A Multilevel Analysis of Classroom Achievement History and Curriculum Condition Effects on Students’ Science Concept Learning

Vasuki Rethinam, Curtis Pyke, & Sharon Lynch
Graduate School of Education and Human Development
The George Washington University

This work was conducted by SCALE-uP: A collaboration between George Washington University and Montgomery County Public Schools (MD); Sharon Lynch, Joel Kuipers, Curtis Pyke, and Michael Szesze, principal investigators. Funding for SCALE-uP was provided by the National Science Foundation, the U.S. Department of Education, and the National Institutes of Health (REC-0228447). Any opinions, findings, conclusions, or recommendations are those of the authors and do not necessarily reflect the position or policy of endorsement of the funding agencies.

Correspondence to: Vasuki Rethinam. E-mail: vasuki@gwu.edu.
Please contact the author for the latest version of the paper.
Abstract

This paper builds on our prior work in applying HLM analysis techniques to multilevel curriculum evaluation data by modeling different classroom context variables as predictors of students’ learning in a series of middle school science curriculum materials quasi-experiments. For this trial, the sample consisted of grade 6 students (n = 1,841; classroom n = 78) selected from intact classrooms in a large metropolitan school district. The null model indicated that 17% of variance exists between classrooms and therefore, HLM is justified. Findings indicate that the full model explained 17% of the variance in students’ outcome scores and that the curriculum intervention was significantly related to students’ learning outcomes. Classroom achievement history variables have different effects on Hispanic students and students eligible for ESOL services. Results show that classroom contexts have an influence on student learning.

Introduction

This work is part of a NSF-funded research study, Scaling-up Curriculum for Achievement, Learning and Equity Project (SCALE-uP). A goal of SCALE-uP is to evaluate and understand the effectiveness of reform-based curriculum materials as a prerequisite to their scale-up in a large and diverse school district. This paper is based on the premise that classroom context variables can interact with individual level demographic variables to predict differences in learning in curriculum materials evaluation research. In other words, classroom level variables may account for a significant portion of variance in outcome measures, and therefore ought to be considered in curriculum studies. This paper uses hierarchical linear modeling (HLM) to account for systematic variability in the data caused by classroom and individual interactions manifest in the data by the non-random assignment of middle school students to schools and to classrooms within schools.

SCALE-Up has relied primarily on ANCOVA techniques at the student level (with posttest scores as the outcome variable and pretest scores as a covariate) to test hypotheses that differences in outcomes are the result of using the innovative curriculum materials (Lynch, Kuipers, Pyke, & Szesze, 2005; Lynch, Taymans, Watson, Ochsendorf, & Pyke, 2007). However, as noted above, there is a nesting problem with these analyses, because SCALE-uP initially used random assignment at the school level; i.e., individual students were not randomly assigned to a curriculum condition. However, in some prior work, we applied HLM analyses to a first trial for the Exploring Motion and Forces (M&F [Harvard Smithsonian Center for Astrophysics, 2001]) curriculum evaluation because of HLM’s potential to address concerns about liberal significance tests derived from ANCOVA analyses with nested data. The purpose of this paper is to refine the classroom context variables using the third year trial for M&F and to test the multilevel (classroom- and student-level) effects on students’ learning.

Background

SCALE-uP Quasi-Experimental Studies

SCALE-uP’s goals are grounded in a commitment to quasi-experimental curriculum unit effectiveness studies conducted in diverse school settings. The evaluation research questions studied by the project are: “does the new curriculum work?” i.e., is the treatment curriculum more effective than comparison curricula and “does the new curriculum close gaps?” i.e., does the new
unit allow all subgroups of students to achieve high standards. Student (single) level ANCOVA techniques on posttest scores with pretests as covariates were initially employed to test for significant curriculum differences (e.g., Lynch et al., 2005) although each student was not randomly selected for curriculum condition; rather they were nested within classrooms and schools. SCALE-uP acknowledges that the choice of student level ANCOVA under these nested conditions is controversial (for a recent discussion, see Lederman and Flick, 2005). For more background information on SCALE-uP studies and analyses, see Rethinam, Pyke, & Lynch (in press) and the Implementation Reports on the SCALE-uP website, www.gwu.edu/~scale-up.

**SCALE-uP HLM Results**

The nesting of the project’s data is a concern because classroom context effects may be seen to affect all students grouped in a classroom and confound overall observations of a curriculum’s effects at the student level. Therefore, we used HLM to conduct a parallel analysis to the ANCOVA’s of M&F’s third implementation (Pyke, Lynch, Kuipers, Szesze, & Watson. 2004). HLM analysis confirmed that: significant classroom variability existed; HLM showed higher effect size estimates for curriculum effects; and, HLM revealed classroom context effects indicated by the ethnic composition (percentage of African Americans) of a classroom (positive relationship). Overall, curriculum effect size increased from small (ES = .10) using ANCOVA to moderate (ES = .60) in the HLM analysis. This increase is attributed in part to the controlling of variation due to individual student factors, while also accounting for variation in outcomes attributed to classroom context, an advantage of HLM (Rethinam, Pyke, & Lynch, in press). One of the limitations of this initial analysis of M&F in the first trial was that the classroom level variables used did not seem the best indicators of classroom context. Individual level demographic variables (SES, special education status, ESOL services, African American, Hispanic, and Asian) were aggregated to class level to create classroom context variables.

The current paper addresses these limitations of our initial HLM analysis, but uses a different data set for a replication of the M&F evaluation, the third trial. An improved model of classroom context and individual effects will be employed in this paper. A new classroom context variable – classroom achievement history – was created for this study. The goal was to better understand and account for the classroom context effects on students’ science concept learning.

**Perspective/Theoretical Framework**

In this paper, we continue to develop a theoretical and practical rationale for multilevel modeling of individual and classroom context effects. We argue that it is not enough to look only at individual student demographic characteristics to characterize the success or failure of an intervention. It is also important to understand the contexts in which interventions are implemented and learning occurs (McDonald, Keesler, Kauffman, & Schneider, 2006). Classroom contexts can influence learning because each student is affected by multiple factors, widely understood to account for variation in learning. There is empirical evidence from sociological and psychological studies of school- and classroom-level influences that classroom conditions can impede or promote learning (McDonald et al., 2006). Assignment of students to classrooms within schools and between schools is an important contextual factor that has not always been examined in curriculum effectiveness studies, although prior studies indicated that instruction in high- and low-track classrooms produces different student outcomes (Callahan, 2005; Van Houtte, 2004). However, in most prior studies, tracking/ability grouping variables were used as school level variables and data were analyzed using single level regression analyses (Carbonaro, 2005; Marks 2006). Other studies that used multilevel analysis included the tracking/ability grouping variable at the school level (McCoach, O’Connell, & Levitt, 2006; Van Houtte, 2004) but have ignored the classroom level analysis of these variables.

Classroom factors are a concern because in practice students are not randomly assigned to classrooms in the middle schools they learn in and naturally occurring placement patterns cause classrooms with distinct norms and values to emerge. Descriptive analyses of demographic classroom composition within the schools of this study illustrate the fact that classrooms evolve
to have unique compositions. In our data, some classrooms have a high concentration of students who generally achieve well in school, i.e., they have a strong record of achievement relative to other classrooms within a given school. These classes tend to have higher proportions of White and Asian American students and fewer students eligible for The Free and Reduced Price Meals System (FARMS, a proxy for SES); English for Speakers of Other Languages (ESOL); and, special education (SPED) services. There are other classrooms with students with low achievement histories relative to other classrooms. These classrooms on average have lower student standardized test scores and grades in science, and tend to have greater proportions of students eligible for FARMS, SPED and ESOL services, and higher proportions of African American and Hispanic students (see Table 1).

The school system in this study does not track middle school students’ performance for class assignment; however, by the time students are in middle school, realities of scheduling in a large, highly diverse suburban school district come into play. Some students, likely those with high levels of social capital and activist parents, are placed in advanced math classes and honors classes. Once placed in a special class, such students may find themselves traveling together throughout the day and also placed together in a science with class that demonstrates a higher achievement history than the school on average presents. In other cases, students with a variety of learning challenges, from identified disabilities to those learning to speak English, or those whose families have not been able to provide all the support that allows students to thrive and achieve optimally in public schools, find themselves placed in classrooms that may be characterized as having, on average, lower achievement history than the school on average presents. Students who are learning English may sometimes, but certainly not always, may also be eligible for FARMS, and in this school district the highest numbers of English language learners are Hispanic. African American students are often disproportionately identified as needing SPED services (REF). Consequently, low achievement history classrooms are formed for students perceived as having special needs and eligible for special services. Such classrooms may be characterized as having smaller class sizes, and may be assigned a teaching aide or additional teacher to work with the science teacher. Sandwiched in between classrooms characterized as having higher and lower achievement histories, of course, are those classrooms with middling achievement histories.

Some might characterize this as de facto tracking. Others may see it as an outcome of scheduling practices and the realities of public school systems from which the public demands a variety of services, from accelerated learning classes for students identified as gifted or talented, to special classes for those who need (and who legally require) extra help. For the purposes of this study, we make no judgments about the organization of middle schools into classes with differing achievement histories. We simply observe that for a given child in a given school, that placement in a science class with high, medium, or low achievement history may perhaps make a difference in student learning outcomes, and this study explores that possibility.

Research Questions

Given the equity focus of SCALE-uP and the ability to conduct multi-level analysis, here we examine the effectiveness and equity implications of a treatment unit, M&F, from a different and realistic standpoint, posing the following research questions after controlling for individual level variables:

1. Does the curriculum intervention, M&F, improve student posttest scores in comparison with students who receive comparison curriculum materials?
2. Which classroom level variables are statistically significant predictors of student posttest scores?
3. Do curriculum condition and achievement history interact with student demographic variables (ESOL, SPED, Asian American, African American, and Hispanic)?

SCALE-uP Participants
The participants in the third year of implementation of M&F trials for this dataset included 1,841 grade 6 students (classroom \( n = 78 \)) from ten large, public middle schools, each selected from one large, Central Atlantic metropolitan school district. The school district has a student population that is diverse in ethnicity, gender, socioeconomic status (SES), students’ status as English language learners, or as having identified disabilities. The ten schools were matched on socio-economic status, the percentage of students in attendance eligible for the Free and Reduced Priced Meals System (FARMS) and math and reading achievement data from 5th grade nationally norm-referenced tests. This resulted in five matched pairs of schools. The schools were divided into low-, medium low-, medium-, medium high- and high-SES schools. However for this study only four matched pairs of schools were used. Outcome data were not collected from a school, and consequently its matched pair was also removed before data were analyzed.

**SCALE-up Procedures**

In each of the SCALE-up quasi-experiments, schools were selected randomly from sets of matched pairs to implement the treatment curriculum units, with the other half of a pair of schools serving as the comparison condition. This current analysis uses data from the grade 6 students indicated above who used the unit M&F (Harvard Smithsonian Center for Astrophysics, 2001). The comparison group experienced the regular curriculum offered by the school district, which focused on the same benchmarks. Students in both the treatment and comparison conditions were given a posttest using an instrument that specifically measured their understanding of the benchmark concepts targeted by M&F.

**Variables for HLM Analysis**

**Independent variables.** Table 2 lists the individual level variables as well as the classroom level variables that were employed in the HLM analyses for this study. All individual level variables, except prior science GPA, were dichotomous. Table 2 presents the coding of all the variables used in the analyses. FARMS and ESOL demographic variables were converted from a three level variable (Never, Prior, and Now eligible for a service) to a dichotomous variable (Never = 0 and Prior-Now = 1. Three dummy variables were created for ethnicity/race (African American was coded 1, others = 0; Asian American = 1, others = 0; Hispanic = 1, others = 0) and White was used as a reference group (‘White’ was used as a reference group because this group was the largest ethnic/race group in the school district, although there was no ethnic majority in this district). Students’ prior science GPA comprised students’ average science grades from the two terms prior to the study and in the same academic year. Grades are awarded on a five point scale with A = 5 and F = 0. For example, a student with a B in the first quarter and an A in the second quarter would have a prior science GPA 4.5.

Classroom level variables consisted of curriculum condition coded as a dichotomous variable (M&F and comparison), and class achievement history (low or high).

The following procedure was used to create the classroom achievement history variable. Classrooms within schools were rank-ordered based on the means of three student-level criteria, (a) prior science GPA in the same academic year, (b) standardized math test scores, and (c) standardized reading test scores obtained from 5th grade tests. The ranked scores for the three criteria were summed to provide two highest achievement classrooms and two lowest achievement classrooms for each school. Classrooms with scores in the middle were coded as a third category (reference category, or medium achievement history). This procedure was conducted for each school. Thus, the class mean achievement history is a naturally occurring classroom composition variable within schools. This variable, mean achievement history has three levels (high, medium, and low) within each school.

The final ranges are based on the ranges across schools. Across all schools, classes with high mean achievement history had scores that ranged from 2.60-3.86 (prior science GPA), 418.71-464.57 (mean math standardized scores), and 411.14-448.93 (mean reading standardized scores). Across all schools, classes with medium mean achievement history had scores that
ranged from 2.25-3.55 (prior science GPA), 386.31-461.27 (mean math score), and 385.05-441.32 (mean reading score). Across all schools, classes with low mean achievement history had scores that ranged from 1.77-3.48 (prior science GPA), 375.75-443.84 (mean math score), and 380.00-429.40 (mean reading score). Dummy variables were created for classroom achievement history. Low Achievement History was coded as 1, others = 0, high achievement history was coded as 1, others = 0, and medium achievement history was used as a reference group (see Figure 1').

Figure 1. Individual and Classroom Level Factors that Influences Student posttest Scores

It is no surprise that students in high achievement history classrooms on average outperform classrooms with low achievement history. However, the concern arises when individual characteristics and school infrastructure, rather than merit inform student placement (Hallinan, 1992). A student’s academic history, which is in itself a result of previous background effects, has been shown to predict placement (Hallinan, 1992). Dauber, Alexander, and Entwisle (1996) found that SES has significant effects on initial middle school course placement, even when previous achievement is controlled. Therefore, in the current study, SES and students’ prior science GPA was used as a control variable at the student level.

Table 2
Individual and Classroom Level Predictor Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Level Variables (Level 1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td>1841</td>
<td>Ethnicity as self-directed by students</td>
</tr>
<tr>
<td>African American</td>
<td>313</td>
<td></td>
</tr>
<tr>
<td>Asian American</td>
<td>277</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>417</td>
<td></td>
</tr>
<tr>
<td>SES (FARMS)</td>
<td></td>
<td>Eligible for free and reduced priced meals; as a proxy for SES</td>
</tr>
<tr>
<td>High = 0</td>
<td>1157</td>
<td></td>
</tr>
<tr>
<td>Low = 1</td>
<td>684</td>
<td></td>
</tr>
<tr>
<td>Special Education</td>
<td></td>
<td>Eligible for Special Education Services</td>
</tr>
<tr>
<td>No = 0</td>
<td>1512</td>
<td></td>
</tr>
<tr>
<td>Yes = 1</td>
<td>302</td>
<td></td>
</tr>
<tr>
<td>English Language Literacy</td>
<td></td>
<td>Qualify for services as a non-native speaker of English in the U.S.</td>
</tr>
<tr>
<td>No = 0</td>
<td>1438</td>
<td></td>
</tr>
<tr>
<td>Yes = 1</td>
<td>403</td>
<td></td>
</tr>
</tbody>
</table>
Prior Science GPA | 1788 | Two quarter science grades prior to the study
--- | --- | ---

### Classroom Level Variables (Level 2)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curriculum Unit</td>
<td>78</td>
<td>M&amp;F and comparison curriculum condition</td>
</tr>
<tr>
<td>Comparison</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Achievement History</td>
<td>78</td>
<td>Classroom prior achievement history based on</td>
</tr>
<tr>
<td>Low</td>
<td>16</td>
<td>prior standardized math and reading scores and</td>
</tr>
<tr>
<td>High</td>
<td>16</td>
<td>prior science GPA.</td>
</tr>
</tbody>
</table>

**Dependent variable.** Student understanding of the target idea was determined by a score on the Motion and Forces Assessment (MFA), a curriculum-independent posttest administered by teachers at the end of instruction. An analysis and development procedure done in collaboration with the AAAS Project 2061 using their assessment analysis standards (Stern & Ahlgren, 2002) was used to create the Motion and Forces Assessment (MFA). The MFA used in the third trial for M&F consisted of 10 items (6 constructed responses and 4 selected responses) that provided the students with 4 different physical phenomena to respond to questions about motion and force. Raters coded student responses using a rating guide that categorized student responses according to their alignment with a scientifically appropriate understanding of the target benchmark (average inter-rater reliability = 0.82).

A standard setting process (Plake, & Hambleton, 2001; Pyke & Hansen, 2005) established cut scores that distinguish among four levels of understanding of the target ideas: 0-20 = no understanding; 21-50 = context-limited understanding; 51-70 = some fluency with ideas; 71-100 = flexible understanding. Cronbach’s Alpha for the ten items of the MFA used in calculating the weighted score indicated that the assessment showed acceptable internal consistency for a multifaceted construct (α = 0.52). Student scores on MFA were transformed into weighted scale scores ranging from 0 to 100.

**HLM as the Multilevel Analysis Technique**

Hierarchical linear modeling (HLM) was used to analyze multilevel effects on outcomes. HLM has several advantages over single level analysis techniques, such as ANOVA and regression. The assumption of independence of cases is not necessary in a multilevel analysis because the probable dependence of students in the same classroom is explored explicitly with nested data (Aitkin & Longford, 1986). Multilevel techniques allow analyses to be conducted simultaneously at multiple levels of data. In the current analyses, both student- and classroom-level sources of variability in outcomes are simultaneously accounted for by specifying a two-level hierarchical model to obtain the best estimates of treatment effects. In this study, student demographic variables (for e.g., ESOL, SPED) were modeled at the individual level and fixed at the classroom level, treating them as covariates to the treatment effect. As fixed, they were predicted only by an intercept (for comprehensive procedures on HLM analysis, see Raudenbush & Bryk, 2002). Classroom level covariates (e.g., low achievement history, high achievement history etc.) were entered to help reduce the unexplained variance attributed to the classroom in the outcomes. This technique helps examine the direct effects of treatment as well as the covariates, modeling both at the individual and classroom level simultaneously on the outcome variable.

The program HLM, version 6.0.3 (Raudenbusch, Bryk, Cheong, & Congdon, 2004) was employed for data analyses. SPSS, version 12.0.1 was used to enter M&F data to obtain composite variables, and to create the SSM (sufficient statistics matrix) file for HLM analysis. All the continuous variables modeled in these analyses were converted to z-scores (M = 0, SD = 1). The categorical variables were dummy coded as 0 and 1 (see Table 2).

HLM analyses had three stages for this study. The first stage was a fully unconditional model. According to Raudenbush and Bryk (2002), the fully unconditional model is the simplest
multilevel model and contains no predictor variables from any level. The fully unconditional model is used to estimate how much variation is attributed to the classroom level and the individual level. The proportion of variance in the dependent variable that is explained at the classroom level is revealed. According to Lee (2000), if the proportion of variance that exists at a higher level of aggregation (classroom level over student level, in this case) is more than 10% of the total variance in the outcome, then it is necessary to consider a multilevel analysis. The analyses in this paper used these guidelines. The second stage of HLM employed is referred to here as the final fitted model, the “intercept as outcomes model,” as indicated by Raudenbush and Bryk (2002).

Effect sizes (ES) are presented as standard deviation (SD) units calculated from HLM coefficients (ES = coefficient/between-classroom SD). For example, a curriculum condition coefficient (.23) is divided by the classroom level SD in the unconditional model (.41) for an ES of .56 (See Lee & Loeb, 2000; Rethinam, Pyke, & Lynch, in press). Because the present multilevel analyses focus on between-classroom differences, ES results are calculated with the classroom level SD units, which is the SD of the classroom means. The third and final stage of the HLM employed a cross-level interaction, i.e., an interaction between classroom level (curriculum condition and achievement history) and individual level (ESOL, SPED, dummy coded ethnicity) variables. Both intercept and individual level demographic variables were modeled as outcomes in a classroom level “slopes as outcomes” HLM model. Power analyses were conducted to detect the power for significant classroom level effects using the “Optimal Design” software published by Spybrooke, Raudenbush, Liu, and Congdon (2006). The power to detect a significant treatment effect was .77. However, the power was very low (power = .42) to detect a significant achievement history effects. The number of clusters that were required to obtain a power of .61 with an effect size of .22 was 100 when other variables were included at the individual and class levels.

Results

Results from HLM Analysis of Third Year of M&F Implementation Data

Table 2 presents sample sizes and average student and classroom level demographics. The variables are presented in their original metric in the descriptive analyses.

The result of the fully unconditional HLM analysis is presented for M&F in Table 3. The unconditional model indicates that the proportion of variance in the individual posttest scores attributed to systematic classroom effects is 17%. The proportion of variance in the posttest score that exists at the individual level accounted for 83% of the variance. It can be concluded that most of the variance exists between students (83%) (within classrooms) and a smaller proportion of variance (17%) exists between classrooms. Because the variance accounted at the higher level is greater than 10%, a multilevel analysis was justified. Also included in Table 3 is the classroom SD, which is the SD of the classroom means, used for ES calculations and reliability. It shows that the student posttest scores have acceptable lambda reliability. Reliability is concerned with the reliability of group mean scores (class mean scores) on the outcome (student posttest scores).

Table 3

<table>
<thead>
<tr>
<th>Fully Unconditional HLM Model for M&amp;F Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students: posttest score</td>
</tr>
<tr>
<td>Within-classroom variance (sigma squared)</td>
</tr>
<tr>
<td>Between-classroom variance (tau)</td>
</tr>
<tr>
<td>Between-classroom SD</td>
</tr>
<tr>
<td>Reliability (lambda)</td>
</tr>
<tr>
<td>Intraclass correlation*</td>
</tr>
</tbody>
</table>

*The intraclass correlation is the proportion of total variance in the outcome that lies systematically between classrooms. It is computed as follows: ICC = tau/(tau + sigma squared).
Table 4 presents only statistically significant main effects at the individual and class level and their interactions. Other measures of classroom context and individual level variables were not statistically significant and are not presented in Table 4. Student SES and prior science GPA were used as control variables. At the classroom level, only one variable had a significant influence on students’ posttest score, and at the individual level, five variables were significant. There were two significant interactions between achievement history and individual level demographic variables.

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>P-Value</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classroom Level Variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-.0003</td>
<td>.04</td>
<td>0.10</td>
<td>---</td>
</tr>
<tr>
<td>Curriculum Condition (CC)</td>
<td>.23</td>
<td>.09</td>
<td>.01</td>
<td>.56</td>
</tr>
<tr>
<td>Low Achievement History (LAH)</td>
<td>-.09</td>
<td>.11</td>
<td>.42?</td>
<td>-.22</td>
</tr>
<tr>
<td>High Achievement History (HAH)</td>
<td>.10</td>
<td>.11</td>
<td>.40?</td>
<td>.24</td>
</tr>
<tr>
<td>Individual Level Variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPED</td>
<td>-.13</td>
<td>.06</td>
<td>.03</td>
<td>-.32</td>
</tr>
<tr>
<td>ESOL</td>
<td>-.22</td>
<td>.07</td>
<td>.002</td>
<td>-.54</td>
</tr>
<tr>
<td>FARMS (SES)</td>
<td>-.16</td>
<td>.06</td>
<td>.01</td>
<td>-.39</td>
</tr>
<tr>
<td>African American</td>
<td>-.15</td>
<td>.07</td>
<td>.04</td>
<td>-.37</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-.07</td>
<td>.07</td>
<td>.30</td>
<td>---</td>
</tr>
<tr>
<td>Prior Science GPA</td>
<td>.24</td>
<td>.03</td>
<td>.001</td>
<td>.59</td>
</tr>
<tr>
<td>Interaction effect a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESOL x LAH</td>
<td>-.42</td>
<td>.15</td>
<td>.01</td>
<td>-.68</td>
</tr>
<tr>
<td>Hispanic x HAH</td>
<td>.25</td>
<td>.14</td>
<td>.06</td>
<td>.40</td>
</tr>
</tbody>
</table>

*a For example, ESOL learning gap is fixed as it did not vary systematically between classrooms. Effect size is computed by dividing each interaction coefficient by the SD of the slope, .62 (for ESOL), calculated by multiplying standard error for that coefficient shown in the main effects HLM model by the square root of the sample size (78). For example, .07 x \sqrt{78} = .62 (for calculations, see Lee, Loeb, & Lubeck, 1998).

**Classroom level predictors.** One classroom context variable, curriculum condition had a significant positive influence on student posttest scores. Controlling for other variables in the model, classroom treatment condition was positively significantly related to the posttest score (ES = .56 SD, p<.01). In other words, students in treatment classrooms who used the M&F curriculum unit were estimated to score 0.56 SD points higher than students in comparison classrooms. A major goal of SCALE-uP was to ascertain the effectiveness of M&F, and this finding shows the unit to effective at the classroom level as well as at the individual student level (cf., Rethinam, Pyke, & Lynch, in press).

**Interaction effects.** There was a significant interaction effect between ESOL status and low achievement history classrooms. The effect of being in a classroom with a low achievement is four times poorer for ESOL students than for non-ESOL students in the same type of classroom. A significant interaction was also found between Hispanic ethnicity and high achievement history classrooms. The effect of being in a classroom with a high achievement history is two and a half times greater for Hispanic students than it is for White students in similar classrooms. There was no significant interaction found between curriculum condition and student
individual level predictors. Controlling for other variables in the HLM model, FARMS status had a significant negative effect on student posttest scores (ES = -.39 SD, p < .01). Low SES (or Prior-Now FARMS) students were estimated to score .39 SD points lower than high SES (or Never FARMS) students. There was a significant association between eligibility for ESOL services and posttest scores (ES = -.54 SD, p < .002). In other words, students eligible for ESOL services were estimated to score .54 SD points lower than students who were native English speakers. Student SPED status was significantly related to the outcome (ES = -.32 SD, p < .03). Students eligible for special education services scored 32 SD points lower than students who were never eligible for special education services. There was a significant relationship between African American students and posttest scores (ES = -.37 SD, p < .04). African American students scored 37 points lower than their White peers. Prior science GPA was significantly related to posttest score (ES = .59 SD, p < .001). Students who had a higher prior science GPA scored .59 SD points higher than the students who come into classrooms with lower prior science GPA scores. These findings were regardless of curriculum conditions. The magnitude of effects for the significant variables was from medium to large.

In addition to the effects, it is also of interest to examine the percent of total available variance explained by the full model. Table 5 presents this information for the fully unconditional model and the final model. It is clear from Table 5 that the overall model explained 17% of the variance in student post test scores.

Table 5. Estimation of Variance Components – Posttest Scores

<table>
<thead>
<tr>
<th>Estimation of Variance Components</th>
<th>Model</th>
<th>Within classrooms (n = 1841)</th>
<th>Between Classrooms (n = 78)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully unconditional model</td>
<td>0.84</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>Final model</td>
<td>0.71</td>
<td>0.12</td>
<td></td>
</tr>
</tbody>
</table>

Variance at each level
- Within Classrooms: 0.84 / (0.17 + 0.84) = .83
- Between Classrooms: 0.17 / (0.17 + 0.84) = .17

Proportion of variance explained by final model
- Within Classrooms: 0.84 - 0.71 / 0.84 = .15
- Between Classrooms: 0.17 - 0.12 / 0.17 = .29

Proportion of total available variance explained by final model
(0.17 x 0.29) + (0.83 x 0.15) = .17 = 17%

Discussion

Curriculum Effectiveness

The main focus of this paper was to examine the classroom level variables to predict learning outcomes in curriculum evaluation research. The curriculum condition was found to be significantly related to students’ science achievement. Students in the M&F treatment condition scored higher than students in the comparison condition in posttest scores. There were no significant interaction effects found between curriculum condition and other demographic variables such as ESOL, special education services, African American, Asian American, and Hispanic ethnic groups. This further confirms that the new treatment unit is effective across groups.

Classroom Achievement History Effects

Significant main effects for classroom achievement history variables were not found in the current study. This, however, may be due to the low power. Interactions between classroom achievement history variables and student level demographic variables were intriguing. Our
findings indicated that there were two significant interaction effects. The results indicated that there was a significant interaction between ESOL and low achievement history classrooms. On an average, being eligible for ESOL services in low achievement history classrooms disadvantages students more than being ESOL eligible in medium achievement history classrooms compared to students who are native English speakers in a low achievement history classroom. Our findings were similar to the findings of other researchers (Callahan, 2005; Wang & Goldschmidt, 1999). Wang & Goldschmidt (1999) found that while students placed in low track courses scored below all other students, the effect of low-track placement was strongest among English learners. They further indicate that the effect of English learner status on academic performance was minimized among those students placed in higher track classes. Harklau (1994) in her analysis of the educational experience of English learners, found that exposure to oral and written style varied by track. English learners who managed to enter into high-track classes became more proficient in complex discourse skills, while those who remained in low-track classes learned to repeat and respond at a superficial level. These findings suggest that English learners will rise to meet the challenges when given access to higher quality curricula (Callahan, 2005). The current study does not test the interaction between curriculum condition, ESOL eligibility, and low achievement history due to power issues: future research should examine this interaction effects.

Another notable interaction effect was found between Hispanic ethnicity and high achievement history classrooms. Hispanic students in high achievement history classrooms outperform Hispanic students in medium achievement history classrooms and compared with White students in high achievement history classrooms. Prior studies on tracking indicate that children from low SES and ethnic minorities are disproportionately placed in low-ability groups early in their education careers (Joseph, 1998; Oakes, 1985).

Individual Level Effects

Our findings indicated that after controlling for SES and prior science GPA, individual variables such eligibility for English language services, special education services, and African American ethnicity were significantly negatively related to posttest scores. In other words, these students scored lower on posttest compared to their counterparts regardless of curriculum condition.

In summary, the M&F curricular intervention did improve students’ posttest scores. The full model explained 17% of the variance in students’ posttest scores. Interaction effects were found between class achievement history and student demographic variables. These results provide further evidence to support multi-level classroom context analyses.

Conclusion

The main purpose of this study was to use multilevel analysis, which is most pertinent for conducting studies of educational context, to account for systematic variation at the classroom level. The multilevel nature of the data has revealed critical classroom level patterns in data that we would not otherwise notice. In this study, we have shown that both curriculum condition as a class level variable and achievement history variables are useful areas to consider for future research. We would recommend other researchers who study educational context to employ multilevel methods at the classroom level. It is not appropriate to ignore the context in studies such as these, where the students are nested within dynamic classroom settings. Our results reiterate that the classroom characteristics have an influence on student learning.

The findings of this study indicated that the curriculum intervention is effective across subgroups. In other words, the curriculum is good for all students after controlling for other variables that might produce initial differences. It is important to consider the achievement history variables. Commonly practices administrative scheduling practices do affect student outcome. Even though a significant main effect was not found for achievement history, the interaction effects suggest these variables may matter in some circumstances. Teachers and school personnel should be concerned about students eligible for ESOL services who are placed in low achievement history classrooms. In general, as a group these students are in a
disadvantaged situation due to their placement in low achievement history classrooms. Hispanic students are at an advantage when they are placed in high achievement history classrooms. Teachers and school personnel should be aware of major factors when placing students into classrooms. Placement should not be based on students’ social/demographic profile. African American students, students requiring special education or ESOL services, and low SES students scored less than their counterparts regardless of treatment condition.

Future research should look more into classroom achievement history variable using more classrooms, especially interaction effects between curriculum condition and classroom achievement history variables. In this study, the variable – eligibility for ESOL services – was used as a dichotomous variable, combining students who had once been eligible for ESOL services and students currently eligible for ESOL. Future study might examine these two variables separately. The present model explained only 17% of the variance in students’ posttest scores and there may be other variables that could be explored to account for more variance in the outcome.

There were several limitations to this study. The use of secondary demographic and prior achievement data limited the examination to variables that were available in the dataset. The cluster size was very small and so the power was low to detect significant main effects for the achievement history variables. In spite of these limitations, this study contributes to curriculum evaluation research by using multilevel analysis and examines the classroom context effects in curriculum evaluation research.

The authors caution the reader against making interpretations or conclusions based on the current preliminary findings. The conclusions are tentative pending the authors’ further analyses and interpretation.

References


Notes

1Dotted lines indicate interaction effect.
Reporting effect sizes are common in education research. The standards suggested by Rosenthal and Rosnow (1984, p. 360) are used. Effects of .5 SD or more in magnitude are large; .3-.5 SD range is moderate; .1-.3 SD is small, and those below .1 SD are trivial. Also, see Lee and Loeb (2000) for more details.
### Table 1. Percentage of Students in each Demographic Category and their Scores by Classroom Achievement History

<table>
<thead>
<tr>
<th>Class</th>
<th>SchSES</th>
<th>%Asian</th>
<th>%AfAm</th>
<th>%White</th>
<th>%Hispanic</th>
<th>%FARMS</th>
<th>%Sped</th>
<th>%Esol</th>
<th>Mpostest</th>
<th>%Math</th>
<th>%Read</th>
<th>MScGPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1T High</td>
<td>7.7</td>
<td>23.1</td>
<td>42.3</td>
<td>26.9</td>
<td>53.8</td>
<td>7.7</td>
<td>23.1</td>
<td>66.97</td>
<td>66.61</td>
<td>65.37</td>
<td>3.02</td>
<td></td>
</tr>
<tr>
<td>B2T High</td>
<td>11.1</td>
<td>22.2</td>
<td>18.5</td>
<td>48.1</td>
<td>63</td>
<td>7.4</td>
<td>25.9</td>
<td>34.83</td>
<td>64.68</td>
<td>63.59</td>
<td>2.91</td>
<td></td>
</tr>
<tr>
<td>C1T Medium</td>
<td>33.3</td>
<td>3.7</td>
<td>37</td>
<td>25.9</td>
<td>37</td>
<td>7.4</td>
<td>18.5</td>
<td>62.13</td>
<td>66.75</td>
<td>64.75</td>
<td>2.93</td>
<td></td>
</tr>
<tr>
<td>C2T Medium</td>
<td>14.3</td>
<td>14.3</td>
<td>67.9</td>
<td>3.6</td>
<td>14.3</td>
<td>25</td>
<td>7</td>
<td>58.66</td>
<td>66.95</td>
<td>64.95</td>
<td>3.39</td>
<td></td>
</tr>
<tr>
<td>A1T Med Low</td>
<td>12.5</td>
<td>12.5</td>
<td>46.9</td>
<td>28.1</td>
<td>46.9</td>
<td>0</td>
<td>12.5</td>
<td>48.81</td>
<td>67.16</td>
<td>63.75</td>
<td>3.86</td>
<td></td>
</tr>
<tr>
<td>A2T Med Low</td>
<td>13</td>
<td>21.7</td>
<td>43.5</td>
<td>21.7</td>
<td>43.5</td>
<td>13</td>
<td>21.7</td>
<td>59.08</td>
<td>67.19</td>
<td>65.4</td>
<td>3.57</td>
<td></td>
</tr>
<tr>
<td>D1T Low</td>
<td>7.4</td>
<td>11.1</td>
<td>77.8</td>
<td>3.7</td>
<td>11.1</td>
<td>7.4</td>
<td>14.8</td>
<td>66.47</td>
<td>71.22</td>
<td>69.06</td>
<td>3.25</td>
<td></td>
</tr>
<tr>
<td>D2T Low</td>
<td>9.7</td>
<td>9.7</td>
<td>80.6</td>
<td>0</td>
<td>16.1</td>
<td>6.5</td>
<td>6.5</td>
<td>68.28</td>
<td>71.47</td>
<td>69.07</td>
<td>3.66</td>
<td></td>
</tr>
<tr>
<td>B1C High</td>
<td>7.1</td>
<td>14.3</td>
<td>64.3</td>
<td>14.3</td>
<td>28.6</td>
<td>14.3</td>
<td>7.1</td>
<td>57.16</td>
<td>64.42</td>
<td>63.25</td>
<td>2.79</td>
<td></td>
</tr>
<tr>
<td>B2C High</td>
<td>4</td>
<td>16</td>
<td>36</td>
<td>44</td>
<td>52</td>
<td>16.7</td>
<td>24</td>
<td>51.46</td>
<td>66.54</td>
<td>64.63</td>
<td>2.6</td>
<td></td>
</tr>
<tr>
<td>C1C Medium</td>
<td>26.1</td>
<td>21.7</td>
<td>39.1</td>
<td>13</td>
<td>30.4</td>
<td>18.2</td>
<td>8.7</td>
<td>51.89</td>
<td>68.87</td>
<td>66.83</td>
<td>3.24</td>
<td></td>
</tr>
<tr>
<td>C2C Medium</td>
<td>15</td>
<td>25</td>
<td>50</td>
<td>10</td>
<td>15</td>
<td>0</td>
<td>20</td>
<td>58.48</td>
<td>70.92</td>
<td>67.05</td>
<td>3.38</td>
<td></td>
</tr>
<tr>
<td>A1C Med Low</td>
<td>14.3</td>
<td>10.7</td>
<td>71.4</td>
<td>3.6</td>
<td>21.4</td>
<td>17.9</td>
<td>7.1</td>
<td>52.93</td>
<td>68.38</td>
<td>66.15</td>
<td>3.19</td>
<td></td>
</tr>
<tr>
<td>A2C Med Low</td>
<td>3.8</td>
<td>7.7</td>
<td>69.2</td>
<td>19.2</td>
<td>33</td>
<td>9.5</td>
<td>37.5</td>
<td>48.41</td>
<td>67.56</td>
<td>66.07</td>
<td>3.63</td>
<td></td>
</tr>
<tr>
<td>D1C Low</td>
<td>23.5</td>
<td>5.9</td>
<td>70.6</td>
<td>0</td>
<td>5.9</td>
<td>11.8</td>
<td>0</td>
<td>65.88</td>
<td>70.57</td>
<td>67.21</td>
<td>3.74</td>
<td></td>
</tr>
<tr>
<td>D2C Low</td>
<td>20.8</td>
<td>4.2</td>
<td>75</td>
<td>0</td>
<td>8.3</td>
<td>16.7</td>
<td>12.5</td>
<td>67</td>
<td>70.53</td>
<td>68.3</td>
<td>3.65</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>SchSES</th>
<th>%Asian</th>
<th>%AfAm</th>
<th>%White</th>
<th>%Hispanic</th>
<th>%FARMS</th>
<th>%Sped</th>
<th>%Esol</th>
<th>Mpostest</th>
<th>%Math</th>
<th>%Read</th>
<th>MScGPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1T High</td>
<td>20.8</td>
<td>29.2</td>
<td>12.5</td>
<td>37.5</td>
<td>91.7</td>
<td>29.2</td>
<td>58.3</td>
<td>49.67</td>
<td>60.11</td>
<td>58.46</td>
<td>2.13</td>
<td></td>
</tr>
<tr>
<td>B2T High</td>
<td>0</td>
<td>23.8</td>
<td>19</td>
<td>57.1</td>
<td>76.2</td>
<td>38.1</td>
<td>33.3</td>
<td>56.74</td>
<td>57.81</td>
<td>58.72</td>
<td>2.23</td>
<td></td>
</tr>
<tr>
<td>C1T Medium</td>
<td>9.5</td>
<td>23.8</td>
<td>23.8</td>
<td>42.9</td>
<td>61.9</td>
<td>42.9</td>
<td>42.9</td>
<td>47.25</td>
<td>59.74</td>
<td>58.5</td>
<td>2.05</td>
<td></td>
</tr>
<tr>
<td>C2T Medium</td>
<td>25</td>
<td>16.7</td>
<td>33.3</td>
<td>25</td>
<td>62.5</td>
<td>9.5</td>
<td>37.5</td>
<td>50.5</td>
<td>62.8</td>
<td>60.79</td>
<td>2.13</td>
<td></td>
</tr>
<tr>
<td>A1T Med Low</td>
<td>3.7</td>
<td>22.2</td>
<td>18.5</td>
<td>55.6</td>
<td>74.1</td>
<td>22.2</td>
<td>51.9</td>
<td>40.26</td>
<td>60.27</td>
<td>59.54</td>
<td>3.13</td>
<td></td>
</tr>
<tr>
<td>A2T Med Low</td>
<td>10</td>
<td>26.7</td>
<td>33.3</td>
<td>30</td>
<td>56.7</td>
<td>24.1</td>
<td>33</td>
<td>51.87</td>
<td>61.21</td>
<td>60.34</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>D1T Low</td>
<td>3.8</td>
<td>7.7</td>
<td>76.9</td>
<td>11.5</td>
<td>7.7</td>
<td>11.5</td>
<td>11.5</td>
<td>65.04</td>
<td>68.28</td>
<td>65.87</td>
<td>3.02</td>
<td></td>
</tr>
<tr>
<td>D2T Low</td>
<td>6.9</td>
<td>13.8</td>
<td>65.5</td>
<td>13.8</td>
<td>13.8</td>
<td>37.9</td>
<td>10.3</td>
<td>53.34</td>
<td>66.6</td>
<td>64.59</td>
<td>2.98</td>
<td></td>
</tr>
<tr>
<td>B1C High</td>
<td>11.5</td>
<td>23.1</td>
<td>11.5</td>
<td>53.8</td>
<td>76.9</td>
<td>38.5</td>
<td>38.5</td>
<td>42.41</td>
<td>61.27</td>
<td>61.46</td>
<td>2.06</td>
<td></td>
</tr>
<tr>
<td>B2C High</td>
<td>8.7</td>
<td>21.7</td>
<td>21.7</td>
<td>47.8</td>
<td>78.3</td>
<td>65.2</td>
<td>47.8</td>
<td>33.29</td>
<td>58.96</td>
<td>58.74</td>
<td>1.77</td>
<td></td>
</tr>
<tr>
<td>C1C Medium</td>
<td>0</td>
<td>40</td>
<td>30</td>
<td>30</td>
<td>55</td>
<td>42.1</td>
<td>20</td>
<td>36.47</td>
<td>60.59</td>
<td>59.07</td>
<td>1.89</td>
<td></td>
</tr>
<tr>
<td>C2C Medium</td>
<td>4.2</td>
<td>45.8</td>
<td>16.7</td>
<td>33.3</td>
<td>58.3</td>
<td>34.8</td>
<td>16.7</td>
<td>35.64</td>
<td>59.29</td>
<td>60.33</td>
<td>2.21</td>
<td></td>
</tr>
<tr>
<td>A1C Med Low</td>
<td>27.8</td>
<td>5.6</td>
<td>61.1</td>
<td>5.6</td>
<td>22.2</td>
<td>11.8</td>
<td>11.1</td>
<td>58.74</td>
<td>65.01</td>
<td>61.08</td>
<td>2.94</td>
<td></td>
</tr>
<tr>
<td>A2C Med Low</td>
<td>5.6</td>
<td>27.8</td>
<td>61.1</td>
<td>5.6</td>
<td>27.8</td>
<td>37.5</td>
<td>11.1</td>
<td>37.8</td>
<td>61.5</td>
<td>59.69</td>
<td>2.47</td>
<td></td>
</tr>
<tr>
<td>D1C Low</td>
<td>47.8</td>
<td>4.3</td>
<td>47.8</td>
<td>0</td>
<td>4.3</td>
<td>4.3</td>
<td>17.4</td>
<td>59.6</td>
<td>67.14</td>
<td>66.06</td>
<td>3.48</td>
<td></td>
</tr>
<tr>
<td>D2C Low</td>
<td>31.8</td>
<td>13.6</td>
<td>36.4</td>
<td>18.2</td>
<td>27.3</td>
<td>27.3</td>
<td>27.3</td>
<td>60.65</td>
<td>65.04</td>
<td>61.69</td>
<td>3.07</td>
<td></td>
</tr>
<tr>
<td>AVG</td>
<td>13.57</td>
<td>21.61</td>
<td>35.57</td>
<td>29.23</td>
<td>49.67</td>
<td>29.81</td>
<td>29.29</td>
<td>48.70</td>
<td>62.23</td>
<td>60.93</td>
<td>2.54</td>
<td></td>
</tr>
</tbody>
</table>