to this problem are present in the research, with Kennedy providing an excellent summary of the various options. Perhaps the most well-received is a method of shuffling residuals, outlined in Beaton (1978) and Freedman and Lane (1983). Consider again Equation 9, but suppose that both X_1 and Z are now continuous predictors. Their method consists of regressing Y on X_2 and Z, computing residuals, randomizing the residuals across individuals and then adding them to the predicted values of Y. The modified Y is then regressed on X_1, X_2, and Z to provide an estimate of B_1 as if X_1 was randomly assigned, conditional upon X_2 and Z.

Kennedy provides another, similar solution. Y and X_1 are independently regressed on X_2 and Z before estimating the residuals, Y^* and X_1^*. Y^* is randomly shuffled to create Y^{*m}, the residualized Y vector under permutation m, before regressing Y^{*m} on X_1^* to estimate B_1. Kennedy shows that this method produces an estimate of B_1 that is identical to the estimate found using the procedure outlined by Freedman and Lane. Anderson and Robinson (2001) note that Kennedy's method is not completely equivalent to Freedman and Lane's, as the correlation coefficients vary by procedure, but confirm that both methods produce identical estimates of partial regression coefficients.

For an exact test of the statistical significance of B_1, estimates of B_1^m are to be found under all M permutations of Y^*. A simple t-statistic, t^m, is then calculated for each B_1^m. Constructing the p-value consists of counting h, the number of per-

mutations in which t^m exceeds t, the t-statistic found using the unpermuted data, and then dividing by the total number of permutations. Noting that the unpermuted case must also be included in the denominator, the p-value of interest is

$$\text{p} = \frac{h}{M+1}$$

If p $< \alpha$, the null hypothesis that B_1=0 can be rejected. Of course, given a large number of cases and thus a very large number of permutations, carrying out an exact test may be infeasible. Instead, past empirical work has generally used 1,000 permutations (Kim, Nelson, and Startz 1991, McQueen 1992), as other studies have indicated that 1,000-10,000 permutations should be sufficient for a powerful, if still approximate, test (Dwass 1957, Keller-McNulty and Higgins 1987, Manly 1991).

The permutation test I develop for statistically significant bias is a basic extension of the methods proposed by Kennedy and others. Suppose now that Z is often unobserved, and one wishes to test whether or not the omission of Z biases estimates of B_1. The estimated bias for B_1 can be written as

$$Bias(\hat{B}_1) = E(\hat{B}_1) - B_1 = B_3\ \pi_1$$

$$Z = \pi_0 + \pi_1 X_1 + \pi_2 X_2 + e$$

$$B_3 = \frac{Cov(Z^{*1,2}, Y^{*1,2})}{Var(Z^{*1,2})}$$

$$\pi_1 = \frac{Cov(Z^{*2}, X_1^*)}{Var(X_1^*)}$$

$$\implies Bias(\hat{B}_1) = \frac{Cov(Z^{*1,2}, Y^{*1,2})}{Var(Z^{*1,2})}\ \frac{Cov(Z^{*2}, X_1^*)}{Var(X_1^*)}$$

Note that $Z^{*1,2} \neq Z^{*2}$. $Z^{*1,2}$ is comprised of the residuals found after regressing Z on X_1 and X_2, while Z^{*2} consists of the residuals generated after regressing Z on only X_2.

Under the null hypothesis that $Bias(\hat{B}_1)$ = 0, one of the following must be true: B_3=0, or π_1=0, or both B_3 = 0 and π_1 = 0. For simplicity, the case of B_3=0 is

considered. A measure of bias that would be observed if the null hypothesis in fact holds true is then

$$Bias(\hat{B_1^m}) = B_3^m \ \pi_1 = \frac{Cov(Z^{*1,2}, Y^{*1,2m})}{Var(Z^{*1,2})} \ \frac{Cov(Z^{*2}, X_1^*)}{Var(X_1^*)}$$

Y and Z are both independently regressed on X_1 and X_2, and the residualized versions of each, $Y^{*1,2}$ and $Z^{*1,2}$, are estimated. $Y^{*1,2}$ is randomly shuffled and regressed on $Z^{*1,2}$ to generate an estimate of B_3^m.

Given the counterfactual provided by each permutation, the only necessity that remains for an appropriate test is a suitable test statistic. For a product of coefficients $B\pi$, prior studies (MacKinnon et al. 2002, Sobel 1982) have noted that the t-statistic can be calculated as

$$t = \frac{B\pi}{\sqrt{B^2\sigma_\pi^2 + \pi^2\sigma_B^2 + \sigma_\pi^2\sigma_B^2}}$$

where the denominator is the standard error of the second order Taylor series approximation of $B\pi$ (Preacher and Hayes, 2008). $\sigma_\pi^2\sigma_B^2$ is typically omitted both because it doesn't appear when using the first order delta method to find the standard error (Sobel 1982) and because it is generally very small (Baron and Kenny 1986, MacKinnon and Dwyer 1993). Of course, both σ_π^2 and σ_B^2 are unknown, and so their estimates are used in their place. As before, an approximate p-value can be found by estimating t for each permutation and calculating the proportion of permutations in which t^m exceeds t, and the null hypothesis of no bias can be rejected if p $< \alpha$. Rejecting the null hypothesis is equivalent to finding that B_1 moves by a significant amount when Z is added as a control, and in this way the permutation test for omitted variable bias functions as a test for parameter stability as well.

The new method is demonstrated as a test only of the bias stemming from the Approaches to Learning score, as this score was responsible for the vast majority

of the overall identified bias. First, A_{it} is regressed on all of the typically-observed right-hand side variables in Equation 3, and the residual, A^*_{it}, is estimated. Then, $L^*_{i(t-1)}$, the residualized Approaches to Learning score, is estimated after regressing the score on the same typically-observed right-hand side variables. A^*_{it} is then shuffled to obtain A^{*m}_{it}, which is regressed on $L^*_{i(t-1)}$. Finally, an estimate of bias in β_j from the Approaches to Learning score under permutation m can be written as

$$Bias^m_{jL} = \delta^m_L \ \pi_{L(je)}$$

The expression above gives an estimate of bias as if the null hypothesis, $Bias_{jL}$ = 0, is true. In the special case of VAM estimates, we are interested not in whether one teacher's estimate is significantly biased, as the general argument would test, but in whether or not VAM estimates are biased across all teachers. In other words, the hypothesis of interest is that the sum of bias across all teachers is zero. Treating the t-value as a standardized distance, and considering that a greater distance corresponds to more bias, I argue that the mean squared t-statistic, $\bar{t^2}$, is an appropriate test statistic in this instance. The mean squared t-statistic thus reflects the mean squared standardized error between the VAM estimates with the Approaches to Learning score and the VAM estimates without the Approaches to Learning score.

If h is now the number of times out of M that $\bar{t^2}^m$, the $\bar{t^2}$ estimated under permutation, is larger than the $\bar{t^2}$ observed in the data, then the p-value is the same as before.

$$\text{p} = \frac{h}{M+1}$$

The $\bar{t^2}$ derived from the actual data was approximately 1.168. None of the 1000 permutations which were carried out resulted in $\bar{t^2}$ exceeding 1.168, meaning that

the p-value of interest was exactly zero. The null hypothesis that VAM estimates are not biased by the omission of the Approaches to Learning score can be rejected. In other words, the identified bias is statistically significant.

How does Omitting Approaches to Learning Result in Misclassification of Teachers?

Though it has been demonstrated that unobserved noncognitive scores bias VAM estimates in a manner that is statistically significant, it isn't yet clear how much such bias actually matters from a practical perspective. The usefulness of value-added models depends on their ability to correctly identify the best and worst teachers so that administrators can accurately determine which teachers should be promoted and which teachers should be fired. Thus, researchers may be interested not just in whether any bias exists but whether or not the bias causes typical value-added models to incorrectly classify teachers according to their VAM estimate.

The most obvious way to investigate the correct classification rate of VAM estimates when noncognitive scores are omitted is to simply compare estimates from models including the Approaches to Learning Score, V_L, to estimates from models in which such scores were left out, V_{conv}. Suppose that an administrator wanted to classify teachers into five equally-sized groups, based on their VAM estimates, for tenure or termination purposes. Teachers would be divided into five quintiles, with the least effective teachers being assigned to the lowest quintile and the most effective teachers assigned to the highest. Of interest is whether or not each teacher is assigned to the same quintile regardless of whether V_L or V_{conv} was used.

Table 2.6 shows how teachers classified according to V_{conv} would have been

classified had the Approaches to Learning score been controlled for. For example, 92% of those teachers identified as the worst teachers were actually the least effective as measured by V_L. According to Table 2.6, if a school district of 500 teachers decided to fire the 100 lowest performing teachers, then about 8 of those terminations would be unjust. Overall, about 12.3% of all teachers are classified into the wrong quintile when V_{conv} is utilized.

Taking these results at face value may be shortsighted, however. One of the key limitations of this analysis is the exceedingly small sample size within classrooms. As mentioned earlier, the mean number of sampled students per classroom by the end of second grade is only about five. To illustrate the effect that the small sample size could have on correct classification rates, consider a classroom of twenty students. Each student has an Approaches to Learning score which is not observed, but can be predicted by a linear function of typically-observed variables, subject to some error. Naturally, the expectation of the function would be the true Approaches to Learning score, providing the error term was independently and identically distributed according to a mean-zero normal distribution. Roughly half of the students would be expected to have their score underestimated, while it would be expected that the score would be overestimated for the other half. Now suppose teachers are typically assigned to the incorrect quintile if at least three-fourths of their students have their score overestimated or if at least three-fourths have their score underestimated. The probability of assignment to an incorrect quintile is then

$$2\sum_{k=15}^{20}\frac{(20!)}{k!(20-k)}(.5^k)(.5^{20-k}) \approx .0414 \tag{14}$$

However, if only four students in the classroom are sampled, the probability be-

comes

$$2\sum_{k=3}^{4}\frac{(4!)}{k!(4-k)}(.5^k)(.5^{4-k}) = .625 \tag{15}$$

Clearly, then, misclassification could occur at an alarming rate if the sample size is small and if the typically-observed variables predict noncognitive scores with error.

Rather than completely discounting the incorrect classifications demonstrated in Table 2.6, though, they should instead be compared to error rates that would be observed if no bias existed. θ_{obs} can be defined as the empirical error rate, while θ_{unb} is the error rate that would be expected if the Approaches to Learning score did not bias VAM estimates. The amount of error due to bias is then

$$\widehat{\theta_{bias}} = \widehat{\theta_{obs}} - \widehat{\theta_{unb}} \tag{16}$$

The process of estimating θ_{unb} is straightforward given the earlier exploration of permutations tests. First, note that

$$V_L = V_{conv} - \delta_L \pi_{(L)je}$$

As before, V_L is the estimate of teacher value added when the Approaches to Learning score is included in the linear regress. V_L can thus be estimated by finding V_{conv} and subtracting the estimated bias from it. Under the null hypothesis that estimates of V_{conv} are unbiased, $\delta_L \pi_{L(je)} = 0$. Thus, θ_{unb} is the error rate when V_L is estimated according to

$$V_L^m = V_{conv} - \delta_L^m \pi_{L(je)}$$

Teachers were placed into a quintile according to V_L^m, and this quintile was compared to the quintile they were classified into using V_{conv}. The error rate, $\widehat{\theta_{unb}}$, was calculated as the percent of teachers placed into a given quintile by V_{conv} who

were placed into a different quintile using the permuted data. This process was repeated 1000 times in order to construct a mean error rate and a 95% confidence interval, which are provided by quintile in Table 2.7.

Mean error rates using V_L^m vary by quintile, but in all cases they are substantially lower than $\widehat{\theta_{obs}}$, and $\widehat{\theta_{obs}}$ falls well outside the 95% confidence interval for $\widehat{\theta_{unb}}$ for all quintiles. The null hypothesis that $\widehat{\theta_{obs}} = \widehat{\theta_{unb}}$ can be rejected, allowing for the conclusion that the omission of the Approaches to Learning score causes some teachers to be incorrectly classified according to their VAM estimate. $\widehat{\theta_{bias}}$, the estimated level of misclassification due to unobserved sorting, is simply the difference between $\widehat{\theta_{obs}}$ and the expectation of $\widehat{\theta_{unb}}$. $\widehat{\theta_{bias}}$ fluctuates across quintiles, ranging from about 7.2% in the first quintile to 17.4% in the middle quintile, and averaging about 11.5% overall. Thus, it can be said that the bias due to the omission of the Approaches to Learning score causes teachers to be incorrectly classified at a rate of about 11.5%.

Discussion

This the first analysis to empirically demonstrate that there exists some typically-unobserved variable that is correlated with both teacher assignment and current achievement, introducing bias into VAM estimates. Though it is perhaps unsurprising given Li-Grining et al.'s (2010) prior work tying it to cognitive gains, the Approaches to Learning score is a statistically-significant predictor of current test scores, even after conditioning on other control variables. Additionally, it is conditionally correlated with classroom assignment, thereby fulfilling the requirements for bias laid out by Goldhaber and Chaplin.

This paper also contributes a permutation test which allows researchers to de-

termine if the bias resulting from an omitted variables is statistically significant. Applying a special version of the test to the study at hand, I show that bias stemming from the omission of the Approaches to Learning score is statistically significant, and so the relationship between the score and the identified bias is indeed causal. The permutation test also offers promise as a replacement for the commonly-used heuristic that results are robust if the inclusion of additional controls moves coefficient estimates by only a small amount. Instead of using an ambiguous term and leaving the interpretation of the heuristic up to the researcher, the permutation test provides a method for determining whether or not the movement in the parameter is statistically significant.

On top of the evidence of bias and methodological contribution, practically, the empirical evidence indicates that a failure to include noncognitive scores in value-added models results in teachers being placed into incorrect value-added quintiles approximately 11.5% of the time. While conventional value-added models perform better when tasked with identifying the best and worst teachers, these models still err in doing so at a rate of about 5.7%. Thus, a policy of firing the lowest-performing teachers would result in erroneous terminations at a non-negligible rate. If teachers were instead classified into deciles, the false classification rate would be higher.

The findings in this paper are especially troubling when considering the proportional selection assumption, which states that selection on observable characteristics is proportional to selection on unobserved characteristics (Altonji, Elder, and Taber 2005). As described in Oster (2013), results are deemed to be robust if coefficients are unaffected by additional controls, but in this instance it is clear that adding the Approaches to Learning score has a significant impact on VAM estimates. Under the proportional selection assumption, then, one cannot reach the conclusion that VAM estimates are robust to unobserved selection, a finding

that has profound ramifications for the appropriateness of using VAM estimates in hiring and firing decisions.

Future research may be aimed at examining whether or not the bias is mitigated by evaluating teachers over a few cohorts of students. It could certainly be the case that VAM estimates from a single year are biased by the omission of the Approaches to Learning score, but that VAM estimates using three years of data for each teacher are not. Since the current body of literature typically doesn't suggest the termination of teachers after just one year of assessment, the problem identified here may be rendered moot in the presence of additional years of data. Unfortunately, the data utilized in this study only contain a single wave of students, and no other known data sets exist to answer this question.

If evaluating teachers over multiple years does not solve the problem, future work should attempt to develop new methods for eliminating the bias without complete reliance on student fixed effects. Otherwise, data limitations may make it possible to fairly evaluate teachers in lower grades. Developing methods to rid VAM estimates of bias could correspond to large cost savings versus the alternative solution of forcing teachers and schools to create and store noncognitive scores for each student.

Finally, future estimates of labor market benefits stemming from VAM-based termination policies that are designed to improve test scores should take this bias into account. If the omission of the Approaches to Learning score introduces additional uncertainty into VAM estimates, then the omission will introduce some variability into benefit estimates as well. Of course, if multiple years of measurement or new techniques are not able to mitigate the bias, then such termination policies might prove to be impractical in light of the demonstrated inaccuracy, anyway.

Tables

Table 2.1: **Coefficients on Observed Characteristics and Typically-Unobserved Noncognitive Scores, School Level Fixed Effects**

	(1)
Lagged Test Score	.679***
	(-0.007)
Minority	-.008
	(0.013)
Female	-0.052 ***
	(0.009)
Free Lunch	-0.038**
	(0.014)
Externalizing Problem Behaviors	0.001
	(0.011)
Internalizing Problem Behaviors	0.003
	(0.010)
Interpersonal Skills	-0.010
	(0.013)
Self Control	-0.006
	(0.015)
Approaches to Learning	0.129***
	(0.011)
Observations	11428

Standard errors in parentheses

Dependent variable is test score from end of first grade

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.2: **Summary Statistics of Component ($Bias_{jk}$) and Overall Bias**

	(1) Mean	(2) Std Dev	(3) Minimum	(4) Maximum
Externalizing Problem Behaviors	.0000138	.0006748	-.0025085	.0031769
Internalizing Problem Behaviors	-.0000847	.0014894	-.0064531	.0071595
Interpersonal Skills	.0001139	.0038291	-.0166674	.0167163
Self Control	.0001439	.0020303	-.0088219	.0093759
Approaches to Learning	.0060094	.0633015	-.2460409	.2415954
Overall	.0061824	.0612254	-.2499989	.226254

Table 2.3: **Coefficients on Observed Characteristics and Approaches to Learning Score, Excluding Other Noncognitive Scores, School Level Fixed Effects**

	(1)
Lagged Test Score	.681***
	(0.007)
Minority	-.008
	(0.012)
Female	-0.052 ***
	(0.009)
Free Lunch	-0.035
	(0.014)
Approaches to Learning	0.120***
	(0.008)
Observations	11428

Standard errors in parentheses

Dependent variable is test score from end of first grade

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.4: **Coefficients on Observed Characteristics and Approaches to Learning Score, When Including 2nd Lagged Score, School Level Fixed Effects**

	(1)
Lagged Test Score	.589***
	(0.011)
2nd Lagged Test Score	.104***
	(0.009)
Minority	-.005
	(0.013)
Female	-0.051 ***
	(0.009)
Free Lunch	-0.024
	(0.014)
Approaches to Learning	0.115***
	(0.008)
Observations	11231

Standard errors in parentheses

Dependent variable is test score from end of first grade

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.5: **Summary Statistics of Bias from Exclusion of Approaches to Learning Score, One vs. Two Lags**

	(1) Mean	(2) Std Dev	(3) Minimum	(4) Maximum
Approaches to Learning - One Lag	.0041528	.0725980	-.2864553	.2865305
Approaches to Learning - Two Lags	-.0033110	.0587389	-.2127328	.2315821

Table 2.6: **Quintiles of Conventional VA by Quintiles of Noncognitive VA**

	Richer VA					
Conv. VA	1	2	3	4	5	Total
1	92	8	0	0	0	100
2	8	83	10	0	0	100
3	0	10	81	9	0	100
4	0	0	9	87	4	100
5	0	0	0	4	96	100
Total	100	100	100	100	100	

Table 2.7: **Summary of Error Rates by Quintile Across 1000 Permutations**

	Quintile				
	1	2	3	4	5
Mean	0.006	0.011	0.011	0.008	0.002
5th Percentile	0.000	0.000	0.000	0.002	0.000
95th Percentile	0.016	0.027	0.022	0.018	0.009
Observed Error	.078	.174	.185	.133	.044
Mean Difference	.072	.163	.174	.125	.042

Figures

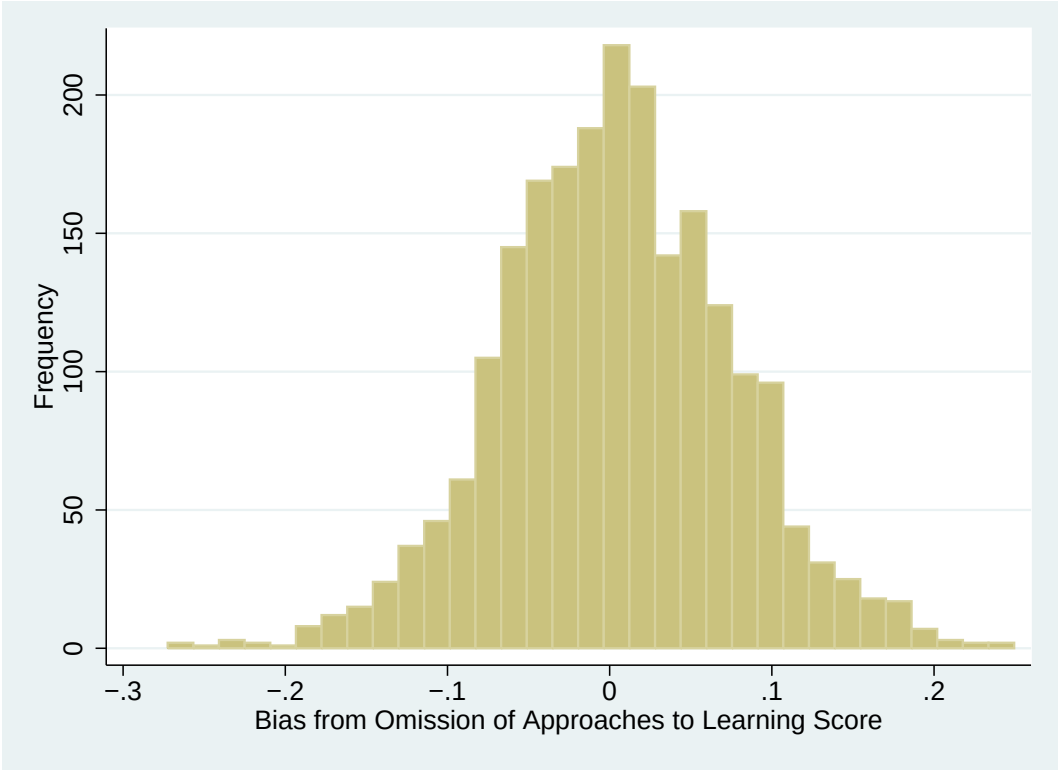

Figure 1: Distribution of Bias from Omission of Approaches to Learning Score

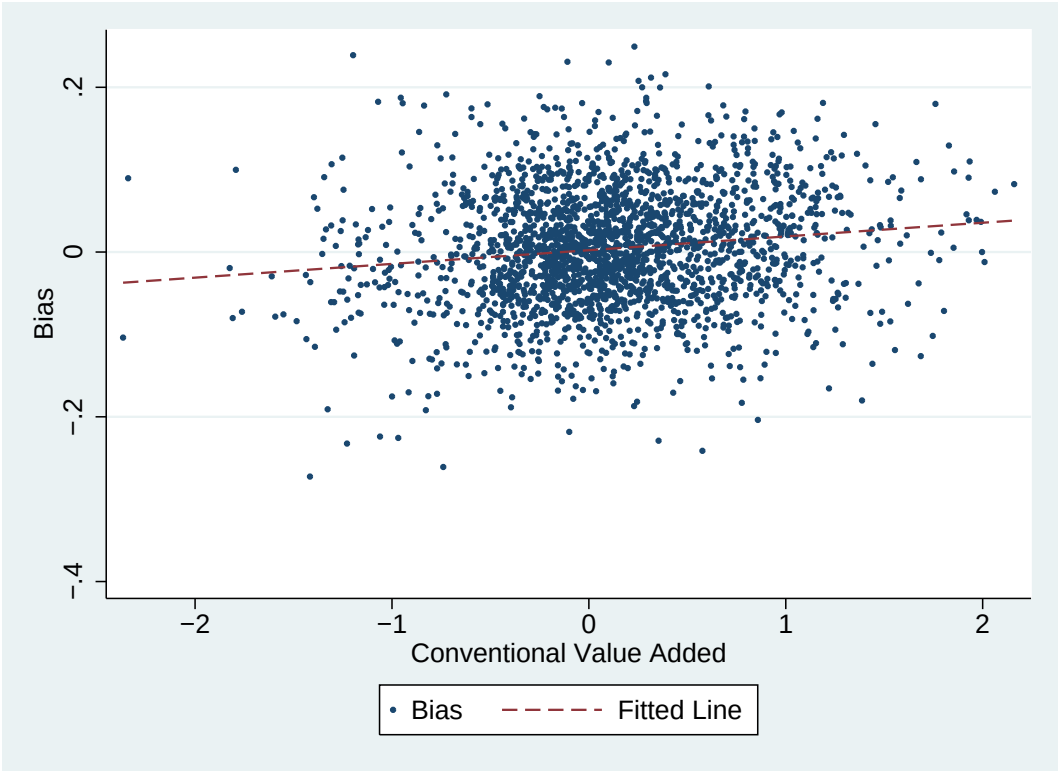

Figure 2: Scatter Plot of Conventional Value Added vs. Bias from Omission of Approaches to Learning Score

Chapter 3

Within-School Sorting on Observables and Unobservables: Evidence from a National Dataset

Introduction

Most prior education research does not assume that students are matched to teachers by a random process. Instead, past research posits that the effects of non-random assignment on the estimation of value-added models, which attempt to quantify the impact a teacher has on students' test scores, can be mitigated by controlling for observable teacher and student characteristics. Recent studies have indicated, though, that this may not be the case. Rothstein (2010) presented compelling evidence of sorting within schools, a finding that was confirmed by Betts and Koedel (2011). While these studies showed the potential for student-teacher matching, they did not empirically determine the factors upon which this sorting might occur.

Kalogrides, Loeb, and Beteille (2013) investigated student sorting within the Miami public school district, finding that less-experienced and minority teachers were typically matched to students with characteristics indicative of lower-achieving students. Dieterle et al. (2012) show that teachers are assigned classes of differing abilities based on their gender, race, experience, certification, and whether or not they have an advanced degree. However, the school districts involved in these studies are not at all representative of the United States as a whole, and therefore the findings might only apply locally. Further, many variables which may be used for sorting were not available in the datasets utilized, and so it is not clear whether or not students might also be sorted on unobservable information.

This analysis builds on prior work in several ways. First, the empirical results confirm some earlier findings, the first nationally-representative study to do so, while also noting relationships that have not been previously documented. Additionally, evidence indicates that the sorting of students to novice or minority teach-

ers may not be based on test scores, as found in prior work, but is instead based on factors that are merely correlated with test scores. Finally, this paper demonstrates that characteristics used to match students to teachers changes over time as more information about each student is ascertained.

This paper will begin with a review of the literature before describing the data and outlining the methods. The paper then shows which observed and typically-unobserved characteristics are used for teacher-student matching in kindergarten and first grade, extending earlier research by using a nationally representative data set containing a much wider array of information on teachers, students, and parents. The data are then analyzed within the framework of earlier studies to assess the impact of model choice on the results. A discussion of the results follows.

Previous Literature

The idea that teacher-student pairing is not completely random is well known to researchers. Typically, the sorting process has been analyzed with respect to how teachers are matched to schools. Several studies have demonstrated that urban schools tend to have a difficult time attracting and retaining high-quality teachers (Krei 1998, Lankford, Loeb, and Wyckoff 2002). Teachers with more experience leave for more desirable schools. These schools generally have a larger proportion of white students, with higher levels of both income and achievement across students (Boyd et al 2005, Clotfelter, Ladd, and Vigdor 2004, Hanushek, Kain, and Rivkin 2004, Horng 2009, Jackson 2009, Scafidi, Sjoquist, and Stinebrickner 2008). Sorting across schools thus leads to a situation in which low-income, low-achieving, and minority students are more likely to be taught by less-qualified and less-experienced teachers.

Such an arrangement is especially disheartening when considering the extent to which teacher quality can impact student outcomes (Taylor et al. 2000). Additionally, new teachers have generally been shown to be less effective than their more-experienced counterparts (Nye, Konstantopoulos, Hedges 2004, Rockoff 2004, Wiswall 2013), though this finding has been subject to some debate (Hanushek, Rivkin, and Kain 2005). Still, the interaction of teacher sorting and student outcomes has been shown to have a detrimental effect on economically-disadvantaged students. Indeed, the fact that minority students tend to have less-experienced teachers has been suggested as one of the driving forces behind the black-white achievement gap (Hanushek and Rivkin 2006).

Though most of the attention paid to teacher sorting has focused on sorting across schools, sorting within schools may occur for various reasons. A few studies have attempted to tackle this issue from an empirical perspective. Research on ability grouping and tracking demonstrates how students may be matched to similar peers based on ability level or other characteristics (Gamoran 1987, Oakes and Guiton 1995). Since higher-quality or more experienced teachers may be more likely to teach high-achieving students who have been tracked together (Kelly 2004), this can lead to a matching of high-quality teachers to high-achieving students within schools. At the same time, this phenomenon is often specific to the post-elementary grades.

Aside from the possibilities of tracking and ability grouping, Kalogrides, Loeb, and Beteille (2013) posit that within-school sorting occurs through power relations inside of schools. In this system, more experienced, white, and male teachers exert their influence to assure themselves of the most desirable classes. Less experienced, minority, and female teachers would therefore be left with less desirable students. Teachers may also be sorted to take advantage of minority teachers

greater effectiveness with or preference for minority students (Dee 2005, Downey and Pribesh 2004). Additionally, Rothstein (2009) suggests that teacher-student matching may take place when teachers who are thought to hold an advantage in teaching a certain set of skills may be matched to those students who need the additional help. Rothstein also believes that principals might assign desirable students to teachers they wish to reward, and misbehaving students to teachers they do not favor. Parents may also try to match their child to a specific teacher, and in doing so create a situation where high-quality teachers end up with students with parents who are more involved in school, on average.

The empirical literature on within-school sorting, outside of grouping and tracking research, is a little more sparse. Using administrative data from Florida, Feng (2010) demonstrates that novice teachers are assigned a greater proportion of low-income students, minority students, and students with behavioral problems. Clotfelder, Ladd, and Vigdor (2004) found that new 7th-grade teachers in North Carolina are assigned more black students relative to more experienced teachers.

One of the most informative works on exactly which factors might be used to match students to teachers comes from Kalogrides, Loeb, and Beteille (2013). In their analysis of data from the Miami public school district, they find that female, minority, and less experienced teachers are assigned classes of students with lower prior average test scores, lower prior attendance rates, and more prior behavioral problems. Additionally, they show that the sorting they identify depends on the characteristics of all teachers in the school. That is, the likelihood that novice teachers are assigned less-desirable students increases with the percent of experienced teachers in the school. Beyond just test scores, they find evidence of racial matching as well. Using a similar model, Dieterle et al. (2012) generally confirm these findings in their study of data from an anonymous state, but instead find that

female teachers are assigned classes with higher previous test scores.

The finding in Kalogrides, Loeb, and Beteille that students may be matched to teachers based on previously-observed behavior is quite important, especially in light of Rothsein's (2009) conclusion that students might be matched to teachers on unobserved student characteristics. After all, the data sets utilized for value-added or peer effects research often do not contain controls for behavior. If teacher-student matching occurs on student behavior or other variables that are typically not observed by the researcher, then such sorting could introduce bias into these models. For example, if lower-achieving students are typically those exhibiting disruptive behavior, then unobserved peer effects could bias estimates relating to test scores.

In fact, some of the most sizable, documented peer effects have been those that result from disruptive behavior. Lazear (2001) was the first to propose a framework under which disruptive behavior might hinder peer achievement, and subsequent empirical work has confirmed his hypothesis (Figlio 2007, Fletcher 2010, Neidell and Waldfogel 2008). Figlio finds that adding one additional disruptive student to a classroom reduces the test scores of all other students in the classroom by roughly four percent. Similarly, Fletcher shows that the presence of a student with serious emotional problems reduces peer test scores by about one-tenth of a standard deviation. Both of these effects are large compared to test score peer effects found in well-controlled studies such as Burke and Sass (2008). Sorting based on prior behavior could additionally be undesirable if the teachers to which misbehaving students are matched are subsequently more inclined to leave the profession. Turnover among new teachers might be higher due to this sorting if new teachers are assigned more poorly-behaving students, for example. Assuming that these teachers are replaced by other new teachers, this would result in a less-

experienced pool of teachers.

Given the lack of understanding regarding teacher-student sorting, especially that which could occur based on unobservable characteristics, the current study adds to the literature by more closely investigating how this matching occurs with a special emphasis on the role of family characteristics typically unobserved in standard administrative data sets.

Data

This study uses data from the Early Childhood Longitudinal Study Kindergarten Class of 1998-99 (ECLS-K), provided by the National Center for Education Statistics (NCES). The ECLS-K is a nationally representative sample of over 20,000 students who entered kindergarten in the fall of 1998 (Tourangeau et al. 2009). These students were spread across roughly 1,000 schools and 2,500 classrooms, and so the sample contains information on about 2.5 classrooms per school and 8 students per classroom, on average. Weights to correct for sampling variability are provided in the data, and the relevant cross-sectional student weights are used in this analysis. Estimates do not appear to be sensitive to the use or choice of weight.

Data such as family and school characteristics, student demographic information, subjective assessments, and standardized test scores was obtained from teachers, parents, and administrators in the fall of kindergarten and the spring of kindergarten, 1st, 3rd, 5th, and 8th grade. The time between the 1st, 3rd, 5th, and 8th grades preclude the post-1st grade years from this analysis, as it is expected that classroom assignment would be conducted mostly on information observed during the immediately preceding year. Additionally, in later grades sometimes

only one or two students are sampled in each classroom, and so classroom-level characteristics would be measured with substantial error.

It is clear that the ECLS-K sample has two distinct advantages over data sets used in earlier studies. First, it is nationally representative, and further, the ECLS-K data set contains many variables that are not included in data sets which have historically been used for value-added research. The breadth of the available variables will allow for the identification of the characteristics on which student-teacher sorting occurs, including those that would commonly be unobserved.

As mentioned earlier, Kalogrides, Loeb, and Beteille showed that students were matched to teachers based on student behavior. While the ECLS-K data set does not contain information on student suspensions as in Kalogrides, Loeb, and Beteille, it does provide well-validated measures of teacher assessments that are thought to serve as appropriate behavioral measures. The Externalizing Problem Behaviors score is constructed from teacher assessments of how often students argued, fought, became angry, acted impulsively, and disturbed ongoing activities (Tourangeau et al. 2009).[5]

The Externalizing Problem Behaviors score measures disruptive behavior with error, but similar noncognitive scores have been used as measures of student behavior in earlier research. Downey and Pribesh (2005) used the Externalizing Problem Behavior scores in their aforementioned study of differences in assessment based on racial mismatches between students and teachers. The same score was used in a similar manner by Finn and Pannozzo (2004), and was employed by Neidell and Waldfogel (2008) in their behavioral peer effects research.

Other noncognitive scores included in the ECLS-K data will also be employed

[5]These assessments were turned into a composite score, with 1 reflecting the lowest amount of the relevant behavior, and 4 reflecting the highest amount of relevant behavior. Thus, a rating of 4 would be considered the worst level of externalized problem behaviors.

in this study, including Self-Control, Approaches to Learning, Internalizing Problem Behaviors, and Interpersonal Skills scores. Waldfogel and Neidell showed that classroom assignment in the first grade may be correlated with these scores, and so it seems prudent to include them in this discussion. At the same time, it is not clear whether or not sorting occurs based on these scores or based on other factors correlated with these scores. For example, black students typically have poorer noncognitive scores, and it is important to distinguish between sorting based on race and sorting based on noncognitive skills.

As noted, previous studies have found within- and across-school teacher sorting based on test scores, race, and behavior. As such, each of these possibilities will be considered. Additionally, family-level characteristics will be included to determine whether or not there may be further sorting on these variables. For example, it may be the case that teachers prefer students who have a father present in the household in the absence of observed information regarding ability or behavior, since the presence of a father may be correlated with both. By incorporating a wider set of family-level information than typically found in administrative data sets, I am able to investigate the role of unobservables in the sorting process.

Before proceeding with the analysis, it is necessary to outline the different classifications of teachers that will be used in this paper. With respect to experience, teachers are separated based on two distinct cutoffs. 'Novice' teachers will be defined as those with seven years or less of experience. This definition is somewhat arbitrary, but it is the same definition used in Kalogrides, Loeb, and Beteille, and additionally is approximately the median years of experience in the ECLS-K data. Thus, it provides a good balance of observations when regressions specific to one type of teacher are estimated.

At the same time, it isn't clear that such a grouping most closely matches any

true criteria used by administrators in assigning students to teachers. In other words, it is not known whether teachers with seven years of experience are sorted to students in a manner that is more similar to teachers with eight years of experience or to teachers with six years of experience Therefore, 'beginner' teachers will refer to those teachers with three years of experience or less. Using multiple definitions will allow for the testing of the sensitivity of these results to the different group designations. For both 'beginning' and 'novice' teachers, 'experienced' teachers will be teachers with more than three or more than seven years experience teaching, respectively. Minority teachers are those who are coded as 'No' for the ECLS-K variable indicating whether or not the teacher is white. White teachers are those who are coded 'Yes'.

To aid in a comparison with previous studies, classroom means by experience or race are presented in Table 3.1. Beginning, novice, and minority teachers tend to be assigned to classrooms with characteristics that are generally less conducive to academic success when compared to experienced teachers (Table 3.1). These classrooms have lower average reading and math scores, a higher proportion of male and minority students, more students who are eligible for free or reduced lunch, fewer students with older siblings, fewer students with moms who have a high school degree, fewer students with fathers in the household, and larger class sizes. Some of the noted differences, such as the proportion of students with older siblings, have not been discussed in earlier research, because prior studies focused only on the top five variables due to a lack of information regarding the bottom three. Overall, though, the summary statistics suggest the type of matching that has been noted in prior works.

Methods

Though the comparison on means is enlightening, many of the characteristics are correlated, so it is difficult to determine the factors on which sorting actually occurs. To examine the extent to which various factors affect classroom placement, simple linear probability models (LPMs) will be used. In all models, school-level fixed effects are used to isolate within-school sorting, as that is the emphasis of this study. Thus, LPMs are appropriate for this analysis as they are often seen as more suitable than logit or probit models for panel data methods. Further, the emphasis of the analysis is not on predicting probabilities, and so some of the shortcomings of LPMs are not as relevant. Because some variables are only available at the classroom level, and because prior work used classroom-level information, the models are constructed at the classroom level. The linear probability models are

$$I_c = \beta_0 + \beta_1 TScore_c + \beta_2 Dem_c + \beta_3 Fam_c + \beta_4 CSize_c + \varepsilon_c$$

I_c represents B_c, N_c, or M_c, which are dummy variables indicating whether or not the classroom was taught by a beginning teacher, a novice teacher, or a minority teacher, respectively. These are regressed on a vector of classroom mean student test score variables, $TScore_c$, classroom mean student demographic variables (percent minority, percent female, percent eligible for free lunch), Dem_c, mean family characteristic variables (percent with older sibling, percent with mothers who have a high school degree, percent with no father in household), Fam_c, and a class size variable, $CSize_c$. The percent female, percent minority, and class size variables are classroom-level variables in the ECLS-K data set, while the other means are constructed using individual student-level variables. All variables were collected at the beginning of the kindergarten year.

This model is slightly different from that used in both Kalogrides, Loeb, and Beteille and Dieterle et al., which instead regressed each classroom characteristic independently on a vector of teacher characteristic variables. The choice of the different model is sensible given the different emphasis of the present study. In this study, the only interest is in sorting on teacher experience or race, which were identified as two important teacher characteristics for sorting in both of these earlier studies. For independent variables, a broader set of student characteristics are considered.

The limited sample size of about eight students per classroom, high prevalence of missing data, and lack of variation in dependent variables within schools present a number of challenges. To reduce the loss of sample size due to missing data, missing classroom mean values were replaced with school means where available. These school means are differenced out in the fixed effects model, but the observations are not omitted from the analysis as they otherwise would have been due to their missing classroom information. In cases where a mean value could not be determined for any classroom in the school, all classrooms in the school were omitted. After carrying out the missing value replacement, each regression equation was estimated independently.

Results

Results in Table 3.3, Table 3.4, and Table 3.5 are provided both with and without school level fixed effects for comparison. It is interesting to note that there is little evidence of inexperienced and minority teachers teaching a greater proportion of economically disadvantaged students, measured by free and reduced lunch eligibility, as in earlier studies. The results from those models without fixed effects

confirm earlier findings. Column 1 of Tables 3.3 and 3.4, for example, show that less experienced teachers are more likely to teach in schools with more minority students. The same columns also show some evidence that teachers are sorted across schools based on student gender, though the majority of this association seems to take place within schools. This may make sense, as teachers might be unlikely to switch schools due to a perceived desirability of students based on the gender composition, which is generally quite homogenous. As a note, the pattern observed here does not significantly change when restricting the analysis to only public schools, which are not likely to be exclusively male or female (results not shown).

Even though the traditional regressions provide interesting results not noted in earlier studies, the focus of this study is on within-school sorting, and so fixed effects must be employed. With fixed effects, it appears as though less experienced teachers are assigned a higher proportion of male students, though this is not quite significant for beginner teachers. Male students are often regarded as exhibiting more disruptive classroom behavior, so it could be the case that experienced teachers prefer a greater proportion of female students. Minority students are also generally judged to be more disruptive, and the evidence indicates that beginner and novice teachers are allotted more minority students as well. These findings would seem to be line with Kalogrides, Loeb, and Beteille, given the less-desirable behavior associated with these students.

Perhaps an even more intriguing finding, though, is that less experienced teachers are assigned within schools to classes of a relatively smaller size. Reduced class sizes are sometimes thought to be associated with improved test scores (Glass and Smith 1979, Hedges, Laine, and Greenewald 1994, Slavin 1989), though this is not universally accepted (Hanushek 1997, Hoxby 2000, Robinson

and Wittebols 1986). Assuming that teachers desire classrooms that are smaller so that they can spend less time dealing with disruptive students and more time on instruction (Finn, Pannozo, and Achilles 2003), one might guess that more experienced teachers would exert their influence to obtain reduced class sizes.

In contrast to earlier research, the results also show that novice teachers are assigned classes with higher average reading scores. The theory behind how this might take place is not clear, since the reading scores are not observed until after the classroom assignments have been made. It seems unlikely that novice teachers would be able to affect reading test scores in the short amount of time before students take the tests. Barring that alternative, the relationship documented here would seem to contradict prior work, though the diverging conclusions may only be a result of the differing model choice, which will be examined in more detail in the next section.

The results in Table 3.5 further demonstrate that minority teachers are matched to a greater proportion of minority students, as found in earlier studies, and that minority teachers teach more students who do not have fathers in the household. Given the lower achievement and more problematic behavior of minority students, it seems reasonable to conclude that some teachers might prefer to avoid them when possible. At the same time, it may be the case that minority teachers prefer minority students, or that minority teachers are thought to be more effective with minority students, both theories that have been put forth in earlier works.

For reference, Table 3.3, Table 3.4, and Table 3.5 also provide the results of the same regressions when the independent variables are limited to those most commonly found in education data sets. When excluding the additional variables, the coefficients on the remaining variables remain generally consistent, though they do change slightly. Though certainly not a statistical test, this provides some

evidence that the coefficients in the sorting model are not measurably altered by the inclusion or exclusion of the typically-unobserved variables.

While these results are interesting, one concern is that assignment of students based on experience and race happens simultaneously, and thus independent estimations of the equations could lead to incorrect conclusions. To test this, seemingly unrelated regression (SUR) models are used, which allow the error terms to be correlated across equations. Again using fixed effects, the results from the SUR models are displayed in Table 3.6. Here we find little evidence of sorting based on experience. Regardless of which classification of teacher experience is used, the results confirm the findings that minority teachers are matched to more minority students and more students without fathers in the household. It should be emphasized that the sorting on the presence of a father in the household is true even when controlling for minority status, which is perhaps the most interesting conclusion that can be reached from these results.

It is important to note that the differences in findings between the independent regressions in Tables 3.2-3.5 and the seemingly-unrelated regressions in Table 3.6 appear to be primarily due to the different, smaller sample. The regressions on experience in Table 3.3 and Table 3.4 include a school if it has both a less-experienced and a more-experienced teacher, while the regressions on race in Table 3.5 include a school if it has both a white and minority teacher. In Table 3.6, schools are only included if they have teachers of both experience levels and both race classifications included in the sample. Results from independent regressions limited to those same schools show results in Table 3.7 that are quite similar to the simultaneous case. This implies that the lack of significance on the class size variable, for example, in the seemingly unrelated regression is not due to a difference

in standard errors but is instead simply due to a change in samples.[6]

To circumvent the issues of both simultaneity and a limited sample, a linear probability model considering schools with only white teachers is estimated. If a school only contains white teachers, then no sorting on race could happen simultaneously with sorting on experience. These results are given in Table 3.8. When conditioning on teacher race, there is a strong correlation indicating that less experienced teachers are assigned to smaller classes. There is once again an indication of matching of novice teachers to students with higher reading scores and smaller classes, while the relationship between less experienced teachers and more male students is marginally insignificant in each case. These findings lend credence to the same findings in the full model that was estimated independently, though it isn't clear if these finding would also apply to schools with a mix of teacher races or only minority teachers. Unfortunately, the sample size is not sufficient to test the latter.

Now that the purposeful sorting of kindergarten teachers to students has been examined, it may be informative to investigate how this matching process might evolve as students enter their second year in school. After all, much more information about each student is known at the beginning of the first grade year. Similar models will be used, but lagged cognitive and noncognitive scores will be employed to help determine the extent to which information obtained during kindergarten is used to match students to teachers in the second year. The linear probability models are now

$$I_c = \beta_0 + \beta_1 TScore_{c_{t-1}} + \beta_2 Dem_c + \beta_3 Fam_c + \beta_4 CSize_c + \beta_5 NonCog_{c_{t-1}} + \varepsilon_c$$

In the revised model, $TScore_{c_{t-1}}$ is a vector of mean cognitive test scores from

[6] It may be the case that teachers are assigned to students differently in these schools than other sampled schools, or this might merely be a result of the lower number of observations and resulting impact on tests of statistical significance.

the spring of the kindergarten year across all students in the current first grade classroom. $NonCog_{c_{t-1}}$ is a vector of mean classroom noncognitive scores from the same time. Summary statistics are presented in Table 3.2.

The means, more or less, mirror those in kindergarten, but there are more pronounced differences in free lunch eligibility and the percent of students with older siblings. The noncognitive scores also show some differences, but these are mostly between minority and non-minority teachers. Since minority teachers apparently have more minority students, and since the noncognitive scores are highly correlated with student race, this is not surprising. To separate the variables truly responsible for sorting from those simply correlated with those variables, regression analysis will again be utilized.

Linear probability models with school-level fixed effects will be used, since the focus is on within-school rather than between-school variation. In the ECLS-K data, a small subset of students who switched districts between the kindergarten and first grade years are followed into their new schools. Thus, some schools become part of the ECLS-K sample due to students moving into them. Because these schools have a limited number of sampled students, and because we would expect schools to have less information about students moving into the district than those who were previously in the district, these schools are omitted from the analysis. Additionally, all students moving into sample schools between kindergarten and 1st grade are omitted, again due to the lower level of information known about them. Results from the first grade models are given in Table 3.9, 3.10, and 3.11.

When not using fixed effects, the evidence in Tables 3.9-3.11 makes it clear that teachers are sorted across schools based on student race. Beginner, novice, and minority teachers all teach classrooms that contain, on average, a higher percent of minority students when compared to their experienced and white peers. Minority

teachers are now shown to be matched to fewer students with older siblings. While such a correlation has not been previously noted, it could be that an older sibling provides additional information about a younger student, and that white teachers prefer students about which they have more information.

As opposed to the kindergarten year, there is no longer any apparent sorting on student gender, and the coefficients are no longer close to being statistically significant. Further, the coefficient that relates the percent of the class that are in a minority group to teacher race is now marginally insignificant at the .05 level. Beginner teachers are demonstrated to be assigned smaller class sizes, though this relationship is not quite significant for novice teachers any longer.

Perhaps the most surprising finding is that little additional sorting is observed based on teacher experience in the first grade. The lack of correlation is especially true of novice teachers, but also is the case when looking at beginners, with the only significant relationship being that beginner teachers are assigned more students who do not have fathers present in the household. Thus, when employing a nationally-representative sample and using the model as specified, the results do not confirm the findings in earlier works that novice teachers are matched to students with lower prior test scores or more minority students.

These results are quite striking, especially with respect to test scores. All previous studies have demonstrated that minority teachers are assigned classrooms with lower average test scores. In this case, the data instead show a positive but insignificant relationship between prior test scores and assignment to a minority teacher. Further, previous studies have shown that novice teachers are also assigned classes of lower-scoring students. While the evidence cannot rule out the matching of students to inexperienced teachers based on test scores, any such relationship is nowhere close to being statistically significant.

As in kindergarten, it is possible that classroom assignment based on race and experience might be simultaneously determined, indicating that a seemingly unrelated regression model is appropriate. When considering such a model, the results in Table 3.12 show that novice teachers are matched to students with higher (less-desirable) Internalizing Problem Behavior scores. Less experienced teachers also appear to be sorted to classrooms with higher prior average reading scores, which again would seem to contradict earlier studies. Suggestive evidence of a relationship between experience and class size remains, but the coefficients are now slightly insignificant.

The results with respect to sorting based on teacher race are interesting as well. Here, there is evidence that minority teachers are matched to students with higher prior average reading scores. Once again, it appears as though minority teachers are assigned fewer students with older siblings. When using the beginner distinction of teacher experience, the evidence shows that these teachers are also assigned students with lower (less-desirable) Approaches to Learning scores. As in Table 3.11, the relationship between minority students and placement with a minority teacher is marginally insignificant. Just like in kindergarten, Table 3.13 suggests that any differences between the SUR and independent models are due mostly to a change in samples rather than due to simultaneous assignment.

As before, an independent model for sorting on experience is estimated while only including schools with only white teachers. The results in Table 3.14 show little evidence of sorting based on experience, with no evidence of this type of sorting for novice teachers. When considering beginner teachers, the results show only that these teachers are assigned to more students who do not have fathers in the household. This may be consistent with Kalogrides, Loeb, and Beitelle, since students without fathers in the household tend to have higher Externalizing Problem

Behavior scores, which are indicative of more disruptive classroom behavior. As in previous iterations, the class size variable falls very close to statistical significance.

Given the sometimes differing results, what can be said about the evidence thus far? Conservatively, it seems fair to conclude that less experienced teachers are assigned smaller classes in kindergarten. There is also some evidence that these teachers are assigned more male students and to students with higher average reading scores. Minority teachers appear to be matched within schools to more minority students and more students who do not have fathers in the household. In first grade, there is some indication that less experienced teachers are given small class sizes, but there is little strong evidence beyond that. Minority teachers are matched to more students who do not have an older sibling. Most surprisingly, the only potential evidence of sorting based on prior test scores is that minority and less experienced teachers might be assigned classrooms with higher prior average reading scores, when controlling for the other factors in this analysis. Additionally, the models using the beginner/experienced groupings generally outperformed the models using the novice/experienced groupings, as measured by the correlation coefficients and F statistics. Thus, for the remainder of this paper, the beginner/experienced grouping will be used in all models.

Taking these away as the main conclusions, a few important questions must now be answered.

Why do these results differ from earlier studies?

There are several discrepancies between this study and prior works. First and foremost, there is little evidence that less experienced or minority teachers are assigned classrooms with lower prior test scores, which directly contradicts findings

by Kalorgrides, Loeb, and Beitelle, by Dieterle et al., and by the Strategic Data Project (2012). Why might this be?

One possibility is that it is the a result of the different sample used. None of the earlier studies used a nationally-representative data set, so it is possible that sorting in the schools selected for those studies differs in some way from the sorting in schools in the ECLS-K sample. At the same time, it could also be a consequence of the different model used in this analysis. As noted earlier, the model in this study predicts teacher experience or minority status as a function of several classroom-level variables. The model thus differs from earlier studies where classroom-level variables were independently modeled as dependent on multiple teacher characteristic variables. Such a model perhaps has the advantage of being better suited to identifying the teacher characteristics associated with each classroom-level variable. However, it is not as adept at noting correlation between classroom-level characteristics. That is, when determining the coefficient relating the teacher characteristics to prior test scores, such a model ignores the possibility that minority students, who tend to have lower test scores, might be sorted to less experienced and minority teachers. One might conclude that sorting occurs on test scores when in fact the principal is only matching students to teachers based on minority status.

To see if the different findings are a result of the data or due to the model, a similar model to earlier studies will be used with the ECLS-K data.

$$Characteristic_c = \beta_0 + \beta_1 B_c + \beta_2 M_c + \varepsilon_c$$

$Characteristic_c$ takes on the value of each classroom characteristic variable in independent regressions. For kindergarten, these characteristics are the percent minority, percent female, and percent of students eligible for free lunch. In first

grade, the average lagged reading and math scores are added. These characteristics were chosen as they are those used in prior studies. B_c and M_c are indicator variables for beginner or minority teacher status.

Table 3.15 provides the results, and the general findings do not represent a huge deviation from the previous model. Minority students are still found to be sorted to minority teachers, and male students appear to be sorted to beginner teachers regardless of model specification. Though interesting, the main departure from earlier studies and the results found in this analysis is in the first grade and with respect to test scores and free lunch eligibility. Those results are given in Table 3.16.

Many of the coefficients are consistent with the previous model with respect to their direction and statistical significance. Minority teachers are still found to be matched to more minority students. The relationship between minority teachers and average prior reading test scores is similar, though when using this model the effect is not quite significant. The most important finding, though, is that beginning teachers are shown to be assigned students with lower prior test scores in both the full sample and in the sample that includes only those schools with only white teachers. This is true of both reading and math scores, though the reading score is not significant in the full sample at the .05 level (p=.063). These results are in stark contrast to the model used in this paper, where no remotely-significant negative correlations were encountered, and where the coefficients on reading scores were almost universally positive.

Thus, the results in this paper do not refute prior studies in which less experienced teachers were found to be matched to students with lower prior test scores. Indeed, the findings are consistent with those earlier works. The ECLS-K data do show that beginning teachers are in fact assigned classes with lower prior test

scores, demonstrating this for the first time at a national level. However, the nature of this model and the additional variables put forth the possibility that these assignments are made not entirely on test scores but perhaps on other variables that are correlated with test scores, such as the presence of a father in the household or race. Noting that the relationship might not be based on test scores but on factors correlated with test scores substantially alters the story that these results can tell us. It may be the case that test scores, race, and presence of a father are all noisy indicators of disruptive behavior, and that the matching process occurs on behavior rather than test scores.

One might think that such a distinction is unimportant. After all, if test scores are perfectly correlated with the unobserved classroom characteristic, then controlling for test scores is enough to control for the unobserved information. If test scores are not perfectly correlated with the unobserved variables, though, then bias would be introduced into estimates of teacher-student sorting, and this bias is apparent in the empirical results. Further, such sorting on unobservables might bias other estimates using student characteristics as controls. For example, if classroom behavior is not observed, and if behavioral peer effects are large, then the systematic assignment of poorly-behaving students to certain types of teachers would make those teachers look less effective than they really are.

Does student sorting change by grade?

While inexperienced teachers might be matched to more male students in kindergarten, this appears not to be the case in the first grade. Similarly, only in the first grade do these teachers appear to be matched to students based on the presence of a father. As for minority teachers, the evidence shows that they are sorted

to more minority students and more students without fathers in the household in kindergarten, while they are matched to more students with older siblings in first grade. Of course, it would not be prudent to assume different effects across years simply because a variable is insignificant in one grade but not the other.

Instead, the consistent variables across kindergarten and 1st grade will be statistically evaluated to determine their differences. First, a dummy variable indicating beginner or minority status was estimated using those variables that are present in each year: percent of students with an older sibling, percent with a father present, percent with a mom who has a high school degree, percent female, and percent minority. Then, the generated coefficients from kindergarten were applied to the first grade characteristics to predict beginner or minority status in that grade. Residuals were calculated by comparing the predicted presence of a beginner or minority teacher to the actual presence. Then, these residuals were regressed on the first grade characteristics. School level fixed effects were used in all cases.

If sorting does not change by grade then the coefficients from kindergarten should predict the beginner or minority status in first grade, and any residuals should be independently and identically distributed according to a normal distribution. A correlation between the residuals and the first grade variables would mean that the residuals are not distributed in this way, and thus we can say that the sorting does in fact change.

The results generally confirm suspicions. There is a large and strongly significant correlation between the residual and both the percent of females in the first grade class and the total class size, indicating that the sorting based on these characteristics to teachers of varying experience levels varies by grade. The sorting of minority teachers to both children without older siblings and students who do not have a father present also appears to change across grades, as both of these coef-

ficients are statistically significant. It does not appear as though minority teachers are necessarily assigned to minority students in a way that differs across grades. This is not overly surprising, though, as the coefficient on the sorting variable was only marginally insignificant in the first grade.

As a whole, the findings imply that the teacher-student matching process definitely changes across grades. This may be due to the additional information about each student obtained during the kindergarten year. Otherwise, though unlikely, it may be the case that some characteristics observable in both years are used for sorting in one grade but not the other.

Discussion

This analysis demonstrates that minority students are sorted to minority teachers within schools, and that less experienced teachers are assigned classrooms with lower average prior test scores. These two findings are consistent with results from previous studies, which were not based on nationally-representative samples. However, other findings in the current study differ from previous results. While the correlation between lower-achieving students and less experienced teachers was observed, the sorting was found to instead be based on other variables that are correlated with test scores. There is no conclusive evidence that less experienced teachers are assigned a greater proportion of minority students, or that minority teachers are purposefully matched to classrooms with lower average prior test scores. It is important to note that this is not to say that these claims are not true; rather, they simply cannot be verified using this data set.

While some factors thought to be used in matching teachers to students were not verified in this analysis, other classroom characteristics proved to be good

predictors of teacher experience or race. Less experienced teachers were found to be matched within schools to more male students in kindergarten and smaller class sizes, neither of which had been previously documented. Further, the results show that minority teachers are matched to fewer students with fathers in the household in first grade and, in kindergarten, to fewer students with older siblings, which also have not been established earlier.

Choosing to use the model employed in this study rather than the model used in earlier studies appears to matter a great deal when assessing teacher-student matching. As noted, when applying the model used in earlier studies to ECLS-K data, the finding that less experienced teachers are assigned students with lower prior test scores is replicated. However, it appears as though sorting may occur based on factors correlated with test scores (like presence of a father) rather on test scores themselves, which substantially changes the implications of these results. Test scores are generally observed and so they can be controlled for, while many other potential sorting factors are not.

There is quite convincing evidence that the classroom assignment process differs between grades, potentially implying that information gleaned during the kindergarten year is used to assign students to teachers in the first grade year. The idea that the differing assignment process occurs as a result of the different information available before each grade is only a working hypothesis, though, and not something that can be tested using the available data.

Aside from model choice, the reasons for the differences between this study and earlier work could be many. In the case of older siblings and fathers in the household, the relationship may not have been found earlier simply because such detailed family information is often not available in similar data sets. Additionally, prior studies used data from school districts that have mixtures of teachers and

students that are not at all nationally representative. Thus, some of the prior findings may be applicable to those school districts, or potentially school districts that are similar, but not to all schools. It is also possible that some evidence of sorting is, in essence, lost in aggregation. If some urban school districts tend to sort students with lower prior test scores to novice teachers, for example, and some rural schools districts sort these students to experienced teachers, then no sorting at all will be observed in the data. Ideally, an analysis based on school type would be carried out, but the sample size is too limited to permit this.

One possible weakness of this analysis is that it uses sample data, so estimated proportions of students with relevant characteristics in each classroom are not as precise as in other studies. Earlier studies contained information on every student, allowing for a better detection of differences across classrooms. However, any overestimation or underestimation of proportions in one classroom should correspond to an offsetting overestimation or underestimation in another, so this factor would only be responsible for the differences if teacher sorting was correlated with irregularities in the ECLS-K sampling or weighting.

Overall, this paper provides external validity for some findings in earlier works. Additionally, the evidence shows that variables truly used for matching may differ from those that had been previously been accepted, and that sorting occurs on variables not generally observed in administrative data sets. Future research should assess the impact of such sorting on models which implicitly assume students are randomly distributed conditional upon observed characteristics, since this is shown not to be the case. Specifically, the impact on value-added models must be assessed, particularly in early grades where prior achievement lags are not available.

Tables

Table 3.1: **Kindergarten Classroom Composition by Teacher Experience or Race**

	≤ 3 **Yrs**	> 3 **Yrs**	≤ 7 **Yrs**	> 7 **Yrs**	**Minority**	**White**
Math	48.33*	49.46	48.75	49.43	46.78*	49.65
Reading	46.78*	48.22	47.38	48.12	46.13*	48.51
% Minority	49.59*	41.64	47.19*	41.00	78.28*	37.53
% Female	48.24	48.64	47.86	49.32	47.66	48.12
% Free Lunch	53.14	51.97	52.67	52.04	69.75*	48.12
% w/ Older Sib	50.44	51.26	50.73	51.27	48.14	51.29
% Mom HS Degree	86.86	88.45	87.10	88.89	84.51*	89.38
% No Dad in HH	24.66	23.41	24.55	22.95	41.37*	21.65
Class Size	19.53	19.50	19.92*	19.04	20.47*	19.40

* indicates significant difference in means between teacher experience levels or between minority and white teachers at .01 level

Table 3.2: **1st Grade Classroom Composition by Teacher Experience or Race**

	≤ 3 **Yrs**	> 3 **Yrs**	≤ 7 **Yrs**	> 7 **Yrs**	**Minority**	**White**
Math, Lagged	49.47*	50.50	49.76*	50.60	47.10*	50.71
Reading, Lagged	48.72*	49.79	48.96*	49.99	47.45*	50.08
% Minority	44.71*	34.41	42.92*	31.84	62.80*	33.22
% Female	46.88	47.36	47.01	47.41	47.40	47.17
% Free Lunch	48.18*	43.84	47.42*	42.77	64.69*	41.45
% w/ Older Sib	46.11*	49.32	46.89	49.80	42.71*	48.77
% Mom HS Degree	84.53	86.54	84.49*	87.62	81.96*	87.56
% No Dad in HH	24.46*	21.45	23.33	21.59	37.51*	20.72
Ext Prob Beh, Lag	1.69*	1.64	1.67	1.64	1.76*	1.65
Int Prob Beh, Lag	1.56	1.55	1.57	1.54	1.58	1.56
Self Control, Lag	3.11	3.15	3.12	3.16	2.98*	3.17
Interpersonal, Lag	3.07	3.09	3.07	3.10	2.96*	3.11
App to Learn, Lag	3.05	3.10	3.06	3.10	2.93*	3.10
Class Size	20.13*	18.73	20.41*	21.28	18.98*	3.10

* indicates significant difference in means between teacher experience levels between minority and white teachers at .01 level

Table 3.3: **Independent LPM Results, Beginner vs. Experienced Teachers, With and Without School-Level FE, Full vs. Limited Variables**

	(1) Beginner	(2) Beginner-FE	(3) Beginner	(4) Beginner-FE
Math	-0.0114	-0.00401	-0.0115	-0.00458
	(-1.49)	(-0.40)	(-1.51)	(-0.46)
Reading	0.0147	0.0209	0.0142	0.0214
	(1.70)	(1.86)	(1.65)	(1.91)
% Minority	0.104**	0.0398	0.102**	0.0207
	(2.65)	(0.44)	(2.68)	(0.23)
% Female	-0.197*	-0.205	-0.198*	-0.210
	(-2.00)	(-1.57)	(-2.01)	(-1.61)
% Free Lunch	-0.0444	0.000897	-0.0429	-0.00420
	(-1.21)	(0.02)	(-1.23)	(-0.08)
Class Size	-0.00978***	-0.0209***	-0.00968***	-0.0205***
	(-3.48)	(-3.92)	(-3.45)	(-3.84)
% w/ Older Sib	-0.00139	-0.0111		
	(-0.03)	(-0.23)		
% Mom HS	0.0269	0.102		
	(0.52)	(1.51)		
% No Dad	0.0316	0.0291		
	(0.72)	(0.52)		
Constant	0.656***	0.839***	0.684***	0.935***
	(7.57)	(6.25)	(10.06)	(8.01)
Observations	1792	1792	1792	1792

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Beginner teachers are those with three years or less of experience.

Table 3.4: **Independent LPM Results, Novice vs. Experienced Teachers, With and Without School-Level FE, Full vs. Limited Variables**

	(1) Novice	(2) Novice-FE	(3) Novice	(4) Novice-FE
Math	-0.0119	-0.0170	-0.0122	-0.0166
	(-1.50)	(-1.65)	(-1.52)	(-1.62)
Reading	0.0376***	0.0463***	0.0362***	0.0449***
	(4.06)	(3.93)	(3.93)	(3.83)
% Minority	0.0995*	0.0799	0.112**	0.0854
	(2.53)	(0.86)	(2.93)	(0.93)
% Female	-0.154	-0.265*	-0.151	-0.258*
	(-1.52)	(-2.03)	(-1.49)	(-1.99)
% Free Lunch	-0.0486	-0.0457	-0.0309	-0.0308
	(-1.32)	(-0.90)	(-0.88)	(-0.62)
Class Size	-0.00551	-0.0144**	-0.00563*	-0.0145**
	(-1.93)	(-2.89)	(-1.98)	(-2.91)
% w/ Older Sib	0.00878	0.0333		
	(0.21)	(0.70)		
% Mom HS	-0.0510	-0.0319		
	(-0.96)	(-0.47)		
% No Dad	0.0677	0.0673		
	(1.51)	(1.20)		
Constant	0.732***	0.945***	0.695***	0.938***
	(8.44)	(7.33)	(10.17)	(8.47)
Observations	1783	1783	1783	1783

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Novice teachers are those with seven years or less of experience.

Table 3.5: **Independent LPM Results, Minority vs. White Teachers, With and Without School-Level FE, Full vs. Limited Variables**

	(1) Minority	(2) Minority-FE	(3) Minority	(4) Minority-FE
Math	-0.00301	0.0130	-0.00598	0.0121
	(-0.27)	(0.84)	(-0.53)	(0.78)
Reading	0.000224	0.00763	-0.00457	0.00600
	(0.02)	(0.41)	(-0.31)	(0.32)
% Minority	0.279***	0.396**	0.296***	0.397**
	(4.43)	(2.71)	(4.73)	(2.72)
% Female	-0.194	-0.162	-0.224	-0.173
	(-1.26)	(-0.77)	(-1.45)	(-0.83)
% Free Lunch	-0.100	-0.0996	-0.0476	-0.0503
	(-1.74)	(-1.28)	(-0.87)	(-0.67)
Class Size	-0.00419	-0.0106	-0.00408	-0.0112
	(-1.02)	(-1.35)	(-0.99)	(-1.43)
% w/ Older Sib	0.0580	0.0622		
	(0.97)	(0.88)		
% Mom HS	-0.000132	-0.0405		
	(-0.00)	(-0.45)		
% No Dad	0.214***	0.192*		
	(3.55)	(2.47)		
Constant	0.297*	0.385*	0.358***	0.429**
	(2.41)	(2.14)	(3.63)	(2.76)
Observations	682	682	682	682

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.6: **Simultaneous LPM Results, With School-Level FE**

	(1) SUR, Beg	(2) SUR, Nov	(3) SUR, Min-Beg	(4) SUR, Min-Nov
Math	-0.0200	-0.00891	0.00897	0.0175
	(-1.37)	(-0.66)	(0.67)	(1.39)
Reading	0.0293	0.0295	0.00265	0.00564
	(1.66)	(1.74)	(0.17)	(0.36)
% Minority	-0.0776	-0.196	0.415**	0.496***
	(-0.55)	(-1.47)	(3.20)	(3.97)
% Females	-0.0393	-0.204	0.139	0.0610
	(-0.20)	(-1.07)	(0.78)	(0.35)
% Free Lunch	0.0283	-0.0578	-0.105	-0.145*
	(0.40)	(-0.88)	(-1.62)	(-2.37)
% w/ Older Sib	0.0528	0.0822	0.0526	0.0514
	(0.85)	(1.36)	(0.92)	(0.91)
% Mom HS	0.132	-0.0259	-0.0228	-0.0316
	(1.73)	(-0.35)	(-0.33)	(-0.45)
% No Dad	-0.0240	-0.00343	0.177**	0.184**
	(-0.35)	(-0.06)	(2.82)	(2.97)
Class Size	-0.0142	-0.0118	-0.00250	-0.00770
	(-1.74)	(-1.52)	(-0.34)	(-1.05)
Constant	0.00	0.00	0.00	0.00
	(1.00)	(1.00)	(1.00)	(1.00)
Observations	551	567	551	567

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Beginner teachers are those with three years or less of experience.
Novice teachers are those with seven years or less of experience.

Table 3.7: **Independent LPM Results, Limited to SUR Sample, With School-Level FE**

	(1) Ind, Beg	(2) Ind, Nov	(3) Ind, Min-Beg	(4) Ind, Min-Nov
Math	-0.0200	-0.00891	0.00897	0.0175
	(-1.08)	(-0.52)	(0.53)	(1.10)
Reading	0.0293	0.0295	0.00265	0.00564
	(1.31)	(1.37)	(0.13)	(0.28)
% Minority	-0.0776	-0.196	0.415*	0.496**
	(-0.43)	(-1.16)	(2.52)	(3.13)
% Females	-0.0393	-0.204	0.139	0.0610
	(-0.16)	(-0.84)	(0.61)	(0.27)
% Free Lunch	0.0283	-0.0578	-0.105	-0.145
	(0.32)	(-0.69)	(-1.27)	(-1.86)
% w/ Older Sib	0.0528	0.0822	0.0526	0.0514
	(0.67)	(1.07)	(0.73)	(0.71)
% Mom HS	0.132	-0.0259	-0.0228	-0.0316
	(1.36)	(-0.27)	(-0.26)	(-0.36)
% No Dad	-0.0240	-0.00343	0.177*	0.184*
	(-0.27)	(-0.04)	(2.21)	(2.35)
Class Size	-0.0142	-0.0118	-0.00250	-0.00770
	(-1.37)	(-1.19)	(-0.26)	(-0.83)
Constant	0.651*	1.064***	0.0304	0.164
	(2.54)	(4.58)	(0.13)	(0.76)
Observations	551	567	551	567

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Beginner teachers are those with three years or less of experience.
Novice teachers are those with seven years or less of experience.

Table 3.8: **Independent LPM Results, Only White Teachers, School-Level FE, Full vs. Limited Variables**

	(1) Beginner	(2) Novice	(3) Beginner	(4) Novice
Math	-0.00491	-0.0283	-0.00489	-0.0285
	(-0.34)	(-1.85)	(-0.34)	(-1.87)
Reading	0.0110	0.0460**	0.0108	0.0436**
	(0.76)	(2.80)	(0.74)	(2.67)
% Minority	0.177	0.250	0.160	0.248
	(1.45)	(1.96)	(1.32)	(1.96)
% Females	-0.325	-0.330	-0.327	-0.321
	(-1.89)	(-1.92)	(-1.90)	(-1.87)
% Free Lunch	-0.0257	-0.0368	-0.0196	-0.0201
	(-0.37)	(-0.54)	(-0.29)	(-0.30)
Class Size	-0.0247***	-0.0141*	-0.0242***	-0.0138*
	(-3.56)	(-2.21)	(-3.50)	(-2.17)
% w/ Older Sib	-0.0202	0.0208		
	(-0.31)	(0.31)		
% Mom HS	0.0699	0.000272		
	(0.61)	(0.00)		
% No Dad	0.100	0.136		
	(1.25)	(1.63)		
Constant	0.926***	0.846***	0.993***	0.869***
	(5.01)	(4.65)	(6.66)	(6.05)
Observations	1073	1068	1073	1068

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Beginner teachers are those with three years or less of experience.
Novice teachers are those with seven years or less of experience.

Table 3.9: **Independent LPM Results, Beginner vs. Experienced Teachers, With and Without School-Level FE, Full vs Limited Variabless, 1st Grade**

	(1) Beginner	(2) Beginner-FE	(3) Beginner	(4) Beginner-FE
Math Lag	-0.00985	-0.0185	-0.0100	-0.0194
	(-1.14)	(-1.62)	(-1.16)	(-1.70)
Read Lag	0.0162	0.0186	0.0132	0.0160
	(1.36)	(1.14)	(1.13)	(0.99)
% Minority	0.132***	0.167	0.130***	0.160
	(3.58)	(1.76)	(3.57)	(1.69)
% Female	0.0341	0.0985	0.0308	0.103
	(0.38)	(0.81)	(0.34)	(0.85)
% Free Lunch	-0.0682*	-0.0734	-0.0446	-0.0504
	(-2.15)	(-1.70)	(-1.48)	(-1.20)
Class Size	-0.00637**	-0.0109*	-0.00673**	-0.0114**
	(-2.61)	(-2.54)	(-2.77)	(-2.69)
% w/ Older Sib	-0.0387	-0.0230		
	(-1.28)	(-0.61)		
% Mom HS	0.0216	0.0415		
	(0.66)	(1.04)		
% No Dad	0.0673	0.0974*		
	(1.84)	(2.16)		
EPB Lag	0.0448	0.0228		
	(1.71)	(0.72)		
IPB Lag	-0.00871	0.0170		
	(-0.29)	(0.46)		
INT Lag	0.0477	0.0236		
	(1.19)	(0.48)		
SC Lag	-0.00891	-0.0109		
	(-0.22)	(-0.22)		
A2L Lag	-0.0327	-0.0125		
	(-1.08)	(-0.34)		
Constant	0.448***	0.491***	0.527***	0.578***
	(4.14)	(3.62)	(9.09)	(7.09)
Observations	2209	2209	2209	2209

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Beginner teachers are those with three years or less of experience.

Table 3.10: **Independent LPM Results, Novice vs. Experienced Teachers, With and Without School-Level FE, Full vs Limited Variables, 1st Grade**

	(1) Novice	(2) Novice-FE	(3) Novice	(4) Novice-FE
Math Lag	-0.00675	-0.0123	-0.00845	-0.0145
	(-0.79)	(-1.09)	(-1.00)	(-1.29)
Read Lag	0.0154	0.0107	0.0137	0.00772
	(1.44)	(0.74)	(1.30)	(0.54)
% Minority	0.143***	0.151	0.144***	0.146
	(3.71)	(1.52)	(3.80)	(1.48)
% Female	0.0107	0.102	-0.00155	0.0924
	(0.12)	(0.82)	(-0.02)	(0.74)
% Free Lunch	-0.00918	-0.0350	-0.00778	-0.0270
	(-0.29)	(-0.78)	(-0.26)	(-0.62)
Class Size	-0.00359	-0.00748	-0.00398	-0.00802
	(-1.44)	(-1.69)	(-1.61)	(-1.82)
% w/ Older Sib	-0.0241	-0.0294		
	(-0.78)	(-0.75)		
% Mom HS	-0.0275	-0.0339		
	(-0.82)	(-0.82)		
% No Dad	-0.00595	0.0414		
	(-0.16)	(0.89)		
EPB Lag	0.0101	-0.000976		
	(0.37)	(-0.03)		
IPB Lag	0.0357	0.0551		
	(1.15)	(1.44)		
INT Lag	0.0744	0.0570		
	(1.79)	(1.08)		
SC Lag	-0.0516	-0.0514		
	(-1.25)	(-0.98)		
A2L Lag	-0.0490	-0.0416		
	(-1.58)	(-1.09)		
Constant	0.598***	0.654***	0.577***	0.622***
	(5.26)	(4.55)	(9.86)	(7.55)
Observations	2186	2186	2186	2186

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Novice teachers are those with seven years or less of experience.

Table 3.11: **Independent LPM Results, Minority vs. White Teachers, With and Without School-Level FE, Full vs Limited Variables, 1st Grade**

	(1) Minority	(2) Minority-FE	(3) Minority	(4) Minority-FE
Math Lag	0.0000622	-0.000697	0.00125	0.000231
	(0.01)	(-0.05)	(0.13)	(0.02)
Read Lag	-0.0189	0.0273	-0.0237	0.0186
	(-1.10)	(1.15)	(-1.40)	(0.80)
% Minority	0.139*	0.227	0.170**	0.259
	(2.41)	(1.59)	(2.99)	(1.83)
% Female	0.0681	0.126	0.0660	0.147
	(0.51)	(0.71)	(0.49)	(0.82)
% Free Lunch	-0.0224	0.0209	0.0101	0.0446
	(-0.47)	(0.32)	(0.22)	(0.71)
Class Size	0.00175	-0.00285	0.000925	-0.00530
	(0.51)	(-0.47)	(0.27)	(-0.88)
% w/ Older Sib	-0.133**	-0.178**		
	(-2.90)	(-3.09)		
% Mom HS	0.00616	0.00248		
	(0.13)	(0.04)		
% No Dad	0.0951	0.0792		
	(1.86)	(1.24)		
EPB Lag	-0.0174	-0.0366		
	(-0.46)	(-0.78)		
IPB Lag	0.0137	0.00819		
	(0.30)	(0.15)		
INT Lag	-0.0107	-0.00970		
	(-0.17)	(-0.13)		
SC Lag	0.0196	0.0474		
	(0.32)	(0.64)		
A2L Lag	-0.0590	-0.0988		
	(-1.34)	(-1.82)		
Constant	0.392*	0.483*	0.181*	0.199
	(2.35)	(2.30)	(2.15)	(1.68)
Observations	848	848	848	848

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3.12: **Simultaneous LPM Results, With School-Level FE, 1st Grade**

	(1) SUR, Beg	(2) SUR, Nov	(3) SUR, Min-Beg	(4) SUR, Min-Nov
Math Lag	-0.0226	-0.0112	-0.00639	-0.00842
	(-1.94)	(-0.95)	(-0.59)	(-0.75)
Read Lag	0.0536*	0.0520*	0.0493*	0.0524*
	(2.40)	(2.23)	(2.35)	(2.42)
% Minority	0.0836	0.115	0.175	0.167
	(0.68)	(0.87)	(1.51)	(1.35)
% Female	0.0864	0.167	0.125	0.0541
	(0.55)	(0.99)	(0.84)	(0.34)
% Free Lunch	-0.118*	-0.0170	0.0711	0.0480
	(-2.08)	(-0.28)	(1.33)	(0.87)
% w/ Older Sib	0.00516	-0.00655	-0.165***	-0.127*
	(0.10)	(-0.13)	(-3.49)	(-2.51)
% Mom HS	0.0193	-0.0307	0.00151	0.0135
	(0.38)	(-0.59)	(0.03)	(0.27)
% No Dad	0.0552	-0.0473	0.0710	0.00967
	(0.99)	(-0.81)	(1.36)	(0.18)
EPB Lag	0.0244	0.00225	-0.0694	-0.0711
	(0.61)	(0.06)	(-1.82)	(-1.75)
IPB Lag	0.0633	0.148*	0.00905	0.0770
	(1.32)	(2.77)	(0.20)	(1.55)
INT Lag	0.0147	0.0701	0.0163	-0.0481
	(0.21)	(0.98)	(0.26)	(-0.72)
SC Lag	-0.0271	-0.0563	0.0352	0.0559
	(-0.43)	(-0.82)	(0.59)	(0.88)
A2L Lag	0.0725	0.0138	-0.115*	-0.0719
	(1.54)	(0.28)	(-2.60)	(-1.54)
Class Size	-0.0108**	-0.0111	0.000273	0.00357
	(-2.05)	(-1.96)	(0.06)	(0.68)
Constant	0.00	0.00	0.00	0.00
	(1.00)	(1.00)	(1.00)	(1.00)
Observations	690	613	690	613

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Beginner teachers are those with three years or less of experience.
Novice teachers are those with seven years or less of experience.

Table 3.13: **Independent LPM Results, Limited to SUR Sample, With School-Level FE, 1st Grade**

	(1) Ind, Beg	(2) Ind, Nov	(3) Ind, Min-Beg	(4) Ind, Min-Nov
Math Lag	-0.0226	-0.0112	-0.00639	-0.00842
	(-1.51)	(-0.73)	(-0.46)	(-0.59)
Read Lag	0.0536	0.0520	0.0493	0.0524
	(1.88)	(1.74)	(1.84)	(1.88)
% Minority	0.0836	0.115	0.175	0.167
	(0.53)	(0.67)	(1.18)	(1.05)
% Female	0.0864	0.167	0.125	0.0541
	(0.43)	(0.77)	(0.66)	(0.27)
% Free Lunch	-0.118	-0.0170	0.0711	0.0480
	(-1.62)	(-0.22)	(1.04)	(0.68)
% w/ Older Sib	0.00516	-0.00655	-0.165**	-0.127
	(0.08)	(-0.09)	(-2.73)	(-1.96)
% Mom HS	0.0193	-0.0307	0.00151	0.0135
	(0.30)	(-0.45)	(0.02)	(0.21)
% No Dad	0.0552	-0.0473	0.0710	0.00967
	(0.78)	(-0.63)	(1.06)	(0.14)
EPB Lag	0.0244	0.00225	-0.0694	-0.0711
	(0.47)	(0.04)	(-1.43)	(-1.36)
IPB Lag	0.0633	0.148**	0.00905	0.0770
	(1.02)	(2.16)	(0.16)	(1.21)
INT Lag	0.0147	0.0701	0.0163	-0.0481
	(0.17)	(0.76)	(0.20)	(-0.56)
SC Lag	-0.0271	-0.0563	0.0352	0.0559
	(-0.33)	(-0.64)	(0.46)	(0.68)
A2L Lag	0.0725	0.0138	-0.115*	-0.0719
	(1.20)	(0.22)	(-2.03)	(-1.21)
Class Size	-0.0108	-0.0111	0.000273	0.00357
	(-1.60)	(-1.53)	(0.04)	(0.53)
Constant	0.318	0.362	0.462*	0.354
	(1.42)	(1.47)	(2.19)	(1.54)
Observations	690	613	690	613

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Beginner teachers are those with three years or less of experience. Novice teachers are those with seven years or less of experience.

Table 3.14: **Independent LPM Results, White Teachers Only, With School-Level FE, 1st Grade**

	(1) Beginner	(2) Novice	(3) Beginner	(4) Novice
Math Lag	-0.0208	-0.0219	-0.0278	-0.0284
	(-1.00)	(-1.10)	(-1.34)	(-1.45)
Read Lag	0.00491	-0.00553	-0.00127	-0.00691
	(0.22)	(-0.29)	(-0.06)	(-0.37)
% Minority	0.235	0.166	0.225	0.167
	(1.78)	(1.20)	(1.70)	(1.21)
% Females	0.159	0.0699	0.137	0.0386
	(0.97)	(0.42)	(0.84)	(0.23)
% Free Lunch	-0.0198	-0.0224	0.00558	-0.00413
	(-0.34)	(-0.38)	(0.10)	(-0.07)
Class Size	-0.0114	-0.00775	-0.0123*	-0.00867
	(-1.92)	(-1.28)	(-2.07)	(-1.43)
% w/ Older Sib	-0.0446	-0.0660		
	(-0.89)	(-1.28)		
% Mom HS	0.0490	-0.0456		
	(0.90)	(-0.81)		
% No Dad	0.148*	0.114		
	(2.32)	(1.73)		
EPB Lag	0.0706	0.0273		
	(1.62)	(0.60)		
IPB Lag	-0.0139	0.00440		
	(-0.28)	(0.09)		
INT Lag	0.0801	0.0354		
	(1.18)	(0.51)		
SC Lag	-0.00725	-0.0221		
	(-0.10)	(-0.31)		
A2L Lag	-0.0916	-0.0657		
	(-1.78)	(-1.26)		
Constant	0.449*	0.758***	0.527***	0.645***
	(2.34)	(3.77)	(4.69)	(5.64)
Observations	1365	1424	1365	1424

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Beginner teachers are those with three years or less of experience.
Novice teachers are those with seven years or less of experience.

Table 3.15: **Coefficients from Model Used in Prior Studies, Kindergarten**

	(1) Pct Minority	(2) Pct Females	(3) Pct Free Lunch
Beg - Full Sample, FE	0.002	-0.014*	0.002
	(0.26)	(-2.40)	(0.10)
Beg - SUR Sample, FE	-0.010	-0.009	0.003
	(-0.77)	(-0.89)	(0.11)
Beg - WT Schools, FE	0.009	-0.016*	-0.004
	(0.95)	(-2.16)	(0.23)
Min - Full Sample, FE	0.033*	0.001	-0.007
	(2.51)	(-.09)	(-0.27)
Min - SUR Sample, FE	0.037**	0.004	0.000
	(2.61)	(0.39)	(0.01)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Each dependent variable is given in the column heading

Beginner teachers are those with three years or less of experience.

Novice teachers are those with seven years or less of experience.

Table 3.16: **Coefficients from Model Used in Prior Studies, 1st Grade**

	(1) Math	(2) Reading	(3) PctMinority	(4) PctFemales	(5) Pct Free Lunch
Beg - Full, FE	-0.754*	-0.718	0.006	-.003	-.009
	(-2.20)	(-1.86)	(1.10)	(-0.65)	(-0.65)
Beg - SUR, FE	-0.121	-0.034	-0.003	-0.000	-0.39
	(-0.18)	(-0.04)	(-0.24)	(-0.04)	(-1.52)
Beg - WT, FE	-1.055**	-1.104*	0.013	-0.002	0.009
	(-2.60)	(-2.47)	(1.96)	(-0.36)	(0.52)
Min - Full, FE	0.024	1.007	0.023*	.010	0.016
	(0.04)	(1.58)	(2.40)	(1.14)	(0.69)
Min - SUR, FE	0.105	0.987	0.024	0.013	0.032
	(0.15)	(1.20)	(1.77)	(1.23)	(1.19)

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Each dependent variable is given in the column heading

Beginner teachers are those with three years or less of experience.

Novice teachers are those with seven years or less of experience.

Table 3.17: **Regression of 1st Grade Experience or Minority Status on Residuals Generated Using Kindergarten Coefficients**

	(1) Beginner	(2) Minority
% Minority	-0.113	0.158
	(-1.20)	(1.13)
% Female	-0.309**	-0.231
	(-2.58)	(-1.31)
% Free Lunch	0.0634	-0.114
	(1.49)	(-1.80)
% w/ Older Sib	0.00247	0.215***
	(0.07)	(3.86)
% Mom HS	0.0566	-0.00294
	(1.45)	(-0.05)
% No Dad	-0.0768	0.136*
	(-1.75)	(2.16)
Class Size	-0.0118**	-0.00702
	(-2.83)	(-1.19)
Constant	0.327***	0.0572
	(3.89)	(0.46)
Observations	2258	873

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Chapter 4

A Closer Look at the Labor Market Benefits of Class Size Reduction

Introduction

A number of recent papers have used data from the Tennessee STAR experiment to estimate the long-term effects of reduced class sizes on various educational outcomes. Krueger and Whitmore (2001) show positive but insignificant effects on eighth-grade test scores and significant effects on both college entrance test-taking and college entrance exam scores. Finn, Gerber, and Boyd-Zaharias (2005) find significant impacts on high school graduation. Chetty et al. (2010) document effects on college attendance, while Dynarksi, Hyman, and Whitmore-Schanzenbach (2011) find positive impacts on both college attendance and degree completion. These and other findings have led some to conclude that class size reduction might be an important tool for improving labor market outcomes and that these labor market benefits outweigh the costs of the intervention (Krueger and Whitmore 2001, Schanzenbach 2006, Chetty et al. 2011). Estimates of class size effects on early test scores also have informed policy that targets funding for smaller class sizes toward schools serving children from lower-income families.

To determine cost effectiveness, costs of lowering class sizes need to be compared to the monetary benefits. These monetary benefits stemming from changes in educational outcomes have not been closely scrutinized. In earlier works the labor market benefits were calculated only by estimating the effect of class size reduction on test scores and then calculating a corresponding labor market benefit using a series of assumptions. Now that more information on other outcomes is available, it seems prudent to compare these estimates based on test scores to those based on other outcomes and to examine their underlying assumptions. For example, Krueger and Whitmore (2001) and Schanzenbach (2006) used a benefit multiplier, which relates improvements in test scores to gains in future wages,

that was larger than later research suggests, and implicitly assumed that effects were homogenous across races and gender. Both the estimated effects on various outcomes and estimated returns to education vary substantially across groups, so these assumptions might overstate the labor market benefits. Moreover, little attention has been paid to the long-term effects of small class size for children from lower versus middle-income households.

Complicating an analysis of several outcomes is the difficulty in determining which benefit estimate is appropriate. Many outcomes have been analyzed, and several, such as college entrance and college degree completion, or test scores and high school graduation, are likely to be correlated with one another. A typical analysis might simply choose the largest benefit under the assumption that this benefit includes all smaller benefits while also including an additional benefit, as this is the generally-accepted approach. This paper puts forth a basic theoretical argument for why such a procedure may systematically overestimate benefits, which was found empirically in Aos et al. (2012). The multiple benefit estimates are assessed using three approaches: the typical approach, a solution proposed in Aos et al., and also in terms of an empirically-driven method derived from prior educational research, so that the various approaches can be compared and assessed for the first time.

The contributions of this paper are several. Most importantly, I show that past estimates of labor market benefits of small class sizes may be too high, and that it is very difficult to conclude that such an intervention generates a positive return on investment when solely considering future wage earnings. While some may argue that these prior estimates capture benefits not accounted for by other outcomes, I caution that it instead may be the case that such estimates were actually subject to the largest error. I also temper earlier research regarding class size reduction

and high school graduation, demonstrating that the previously-documented effect might have been an artifact of issues in study implementation and selection of an inappropriate control group. Finally, I show that labor market benefits accrue almost exclusively to males.

After a review of the literature and a discussion of the data, prior studies examining long-term educational benefits will be closely explored. For each outcome, new estimates of labor market benefits are generated, and the sensitivity to various assumptions is tested. A theory outlining why the typical treatment of multiple, potentially-overlapping benefits may be inappropriate is put forth and the multiple estimates are then analyzed in light of this theory and Aos et al.'s recent empirical work. A discussion of the findings follows.

Research on Class Size Reduction

Theories abound as to why students might perform at a higher level if placed into smaller classes. Some studies suggest that teachers are able to offer better instruction (Zahorik 1999), but others (Shapson et al. 1980) have refuted this concept. Generally, in fact, class size reductions do not seem to induce detectable differences in teaching practices (Stasz and Stecher 2002). While other studies argue that teachers who are responsible for fewer students are able to spend more time on instruction and less time dealing with behavioral issues (Molnar, Smith, and Zahorik 1999), perhaps the most convincing research finds that test score gains may result from teachers being more able to incentivize those students who are less engaged (Babcock and Betts 2009).

Though the mechanism through which small class sizes act as a catalyst for improved performance seems to have been identified, the question of whether or

not reduced class sizes actually are related to an increase in test scores remains surprisingly unsettled. As Rockoff (2009) notes, class size was the subject of much inquiry in the early 20th century, with school districts seeking to reduce costs per pupil. The vast majority of these studies found no negative effect of increasing class sizes, informing education policy for years to come. Results of later analyses were decidedly mixed. Glass and Smith (1979) and Slavin (1989) both found positive impacts of class size reduction, while Robinson and Wittebols (1986) showed that reduced classes do not necessarily result in improved test scores in general, even if there are effects in earlier grades and among lower-income students.

Researchers subsequently turned to modern econometric techniques in an attempt to settle the class size debate, with some finding at least some indication that reduced class sizes may result in test score gains. In their study of Israeli public schools, Angrist and Lavy (1999) noted a significant increase in test scores due to small class sizes among fourth and fifth graders, even if the same was not true of third graders. Jepsen and Rivkin (2009) found positive effects on test scores, but these were offset by decreases in teacher quality which followed from the hiring of more teachers in order to reduce class sizes. Dee and West (2011) did not detect a significant effect overall, though they did note a positive impact on test scores among students in urban schools.

Still other research found no evidence that class size reduction provided any significant gain. Hanushek (1997) questioned earlier studies on class size, given that negative effects of larger teacher-pupil ratios are found nearly as often as positive effects. Exploiting longitudinal population variations, Hoxby (2000) did not detect significant effects of reduced class sizes on achievement either. Chingos (2010) examined the class size reduction initiative in Florida, finding no significant relationship between class size and test scores. Thus, arguments regarding the

impact of smaller classes on student test scores remain unsettled.

Perhaps at least partially to blame for the inconsistent findings is that nearly all of the modern studies were either nonexperimental or, in the case of the meta-analyses, relied almost exclusively on nonexperimental studies. Due to the reliance on this framework, much weight has been given to the study with the best design and implementation. After all, as Krueger (2003) notes, "...one good study can be more informative than the rest of the literature (p.35)." That study, according to Krueger, is Tennessee's Project STAR, an experiment in which students and teachers were randomly assigned to different sizes of classrooms. In recent years studies based on STAR data have come to dominate the discussion about small class sizes due to the experiment's preeminent status as a controlled, randomized experiment in class size reduction.

The earliest analyses of STAR data showed a strong relationship between smaller class sizes and an improvement in student test scores (Word et al. 1990, Finn and Achilles 1990). While subject to some issues, the general findings have been confirmed in later studies (Ding and Lehrer 2010, Nye, Hedges, and Konstantopoulos 2000, Krueger 1999) which attempted to mitigate earlier problems. Though Hanushek, one of the most prominent doubters of class size reduction as effective policy, cites errors in design and implementation as reason to discount other studies, even he concludes that there are significant effects from reduced class sizes in kindergarten and perhaps first grade (1999). Overall, the evidence from the sole large-scale experiment in this area seems to strongly indicate that students assigned to smaller classrooms see an improvement in achievement.

It is less clear, though, how the effects of small class sizes might be experienced in later grades. Some might assume that all learning from one grade persists into later years (Dieterle et al. 2012), but this isn't necessarily the case. For exam-

ple, Jacob, Lefgren, and Sims (2010) show that at least three-quarters of teacher-induced gains fade out within one year. Additionally, some educational interventions have been demonstrated to produce short-term but not long-term gains, while others lead to substantial gains that carry over well into the future (Barnett 1995). As mentioned earlier, a number of studies have used Project STAR data to determine the long-term effects of small class sizes. Each of these will be discussed in more detail following a discussion of the Tennessee STAR data.

Data

Tennessee's Project STAR was a four-year longitudinal experiment carried out in selected Tennessee schools starting in 1985. The Project followed a single cohort of students beginning in kindergarten through their third grade year. Intending to measure the effects of class size on students, each student was randomly assigned to either a small classroom (13-17 students), a regular classroom (22-25 students), or a regular classroom with an aide (22-25 students) (Krueger 1999). Krueger explores the validity of the random assignment procedure and finds that the randomization was carried out quite well, with no evidence of sorting of students to class type based on observable characteristics. Students were supposed to be placed into the same type of classroom in all four grades, but this didn't always prove to be the case, as will be discussed. To ensure that the study could not be contaminated by principals' assignment of teachers to classrooms, this assignment was performed randomly as well.

Every school in Tennessee was invited to be a part of the experiment, of which 180 schools across 50 districts responded with a desire to participate. However, only 100 of these schools had a sufficient number of kindergarten students to have

one class of each type, so the remainder of the schools could not be included. The Tennessee State Department of Education, who provided the funding for the project, mandated that schools from inner city, suburban, urban, and rural areas were all included. After taking this direction into account, 79 schools in 42 districts were ultimately included in the experiment. There was some attrition in the schools, with only 75 remaining by the conclusion of the experiment (Finn et al. 2007).

The end result was approximately 6,300 students being included in the experiment in kindergarten. Students beginning school in first grade were also randomized into the various class types (Finn et al. 2007). The number of these students (2,314 (Hanushek 1999)) was quite large, since kindergarten was not mandatory in Tennessee at the time of the study (Sojourner 2013). Though less than in first grade, a sizable number of students moved into STAR schools and were randomized into classrooms in second and third grade, too (1,791 and 1,389 (Hanushek 1999), respectively). The sample was also affected by outward mobility, with some students moving from one STAR school to another, and with other students moving out of the study entirely. Students moving between STAR schools were assigned to the same type of classroom, space permitting. Students moving from a STAR school to a non-STAR school were not followed. This movement out of STAR schools did have an effect on class sizes, sometimes resulting in the 'regular' classrooms actually containing between thirteen and seventeen students.

A relatively limited amount of information was recorded during the study, such as student demographic variables, achievement test scores, motivation and self-concept scores, and school and teacher information. Crucial to this study, a substantial amount of data was added to the student records later on, including information on whether or not the student graduated from high school and if they took a college entrance exam. The high school graduation variable was determined from

high school transcripts and the Tennessee State Department of Education (Finn et al. 2007). For about 57% of students, graduation status could not be ascertained, and students for which this is true were excluded from the analysis of effects on high school graduation. Information about college entrance test-taking was added through a matching process carried out by the organizations which administered the ACT and SAT tests, as described in Krueger and Whitmore. Some studies on the long-term effects utilized additional, external data sets which will be described as part of the in-depth discussion of each paper.

Labor Market Benefit Estimates Based on Eighth Grade Test Scores

To begin, the labor market benefits derived from gains in eighth grade test scores will be considered. Krueger and Whitmore (2001) find a positive but insignificant impact on eighth grade test scores when using school-by-entry-year fixed effects. In other words, students are classified based on their initial classroom size assignment and are compared to the other students in their school cohort. This is quite important, since Hanushek (1999) noted that 9-12% of students switched between class size types each year. It could be the case that students switching from regular classes into smaller classes in first grade have parents who differ from the parents of children who did not switch, for example, so this switching could bias results. However, the use of these school-by-entry-year fixed effects helps to mitigate this problem. Such a technique has come to be accepted as perhaps the best for analyzing the STAR data set, with most recent influential papers utilizing it (see Schanzenbach 2006, Chetty et al. 2011, Dynarski, Hyman, and Whitmore-Schanzenbach 2012, for example).

After correcting for this switching issue by using the fixed-effects technique, Krueger and Whitmore (2001) find that a reduction in class size is associated with an improvement in eighth-grade test scores of about 1.5%. This figure is not homogenous across race, with a gain of just over 2% for black students compared to about 1% for white students. Because numerically-precise estimates were not given[7], and because labor market earnings from test score gains are most often estimated using standardized test scores, their analysis was replicated using standardized scores. The model is:

$$A_i = \beta_0 + \beta_1 X_i + \beta_2 T_i + \epsilon_i$$

The vector X_i contains student demographic characteristics including race, gender, and free lunch eligibility. The dummy variable T_i indicates whether or not the student was initially assigned into a small class, while A_i is the student's standardized test score in either reading or math, or, as in Krueger and Whitmore, the average of the two scores. Each regression was carried out for all students (Table 4.1), by race (Table 4.2 and Table 4.3), by gender (Table 4.4 and Table 4.5), and by free lunch eligibility (Table 4.6 and Table 4.7).

The results are not overly surprising. Small class sizes boost test scores by about one-tenth of a standard deviation. These gains are driven primarily by improvements in reading scores. In a departure from Krueger and Whitmore, who found larger impacts for black students than for whites, here the effect appears to be roughly constant across races, or perhaps even stronger among white students. On the other hand, students who were eligible for free lunch experienced larger test score gains than did their higher-income peers. Also interesting is the extent to which boys drive the improvement in math scores, with girls exhibiting only a tiny, insignificant gain.

[7] Krueger and Whitmore expressed these results only in graphs

Of course, primary interest is in the labor market benefits that students might derive from these test score gains. Following Aos et al., such estimates can be generated using the formula below:

$$PV = \sum_{y=age}^{65} \frac{1.01^{y-age}[(\alpha \times .74 \times \beta_2 \times ExpectedEarnings) - ExpectedEarnings]}{1.03^{(10+y-age)}}$$

First, using a method outlined in Aos et al., expected labor market earnings are calculated by age for all persons aged 18-65 who attained at least the seventh grade. This is done both across all students and within race and gender groups, using data from the 2000 March Current Population Survey, which is when these students would have graduated from high school. The expected labor market earnings are subsequently inflated by 1% per annum.

Then, test score gains are linked to wage gains by multiplying the inflated expected labor market earnings by (α + 1), which represents the percent gain in annual earnings stemming from a one standard deviation increase in standardized test scores at the conclusion of high school. As one might surmise, the α value chosen can have a large impact on estimates. If too high of a value is chosen, the estimated labor market gains will be too high, and if too low of a value is chosen, the estimated labor market gains will subsequently be too low. Both Krueger and Whitmore (2001) and Schanzenbach (2006) set α = .2, based on Neal and Johnson's (1996) work relating AFQT test scores to future labor market earnings. However, later analyses have suggested much smaller values. Hanushek's (2009) survey of the literature led him to conclude that a one standard deviation increase in math scores corresponds to a 12% increase in earnings, while Aos et al. came to a similar conclusion, finding a median estimate in the literature of 11.8%. Thus, setting α = .118 might also be appropriate, perhaps even more so than the .2 value used in earlier works.

Since the test scores were observed at the end of eighth grade, and since induced test score gains often fade over time, they must be adjusted downward to reflect fadeout in scores between eighth grade and the end of high school. Aos et al. suggest using a fadeout multiplier of .74, since they find that, on average, 74% of gains realized in eighth grade persist to the end of twelfth grade.

At this point, the figure described reflects the wage gains from a one standard deviation increase in test scores observed at the end of eighth grade. To find the wage gains from small class sizes, then, the figure must by multiplied by β_2, which gives the relationship between assignment into a small class and eighth grade test scores. Finally, the expected wage gains from the small class assignment are discounted at a rate of 3% per year back to the point of intervention.

This procedure was carried out across all students and then separately by race and gender groups. That is, to calculate the expected wage gains for males, for example, the estimated coefficient from Table 4.4 was multiplied by wage gains experienced by males for a one standard deviation increase in test scores. Since an identifier for prior free lunch eligibility is not available in the Current Population Survey, labor market gains stemming from class size reduction were calculated by multiplying the coefficients from Tables 4.6 and Tables 4.7 by the overall wage gains across all student types.

Table 4.8 provides estimates of labor market benefits resulting from class size reduction's impact on eighth grade test scores using α values of .118 in column 1 and .20 in column 2. As expected, the estimates using the larger α are substantially larger. However, even using the higher α value, the estimated benefits are just under the costs of the program, which Chetty et al. (2011) estimated at $7,508 (adjusted to 2000 dollars)[8]. Regardless of the specification, it appears as though

[8]The collection of cost data was not part of the Tennessee STAR design. Chetty et al. estimated

white students benefit more than blacks, male students more than females, and low income students more than high income students.

Hypothetically, using the overall measures, found by multiplying the overall labor market gains by the overall effect, could produce inaccurate estimates. For example, if the effect is largest among males, and if labor market returns are largest among males, then the overall effect might underestimate benefits. An alternative estimate can be created by calculating a weighted average across race, gender, and free lunch eligibility based on the proportion of the sample in the respective group. Carrying this out across races or genders yields estimates that are lower than the overall estimate for each α value. Performing this calculation by free lunch eligibility pushes the estimated benefits a bit larger, but they are still lower than the program costs, as demonstrated in Table 4.9.

Though these measures are all lower than estimates in earlier research, they may in fact still be too high. Returning to Table 4.1, it is apparent that most of the test score gains are in reading, with only a smaller improvement in math. Hanushek's (2009) estimate of a 12% gain in annual earnings was actually specific to math scores, based partly on the work of Murnane et al. (2000) who found that reading scores had no explanatory power on future earnings after controlling for math scores. As a result, creating a labor market benefit estimate based on the composite score might grossly exaggerate benefits. Column 3 of Table 4.8 provides estimates of benefits for all students and by race and gender when calculated based only on math scores.

Total estimated benefits are now much lower, at only $2,495, as shown in column 3 of Table 4.8. Perhaps the most striking result is that girls receive almost no benefit. Calculating benefits by race, gender, or free lunch eligibility and then tak-

the cost of class size reduction to be $9,355 in 2009 dollars

ing a weighted average yields results that are slightly larger, but still far, far below earlier estimates, and far below the cost threshold. Thus, though several estimates of total labor market benefits were generated, none of those based on eighth grade test scores were larger than the costs of the intervention.

Labor Market Benefit Estimates Based on Entrance Exam Scores

Krueger and Whitmore (2001) also looked at the effect of class size reduction on college entrance exam scores, which they ultimately used for estimating labor market benefits. For this model, they used school fixed effects instead of school-by-entry year fixed effects. However, in this analysis, school-by-entry-year fixed effects will still be used for consistency. Cognizant of the fact that small classes apparently spurred more students to take a college entrance test, and that these marginal test-takers might be weaker than the average student, they used both a linear truncation and Heckman correction method to try to compensate for potential selection. Results were fairly consistent under either specification.

The Heckman procedure may be problematic since it is not possible to estimate a probit model with fixed effects. The linear truncation method rests on the questionable assumption that the additional students taking the exam due to small classes are exclusively the lowest-scoring students. Thus, the linear truncation method provides sort of an upper bound. The model is similar to before:

$$C_i = \beta_0 + \beta_1 X_i + \beta_2 T_i + \epsilon_i$$

The college entrance score is denoted by C_i. For most students, this is their standardized ACT score, but for a small number who took only the SAT exam, it is their SAT score converted to the ACT scale, using a table provided by The College Board, and subsequently standardized (Krueger and Whitmore 2001). When using

linear truncation, the equation is estimated excluding the lowest-scoring students who were randomized into a small class such that the proportion of students taking the exam is equal across the treatment and control groups.

I am mostly able to replicate Krueger and Whitmore's results, as demonstrated in Table 4.12, Table 4.13, Table 4.14, and Table 4.15. Small classes appear to lead to an increase of .18 standard deviations in college exam scores across all students. This effect is quite heterogenous, however. Male, black, and low-income students derive substantial gains, while female, white, and higher-income students receive a smaller benefit. Small discrepancies may arise from the fact that I continued to utilize school-by-entry year fixed effects rather than switching to the school fixed effects they used.

Krueger and Whitmore then performed a procedure similar to what I carried out with eighth grade test scores, using an α value of .2, as noted earlier. One important deviation from before is that a fadeout multiplier is not necessary, as the scores were observed at the conclusion of high school. Table 4.8 presents results from ACT scores using an α value of .118 in column 4, and of .2 in column 5. When using ACT scores, the overall benefit is substantially above the estimated cost, regardless of which α value is used. Due to the large labor market gains exhibited among male, black, and low-income students, column 4 of Table 4.10 shows that the weighted averages exclusively show a positive return on investment.

But, as with eighth grade test scores, it may not be appropriate to apply the composite ACT score to expected earning gains. While Neal and Johnson's derivation of the relationship between score gains and earnings used the AFQT, which is a similarly-composite score, other studies have failed to find that gains in subjects other than math actually matter. Table 4.15 shows that gains in ACT test scores are fairly constant across subjects, with math gains actually being nearly tied with

science gains as the largest. Still, though, it isn't clear whether or not these gains in other subjects actually matter.

Column 6 of Table 4.8 presents estimates from using only the ACT math scores and an α value of .118. Of course, given the much smaller effect size, the estimated benefits are much lower than before. Driven by the large math score improvements by males and their high returns to education, the weighted average based on gender is slightly below the estimated costs. Based on gains in ACT scores, then, there is some evidence that the labor market benefits of small class sizes exceed the cost, but overall the results are mixed. This is interesting considering that these mixed results were calculated when assessing the benefits at the upper bounds of the effect sizes. In other words, the benefits calculated at the upper bounds are still only sometimes larger than the costs.

Labor Market Benefit Estimates Based on High School Graduation

High school graduation is another outcome that has been linked to small class size. While this link should be treated with caution due to the prevalence of missing data with respect to the high school graduation outcome, perhaps the most significant contribution in this area comes from Finn, Gerber, and Boyd-Zaharias (2005). Using hierarchical linear models to analyze Tennessee Project STAR data, they find that three or more years of smaller classes increases the chances that a student graduates high school. However, they did not account for the issues in implementation noted by Hanushek (1999). Most importantly, Finn, Gerber, and Boyd-Zaharias did not consider that 9-12% of students switched between class size types each year, and perhaps didn't use an appropriate control group. Thus,

the observed increase in graduation might truly only be reflective of unobserved differences in family or other characteristics.

I replicated this analysis both with and without the school-by-entry-year fixed effects designed to alleviate the switching problem using a simple logit model:

$$HS_i = \beta_0 + \beta_1 X_i + \beta_2 3YRS_i + \epsilon_i$$

HS_i is a dummy variable indicating whether or not the student graduated high school, X_i is again a vector of individual-level characteristics, and $3YRS_i$ is a dummy variable set to 1 if a student was randomized into a small classroom and remained in one for at least three years.

Column 1 of Table 4.19 shows positive impact of three years of small class sizes on high school graduation noted in Finn, Gerber, and Boyd-Zaharias. This model, though, compares students who experienced three years of small classes, and thus typically didn't switch schools or otherwise exit the study, to all other students who were assigned to regular classrooms. Column 2 restricts the sample to only those students who were in a Project STAR school for at least three years. This greatly reduces the coefficient, leading one to believe that differences in student mobility may be partly behind the earlier findings.

In column 3, the sample is no longer restricted to only those students who were in a Project STAR school for at least three years, but school-by-entry-year fixed effects are added in. Compared to column 1, it appears as though a substantial portion of this effect may be due to between-school variation, too. The coefficient further shrinks in column 4, when the sample is again restricted and fixed effects are imposed. The effect all but disappears using the preferred specification based on initial assignment and using fixed effects in column 5. Indeed, the p-value (.56) is nowhere near statistical significance.

As with test scores, these effects are broken down by race, gender, and income. For the purpose of comparison with other outcomes to be discussed, linear probability models will be used here instead of a logistic regression as in Table 4.19. Table 4.20, Table 4.21, and Table 4.22 provide the LPM results for all students as well as broken down by student type. As one would expect, the effect is still far from significance when utilizing the linear probability model. When breaking down the students by gender, race, or income, the effect on high school graduation is still not significant for any group. A pattern is perhaps beginning to emerge, though, as males again appear to receive a much larger benefit than females.

Benefits from high school graduation or other attainment outcomes were calculated using data from the 2000 March Current Population Survey, similar to the benefits from test scores. Mean earnings for both those attaining a certain level and those not attaining that level but at least completing seventh grade were calculated separately for each age 18-65. Wages were again inflated by 1%, and the difference in wages was discounted at a 3% annual rate. These discounted sums of the growth-inflated differences were then added together to determine the present value. Expressed mathematically,

$$PV = \sum_{y=age}^{65} \frac{InflatedAttainedEarnings - InflatedNotAttainedEarnings}{1.03^{(10+y-age)}}$$

This present value was calculated for all persons, and separately by both race and gender, and then the relevant present values were multiplied by the corresponding group-specific coefficients. This allows for the calculation of labor market benefits for males and females independently, for example. Following this method the overall labor market benefit estimate is slightly less than half of the estimated cost, as demonstrated in column 7 of Table 4.8. The weighted average by race is lower, while the weighted average by gender and income are both slightly larger.

However, even these are well below the cost. Thus, there is little evidence to suggest that randomization into a small classroom generates a positive return on investment when measured through high school graduation.

Labor Market Benefit Estimates Based on College Attendance

Another major study looking at a long-term outcome is that of Chetty et al. (2011). They determined whether or not a Project STAR student attended college by matching students to 1098-T forms provided to them by the IRS. Such forms report tuition payments and scholarships received, and so every student attending a college, university, vocational school, or other postsecondary institution receives a 1098-T. They find that black and free lunch eligible students are more likely to attend college on time and attend college by age 27. White students are found to be more likely to attend college on time but less likely to attend college by 27, though neither of the results with respect to white students are statistically significant. As with math test scores and high school graduation, males experience a much larger positive impact from class size reduction, as they are 2.3% more likely to attend college by 27, compared with about .5% for females.

The labor market benefits from college attendance were estimated in a way that is quite similar to how the benefits from high school were estimated. Again utilizing the 2000 March Current Population Survey, mean earnings for those having attended some college and for those who never attended college were computed by age for all persons 18-65. After inflating wages and discounting the differences as before, the differences were again summed together to find the present value of college attendance, and this figure was multiplied by the corresponding regression coefficient to estimate the gains resulting from class size reduction. The results

from this exercise are presented in column 2 of Tables 4.9 and 4.11. Using the typical method, the overall benefit is only $4,685. Continuing the pattern, males appear to receive a much larger benefit in terms of future wage gains than do females.

When carrying out the estimation based on the weighted average approach by income, an estimate of $9,191 is generated. Thus, such an analysis provides another avenue for finding a benefit estimate that exceeds the costs. However, the weighted averages based on race or gender are only $4,642 and $4,815, respectively.

Labor Market Benefit Estimates Based on Noncognitive Scores

Chetty et al. generate an estimate of the benefits of class size reduction based on effects on kindergarten test scores. They consider the relationship between test score gains from improvements in classroom quality and future earnings, and then assume that the relationship between score gains from class size reduction and future earnings is similar. This is a perfectly reasonable assumption, but perhaps could be examined in light of their findings with respect to fadeout and noncognitive scores.

Given that the cognitive gains from reduced class sizes start to diminish quite quickly, and yet students in small class sizes exhibit improvements in a variety of later life outcomes, Chetty et al. hypothesize that gains in noncognitive areas may play a role. It could be the case that noncognitive scores do not fade out and are simply not observed until they arise again through outcomes such as college attendance. Luckily, the STAR data contain information on fourth grade noncognitive scores, and Chetty et al. provide information about the relationship between these

scores and future earnings. It is therefore possible to estimate the impact of small classes on noncognitive scores and then to generate an estimate of labor market benefits based on this effect.

Fourth grade teachers in STAR classrooms rated a subsample of students with respect to their behavior. These scores were then combined into standardized scores in four areas: effort, initiative, nonparticipatory behavior, and how much the student values the class. Chetty et al. generated percentile ranks for each of the four scores and then calculated the average to create a noncognitive index. The same procedure was followed here. The effect of assignment into a small class on the noncognitive index is given in Tables 4.23, 4.24, and 4.25.

Small class sizes appear to have significant and positive impact on fourth grade noncognitive scores. Male and female students experience similar gains, but black students receive much larger benefits than do white students. Especially in light of the disparity in effect sizes by race, it may be surprising that the effect is quite homogenous by income.

To convert these scores to a lifetime earnings benefit, I note that Chetty et al. found that a one percentile increase in the noncognitive index relates to a $106 gain in age 27 earnings. Adjusting earnings up or down by 1% annually and then discounting by 3% yields a lifetime earnings benefit of $2,264 per percentile increase in the noncognitive scores. Multiplying the per-percentile gain by the overall gain of roughly 1.5% results in a labor market benefit of $3,596 coming from noncognitive scores, again much smaller than the costs. Gains by race, gender, and income are provided in column 3 of Table 4.9, and the weighted average results are presented in column 3 of Table 4.11.

Of course, some might argue that gains from noncognitive scores are at least somewhat independent of gains from cognitive scores. Fortunately, evidence in

Chetty et al. allows one to determine how much the combination of noncognitive and cognitive scores might impact future earnings. When regressing both types of scores on future wages they find that a one percentile increase in noncognitive scores relates to about an $88 increase in earnings, and a one percentile increase in math and reading scores corresponds to a gain of just over $34. Table 4.26, Table 4.27, and Table 4.28 shows the relationship between the percentile combined math and reading score and assignment to small classes.

Under this formulation, the labor market benefits from a one percentile increase in noncognitive scores is worth about $1,874, and the same increase in cognitive scores equates to about $735. Multiplying these amounts by the respective effect sizes and then summing them together still provides an estimate that is far below the estimated costs, though the weighted average based on race does come a bit closer to the costs.

Labor Market Benefit Estimates Based on Degree Completion

Dynarski, Hyman, and Whitmore Schanzenbach (2011) carry out an analysis similar to Chetty et al.'s, but with a different external data set. They instead match the STAR sample to data from the National Student Clearinghouse, which is an organization that keeps track of college student enrollment for the purpose of determining whether or not student loans should be in deferment or not. Their findings are similar to those in Chetty et al. in that black students placed into small classes are more likely to attend college then those placed into regular classes. Further, they find a significant positive impact on degree completion, though this effect does not vary substantially by race. As one might have come to expect, the impact on college attendance is larger for males than females. Results for degree completion

by gender and income are not available.

Column 5 of Table 4.9 provides estimates of labor market benefits derived from obtaining a college degree. Utilizing a procedure much in line with that used to estimate the labor market gains from college attendance, mean earnings by age for persons 18-65 were calculated for persons attaining a college degree and for those not attaining a college degree using data from the 2000 March Current Population Survey. Wages were inflated by 1% annually and the differences were again discounted by 3%. Once the estimated gains from degree completion were computed, they were multiplied by the effect size found in Dynarski, Hyman, and Whitmore Schanzenbach to estimate the labor market benefits from small class size. The estimated impact here is slightly larger than when looking only at college attendance, at $5,111. Notably, the benefit is pretty equally distributed across races, in contrast to gains from college attendance. Once again, there is little evidence that the labor market benefits from class size reduction exceed the cost.

Evaluating Labor Market Benefits Experienced Through Multiple Avenues

There are clear and positive impacts on wages from small class sizes. However, we are interested not in whether the intervention produces results but whether or not the benefits outweigh the costs. Chetty et al. (2011) calculated the costs of 2.14 years[9] of small classes to be $7,508 per student, so developing an accurate estimate is very important in determining whether or not the policy generates a positive return on investment.

Following conventional methods, a researcher might choose $19,016 as the

[9]The average number of years of small classes experienced by students initially randomized into a small class was 2.14.

estimate of labor market gains from class size reduction, as it is the largest benefit calculated in the typical way of multiplying the overall effect by the overall labor market gains. The reasoning behind such an approach is motivated by the idea that a gain in one outcome, such as college entrance exam test scores, might manifest itself as an improvement in another outcome, like college attendance, and so summing the two benefits together would result in a double-counting. Since the benefit from the exam scores is largest, it may be assumed that this benefit includes both those benefits from college as well as some marginal benefits. Aos et al. (2012), though, maintain that this strategy overestimates benefits. While they established their finding empirically through Monte Carlo simulations, no known theoretical basis for the systematic overestimation exists. Though straightforward, it may be prudent to put forth a basic argument.

Suppose that we wish to consider N different outcomes, (O_1,...,O_N) and suppose that the true labor market benefit of each is (B_1,...,B_N), respectively. Assume that the benefits are ordered smallest to largest. Further suppose that the benefits of smaller outcomes are always entirely captured by all larger outcomes. Of course, we don't know (B_1,...,B_N), and instead can only find the estimate of each, ($\hat{B_1}$,...,$\hat{B_N}$). For illustration, assume that the estimated benefits are normally distributed, though this argument can easily be extended to other symmetric distributions.

Under this assumption, $P(\hat{B_N} > B_N)$ = .5, and $P(\hat{B_N} < B_N)$ = .5. This exhausts the possibilities for underestimation, but not for overestimation. After estimating benefits for each outcome, it could be the case that $\hat{B_{N-1}}$ exceeds $\hat{B_N}$. A researcher would then choose $\hat{B_{N-1}}$ as the largest, viewing it as an estimate of B_N. Formally, we must add $P(\hat{B_{N-1}} > B_N)$ to our probability of overestimation. Under the normality assumption, this probability is

$$1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{B_N - B_{N-1}}{\sigma_{N-1}}} e^{\frac{t^2}{2}}\, dt$$

The expression is always positive, leading to overestimation bias whenever more than one alternative is estimated. It is decreasing in $B_N - B_{N-1}$ and increasing in σ_{N-1}. Thus, as the true benefits grow closer together, and as the error of the smaller benefit increases, the probability of overestimation increases.

Of course, the probability is also positive for all other alternatives, and so the total probability of overestimation is

$$.5 + P(\hat{B_1} > B_N) + ... + P(\hat{B_{N-1}} > B_N)$$

The smaller the distance between true benefits and the larger the error of each, the greater the probability of overestimation. Further, since each term is positive, the probability is also increasing in N, the number of outcomes evaluated. It is apparent, then, that such an approach might routinely overestimate benefits, which is exactly what Aos et al. found. In the working example, it seems at least possible that the estimate based on the composite college entrance exam scores is measured with considerable error, and so the probability of overestimating benefits may be substantially greater than .5.

To correct for the overestimation bias, Aos et al. propose using an average of benefit estimates weighted by sample size.

$$\sum_1^N \frac{S_n}{S_1 + ... + S_N} \hat{B}_n$$

In the formula above, S_n is the effective sample size for outcome n. Even though all of the estimates in this case use the same sample, the sample sizes differ considerably. For example, many fewer students are in the sample of those taking a college entrance exam. Effective sample sizes for each outcome are given in

Table 11. Since sample sizes are not available by race and gender in Chetty et al., this approach can only be applied to the overall effects.

The correction in Aos et al. was applied to the multiple benefit estimates found in the present study. First, the correction was carried out using estimates that were calculated with an α value of .12 to generate a more conservative estimate. Doing so yields an estimate of $4,659. To provide a more liberal estimate of benefits while still utilizing the correction methodology, the procedure was subsequently carried out using those estimates calculated with an α value of .20. Allowing for the higher α value results in an estimate of $6,061. Regardless of the α value utilized, the corrected benefit estimates are still far below the estimated costs of the intervention.

A Marginal Effects Approach to Evaluating Multiple Labor Market Benefits

While the method presented in Aos et al. may help guard against estimation errors, it is not clear that it is necessarily the best method. Both the conventional method and the Aos et al. method suppose that there is some sort of basic relationship between two outcomes. For example, both assume that labor market earnings will increase based on college attendance primarily because many of those beginning college will earn a degree. Thus, both college attendance and college completion measure similar outcomes, although one does so more directly. In the conventional method, we assume both are the most accurate estimates possible and that the larger benefit essentially contains the smaller. In the Aos et al. method, we suppose that both are measured with error and therefore simply take the average of the two in order to reduce the potential for that error.

But these approaches may be missing the greater story that these various outcomes tell. In some cases, such as this one, we have additional information about marginal effects. For example, we have information about both college attendance and degree completion. One must attend college to complete a degree, and so obtaining a degree provides a marginal benefit, distinct from that derived from simply attending college.

It could be the case that small class sizes push certain groups of students to begin college but then drop out. In fact, that's exactly what the evidence indicates. Black students assigned to small classes are 5.3% more likely to attend college but only about 1.6% more likely to attain a degree. In other words, small class sizes have the effect of inducing about 3.7% more black students to attend college but not complete a degree. Therefore, we should perhaps not average these effects together but instead sum the benefit from being a college drop out with the benefit from completing a degree.

For all students, and for each group, the increase in the probability that a student attends college but does not complete a degree can be calculated by taking the effect with respect to college attendance and then subtracting off the effect of obtaining a degree. Labor market benefits can be calculated in a way similar to before, but while comparing persons with a given level of attainment to the relevant step below. For example, the wage benefit from obtaining some college would be derived from looking at the wages of persons attaining some college compared to persons who have graduated high school, rather than comparing these persons to the entire population that has attained at least the seventh grade.

While exploring the marginal effect of college degree completion conditional on attending college is conceptually straightforward, the marginal effect of degree completion conditional on test score gains perhaps isn't so simple. Fortunately,

Murnane et al. (2000) show that about one-third of the earnings gains from increases in test scores come through the additional probability of obtaining a college degree, conditional upon graduating high school. We are still left with the puzzle, though, of determining how test scores relate to high school graduation and attending some college.

The high school graduation part is straightforward. The results in Table 4.30, Table 4.31, Table 4.32 show that reduced class sizes have no impact on high school graduation after conditioning upon eighth grade test scores, which also would have been impacted by the class size reduction. While it could be argued that high school graduation is impacted by skills that don't show up in cognitive tests, the best assumption to be made while working with the available information is to disregard any gains from graduation and instead count the test score gains fully.

The effect on college attendance conditional upon eighth grade test scores is still unknown. To be overly generous, I will assume that the labor market benefits from attending some college are completely independent of both test scores and degree completion so they were not captured by Murnane et al.

Labor market benefits can thus be calculated as

$$(2 / 3) \times M_1 + (1 / 3) \times M_2 + M_3$$

where M_1 are the labor market gains from test scores, M_2 from a college degree, after controlling for test scores, and M_3 from attending some college. Table 4.33 provides estimates for each different test score outcome both by using the conventional method and the weighted average procedure. Only those estimates using the composite ACT score, evaluated at the upper bound and using the α value of .2 or above exceed the cost of the intervention.

Discussion

All told, 41 estimates of the overall labor market benefits from class size reduction based on individual educational outcomes were generated throughout the course of this study. Two estimates were calculated using the method in Aos et al., and another 12 were derived from a new approach developed from past research on the relationship between test scores, college degree completion, and labor market earnings. Of these, only ten of them were larger than $7,508, which is the cost of the program estimated by Chetty et al. (2011). Of the ten benefit estimates that exceeded the costs , nine of them used ACT scores evaluated by way of a linear truncation procedure which produces an upper bound. Beyond estimates generated using ACT scores, only a weighted average approach based on free lunch eligibility and using college attendance as an outcome resulted in an estimate larger than the costs. The estimated benefits do not exceed the costs when measured by any other method or outcome.

Certainly, the results indicate that researchers should be hesitant to accept the idea that class size reduction delivers a positive return on investment. After all, it is only possible to arrive at such a conclusion when using the most liberal assumptions possible. While it could be the case that the most liberal assumptions result in the most accurate assessment of labor market gains, such a scenario seems a bit unlikely.

It is absolutely crucial to point out that this does not mean that small class sizes do not generate a positive return. Rather, this study only indicates that the benefits in terms of wages alone might not outweigh the cost. Though limited in scope, Schanzenbach (2006) found significant reductions in crime from small class sizes, which creates substantial savings in terms of incarceration and other costs.

Further, this analysis did not consider non-cash benefits. Aos et al. estimate non-cash benefits at 33% of annual wages. The inclusion of these benefits in a similar cost-benefit analysis of class size reduction would shift benefit estimates upward, with more of the estimates being greater than the costs. Nonetheless, many of the estimates would still be below the costs, so there would still be considerable uncertainty regarding whether or not the intervention generates a positive return.

Conclusion

This paper has contributed to the existing literature in a number of ways. First and foremost, this study has illustrated the uncertainty regarding the labor market benefits of class size reduction, and forces researchers to approach estimates of high labor market returns with some degree of skepticism. Second, this inquiry has noted that the relationship between class size and high school graduation is actually quite weak, and that previous findings resulted from problems in the implementation of the STAR experiment and subsequent analysis. Third, a theoretical argument as to why choosing the largest benefit in a cost-benefit analysis will systematically overestimate benefits was presented. Finally, this is the first paper to point out that, even if there are large-enough gains from class size reduction, they appear to accrue primarily to males. Taking this evidence as a whole, it seems as though further research should be considered before class size reduction initiatives are carried out in the future.

Tables

Table 4.1: **The Effect of Small Class Sizes on Eighth-Grade Test Scores, All Students**

	(1) Combined	(2) Math	(3) Reading
Small Class	0.0946***	0.0550	0.138***
	(3.36)	(1.75)	(4.39)
Male	-0.0835***	-0.0958***	-0.0786**
	(-3.42)	(-3.51)	(-2.87)
Black	-0.377***	-0.312***	-0.426***
	(-7.26)	(-5.40)	(-7.32)
Free Lunch	-0.374***	-0.383***	-0.370***
	(-10.04)	(-9.16)	(-8.85)
Constant	0.341***	0.350***	0.325***
	(9.18)	(8.37)	(7.78)
Observations	4439	4473	4481

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.2: **The Effect of Small Class Sizes on Eighth-Grade Test Scores, Black Students**

	(1) Black - Combined	(2) Black - Math	(3) Black - Reading
Small Class	0.0773	0.0636	0.101
	(1.69)	(1.26)	(1.96)
Male	-0.0593	-0.0861*	-0.0455
	(-1.56)	(-2.07)	(-1.06)
Free Lunch	-0.329***	-0.276**	-0.382***
	(-3.49)	(-2.66)	(-3.59)
Constant	-0.157	-0.154	-0.166
	(-1.68)	(-1.50)	(-1.57)
Observations	1719	1744	1743

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.3: **The Effect of Small Class Sizes on Eighth-Grade Test Scores, White Students**

	(1) White - Combined	(2) White - Math	(3) White - Reading
Small Class	0.102**	0.0529	0.155***
	(2.80)	(1.28)	(3.80)
Male	-0.0913**	-0.0951**	-0.0907*
	(-2.81)	(-2.60)	(-2.50)
Free Lunch	-0.377***	-0.391***	-0.368***
	(-8.93)	(-8.19)	(-7.79)
Constant	0.390***	0.408***	0.370***
	(11.06)	(10.24)	(9.38)
Observations	2720	2729	2738

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.4: **The Effect of Small Class Sizes on Eighth-Grade Test Scores, Boys**

	(1) Boys - Combined	(2) Boys - Math	(3) Boys - Read
Small Class	0.0975*	0.0834	0.122*
	(2.09)	(1.60)	(2.37)
Black	-0.396***	-0.319***	-0.445***
	(-4.65)	(-3.36)	(-4.69)
Free Lunch	-0.388***	-0.383***	-0.414***
	(-6.34)	(-5.57)	(-6.10)
Constant	0.260***	0.235***	0.276***
	(4.57)	(3.68)	(4.35)
Observations	2093	2112	2110

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.5: **The Effect of Small Class Sizes on Eighth-Grade Test Scores, Girls**

	(1) Girls - Combined	(2) Girls - Math	(3) Girls - Read
Small Class	0.0697	0.00948	0.128**
	(1.96)	(0.24)	(3.16)
Black	-0.347***	-0.290***	-0.393***
	(-5.14)	(-3.81)	(-5.12)
Free Lunch	-0.361***	-0.377***	-0.341***
	(-7.89)	(-7.31)	(-6.53)
Constant	0.339***	0.360***	0.306***
	(7.62)	(7.16)	(6.03)
Observations	2346	2361	2371

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.6: **The Effect of Small Class Sizes on Eighth-Grade Test Scores, Free Lunch Eligible**

	(1) Eligible - Combined	(2) Eligible - Math	(3) Eligible - Read
Small Class	0.101**	0.0489	0.161***
	(2.86)	(1.24)	(4.08)
Black	-0.347***	-0.271***	-0.408***
	(-5.89)	(-4.15)	(-6.22)
Male	-0.0922**	-0.0959**	-0.0994**
	(-3.08)	(-2.89)	(-2.98)
Constant	-0.108**	-0.123**	-0.107**
	(-2.95)	(-3.00)	(-2.61)
Observations	3254	3286	3291

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.7: **The Effect of Small Class Sizes on Eighth-Grade Test Scores, Free Lunch Ineligible**

	(1) Ineligible - Combined	(2) Ineligible - Math	(3) Ineligible - Read
Small Class	0.0757	0.0728	0.0779
	(1.61)	(1.36)	(1.45)
Black	-0.519***	-0.496***	-0.538***
	(-4.05)	(-3.41)	(-3.66)
Male	-0.0738	-0.107*	-0.0400
	(-1.73)	(-2.20)	(-0.82)
Constant	0.527***	0.558***	0.492***
	(15.57)	(14.50)	(12.67)
Observations	1185	1187	1190

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.8:
Labor Market Benefit Estimates Based on Test Scores, Discounted to Point of Intervention (in 2000 dollars)

	(1) 8th, All α=.118	(2) 8th, All α=.20	(3) 8th, Math α=.118	(4) ACT, All α=.118	(5) ACT, All α=.20	(6) ACT, Math α=.118
All	4,291	7,274	2,495	11,219	19,016	1,937
Black	2,878	4,879	2,368	15,700	26,610	7,598
White	4,832	8,189	2,506	2,089	3,540	6,000
Male	5,884	9,972	5,033	20,469	34,693	17,125
Female	2,311	3,917	314	5,253	8,886	-3,781
Eligible	4,582	7,766	2,218	13,488	22,861	6,000
Ineligible	3,434	5,820	3,302	4,604	7,804	49

Table 4.9:
Labor Market Benefit Estimates Based on Other Outcomes, Discounted to Point of Intervention (in 2000 dollars)

	(1) HS Grad	(2) Col Att	(3) 4th, Noncog	(4) 4th, Cog +Noncog	(5) Col Deg
All	3,634	4,686	3,595	5,288	5,111
Black	3,311	13,513	5,784	10,613	4,976
White	2,337	-508	2,652	4,507	4,919
Male	11,283	8,216	3,920	5,863	
Female	-1,391	1,002	4,127	6,538	
Eligible	4,013	11,664	5,617	7,676	
Ineligible	2,550	-2,728	1,330	4,150	

Table 4.10:
Labor Market Benefits Based on Test Scores, Weighted Average Approach, Discounted to Point of Intervention (in 2000 dollars)

	(1) 8th, All α=.118	(2) 8th, All α=.20	(3) 8th, Math α=.118	(4) ACT, All α=.118	(5) ACT, All α=.20	(6) ACT, Math α=.118
All - by Race	4,114	6,973	2,455	8,006	13,570	2,033
All - by Gender	4,199	7,117	2,808	13,289	22,525	7,267
All - by Eligibility	4,385	7,432	2,404	10,380	17,593	4,970

Table 4.11:
Labor Market Benefits Based on Other Outcomes, Weighted Average Approach, Discounted to Point of Intervention (in 2000 dollars)

	(1) HS Grad	(2) Col Att	(3) 4th, Noncog	(4) 4th, Cog +Noncog
All - by Race	2,694	4,642	3,802	6,750
All - by Gender	5,407	4,815	4,017	6,182
All - by Eligibility	3,762	9,192	4,881	5,576

Table 4.12: **The Effect of Small Class Sizes on ACT Scores by Gender**

	(1) All	(2) Male	(3) Female
Small Class	0.183***	0.251***	0.117*
	(4.45)	(3.58)	(2.23)
Male	0.0546		
	(1.51)		
Black	-0.501***	-0.502***	-0.452***
	(-6.37)	(-3.94)	(-4.28)
Free Lunch	-0.265***	-0.150	-0.332***
	(-5.32)	(-1.73)	(-5.39)
Constant	0.226***	0.190**	0.272***
	(4.98)	(2.77)	(4.92)
Observations	2421	989	1432

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.13: **The Effect of Small Class Sizes on ACT Scores by Race**

	(1) All	(2) Black	(3) White
Small Class	0.183***	0.312***	0.0553
	(4.45)	(4.87)	(1.05)
Male	0.0546	0.0441	0.0463
	(1.51)	(0.82)	(0.97)
Black	-0.501***		
	(-6.37)		
Free Lunch	-0.265***	-0.409***	-0.229***
	(-5.32)	(-4.05)	(-3.79)
Constant	0.226***	-0.194*	0.239***
	(4.98)	(-1.97)	(5.65)
Observations	2421	845	1605

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.14: **The Effect of Small Class Sizes on ACT Scores by Free Lunch Eligibility**

	(1) All	(2) Eligible	(3) Ineligible
Small Class	0.183***	0.220***	-0.0751
	(4.45)	(4.12)	(-1.12)
Male	0.0546	0.0772	-0.0133
	(1.51)	(1.64)	(-0.22)
Black	-0.501***	-0.497***	-0.522**
	(-6.37)	(-5.64)	(-2.80)
Free Lunch	-0.265***		
	(-5.32)		
Constant	0.226***	-0.0998	0.317***
	(4.98)	(-1.77)	(6.83)
Observations	2421	1443	1029

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.15: **The Effect of Small Class Sizes on ACT Subject Scores, All Students**

	(1) ACT Math	(2) ACT English	(3) ACT Reading	(4) ACT Science
Small Class	0.0814	0.0653	0.0622	0.0821
	(1.91)	(1.91)	(1.91)	(1.91)
Male	0.253***	0.203***	0.193***	0.255***
	(6.68)	(6.68)	(6.68)	(6.68)
Black	-0.518***	-0.415***	-0.396***	-0.522***
	(-6.23)	(-6.23)	(-6.23)	(-6.23)
Free Lunch	-0.153**	-0.123**	-0.117**	-0.154**
	(-2.93)	(-2.93)	(-2.93)	(-2.93)
Constant	0.0611	-0.0762*	-0.170***	-0.145**
	(1.29)	(-2.00)	(-4.68)	(-3.02)
Observations	2378	2378	2378	2378

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.16: **The Effect of Small Class Sizes on ACT Math Scores by Gender**

	(1) All	(2) Male	(3) Female
Small Class	0.0814	0.243**	-0.0299
	(1.91)	(3.25)	(-0.56)
Male	0.253***		
	(6.68)		
Black	-0.518***	-0.695***	-0.312**
	(-6.23)	(-5.04)	(-2.86)
Free Lunch	-0.153**	0.0229	-0.246***
	(-2.93)	(0.24)	(-3.90)
Constant	0.0611	0.225**	0.0804
	(1.29)	(3.05)	(1.41)
Observations	2378	971	1409

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.17: **The Effect of Small Class Sizes on ACT Math Scores by Race**

	(1) All	(2) Black	(3) White
Small Class	0.0814	0.216**	-0.0187
	(1.91)	(3.24)	(-0.34)
Male	0.253***	0.120*	0.294***
	(6.68)	(2.15)	(5.92)
Black	-0.518***		
	(-6.23)		
Free Lunch	-0.153**	-0.332**	-0.113
	(-2.93)	(-3.19)	(-1.80)
Constant	0.0611	-0.314**	0.0607
	(1.29)	(-3.09)	(1.38)
Observations	2378	834	1572

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.18: **The Effect of Small Class Sizes on ACT Math Scores by Free Lunch Eligibility**

	(1) All	(2) Eligible	(3) Ineligible
Small Class	0.0814	0.0959	0.000795
	(1.91)	(1.84)	(0.01)
Male	0.253***	0.277***	0.208***
	(6.68)	(5.95)	(3.36)
Black	-0.518***	-0.557***	-0.483*
	(-6.23)	(-6.42)	(-2.50)
Free Lunch	-0.153**		
	(-2.93)		
Constant	0.0611	-0.140*	0.154**
	(1.29)	(-2.51)	(3.21)
Observations	2378	1411	1005

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.19: **The Effect of Small Class Sizes on High School Graduation**

	(1) All	(2) 3 Yrs STAR	(3) All, FE	(4) 3 Yrs STAR, FE	(5) Preferred
3Yrs+ Small	0.425***	0.141	0.215	0.0762	
	(3.58)	(1.10)	(1.66)	(0.55)	
Male	-0.601***	-0.716***	-0.597***	-0.725***	-0.596***
	(-7.36)	(-6.18)	(-6.71)	(-5.73)	(-6.71)
Black	-0.272**	-0.280*	0.344	0.510	0.356
	(-3.24)	(-2.30)	(1.88)	(1.74)	(1.94)
Free Lunch	-2.013***	-1.772***	-1.760***	-1.610***	-1.781***
	(-13.71)	(-11.34)	(-10.30)	(-8.54)	(-10.44)
Small Class					0.0591
					(0.56)
Constant	3.064***	3.217***			
	(20.62)	(19.98)			
Observations	3580	2357	3403	2208	3403

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Logit models; see column headings for specification

Table 4.20: **The Effect of Small Class Sizes on High School Graduation by Gender**

	(1) All	(2) Male	(3) Female
Small Class	0.0115	0.0298	-0.00560
	(0.72)	(1.19)	(-0.26)
Male	-0.0954***		
	(-6.89)		
Black	0.0620*	0.0618	0.0490
	(2.08)	(1.36)	(1.16)
Free Lunch	-0.198***	-0.232***	-0.169***
	(-9.90)	(-7.66)	(-6.39)
Constant	0.904***	0.827***	0.894***
	(45.27)	(29.79)	(34.30)
Observations	3580	1735	1845

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Linear Probability Models with School-by-Entry-Year Fixed Effects

Table 4.21: **The Effect of Small Class Sizes on High School Graduation by Race**

	(1) All	(2) Black	(3) White
Small Class	0.0115	0.0119	0.00739
	(0.72)	(0.37)	(0.41)
Male	-0.0954***	-0.126***	-0.0796***
	(-6.89)	(-4.69)	(-5.02)
Black	0.0620*		
	(2.08)		
Free Lunch	-0.198***	-0.144*	-0.204***
	(-9.90)	(-2.37)	(-10.21)
Constant	0.904***	0.842***	0.952***
	(45.27)	(13.88)	(57.28)
Observations	3580	1326	2254

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Linear Probability Models with School-by-Entry-Year Fixed Effects

Table 4.22: **The Effect of Small Class Sizes on High School Graduation by Free Lunch Eligibility**

	(1) All	(2) Eligible	(3) Ineligible
Small Class	0.0115	0.0127	0.00807
	(0.72)	(0.57)	(0.50)
Male	-0.0954***	-0.120***	-0.0409**
	(-6.89)	(-6.27)	(-2.82)
Black	0.0620*	0.0705	0.0113
	(2.08)	(1.85)	(0.26)
Free Lunch	-0.198***		
	(-9.90)		
Constant	0.904***	0.681***	0.961***
	(45.27)	(29.21)	(82.18)
Observations	3580	2544	1036

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Linear Probability Models with School-by-Entry-Year Fixed Effects

Table 4.23: **The Effect of Small Class Sizes on 4th Grade Noncognitive Scores by Gender**

	(1) All	(2) Male	(3) Female
	(4.23)	(3.03)	(2.60)
Male	-0.0450***		
	(-5.41)		
Black	-0.0774***	-0.106***	-0.0610*
	(-4.45)	(-4.20)	(-2.31)
Free Lunch	-0.105***	-0.0862***	-0.115***
	(-9.17)	(-5.06)	(-7.18)
Constant	0.617***	0.564***	0.622***
	(62.27)	(42.90)	(49.34)
Observations	3034	1520	1514

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.24: **The Effect of Small Class Sizes on 4th Grade Noncognitive Scores by Race**

	(1) All	(2) Black	(3) White
Small Class	0.0401***	0.0793***	0.0315**
	(4.23)	(3.84)	(2.91)
Male	-0.0450***	-0.0376*	-0.0480***
	(-5.41)	(-2.18)	(-5.00)
Black	-0.0774***		
	(-4.45)		
Free Lunch	-0.105***	-0.0585	-0.106***
	(-9.17)	(-1.74)	(-8.51)
Constant	0.617***	0.447***	0.632***
	(62.27)	(13.60)	(64.52)
Observations	3034	702	2332

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.25: **The Effect of Small Class Sizes on 4th Grade Noncognitive Scores by Free Lunch Eligibility**

	(1) All	(2) Eligible	(3) Ineligible
Small Class	0.0401***	0.0412**	0.0415**
	(4.23)	(3.14)	(2.94)
Male	-0.0450***	-0.0392***	-0.0517***
	(-5.41)	(-3.51)	(-4.07)
Black	-0.0774***	-0.0729***	-0.102**
	(-4.45)	(-3.63)	(-2.63)
Free Lunch	-0.105***		
	(-9.17)		
Constant	0.617***	0.480***	0.660***
	(62.27)	(44.13)	(65.52)
Observations	3034	1818	1216

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.26: **The Effect of Small Class Sizes on 4th Grade Cognitive Scores by Gender**

	(1) All	(2) Male	(3) Female
Small Class	0.0159*	0.0173	0.0182
	(2.10)	(1.55)	(1.71)
Male	-0.0449***		
	(-6.43)		
Black	0.0120	0.00662	0.0264
	(0.71)	(0.27)	(1.09)
Free Lunch	-0.0713***	-0.0732***	-0.0676***
	(-7.79)	(-5.47)	(-5.23)
Constant	0.553***	0.507***	0.548***
	(64.48)	(45.73)	(49.30)
Observations	1920	953	967

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.27: **The Effect of Small Class Sizes on 4th Grade Cognitive Scores by Race**

	(1) All	(2) Black	(3) White
Small Class	0.0159*	0.0255	0.0117
	(2.10)	(1.57)	(1.36)
Male	-0.0449***	-0.0446**	-0.0466***
	(-6.43)	(-2.97)	(-5.84)
Black	0.0120		
	(0.71)		
Free Lunch	-0.0713***	-0.0644*	-0.0729***
	(-7.79)	(-2.30)	(-7.48)
Constant	0.553***	0.554***	0.556***
	(64.48)	(20.36)	(70.69)
Observations	1920	498	1422

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.28: **The Effect of Small Class Sizes on 4th Grade Cognitive Scores by Free Lunch Eligibility**

	(1) All	(2) Eligible	(3) Ineligible
Small Class	0.0159*	0.0248*	0.00587
	(2.10)	(2.34)	(0.53)
Male	-0.0449***	-0.0480***	-0.0388***
	(-6.43)	(-4.93)	(-3.79)
Black	0.0120	0.0133	-0.0186
	(0.71)	(0.65)	(-0.56)
Free Lunch	-0.0713***		
	(-7.79)		
Constant	0.553***	0.477***	0.559***
	(64.48)	(42.65)	(64.46)
Observations	1920	1099	821

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.29:
Sample Size for Each Study

	(Sample Size)
8th Grade Test Scores, .118	4,439
8th Grade Test Scores, .2	4,439
ACT Scores, .118	2,428
ACT Scores, .2	2,428
HS Graduation	3,580
Some College	10,992
College Degree	11,269
4th Grade Cog + Noncog	747

Table 4.30: **The Effect of Small Class Sizes on High School Graduation by Gender, Conditional On Test Scores**

	(1) All	(2) Male	(3) Female
Combined Scores	0.0883***	0.0904***	0.0821***
	(9.70)	(6.49)	(6.32)
Small Class	-0.00199	-0.0123	0.00933
	(-0.13)	(-0.47)	(0.46)
Male	-0.0489***		
	(-3.54)		
Black	0.0728*	0.0325	0.123**
	(2.34)	(0.65)	(2.86)
Free Lunch	-0.102***	-0.123***	-0.0899***
	(-5.06)	(-3.85)	(-3.52)
Constant	0.876***	0.852***	0.851***
	(44.80)	(30.54)	(34.01)
Observations	2849	1302	1547

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.31: **The Effect of Small Class Sizes on High School Graduation by Race, Conditional On Test Scores**

	(1) All	(2) Black	(3) White
Combined Scores	0.0883***	0.113***	0.0767***
	(9.70)	(5.79)	(7.71)
Small Class	-0.00199	-0.00320	-0.0122
	(-0.13)	(-0.10)	(-0.71)
Male	-0.0489***	-0.0822**	-0.0389*
	(-3.54)	(-2.91)	(-2.52)
Black	0.0728*		
	(2.34)		
Free Lunch	-0.102***	-0.0449	-0.112***
	(-5.06)	(-0.76)	(-5.58)
Constant	0.876***	0.865***	0.912***
	(44.80)	(15.03)	(56.71)
Observations	2849	1007	1842

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.32: **The Effect of Small Class Sizes on High School Graduation by Free Lunch Eligibility, Conditional On Test Scores**

	(1) All	(2) Eligible	(3) Ineligible
Combined Scores	0.0883***	0.107***	0.0403***
	(9.70)	(8.30)	(4.14)
Small Class	-0.00199	-0.00840	0.00822
	(-0.13)	(-0.35)	(0.54)
Male	-0.0489***	-0.0686***	-0.0218
	(-3.54)	(-3.35)	(-1.58)
Black	0.0728*	0.0820	0.0181
	(2.34)	(1.94)	(0.43)
Free Lunch	-0.102***		
	(-5.06)		
Constant	0.876***	0.757***	0.941***
	(44.80)	(30.09)	(77.41)
Observations	2849	1888	961

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimated using OLS with School-by-Entry-Year Fixed Effects

Table 4.33:
Labor Market Benefits, Alternative Approach

	(1) 8th, All α=.118	(2) 8th, All α=.20	(3) 8th, Math α=.118	(4) ACT, All α=.118	(5) ACT, All α=..20	(6) ACT, Math α=.118
All	4,261	6,249	3,064	6,293	9,694	2,691
All-Weighted	4,609	6,515	3,503	4,452	6,248	3,222

Chapter 5

Conclusion

Can value-added measures be utilized to accurately assess teacher quality so teachers can be paid according to their worth? Are certain types of teachers assigned certain types of students, clouding analyses that fail to account for these differences? Do the returns to class size reduction initiatives outweigh the substantial costs? This dissertation has provided key insights into answering all three of these questions.

The second chapter in this dissertation showed that a failure to control for the Approaches to Learning noncognitive score introduces bias into value-added models, and that this bias is statistically significant. Such a finding is crucial to an understanding of whether or not econometric techniques can mitigate the bias. Now that researchers finally know of an actual student-level characteristic that can generate bias when omitted, it will be possible to test whether or not current techniques sufficiently eliminate the bias. Perhaps student-level fixed effects will render the Approaches to Learning score harmless, or maybe the year-to-year variability of noncognitive scores is random enough to ensure that teachers observed over multiple waves of students are still evaluated in an unbiased manner. Future research must be aimed at addressing these questions

The third chapter showed that, while our existing understanding of how teachers are matched to students is technically correct, the assignment process might be more complicated than initially suspected. Assignment on unobservables carries important ramifications for empirical research. Many data sets do not contain the rich set of variables found in the ECLS-K, and so analyses of these other data sets may be unable to account for confounding relationships. If researchers are unable to control for the unobserved information in their model, it may be prudent to interpret results with caution until more research can further establish exactly how the non-random assignment of students to teachers is carried out.

The fourth chapter demonstrated that further research is required before one can conclude that the labor market benefits from class size reduction definitely exceed the costs. This may be an especially urgent matter, considering that many schools have implemented class sized reduction initiatives under the assumption that the net return is positive. It may very well be that there are additional benefits from small classes, such as noncash benefits or reductions in crime. Before researchers can go back to assuming class size reduction delivers a positive return on investment, though, further research will have to determine whether or not that is actually the case.

Bibliography

Alexander, K., D. Entwisle, and S. Dauber. (1993). “First-Grade Classroom Behavior: Its Short- and Long-Term Consequences for School Performance.” Child Development 64(3): 810-814.

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

Anderson, T, and C. Hsiao. (1981). “Estimation of dynamic models with error components.” Journal of the American Statistical Association 76(375): 598-609.

Anderson, M., and J. Robinson. (2001). “Permutation Tests for Linear Models.” Australia and New Zealand Journal of Statistics 43(1): 75-88.

Angrist, J. and V. Lavy. (1999). “Using Maimonides’ Rule to Estimate the Effect of Class Size on Scholastic Achievement.” The Quarterly Journal of Economics 114(2): 533-575.

Aos, S., S. Lee, E. Drake, A. Pennucci, M. Miller, L. Anderson, and M. Burley. (2012). “Return on Investment: Evidence Based Options to Improve Statewide Outcomes, Technical Appendix II Methods and User-Manual.” Washington State Institute for Public Policy.

Babcock, P., and J. Betts. (2009). “Reduced-class distinctions: Effort, ability, and the education production function.” Journal of Urban Economics 65:314-322.

Barnett, S. (1995). “Long-Term Effects of Early Childhood Programs on Cognitive and School Outcomes.” The Future of Children 5(3): 25-50.

Baron, R. and Kenny, D. (1986). “The moderator-mediator variable distinction in social psychological research: Conseptual, strategic, and statistical considerations.” Journal of Personality and Social Psychology 51(6): 1173-1182.

Beaton, A. (1978). “Salvaging Experiments: Interpreting Least Squares in Non-random Samples.” in Computing Science and Statistics: Tenth Annual Symposium on the Interface. D. Hogben and D. Fife, Eds. Washington: U.S. Department of Commerce, pp. 137-145.

Boyd, D., Lankford, H., Loeb, S., and J. Wyckoff. (2005). “The Draw of Home: How Teacher’s Preferences for Proximinty Disadvantage Urban Schools.” Journal of Policy Analysis and Management 24(1): 113-123.

Brown, B. and J. Maritz. (1982). “Distribution-Free Methods in Regression.” Australian Journal of Statistics 24(3): 318-331.

Burke, M., and T. Sass. (2008). “Classroom Peer Effects and Student Achievement.” Urban Institute CALDER Working Paper 18.

Chetty, R., J. Friedman, N. Hilger, E. Saez, D. Whitmore Schanzenbach, and D. Yagan. (2011). “How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star.” The Quarterly Journal of Economics 126(4): 1593-1660.

Chetty, R, J. Friedman, and J. Rockoff. (2013). “Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates.” NBER Working Paper No. 19423. National Bureau of Economic Research.

Chingos, M. (2010). “The Impact of a Universal Class-Size Reduction Policy: Evidence from Florida’s Statewide Mandate.” Harvard University Program on Education Policy and Governance Working Paper 10-03.

Claessens, A., Duncan, G., and M. Engel. (2009). “Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K.” Economics of Education Review 28(4): 415-427.

Clotfelter, C., Ladd, H., and J. Vigdor. (2004). “Who Teaches Whom? Race and the Distribution of Novice Teachers.” Economics of Education Review 24(4): 377-392.

Corcoran, S., and D. Goldhaber. (2013). “Value Added and Its Uses: Where You Stand Depends on Where You Sit.” Education Finance and Policy 8(3): 418-434.

Dee, T. (2005). “A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?” The American Economic Review 95(2): 158-165.

Dee, T., and M. West. (2011). “The Non-Cognitive Returns to Class Size.” Educational Evaluation and Policy Analysis 33(1): 23-46.

Dieterle, S. Guarino, C., Reckase, M., and J. Wooldridge. (unpublished draft) “How do Principals Group and Assign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value-added.”

Ding, W. and S. Lehrer. (2010). “Estimating Treatment Effects from Contaminated Multiperiod Education Experiments: The Dynamic Impacts of Class Size Reductions.” The Review of Economics and Statistics. 92(1): 31-42.

Downey, D., and S. Pribesh. (2004). “When Race Matters: Teachers’ Evaluations of Students’ Classroom Behavior.” Sociology of Education 77(4): 267-282.

Dwass, M. (1957). “Modified Randomization Tests for Nonparametric Hypotheses.” The Annals of Mathematical Statistics 28(1): 181-187.

Dynarski, S., Hyman, J., and D. Whitmore Schanzenbach. (2011). “Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion.” NBER Working Paper No. 17533.

Fantuzzo, J., Bulotsky-Shearer, R., McDermott, P., McWayne, C., Frye, D., and S. Perlman. (2007). “Investigation of Dimentions of Social-Emotional Classroom Behavior and School Readiness for Low-Income Urban Preschool Children.” School Psychology Review 36(1): 44-62.

Feng, L. (2010). “Hire Today, Gone Tomorrow: New Teacher Classroom Assignments and Teacher Mobility.” Education Finance and Policy 5(3): 278-316.

Figlio, D. (2007). “Boys Named Sue: Disruptive Children and Their Peers.” Education Finance and Policy 2(4): 376-394.

Finn, J. and C. Achilles. (1990). “Answers and Questions About Class Size: A Statewide Experiment.” American Education Research Journal 27(3):557-577.

Finn, J., Boyd-Zaharias, J., Fish, R., and S. Gerber. (2007). “Project STAR and Beyond: Database User’s Guide.” HEROS Incorporated.

Finn, J., Gerber, S., and J. Boyd-Zaharias. (2005). “Small Classes in the Early Grades, Academic Achievement, and Graduating from High School.” Journal of Educational Psychology 97(2):
214-223.

Finn, Jeremy D. and G. Pannozzo. (2004). “Classroom Organization and Student Behavior in Kindergarten.” The Journal of Education Research 98(2): 79-92.

Finn, J., Pannozzo, G., and C. Achilles. (2003). “The Why’s of Class Size: Student Behavior in Small Classes.” Review of Educational Research 2003(73): 321-368.

Fisher, R. (1935). “The Design of Experiments.” Edinburgh: Oliver and Boyd.

Fletcher, J. (2010). “Spillover effects of inclusion of classmates with emotional problems on test scores in early elementary school.” Journal of Policy Analysis and Management 29(1): 69-83.

Freedman, D., and Lane, D. (1983). “A Nonstochastic Interpretation of Reported Significance Levels.” Journal of Business and Economic Statistics 1(4): 292-298.

Fryer, R., and S. Levitt. (2004). “Understanding the Black-White Test Score Gap in the First Two Years of School.” The Review of Economics and Statistics 86(2): 447-464.

Gail, M, Tan, W., and D. Piantadosi. (1988). “Tests for No Treatment Effect in Randomized Clinical Trials.” Biometrika 75(1):57-64.

Gamoran, A. (1987). “The Stratification of High School Learning Opportunities.” Sociology of Education 60(3): 135-155.

Glass, G. and M. Smith. (1979). “Meta-analysis of Research on Class Size and Achievement.” Educational Evaluation and Policy Analysis 1(1): 2-16.

Goldhaber, D., and M. Hansen. (2010). “Using Performance on the Job to Inform Teacher Tenure Decisions.” American Economic Review 100(2):250-255

Goldhaber, D., and D. Chaplin. (2011). “Assessing the “Rothstein falsification test.” Does it really show teacher value-added models are biased?” CEDR Working Paper #2012-6. University of Washington.

Gordon, R., Kane ,T., and D. Staiger. (2006). “Identifying Effective Teachers Using Performance on the Job.” Path to Prosperity: Hamilton Project Ideas on Income Security, Education

Hair, E., Halle, T., Terry-Humen, E., Lavelle, B., and J. Calkins. (2006). “Children’s school readiness in the ECLS-K: Predictions to academic, health, and social outcomes in first grade.” Early Childhood Research Quarterly 21(4): 431-454.

Harris, D., and T. Sass. (2006). “Value-added models and the measurement of teacher quality.” Florida State University.

Hanushek, E. (1986). “The Economics of Schooling: Production and Efficiency in Public Schools.” Journal of Economic Literature 24(3): 1141-1177.

___. (1997). “Assessing the Effects of School Resources on Student Performance: An Update.” Educational Evaluation and Policy Analysis 19(2): 141-164.

___. (1999). “Some Findings from an Independent Investigation of the Tennessee STAR Experiement and from Other Investigations of Class Size Effects.” Educational Evaluation and Policy Analysis 21(2): 142-163.

___. (2009). “The economic value of education and cognitive skills.” In G. Sykes, Schneider, B., and D. Plank (Eds). Handbook of Education Policy Research (p. 39-56). New York: Routledge.

___. (2011). “The economic value of higher teacher quality.” Economics of Education Review 30(3): 466-479.

Hanushek, E., Kain, J., and S. Rivkin. (2004). “Why Public Schools Lose Teachers.” Journal of Human Resources 39: 326-354.

___. (2005). “Teachers, Schools and Academic Achievement.” Econometrica 77: 417-458.

Hanushek, E. and S. Rivkin. (2006). “School Quality and the Black-White Achievement Gap.” NBER Working Paper #12651.

___. (2010). “Generalizations about Using Value-Added Measures of Teacher Quality” American Economic Review 100(2): 267-271

Heckman, J. (2000). “Policies to foster human capital.” Research in Economics 54: 3-56.

Hedges, L., Laine, R., and R. Greenewald. (1994). “Does Money Matter? A Meta-analysis of Studies of the Effects of Differential School Inputs on Student Outcomes.” Educational Researcher 23(3): 5-14.

Horng, E. (2009). “Teacher Tradeoffs: Disentangling Teachers’ Preferences for Working Conditions and Student Demographics.” American Educational Research Journal 46: 690-717.

Hoxby, C. (2000). “The Effects of Class Size on Student Achievement: New Evidence from Population Variation.” The Quarterly Journal of Economics 115(4): 1239-1285.

Jackson, C. (2009). “Student Demographics, Teacher Sorting, and Teacher Quality Evidence From the End of School Desegregation.” The Journal of Labor Economics 27: 213-256.

Jacob, B., Lefgren, L., and D. Sims. (2010). “The Persistence of Teacher-Induced Learning.” Journal of Human Resources 45(4): 915-943.

Jepsen, C., and S. Rivkin. (2009). “Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size.” Journal of Human Resources 44(1): 223-250.

Kalogrides, D, Loeb, S., and T. Beteille. (2013). “Systematic sorting: Teacher characteristics and class assignments.” Sociology of Education 86(2): 103-123.

Kane, T., and D. Staiger. (2008). “Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation.” NBER Working Paper No. 14607. National Bureau of Economic Research.

Keller-McNulty, A. and J. Higgins. (1987). “Effect of Tail Weight and Outliers and Power of Type-I Error of Robust Permutation Tests for Location.” Communication in Statistics-Computation and Simulation 16(1): 17-35.

Kelly, S. (2004). “Are Teachers Tracked? On What Basis and With What Consequences.’ Social Psychology of Education 7: 55-77.

Kennedy, P. (1995). “Randomization Tests in Econometrics.” Journal of Business and Economic Statistics 13(1): 85-94.

Kempthorne, O. (1952). “The Design and Analysis of Experiments.” New York: Wiley.

Kinsler, J. (2012). “Assessing Rothstein’s critique of teacher value-added models.” Quantitative Economics 3(2): 333-362.

Kim, M.J., Nelson, C.R., and R. Startz. (1991). “Mean Revision in Stock Prices? A Reappraisal of the Empirical Evidence.” Review of Economic Studies 58(3): 515-528.

Koedel, C. (2009). “An empirical analysis of teacher spillover effects in secondary school.” Economics of Education Review 28(6):682-692.

Koedel, C., and J. Betts. (2007). “Re-examining the role of teacher quality in the educational production function” Working Paper No. 07-08, University of Missouri, Columbia.

___. (2011). “Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique.” Education Finance and Policy 6(1): 18-42.

Krei, M. (1998). “Intensifying the Barriers: The Problem of Inequitable Teacher Allocation in Low-Income Urban Schools.” Urban Education 33: 71-94.

Krueger, A. (1999). “Experimental Estimates of Education Production Functions.” Quarterly Journal of Economics 114(2): 497-532.

___. (2003). “Economic Considerations and Class Size.” The Economic Journal 113(485): F34-F63.

Krueger, A. and D. Whitmore. (2001). “The Effect of Attending a Small Class in the Early Grades on College-test Taking and Middle School Test Results: Evidence from Project Star.” The Economic Journal 111(468): 1-28.

Lankford, H., Loeb, S., and J. Wyckoff. (2002). “Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis.” Educational Evaluation and Policy Analysis 24: 37-62.

Lazear, E. (2001). “Educational Production.” Quarterly Journal of Economics 777-803.

Levin, B. and H. Robbins. (1983). “Urn Models for Regression Analysis, With Applications to Employment Discrimination Studies.” Law and Contemporary Problems 46(4):247-267.

Li-Grining, C., Votruba-Drzal, E., Maldonado-Carreno, C., and K. Haas. (2010). "Children's early approaches to learning and academic trajectories through fifth grade." Developmental Psychology 46(5): 1062-1077.

MacKinnon, D., and J. Dwyer. (1993). "Estimating mediated effects in prevention studies." Evaluation Review 17(2): 144-158.

MacKinnon, D., Lockwood, C., Hoffman, J., West, S., and V. Sheets. (2002). "A comparison of methods to test mediation and other intervening variable effects." Psychological Methods 7(1): 83-104.

Manly, B. (1991). "Randomization and Monte Carlo Methods in Biology." London: Chapman and Hall.

McClelland, M., Cameron, C., McDonald Connor, C., Farris, C., Jewkes, A., and F. Morrison. (2007). "Links between behavioral regulation and preschoolers' literacy, vocabular, and math skills." Development Psycholody 43(4): 947-959.

McClelland, M., Morrison, F., and D. Holmes. (20000. "Children at risk for early academic problems: the role of learning-related social skills." Early Childhood Research Quarterly 15(3): 307-329.

McQueen, G. (1992). "Long-Horizon Mean-Reverting Stock Prices Revisited." Journal of Financial and Quantitative Analysis 27(1): 1-18.

Molnar, A., Smith, P., and J. Zahorik. (1999). 1998-1999 Evaluation Results of the Student Achievement Guarantee in Education (SAGE) Program. Milwaukee: University of Wisconsin, School of Education.

Murnane, R., Willett, J., Duhaldeborde, Y., and J. Tyler. (2000). "How Important Are the Cognitive Skills of Teenagers in Predicting Subsequent Earnings?" Journal of Policy Analysis and Management 19(2): 547-568.

Neal, D. and W. Johnson. (1996). "The Role of Premarket Factors in Black-White Wage Differences." The Journal of Political Economy 104(5): 869-895.

Neidell, M., and J. Waldfogel. (2010). "Cognitive and Noncognitive Peer Effects in Early Education." The Review of Economics and Statistics 92(3): 562-576.

Neyman, J. (1923). "On the application of probability theory to agricultural experiments: principles." Roczniki Nauk Rolnicyzch 10: 1-51.

Nye, B., Hedges, L., and S. Konstantopoulos. (2000). “Do the Disadvantaged Benefit More from Small Classes? Evidence from the Tennessee Class Size Experiment.” American Journal of Education 109(1): 1-26.

Oakes, J., and G. Guiton. (1995). “Matchmaking: The Dynamics of High School Tracking Decisions.” American Educational Research Journal 32(1): 3-33.

Oja, H. (1987). “On Permutation Tests in Multiple Regression and Analysis of Covariance Problems.” Australian Journal of Statistics 29(1): 91-100.

Oster, E. (2013). “Unobservable Selection and Coefficient Stability: Theory and Validation.” NBER Working Paper No. 19054, National Bureau of Economic Research.

Preacher, K., and A. Hayes. (2008). “Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models.” Behavior Research Methods 40(3):879-889.

Ready, D., LoGerfo, L., Burkam, D., and V. Lee. (2005). “Explaining Girls’ Advantage in Kindergarten Literacy Learning: Do Classroom Behaviors Make a Difference?” The Elementary School Journal 106(1): 21-38.

Robinson, G. and J. Wittebols. (1986). Class Size Research: A Related Cluster Analysis for Decision Making. Arlington, Virginia: Educational Research Service.

Rockoff, J. (2004). “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” American Economic Review Papers and Proceedings, 94(2): 247-252.

Rockoff, J. (2009). “Field Experiments in Class Size from the Early Twentieth Century.” Journal of Economic Perspectives 23(4): 211-30.

Rothstein, J. (2009). “Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables.” Education Finance and Policy 4(4): 537-571.

___. (2010). “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement.” The Quarterly Journal of Economics 125(1): 175- 214.

Scafidi, B., Sjoquist, D., and T. Stinebrickner. (2008). “Race, Poverty, and Teacher Mobility.” Economics of Education Review 26(2): 145-159.

Schanzenbach, D. W. (2006). “What Have Researchers Learned From Project STAR?” Brookings Papers on Education Policy 205-228.

Scheffé, H. (1959). “The Analysis of Variance.” New York: John Wiley.

Shapson, S., Wright, E., Eason, G., and J. Fitzgerald. (1980). “An Experimental Study of the Effects of Class Size.” American Educational Research Journal 17(2): 141-152.

Slavin, R. (1989). “Class Size and Student Achievement: Small Effects of Small Classes.” Educational Psychologist 24(1): 99-110.

Smith, A. (1776). “An Inquiry into the Nature and Causes of the Wealth of Nations.” Edwin Cannan, (Ed.). London: Methuen and Company, Ltd.

Sobel, M. (1982). Asymptotic confidence intervals for indirect effects in structural equations models. In Sociological Methodology 1982, S. Leinhart (Ed.) San Francisco: Jossey-Bass, pp. 290-312.

Sojourner, A. (2013). “Identification of Peer Effects with Missing Peer Data: Evidence from Project STAR.” The Economic Journal 123(569): 574-605.

Staiger, D., and J. Rockoff. (2010). “Searching for effective teachers with imperfect information.” Journal of Economic Perspectives 24(3): 97-118.

Stasz, C. and B. Stecher. (2002). Before and After Class-Size Reduction: A Tale of Two Teachers. In M. Wang and J. Finn (Eds.), Taking Small Classes One Step Further. pp. 19-50. Greenwich, CT: Information Age.

Strategic Data Project. (2012). “The Novice Teacher Placement Pattern: Do Low-Performing Students Get Placed with Novice Teachers?” The SDP Human Capital Diagnostic Strategic Performance Indicator.

Swartz, J., and D. Walker. (1984). “The relationship between teacher ratings of kindergarten classroom skills and second-grade achievement scores: An analysis of gender differences.” Journal of School Psychology 22(2): 209-217.

Taylor, B., Pearson, D., Clark, K., and S. Walpole. (2000). “Effective Schools and Accomplished Teachers: Lessons about Primary-Grade Reading Instruction in Low-Income Schools.” The Elementary School Journal 101(2): 121-165.

Tourangeau, K., Nord, C., Thanh, L., Sorongon, A., Najarian, M., and E. Hausken. (2009). “Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K): Combined User’s Manual for the ECLS-K Eighth-Grade and K8 Full Sample Data Files and Electronic Codebooks.” National Center for Education Statistics.

Wang, M. and J. Stull. (2000). “School Characteristics and Classroom Practice: Smaller Versus Larger Classes.” In M. Wang and J. Finn (Eds.), How Small Classes Help Teachers do Their Best: pp. 175-198. Philadelphia: Temple University Center for Research in Human Development and Education.

Welch, W. (1990). “Construction of Permutation Tests.” Journal of the American Statistical Association 85(411): 693-698.

Wiswall, M. (2013). “The Dynamics of Teacher Quality.” Journal of Public Economics 100: 61-78.

Word, E., Johnston, J., Bain, H., Fulton, B., Boyd-Zaharias, J., Lintz, M., Achilles, C., Folger, J., and C. Breda. (1990). Student/Teacher Achievement Ratio (STAR), Tennessee’s K-3 Class Size Study: Final Summary Report, 1985-1990. Nashville: Tennessee State Department of Education.

Zahorik, J. (1999). “Reducing Class Size Leads to Individualized Instruction.” Educational Leadership 57(1): 50-53.