WORKING
PAPER

Achievement of Students in Multigrade Classrooms

Evidence from the Los Angeles Unified School District

LOUIS T. MARIANO, SHEILA NATARAJ KIRBY

WR-685-IES

June 2009

Prepared for the Institute of Education Sciences

EDUCATION

ACKNOWLEDGMENTS

The authors are grateful to Harold Himmelfarb of the Institute of Education Sciences for his support of the larger study under which this work was performed. Cynthia Lim and Glenn Daley of the Los Angeles Unified School District (LAUSD) provided access to student achievement data. Eva Pongmanopap of LAUSD was helpful in building the student achievement files and in clarifying issues related to the data. We are also grateful to Richard Buddin for providing the data and for his support of the study. RAND Education provided additional support for carrying out the analyses.

This paper is part of a larger research project "Teacher Licensure Tests and Student Achievement" that is sponsored by the Institute of Education Sciences in the United States Department of Education under grant number R305M040186.

1. INTRODUCTION

The dominant classroom organization in U.S. schools is the monograde classroom, containing students of a similar age range, assigned to a single grade level, but with a range of abilities. This is also sometimes referred to as the "single age class," because it contains students of a specified age range congruent with grade level. Advocates of alternative grouping practices—in which children of different ages are grouped together—suggest that multi-age groupings are "more aligned with children's natural groupings and learning tendencies" (Ong, Allison, and Haladyna, 2000: 206) and point to research that shows non-cognitive and cognitive benefits to children in these multi-age classrooms (Katz, Evangelou, and Hartman, 1990; Pavan, 1992; Veenman, 1995, 1996, 1997; Allison and Ong, 1996; Kelley and Fitterer, 1998; Ong, Allison, and Haladyna, 2000).

Interest in multi-age education peaked in the early 1990s, and a growing number of school districts put such programs in place, attracted by their emphasis on developmentally appropriate practices (Pardini, 2005). In 1990, the Kentucky Education Reform Act "embraced the multi-age philosophy and mandated that every school in the state provide an ungraded primary program. Children were to be given the opportunity to progress from kindergarten through 3rd grade at their own pace" (p. 3). However, by 1998, Kentucky relaxed its mandate in the face of growing dissatisfaction of teachers and administrators who found the ungraded programs difficult to implement and of parents who did not quite understand the workings of multi-age classroom. With the onset of No Child Left Behind, the interest in multi-age education declined still further, because of the very specific grade-level standards and testing requirements.

However, to some extent, students continue to be grouped together for instructional purposes if perhaps largely for administrative rather than philosophical reasons, in what are called multigrade or combination classrooms. This may be due, for example, to having fewer teachers than grade levels or uneven pupil enrollment (Veenman, 1995; Mason and Burns, 1997). These multigrade classrooms are very different in nature from multi-age classrooms where students are deliberately organized across grade levels by choice and for pedagogical or philosophical reasons (Veenman, 1995; Bacharach, Hasslen, and Anderson, 1995; Mulcahy, 1999). In multigrade classrooms, grade levels remain distinct and students remain linked with their grade level as opposed to students in multi-age classrooms who tend to remain ungraded and to be integrated into one learning community (Mulcahy, 1999: 5).

There is mixed evidence regarding the effects of multigrade classrooms on student achievement and much of it is dated. The constrained fiscal environments facing many of the nation's districts may lend fresh impetus to this practice; as such, it is important to understand how students placed in these classrooms perform relative to their peers. This paper presents new evidence from the Los Angeles Unified School District (LAUSD) schools on the achievement of students in multigrade classrooms and uses a quasi-experimental method to define a plausible comparison group of peers in a monograde classroom. It seeks to examine the following counterfactual: how would these students have performed had they been in a monograde classroom?

This paper is organized into several sections. The next section briefly reviews the literature on the effects of multigrade/multi-age classes on students. Section 3 presents an overview of the data and methods used in the analysis. Section 4 presents findings from our analysis. A final section presents our conclusions.

2. REVIEW OF THE LITERATURE

Almost all of the reviews and studies done in the late 1990s point to the importance of distinguishing between multi-age and multigrade classrooms and suggest that mixed results often found in the literature on effects of such groupings on student achievement are largely attributable to inconsistent definitions of different types of multi-age and multigrade groupings. In spite of these issues, there appear to be some consistent findings. Veenman (1995) reviewed 56 studies and concluded that:

(a) "students in the multigrade classes do not appear to learn more or less than their counterparts in the single-grade classes. No consistent differences were found with respect to reading, mathematics, language, or composite scores…The median effect size across the 34 studies for which effect sizes could be computed was essentially zero" (p. 367)

(b) "students in the multi-age classes did not learn more or less than students in the single-age classes. The median effect size for the 8 studies for which effect sizes could be computed was again essentially zero" (p. 367)

(c) However, with respect to noncognitive outcomes, students in both the multi-age and multigrade classes tended to score as well as or higher on attitudes towards school, personal adjustment, and self-concept than students in the single-age classes, although the differences in both cases were rather small.

Mason and Burns (1996), however, pointed out that the finding of no difference with respect to student achievement actually translates into a small negative effect because "There is considerable

evidence that principals, in an effort to reduce the burden on multigrade teachers, place more able, more independent, and more cooperative students in multigrade classes" (p. 311) and also some (admittedly "sketchy") evidence that better teachers are assigned to these classes (p. 312). They suggested that the need to cover two different curricula, greater classroom demands, less attention to individual students, and greater teacher stress lead to lower quality of instruction in multigrade classes.

This suggestion was refuted by Veenman in his 1996 rejoinder to Mason and Burns. He noted that students in the multigrade classes did not spend their learning time differently than students in single-grade classes (a statement that runs counter to his earlier statement that not much was known about the instructional practices employed in these classrooms). He also concluded that the effects of multigrade grouping on student achievement were small and positive for lower grades (K-2) essentially zero for intermediate grades (3-4), and small and negative for higher grades (5-6). Veenman's 1997 article also reiterated that there was really not much evidence of purposeful assignment, as claimed by Mason and Burns.

Russell, Rowe, and Hill (1998) used data from the Victorian Quality Schools Project, a large, comprehensive, three-year, longitudinal study of school and teacher effectiveness in Victoria, Australia, to examine the effect of multigrade classrooms on student achievement. Their analyses showed some significant negative effects on achievement associated with multigrade classes and some non-significant effects, although the results differed between years (1993 and 1994) and between subject areas: literacy and numeracy. The qualitative phase of the Project in 1995 focused on multigrade classes, conducting extensive interviews with principals and teachers. Their case studies provide "strong support to the conclusions drawn from other research that the multigrade class structure is a more difficult, complex and challenging one than that provided by the single-grade structure…Repeated emphasis is placed on the importance of having the strongest, best, most experienced teachers in multigrade classes, the amount and quality of organization and planning needed, the exaggerated range of students (ability, achievement, maturity, behavior) in the classroom, and the importance of having a proportion of independent learners who will continue to work on their own when the teacher is occupied with another group. A further indication is the strong majority preference to place children in a single-grade class in the year following a multigrade placement" (no page numbers).

Burns and Mason (1998; 2002) examined the class distributional properties of 200 elementary school classes in two districts, 56 of which were combination or multigrade classes. They found evidence that principals intentionally manipulated class composition for instructional purposes,

assigning higher ability and more independent students to multigrade classes and that these assignment procedures affected the class distributional properties and achievement variation within and between classes. Their study underscored "the importance of considering the link between class formation and class composition, on the one hand, and class composition and student achievement, on the other hand" (p. 229).

In contrast, Wilkinson (2003) conducted in-depth case studies of single-grade and multigrade classrooms serving 2nd through 6th grade students, focusing on the distributional properties of classrooms and in particular, the reading abilities of students. The study did not find a difference in the ranges of abilities in the two types of classrooms or evidence to suggest heavier demands on teachers in combination classes. To avoid selection bias, Ong, Allison, and Haladyna (2000) used Title I status for a group of 3rd graders from three school districts to examine whether the type of student taught seemed affected by classroom organization. They concluded that non-Title I students in multi-age classrooms benefited from such a grouping but this was not true for Title I and other traditionally lower-achieving students.

Lloyd (1999) summarized the literature with respect to multi-age classes and high ability students. He concluded that most studies appeared to be positive. However, he did warn against overselling the idea, calling multi-age classes "one potentially effective option, especially where they are formed as a permanent option and taught by teachers committed to and able to support each other in this form of organization" (p. 204). In such a case, he argued, "it is likely that the sorts of activities which are carried out in the class are different from those in a single-grade class or multi-grade class where the teacher approaches the students as members of a particular grade and with expectations of similarity rather than difference."

3. DATA AND METHODS

Data

This study is based on panel data from the Los Angeles Unified School District (LAUSD) for students in grades 2 through 5 for six consecutive school years from 2002 to 2007. We refer to a school year by the calendar year in which the school year ends; e.g., we refer to the 2001–2002 school year as 2002. The students are enrolled in self-contained classrooms taught by a single teacher, where the student and teacher data are linked by an identifying variable, although the identifiers for each are scrambled.

LAUSD is a large, diverse urban school district. Annual enrollment is about 730,000 students in over 800 schools. The available data consist of over 1.3 million yearly individual records from

approximately 560,000 unique LAUSD students. For our multigrade classroom analyses, we included only classrooms with between 15 and 35 students; we also exclude students in handicapped/special education schools and early primary schools. From these records, we identified all classrooms that served students in more than one grade. Among these, we classified as classroom as "multigrade" if:

- All students in the classroom were in one of two adjacent grades, and
- Students in each grade comprised at least 25% of the classroom.

Approximately 3.8 percent of classrooms met this definition. Eighteen percent of schools have at least one multigrade classroom, but fewer than 2 percent have more than a quarter of their students in multigrade classrooms.

Once the multigrade classrooms were identified, we further restricted the analytic dataset to include only those students with prior year assessment scores. Prior year scores are critical for identifying the monograde classroom students who are most similar to their multigrade counterparts; they are also the strongest available predictor of future performance. Students are first tested in 2nd grade so there are no prior scores available for 2nd graders. Thus, the data from 2002 was used as baseline information only, and we concentrated on the outcomes of 3rd, 4th, and 5th grade students. The final analytic dataset consisted of approximately 730,000 individual records from just over 380,000 unique students.

Table 1 displays the cumulative number of student records in the analytic dataset and percentage of students in each grade assigned to multigrade classrooms.[1] Approximately two-and-a-half to five percent of students were assigned to multigrade classrooms. Among multigrade classroom students, the grade 3 – 4 configuration was the least common, with 3rd graders more likely to be in a grade 2 – 3 configuration and 5th graders more likely to be in a grade 4 – 5 configuration.

[1] If students without prior year scores were included, the percentages of multigrade students in Table 3.1 would be slightly higher, between one tenth and one half of one percentage point.

Table 1. Percentage of Students in Multigrade Classrooms by Grade Configuration, 2003 through 2007

Grade	Total Records	% Monograde	% Multigrade in Grade Range		
			2 – 3	3 – 4	4 – 5
3	245,199	97.62	1.87	0.50	–
4	244,624	94.89	–	0.59	4.52
5	239,091	95.06	–	–	4.94

As displayed in Table 2, the percentage of students in multigrade classrooms in LAUSD declined over the period examined.[2] In 2003, over 7.5 percent of 4th and 5th graders and over 3 percent of 3rd graders were in multigrade classrooms; by 2007, the multigrade assignments had dropped to under 4 percent and under 2 percent respectively.

Table 2. Percentage of Students in Multigrade Classrooms by Year, 2003 through 2007

Grade	% of Students in Multigrade Classrooms				
	2003	2004	2005	2006	2007
3	3.07	2.32	2.24	2.20	1.90
4	7.59	5.47	4.79	3.45	3.85
5	7.72	5.11	4.72	3.40	3.67

Table 3 displays the distribution of several key student demographic variables among all students, students with prior year scores retained into the analytic sample, and students in multigrade classrooms. Although a significant portion of the full sample was not qualified for the analytic sample, the distributions among these variables are generally comparable between the full and analytic versions of the sample. Among students in the analytic sample, 75 percent were Hispanic, 10 percent were black, and 6 percent were Asian/Pacific Islander. Just under half of the students were classified as Limited English Proficient (LEP). The share of Hispanic, Asian/Pacific Islander, and black students classified as LEP was 59, 40, and 1 percent, respectively (not shown). About 75 percent of students were eligible for the free/reduced lunch

[2] If students without prior scores were included, the percentages in Table 3.2 would very slightly higher, typically about one tenth of one percentage point.

program. Among students who had parental education data, 28 percent of students came from households where the highest level of parental education was a high school degree, and 19 percent of students had a parent with a college or graduate school degree.

Table 3. Distribution of Key Student Demographic Variables

Variable	% of Full Sample	% of Analytic Sample	% of Multigrade Students
Female	49.56	49.57	49.29
Race/Ethnicity			
Black	10.24	10.04	10.93
Hispanic	71.26	75.29	72.12
Asian/Pacific Islander	5.89	6.03	5.72
White/non-Hispanic/Other	12.61	8.64	11.23
Eligible for free/reduced lunch	75.45	74.63	73.53
Limited English Proficient (LEP) student	48.48	46.70	46.57
Special Education student	7.35	7.77	8.61
Gifted student	9.05	12.06	14.81
Highest Parental Education			
Graduate school	4.24	4.43	5.93
College degree	9.01	9.32	10.58
Some college	12.04	12.66	12.67
High school degree	19.51	20.89	20.6
Less than high school diploma	24.53	26.57	25.37
Data not available	30.67	26.13	24.85

As seen in Table 3, both special education and gifted students were present in multigrade classrooms. Special education students in multigrade classrooms were 1.6 to 2.0 times more

likely to be found in a configuration that featured their grade and the grade below; conversely, gifted students in multigrade classrooms were 1.6 to 2.3 times more likely to be found in a configuration that featured their grade and the grade above. It was not uncommon for special education and gifted students to be found in the same multigrade classroom; 29 percent of multigrade classrooms contained at least one student from each classification. Only 17 percent of multigrade classrooms had neither special education nor gifted students, and a very small number, less than 1 percent, consisted only of gifted students.

In addition to the student data, the analytic dataset also contained teacher background variables. The elementary LAUSD teacher workforce is diverse and experienced. In our sample, the average teaching tenure was 10 years, but the distribution was skewed with about 20 percent of teachers in their first three years of teaching. Three-fourths of the teachers were women. About 32 percent of teachers were Hispanic, 12 percent were black, and 12 percent were Asian. Approximately 20 percent of the teachers had a master's degree, and 1 percent had a doctorate.

We examine student achievement measured on the California Standards Test (CST) in English language arts (ELA) and mathematics. The CST is a criterion-referenced test measuring student proficiency relative to the California curriculum standards and is administered to students starting in grade 2. Scores for each grade and subject are reported on a common scale that ranges from 150 to 600, with proficiency indicated by a score of at least 350. The average scores for LASUD students in our analytic sample were 325.9 in ELA and 348.5 in mathematics.

Within each subject outcome, we calculated the standard deviations for each grade separately, using all available LAUSD observations pooled across the years of analysis. Although the scores from each subject are placed on the same reporting scale between 150 and 600, the empirical standard deviations of the outcomes tended to be higher for mathematics, where they ranged from 70 to 80 points, than for ELA, which fell roughly between 50 and 55 points. These standard deviations were used to calculate the effect sizes reported later in the paper.

Methods

The literature suggests that effects might differ across grades and subjects (Veenman, 1996; Russell et al., 1998). Thus, for example, 4th graders in a grade 3–4 configuration may not have the same experiences as 4th graders in a grade 4–5 configuration, leading to different outcomes with respect to student achievement. As a result, we examined the effect of being in a multigrade classroom separately for the ELA and mathematics CST outcomes within each combination of grade and multigrade classroom configuration found in the data. For each subject, we examined the effects on five different multigrade outcomes: 3rd graders in a grade 2 – 3 configuration; 3rd

and 4th graders in a grade 3 – 4 configuration; 4th and 5th graders in a grade 4 – 5 configuration. Monograde students in the same grade as the multigrade students served as a comparison group.[3]

We use a doubly robust regression (Bang and Robins, 2005) approach to estimate the treatment effect of being placed in a multigrade classroom. This is a two-stage process. In the first stage, we use propensity scoring techniques to weight the monograde classroom students (i.e., the control group) so that the distribution of their characteristics matches the distribution of characteristics of the multigrade classroom subjects (i.e., the treatment group). We then use these weights in the second stage, where we estimate the treatment effect via a weighted multiple linear regression model. We discuss these stages in greater detail below.

As this is an observational study, the characteristics of monograde and multigrade students may not be in balance. For example, the distribution of prior assessment scores may be different across these two groups. The analytic goal of the analysis is to predict how multigrade classroom students would have fared if they were placed instead in a monograde classroom. We use the available data from the monograde students in a regression framework to estimate this counterfactual. However, if the monograde students do not look like their multigrade counterparts, then the regression must rely upon extrapolation in predicting the multigrade classroom effect. To avoid the problem of imbalance, we use a propensity scoring technique to weight the monograde students so that the distribution of their characteristics matches the distribution of characteristics of the multigrade students as closely as possible.

In the propensity weighting approach, we first estimated $P(\text{treated} \mid \mathbf{x}_i)$, the probability that a student with features $\mathbf{x}_i$ would be in a multigrade classroom (i.e., the propensity score). We then assigned a set of propensity weights: each multigrade student received a weight of $w_i = 1$, while the monograde students received a weight equal to their estimated log odds of being in the treatment group, $w_i = P(\text{treated} \mid \mathbf{x}_i)/(1 - P(\text{treated} \mid \mathbf{x}_i))$. Thus, the monograde students who most resemble the multigrade students received higher weights than those monograde students who were more dissimilar. McCaffrey, et al. (2004) showed that by assigning each untreated subject *i* a weight as specified above, the distribution of the characteristics of the untreated group would be the same as the treated group. Note that the propensity scores are estimated independently of

[3] In some cases, classrooms were identified with more than one grade, but did not fit our definition of a multigrade classroom. In those cases where more than one grade was present but one of the grades accounted for over 75 percent of the classroom, students from that modal grade in the classroom were included in the set of monograde students serving as the control group, but the remaining students were excluded.

the outcomes, so that the outcomes do not influence how the propensity weights adjust for **x**. Thus, the bias that may be introduced by fitting a variety of regression models to the outcome is avoided.

We estimated the propensity scores using a flexible, non-parametric generalized boosting method, including the method described in McCaffrey, Ridgeway, and Morral (2004) and implemented in the Toolkit for Weighting and Analysis of Nonequivalent Groups (twang) package for the R statistical environment. The student characteristics accounted for in the vector of covariates **x** included the set of variables in Table 3.3, along with the prior year CST scores in both ELA and mathematics. We did not constrain comparisons between multigrade and monograde classroom students to be within the same school. It is possible that the mechanism for classroom assignment was not random and that meaningful differences between multigrade and monograde students existed among students in the same school that were not captured in our available characteristic variables. Instead, we allowed comparisons across schools within each of the eight local districts in LAUSD. To complement the student characteristics and better ensure a balance among classroom experiences, we also included a set of teacher characteristic variables in the set of covariates **x**, including teacher length of experience, highest degree attained, race/ethnicity, and gender.

Once the propensity scores were estimated, we examined the balance among the multigrade and monograde students to ensure proper model specification. For each covariate in **x**, we calculated a Kolmogorov-Smirnov (K-S) statistic to examine the equality of the distribution between the monograde and multigrade classroom students, with the goal of choosing propensity scores that minimized the set of K-S statistics for all covariates as well as possible. Where necessary, we adjusted the tuning parameters of the generalized boosting method and refit the propensity scores to achieve better distributional matches between the two groups.

We calculated a separate set of propensity scores for each of the outcomes examined. A monograde classroom student had a different probability of treatment depending on the set of treated students being examined. For example, a 4th grade student would have a different probability of being assigned to a multigrade classroom with a grade 3 – 4 configuration versus a 4 – 5 configuration, and different weights need to be applied to that student in achieving distributional balance for each case. Thus, by fitting separate propensity scores, we were able to optimize distributional balance for each outcome examined.

With a final set of propensity scores providing balance between the multigrade and monograde student groups, we fit a weighted multiple regression model to estimate the multigrade treatment effect, with weights w_i as identified above. The regression model included all the available

student, teacher, and district variables (**x**) that were included in estimating the propensity scores, as well as a treatment indicator variable (Z_i) denoting whether the student was from a multigrade (Z_i=1) or monograde (Z_i=0) classroom.

The fitted weighted linear regression has the form:

$$Y_i = \mu_o + \beta_w Z_i + \Psi_w X_i + \varepsilon_i. \qquad (1)$$

where Y_i is the assessment score for student *i*. The estimate of the coefficient of the treatment indicator, $\hat{\beta}_w$, is then an estimate of the treatment effect of being in a multigrade classroom. The inclusion of the available covariates in both the propensity scoring and regression models makes the treatment effect estimate "doubly robust" in the sense that if either the propensity score model is correct or the regression model is correct then the treatment effect estimator will be consistent (Bang and Robins, 2005). Note that the form of the propensity weights, a weight of 1 for the treated students and the log odds of treatment for the control students is consistent with the counterfactual "How would the treated students have performed, had they not been treated?" Thus we are estimating the treatment effect on the treated.

As discussed, the doubly robust regression approach accounts for all available covariates in both the propensity scoring and multiple regression stages of the analysis. Thus, the regression includes all the student covariates listed in Table 3, as well as teacher characteristics (gender, race/ethnicity, educational attainment, and years of teaching experience). In addition, because we restricted comparisons to within local districts, the regression also contains dummy variables to control for the local district.[4]

However, this design does not account for unobserved potential confounding variables, particularly those that are weakly correlated or independent of the available covariates. For example, parental motivation may be a potentially omitted factor—more motivated parents of younger children may want them to be placed in a combined classroom with higher grade students and may also work with them to improve their performance and help them keep up with the class. As such, we need to exercise care in interpreting the coefficient of the treatment indicator, as it may not necessarily indicate a causal relationship between multigrade assignment and the outcome. As we discuss the "treatment effects" estimated from the model below, the reader should keep this caveat in mind.

---

[4] As noted above, we did not constrain comparisons between multigrade and monograde classroom students to be within the same school. After including all the available student and teacher variables, school effects would account for only an additional 1 to 2 percentage points of the unexplained variance in the outcome variables.

4. RESULTS

Table 4 displays the doubly robust regression estimates of the impact of being assigned to a multigrade classroom among 3rd, 4th, and 5th graders in LAUSD over the 2003 through 2007 school years. Here we scaled the treatment estimates by the empirical standard deviation of the outcome among LAUSD students, so that they could be interpreted as effect sizes. Also displayed in the table are 95 percent confidence intervals for the effect size and the number of multigrade students included in the analyses.

Table 4. Impact of Being Assigned to a Multigrade Classroom in 3rd, 4th, or 5th Grade, 2003 through 2007

Assessment & Grade	Multigrade Configuration	Number of Multigrade Students	Treatment Estimate[1]	95% Confidence Interval
ELA				
3rd	2 – 3	4,439	-0.030*	(-0.049, -0.011)
3rd	3 – 4	1,171	-0.014	(-0.051, 0.022)
4th	3 – 4	1,367	-0.016	(-0.057, 0.024)
4th	4 – 5	10,827	-0.037*	(-0.048, -0.026)
5th	4 – 5	11,572	-0.011	(-0.023, 0.000)
Mathematics				
3rd	2 – 3	4,425	-0.042*	(-0.061, -0.024)
3rd	3 – 4	1,165	-0.062*	(-0.100, -0.025)
4th	3 – 4	1,366	-0.033	(-0.079, 0.012)
4th	4 – 5	10,817	-0.091*	(-0.103, -0.079)
5th	4 – 5	11,564	-0.055*	(-0.066, -0.043)

Note: [1] The estimates are reported as effect sizes. We standardized using the empirical standard deviation for the assessments among students in LAUSD.
*Significant at .05 level.

Across both subjects and all grades and multigrade configurations, treatment estimates were consistently very small and negative. For ELA, the treatment estimates ranged from -0.04 to -0.01 standard deviations. Only two of these effects were statistically significant, 3rd graders in grade 2 – 3 configurations and 4th graders in grade 4 – 5 configurations, with estimated effect sizes of

-0.03 and -0.04 respectively. In both these instances, large sample sizes had the power to identify very small significant effects. However, the effect sizes, while statistically significant, are not of substantive significance. Overall, 3rd, 4th, and 5th grade students in multigrade classrooms performed on average between 1/25th and 1/100th of a standard deviation lower on the CST ELA assessment than expected had they been in a monograde classroom.

The multigrade treatment estimates for mathematics outcomes ranged from -0.09 to -0.03, and the results were statistically significant for 3rd graders in both the grade 2 – 3 and 3 – 4 configurations, and for both 4th and 5th graders in the grade 4 – 5 configuration. With the exception of the effect size of 0.09 for 4th graders in a grade 4 – 5 configuration, the statistically significant effect sizes were again too small to be of practical significance. Overall, 3rd, 4th, and 5th grade students in multigrade classrooms performed on average between 1/10th and 1/30th of a standard deviation lower on the CST mathematics assessment than expected had they been in a monograde classroom. To put these findings into context, even for the 4th graders in a grade 4 – 5 configuration, the magnitude estimated effect sizes are considerably smaller than the 0.20 effect size typically considered to be "small" according to the guidelines set forth by Cohen (1988). A more recent paper by Hill et al. (2007) argues against using such rules of thumb because they "ignore the context that produces a particular estimate of program impact and that better guidance for interpreting impact estimates can be obtained from empirical benchmarks." The authors point to "the importance of interpreting the magnitude of an intervention effect *in context:* of the intervention being studied, of the outcomes being measured, and of the samples or subgroups being examined. Indeed, it is often useful to use multiple benchmarks when assessing the observed impacts of an intervention" (p. 11). Following their third benchmark—observed effects from similar interventions—they summarize the results of 76 prior meta-analyses of educational interventions and find a mean effect size of 0.22-0.23 standard deviations. Even using these guidelines, the effect sizes we report are quite small and only one—4th graders in the grade 4-5 configuration in mathematics—approaches any practical significance.

For additional context, we may also consider the impact of multigrade classroom assignment in terms of the performance levels mapped to the reported assessment scale. As noted above, the "proficient" performance level on the CST begins at a scale score of 350. The level immediately below, "basic", is consistently defined across grades to be in the 50-point range from 300 to 349, for both subjects. As standard deviations of the ELA outcomes are very near 50, the estimated effect sizes displayed in Table 4 for ELA translate directly into the width of the "basic" performance level, with average expected point differences ranging from 0.5 to just under 2 points lower on the CST reporting scale for multigrade students (i.e., between 1/100th and 1/25th

of the basic CST range). Given the higher standard deviations among the mathematics outcomes, the treatment effect sizes displayed in Table 4 translate into average expected point differences for multigrade students ranging from just over 2 to just over 6 points lower on the CST scale (i.e., between approximately 1/25th and 1/12th of the basic CST range).

5. CONCLUSIONS

This paper used a quasi-experimental approach to examine the effect of being assigned to multi-grade classrooms on students' achievements. We used propensity scoring techniques and doubly robust regression to estimate the counterfactual: How would these students have performed, had they been placed in monograde classrooms. The approach allows us to define plausible comparison groups that could be used as controls in the model and provides a robust estimate of the effect of the treatment on all treated students.

In keeping with previous literature, we found consistently small and negative effects on student achievement, regardless of grade or subject, even controlling for teacher characteristics (as suggested by Mason and Burns, 1996). Overall, 3rd, 4th, and 5th grade students in multigrade classrooms performed on average between 1/25th and 1/100th of a standard deviation lower on the CST ELA assessment than expected had they been in a monograde classroom. Similarly, 3rd, 4th, and 5th grade students in multigrade classrooms performed on average between 1/10th and 1/30th of a standard deviation lower on the CST mathematics assessment than expected had they been in a monograde classroom. While several of these effects were statistically significant, with one exception (4th graders in grade 4-5 configuration in mathematics), none was large enough to be substantively significant.

Veenman (1995) offered several factors to explain why student learning in alternative groupings—such as multi-age and multigrade classrooms—did not differ from student learning in single-grade or single-age classes. We were able to address the issue of bias mentioned both by Mason and Burns (1996) and by Veenman (1995) that arises when more capable students are selected into multigrade classes, producing non-equivalent samples for comparison. However, we lack information on the three other factors he discussed:

- First, it is unlikely that grouping alone can affect student learning because the latter depends more on the quality of instructional practices than organizational structure and on the ability of the teacher to adapt instructional strategies to foster more collaborative learning.
- Second, teachers may be "ill-prepared to teach two or more grades at the same time and may not have teaching materials that are adequately suited to multigrade teaching at their disposal" (p. 371).

- Third, as noted by both Veenman and Russell et al. (1998), multigrade teaching is demanding—teachers may have little energy to pursue potentially more effective grouping strategies in their teaching and may end up using many of the same practices as in single-grade classes.

It is possible that our small, negative findings may be due to the lack of teacher preparation and training to teach in these alternative classrooms and their inability to take advantage of these grouping strategies rather than some inherent characteristic of the multigrade classroom structure. If so, these findings underscore the importance of realizing that the benefits of multigrade classroom, if they exist, are unlikely to accrue unless teachers are trained and adequately supported when placed in such classrooms. This is especially important today when districts may consider adopting more multigrade classrooms as one way of dealing with their constrained budgets. Our findings also suggest that researchers need to more fully understand what happens within these classrooms in terms of instructional practice in order to come to more definitive conclusions regarding the pros and cons of multigrade classrooms.

REFERENCES

Allison, J., & Ong, W. (1996). Advocating and implementing multiage grouping in the primary years. Dimensions of Early Childhood, 24(2), 18-24.

Bacharach N., Hasslen R. C., Anderson J.,(1995). Learning together: A manual for multiage grouping. Corwin Press Inc, Thousand Oaks, California.

Bang, H. and Robins, J.M. "Doubly Robust Estimation and Missing Data and Causal Inference Models." *Biometrics*. 2005;61:962-972.

Burns, R. B. and D. A. Mason (1998). "Class Formation and Composition in Elementary Schools." American Educational Research Journal **35**(4): 739-772.

Burns, R. B. and D. A. Mason (2002). "Class Composition and Student Achievement in Elementary Schools." American Educational Research Journal **39**(1): 207-33.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Katz, L. W., Evangelou, D., & Hartman, J. A. (1990). The case for mixed-age grouping in early education. Washington, DC: National Association for the Education of Young Children.

Kelley, M. F., & Fitterer, H. (April 1998). Multiage and traditional classroom programs: A comparison of standardized test score data, group cooperation and problem-solving performance. Paper presented at the Annual Conference of the Association for Childhood Education International, Tampa, FL.

Mason, D. A. and R. B. Burns (1996). "'Simply No Worse and Simply No Better' May Simply be Wrong: A Critique of Veenman's Conclusion about Multigrade Classes." Review of Educational Research 66(3): 307-322.

McCaffrey, D, G. Ridgeway, and A. Morral. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4), 403-425.

Mulcahy, D. (1999). From multigrade to multiage: A journey of discovery, affirmation and transformation. Paper presented at The sustainability of small rural schools across the North Atlantic Rim: 50th anniversary symposium (August 11-15, 1999): St Anthony, Newfoundland, Canada.

Ong, W., J. Allison, and T. M. Haladyna. (2000). "Student Achievement of 3rd-Graders in Comparable Single-Age and Multiage Classrooms." Journal of Research in Childhood Education **14**(2): 205-15.

Pardini, P. (2005). "The Slowdown of the Multiage Classroom: What Was Once a Popular Approach Has Fallen Victim to NCLB Demands for Grade-Level Testing." School Administrator **62**(3): 22.

Pavan, B. N. (1992). The benefits of nongraded schools. Educational Leadership, 50, 22-25.

Russell, V. J., K. J. Rowe, and P. W. Hill. (1998). "Effects of Multigrade Classes on Student Progress in Literacy and Numeracy: Quantitative Evidence and Perceptions of Teachers and School Leaders." Paper presented at the 1998 Annual Conference of the Australian Association for Research in Education, Adelaide, 29 November - 3 December, 1998.

Veenman, S. (1995). Cognitive and noncognitive effects of multigrade and multi-age classes: A best-evidence synthesis. Review of Educational Research, 65, 319-381

Veenman, S. (1996). Effects of multigrade and multi-age classes reconsidered. Review of Educational Research, 66(3), 323-340.

Veenman, S. (1997). "Combination Classrooms Revisited." Educational Research and Evaluation **3**(3): 262-76.

Wilkinson, I. A. G. and R. J. Hamilton (2003). "Learning To Read in Composite (Multigrade) Classes in New Zealand: Teachers Make the Difference." Teaching and Teacher Education **19**(2): 221-35.