Research in Mathematics Education, 2013
Vol. 15, No. 2, 177–188, http://dx.doi.org/10.1080/14794802.2013.797749

HLM in cluster-randomised trials – measuring efficacy across diverse populations of learners

Stephen Hegedus[a]*, John Tapper[b], Sara Dalton[a] and Finbarr Sloane[c]

[a]*Kaput Center for Research and Innovation in STEM Education, University of Massachusetts Dartmouth, USA;* [b]*Department of Education and Human Services, University of Hartford, USA;* [c]*National Science Foundation, USA*

We describe the application of Hierarchical Linear Modelling (HLM) in a cluster-randomised study to examine learning algebraic concepts and procedures in an innovative, technology-rich environment in the US. HLM is applied to measure the impact of such treatment on learning and on contextual variables. We provide a detailed description of such methods, methodically analysing nested classroom data with respect to various outcome measures through HLM.

Keywords: cluster-randomised trials; HLM; algebra and learning; SimCalc

Introduction: overview of study

The SimCalc Program of Study aims to address various issues of long-term national concern in the US: the problem of student motivation and alienation in the nation's high schools, especially urban high schools (National Research Council 2003), and the widely acknowledged unfulfilled promise of technology in education, especially mathematics education (Cuban 2001). We also address the need to retain more students in mathematics beyond algebra, which increasingly means past high-stakes state examinations, into fundamental courses such as pre-calculus and calculus, to help prepare them for higher education and careers in STEM programmes. It has been stated that the introduction of algebra in high schools in the US is too late and should be done earlier to avoid such problems (Kaput, Carraher, and Blanton 2007). Traditionally, *Algebra 1* in US high schools focuses on the introduction of linear functions, linearity, co-variation, rate, and systems of linear equations. It is introduced in the Freshman year of high school (ages 14–15). The SimCalc project has focused on addressing these topics under the broader theme of the mathematics of change and variation, a core school mathematics strand that is representationally demanding and that is studied at many levels by all students, from pre-algebra through calculus (Kaput 1994).

SimCalc MathWorlds® combines two innovative technological ingredients to address core mathematical ideas in deep and sustainable ways for mathematics learners (Beatty and Geiger 2010). First, the software addresses content issues through dynamic representations integrating simulations with editable graphs, equations and tables; secondly, it uses wireless networks to enhance student participation in the classroom. The algebra course focused on in this paper was

*Corresponding author. Email: shegedus@umassd.edu

© 2013 British Society for Research into Learning Mathematics

developed at the Kaput Center; it fuses these two ingredients through new curriculum materials that replace core mathematical units in high school algebra courses and through non-traditional classroom arrangements. Activities in a SimCalc learning environment are structured to create meaningful mathematical variation across students' activities, e.g. students can be assigned to a group that has a number and have their own number within the group; both of these numbers can be used as parameters in a linear equation.

We studied the impact on learning core algebra concepts and skills that were aligned with Massachusetts state frameworks and deepened student understanding by building skill sets that crossed algebraic topics and reasoning. We focused on the concept of slope as rate, and not just slope of a function. In addition, we measured the impact of such a technology-enhanced approach to learning algebra on other important longitudinal factors, such as a student's confidence in their mathematical ability and their attitudes towards learning mathematics over time.

Our research program is the result of 15 years of prior work (Hegedus and Roschelle, 2013) focused on the development and implementation of technology-enhanced learning environments that have proved to be effective through quasi-experimental research studies and randomised controlled trials at scale (Roschelle et al. 2010). Here, we report on the results of an efficacy study conducted in Massachusetts that continues to examine under what conditions such materials can be effective in improving learning of core algebra concepts across a wide, diverse population of students in school districts of varying achievement.

Research design

Our program of study was funded by the US Department of Education's Institute of Education Sciences to investigate the impact of SimCalc materials on a large sample of classrooms that differ in various ways. In order to control for various confounding and explanatory variables and to account for clustering at the classroom level, we designed a cluster-randomised trial (where classes are randomly assigned SimCalc materials or the existing algebra curriculum). As our sample was randomised at the classroom level, we were interested in effects on classrooms and for individual students. The desire to locate effects at both the classroom and student level led us to adopt a multi-level approach using HLM. In this paper we review the findings of our HLM analysis, providing a description of the process for researchers who may be new to multilevel modelling. We will examine the construction and application of the following models to analyse our data and address four research questions:

(1) The *Empty* or *Open Model* to determine the variance accounted for at each of our two levels of interest – classroom level and student level. *How much variance exists between classes and within classes on our outcome measure, a mathematics content post-test?*
(2) The *Random Intercepts Model* to investigate the effect of a variable at the student level. *How much effect do student pre-test scores have on student post-test scores at the end of the Freshman school year?*
(3) The *Means as Outcomes Model* to test the impact of a classroom contextual variable on learning. *Can group membership in an honors class predict a significantly higher post-test score?*

(4) The *Random Slopes Model* to look at the impact of level 2 variables on level 1 variables. *For students in honors classes, is treatment group a predictor for post-test score when controlling for pre-test score? Is the variance in slopes across groups (classes) significantly related to the experimental group of the students when controlling for the pre-test?*

A power analysis prior to the study confirmed the sample size and numbers of classrooms (clusters) necessary for this study. Our research program made use of various cluster-randomised trials over a 4-year longitudinal study. We completed two pilot studies in Years 1 and 2, which guided our Algebra 1 and 2 curriculum and instrument development. We conducted two main studies in Year 2 (Algebra 1) and Year 3 (Algebra 2) following our pilot studies, and a replication study in Year 4 (Algebra 2). This report focuses on the data collection from our Year 2 Algebra 1 main study.

Sample

From a total sample size of 60 eligible classes, 28 classrooms were selected to participate in the study by randomly selecting a pair (ensuring each pair was from one school), randomly assigning one of the pair red or black, and with 14 pairs selected, flipping a coin to assign whether red was Treatment or Control (and hence black the other condition).

Approximately 92% (568 students of 615 students listed on the class rosters) of our entire student sample agreed to participate in the study. In terms of attrition, approximately 2.3% (13/568) of the student sample transferred out of the participating classroom or transferred out of the school, and we could not collect complete data for approximately 5.5% (31/568) of our participants because their teacher left the study for medical reasons. Our final sample (Treatment = 257; Control = 195 students) includes only students who took a pre-test and a corresponding post-test. The pre-test and post-test were identical. This restricted sample is used in the analyses in this paper. We compared this restricted sample with our initial sample, and found no significant differences on their pre-test scores.

Treatment teachers implemented an 8–12 week SimCalc Algebra 1 replacement curriculum consisting of 12 units in conjunction with the SimCalc MathWorlds® software[1]. The variation in implementation length was due to differences in school schedules – some schools' class periods lasted 90 minutes while others were only 50 minutes. The pre-test was administered prior to the intervention at various times in each school due to differences in schedules and at similar times in the control classrooms when the relevant content was being covered. The post-test was administered at the conclusion of the unit/relevant content, which was towards the end of the Freshman year.

The study included rural, urban and suburban schools of varying proficiency levels (from high to low) and ethnic diversity. There were no significant differences for participating teachers in terms of gender (Treatment = 42.86% female, Control = 50% female), holding a mathematics teaching licence (Treatment = 100%, Control = 87%), or years of teaching experience (average for Treatment = 10.65 yrs, average for Control = 6.87 yrs).

Design of instrument

The *Algebra 1* content test focused on linear functions, co-variation, slope as rate versus slope as m in $y = mx + b$, simultaneous equations, and systems of linear functions. The test instrument was constructed following the principled assessment design approach (Mislevy et al. 2003). In this approach, we identified the specific knowledge or skills to be tested, and then articulated an evidence model that specified the kinds of tasks (problem solving, application, open-ended, multiple-choice) that were likely to reveal whether students had mastered the target knowledge base and skill set. Next, we created specific problems. The evidence model was refined through piloting and procedures to ensure the technical quality of assessments. Specific details on the instrument construction and validation can be found in our technical report series at http://www.kaputcenter.umassd.edu/products/technical_reports/

Items in the content test were focused on core algebraic concepts and skills, including: graphical interpretation, rate and proportion, computational/procedural operations and making connections across representations.

Analytical framework

HLM analyses are most useful for nested data such as students within classes, students within schools, etc. We conducted cluster-randomised trials, so nesting was important and needed to be accounted for in the analysis. Our research questions focus on measuring ability to learn in differently achieving classrooms. This structures the analytical framework. In our study, the unit of analysis was the student and the unit of randomisation was the classroom (the cluster). If the outcome variable, Y, is, say, students' aggregate post-test scores, this outcome has an individual *and* a group aspect to it (it might differ across classes). Similarly, the independent variable X, say, a student's pre-test score, has a group aspect to it. In our datasets, the nesting structure (where the groups are classes of individual students) is meaningful, and it is unfounded to assume that the group structure is completely represented by the explanatory variables. Nesting can be represented by several mathematical equations that interact with each other and, through substitution, can be written as one mixed model. For example:

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j} * X_{ij} + R_{ij}$$

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * Z_j + U_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} * Z_j + U_{1j}$$

represents a 2-level model, where level 1 accounts for individual or micro-unit variation and level 2 for group or macro-unit. Between-group variability can be modelled by letting the intercepts and slopes vary between groups; for instance, some classrooms have higher scores on the post-test than other classrooms. The outcome variable, Y_{ij}, is a student score on a post-test where i is the index for each student ($i = 1 \ldots n_j$) and j is the index for each classroom, $j = 1 \ldots N$ (note we use capital N for groups). For individual *i* in group *j* there are two variables, the dependent variable, Y, and an explanatory variable X, denoted by X_{ij}. There is also an explanatory variable at the group level Z, denoted by Z_j. Note the subscript for Z is just *j* since there is no variation by student (level 1); throughout, the first subscript (if any) refers to variation at level 1, the second subscript refers to variation at level 2, and a zero corresponds to no variation at that level. So our level 1 equation has group-dependent intercepts described by the level 2 equation $\beta_{0j} = \gamma_{00} + \gamma_{01} * Z_j + U_{0j}$ and group-dependent slopes described by the level 2 equations $\beta_{1j} = \gamma_{10} + \gamma_{11} * Z_j + U_{1j}$. If there were two level 1 predictors, then we would have another equation for β_{2j}, and so on.

Finally, γ_{00} is the grand mean for X, U_{0j} is the random effect (or unexplained variability) at the group level, and R_{ij} the random effect at the individual level. γ_{01} is a regression coefficient describing the effect of group level predictors (Z), and γ_{10} describes the main effects of individual level predictors (X). γ_{11} is a cross-interaction effect, to which we return later. When the level 2 equations are substituted into the level 1 equation, we get the full model:

$$Y_{ij} = \gamma_{00} + \gamma_{01} * Z_j + \gamma_{10} * X_{ij} + \gamma_{11} * Z_j * X_{ij} + U_{1j} * X_{ij} + U_{0j} + R_{ij}$$

We have rearranged the terms to illustrate the fixed effects part of the model (first 4 terms) and the random effects part of the model (last 3 terms). The aim of HLM is to explore whether such explanatory variables (when nested) can significantly predict the dependent variable, and how much variance can be accounted for by the random effects. Our results will be tabulated by such effects. Including random effects allows there to be unexplained between-group variability that can be explained by group level variables. This approach has benefits over ANCOVA (especially with small group sizes < 50) as the model assumes independent and identically distributed group and individual effects, i.e., $U_{0j} \sim (0, \tau_{00})$ and $R_{ij} \sim (0, \sigma^2)$. Unexplained group effects are controlled by similar mechanisms across all groups, and operate independently between groups (Snjiders and Bosker 1999). We discuss tests for these assumptions later.

We focus only on 2-level models in this paper where our micro-units are students (scores on our pre-test) and macro-units are classrooms of students with an average class size (cluster size) of 17 students. Dependency of observations on the micro-units within the macro-units is of primary interest. For example, success of a student within a particular class may derive from interactions with other students sharing the same class environment, or from the effect of sharing the same teacher, or from numerous other classroom level conditions.

Results

We now present the results for our four research questions, using progressively more complex modelling. We present each analysis in the context of our student performance data in algebra classrooms using a threefold structure: (1) Theoretical model from HLM; (2) Application of the model to our dataset; and (3) Results and interpretation.

Research Question 1 (Empty Model): *How much variance exists between classes and within classes on our outcome measure, a mathematics content post-test?*

Model: This question can be addressed using the Empty model, which is the simplest form of HLM, and can also described as a 1-way random effects ANOVA. Contrast this with a full 2-level model (described above):

$$Y_{ij} = \gamma_{00} + U_{0j} + R_{ij}$$

Here, the dependent variable γ_{00} is the population grand mean, U_{0j} is a random effect at the group level (macro-unit), and R_{ij} is a random effect at the individual level (micro-unit). The macro-unit has a true mean $\gamma_{00} + U_{0j}$ and each micro-unit *i* with this macro-unit *j* deviates from this mean by R_{ij}. From this model, a parameter called the intraclass correlation coefficient (ICC) can be calculated by comparing the level 1 (σ^2) and level 2 (τ_{00}) variances:

$$\rho = \frac{Var\left(U_{0j}\right)}{Var\left(U_{0j}\right) + Var\left(R_{ij}\right)} = \frac{\tau_{00}}{\tau_{00} + \sigma^2}$$

Application: The ICC can be interpreted in two ways: (1) The proportion of total variability due to the group level; or (2) The correlation between two randomly drawn micro-units from one randomly drawn macro-unit. Note γ_{00} should be close to, but not the same as, the raw arithmetic mean due to weighting of various groups. *Results and interpretation*: Table 1 outlines the relevant output from HLM6 software (http://www.ssicentral.com/hlm/) for data collected in our study. The ICC can be calculated from this output as .515. This demonstrates a large proportion of variance due to the group, hence nesting individuals in classes is relevant and HLM should be used. Normal levels for ρ are between .05 and .2.

Research Question 2 (The Random Intercepts Model): *How much effect do student pre-test scores have on student post-test scores at the end of the Freshman school year?*

Model: The Random Intercepts (or Random Coefficients) model is a simple form of HLM, as it does not involve random slopes (see later). Here, we introduce one explanatory variable X at level 1:

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j} * X_{ij} + R_{ij}$$

Level 2:

$$\beta_{0j} = \gamma_{00} + U_{0j}$$
$$\beta_{1j} = \gamma_{10}$$

Table 1. Fixed effects and random effects.

Fixed Effect	Parameter	Coefficient	SE	t	P
Intercept	γ_{00}	9.551	.731	13.058	$p < .001$
Random Effect	Parameter	Variance component	df	χ^2	
Level 2 variance	$\tau_{00} = Var(U_{0j})$	14.509	28	524.247	$p < .001$
Level 1 variance	$\sigma^2 = Var(R_{ij})$	13.662			

Deviance = 2541.797

The simple model has four parameters, regression coefficients γ_{00} and γ_{10} and variance components U_{0j} and R_{ij}, so the model includes unexplained variability at both levels. γ_{00} is the intercept for the average group and γ_{10} is an unstandardised within-class regression coefficient, or slope of the X variable, which as a constant is fixed across all groups (hence this is a random intercepts model). It describes the relationship between X and Y; increasing one unit of X leads to an average increase of Y by β_{1j}.

Application: With such variables, our model tests the question of whether pre-test scores (X) predict post-test scores (Y) accounting for groups (classes).

Results and interpretation: The results (Table 2) indicate that the intercept and slope for our content test at the end of the school year are significantly different from 0. The model is Y = 10.104 + .667 * (PRETEST). For every one-point increase on the pre-test, we could expect an increase of almost .67 points on the post-test for the end-of-school year content test. The Chi-squared test indicates that there is significant between-group (class) variance in the intercept parameter across groups (classes). This indicates that there is a significant amount of variance in pre-test scores between groups (classes).

Comparing the output for the random effects with the Empty Model will illustrate lower variance components for τ_{00} and σ^2. This is because the between-group differences are partially explained and include the effect of X. A new ICC can be calculated using the same formula as above, but this is referred to as the residual ICC because it controls for X (see Table 2 for variance values). For our dataset $\rho(Y|X) = 2.646/(9.677 + 2.646) = .215$. This is lower than the raw ICC since the residual variances are lower. This is because the classes (groups) differ by average pre-test score, so this level-1 student variable (X_{ij}) also explains part of the differences between the groups.

What happens if $\tau_{00} = 0$? We briefly consider the difference between HLM and Ordinary Least Squares (OLS) regression methods for the same set of data to support our approach for using multi-level analysis based on our research design of cluster-randomised trials. OLS is a method for estimating the unknown parameters in a linear regression model. As τ_{00} tends to zero, there is less between-group variance and hence less call for nesting data and an HLM analysis. When $\tau_{00} = 0$ the HLM6 software produces OLS estimates for regression coefficients and standard errors (see Table 3). This output illustrates that the OLS coefficients overestimate the effect in contrast to the random intercepts model (10.482 in contrast to 10.104), but more importantly under-estimate the standard error (0.249 in contrast to 0.345). Hence,

Table 2. Fixed effects and random effects for simple HLM.

Fixed Effect	Parameter	Coefficient	SE	t	p
Intercept	γ_{00}	10.104	.345	29.302	$p < .001$
Slope	γ_{10}	.667	.060	11.190	$p < .001$
Random Effect	**Parameter**	**Variance component**	**df**	χ^2	
Level 2 variance	τ_{00}	2.646	28	146.923	$p < .001$
Level 1 variance	σ^2	9.677			

Deviance = 2358.441

Table 3. Estimators for OLS and HLM.

Least Squares Estimates of fixed effects (OLS)		Random Intercepts	
γ_{00}	10.482	γ_{00}	10.104
SE	.249	SE	.345
γ_{10}	.829	γ_{10}	.667
SE	.045	SE	.060

standard OLS techniques aggregate classroom effect and do not distinguish the effect of clustering.

Research Question 3 (Means as Outcomes Model): *Can group membership in an honors class predict a significantly higher post-test score?*

Model. Here we introduce one explanatory variable Z at level 2:

Level 1:

$$Y_{ij} = \beta_{0j} R_{ij}$$

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * Z_j + U_{0j}$$

In this application we do not have a level 1 explanatory variable, but we introduce a level 2 explanatory variable to aggregate the effect at the group level. So Z_j is a level 2 predictor with γ_{01} as a between-group regression coefficient. Hence this is traditionally called a Means as Outcomes model.

Application: Classes within each school varied by school-defined class level – there were honors classes and non-honors classes within each school. We chose to investigate whether there was a significant difference between groups when predicting post-test outcomes using a Means as Outcomes model. Y_{ij} is the post-test score for an individual *i* in class *j*, and γ_{01} * *(HONORS)* is a dichotomous variable that identifies students as either being in an honors class (1) or in a non-honors class (0).

Results and interpretation: Since the code for non-honors students is 0 and the code for honors students is 1, the coefficient of 8.65 indicates the average score of a non-honors student (see Table 4). The coefficient for the honors classes is the slope of the line from zero (non-honors) to one (honors). This slope is significant ($p < .001$) and suggests a predicted advantage for honors students of more than five points on the post-test.

Research Question 4 (The Random Slopes Model): *For students in honors classes, is treatment group a predictor for post-test score when controlling for pre-test score? Is*

Table 4. Honors classes predicting post-test results.

Fixed Effect	Parameter	Coefficient	SE	t	p
Intercept	γ_{00}	8.65	0.69	12.56	p <.001
Intercept	γ_{01}	5.18	1.53	3.38	p <.001
Random Effect	Parameter	Variance component	df	χ^2	p
Level 2 variance	τ_{00}	10.84	27	355.81	p <.001
Level 1 variance	σ^2	13.66			

the variance in slopes across groups (classes) significantly related to the experimental group of the students when controlling for the pre-test?

Model: This is the most commonly used general HLM for 2-level models where several explanatory variables are added at level 2, including random effects.

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j} * X_{ij} + R_{ij}$$

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * Z_j + U_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} * Z_j + U_{1j}$$

This model involves more level 2 random effects, i.e., $\tau_1 = \text{var}(U_{1j})$ and $\tau_{01} =$ covariance (U_{0j}, U_{ij}), and these need careful examination. It is customary to add level 2 explanatory variables in both the intercept and slope equations, but we express caution here. More explanatory variables can be added in each equation, but a good rule of thumb is to think of the intercept equation as an initial/entry stage for an intervention/experiment, so that any added variables should theoretically be considered as dependent variables of such a state at that time. Similarly, the slope equation, β_{1j}, should be considered as a growth coefficient, and so level 2 explanatory variables should be added that are theoretically considered as potential contributors to change/growth over the course of the intervention. When there are q level 1 predictors, there will be an equivalent number of level 2 slope equations, denoted β_{pj} where $p = 1 \ldots q$.

Application[2]:

Level 1:

$$(\text{POSTTEST})_{ij} = \beta_{0j} + \beta_{1j} * (\text{PRETEST})_{ij} + R_{ij}$$

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} * (\text{TREAT})_j + \gamma_{02} * (\text{TREAT} * \text{HONORS})_j + U_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} * (\text{TREAT} * \text{HONORS})_j$$

TREAT refers to the experimental group (0: Control; 1: SimCalc). We first ran a random slopes model with TREAT and HONORS as explanatory variables at level 2. The main effect of TREAT was not significant ($p = .853$) but the main effect of HONORS was significant ($p = .003$) Because of this, we wanted to force an interaction between TREAT and HONORS (i.e., Z = TREAT*HONORS) to see if this could account for such significance in the HONORS group.

Results and interpretation: The test associated with γ_{01} is the main effect for the treatment group (see Table 5), which is not significant. However, the main effect for the interaction between TREAT and HONORS is significant, $\gamma_{02} = 1.89$, $p = .044$. This indicates that the post-test scores for SimCalc honors students are significantly greater than those for Control students, and than those for non-honors students at the class level. Additionally, we see in γ_{10} that there is a significant difference between groups (classes) on the post-test when controlling for pre-test scores. The TREAT * HONORS interaction, however, is not mediating a cross-level interaction between

Table 5. Fixed effects and random effects for full HLM.

Fixed Effect	Parameter	Coefficient	SE	t	p
Intercept	γ_{00}	10.01	0.43	23.69	p <.001
TREAT	γ_{01}	−0.18	0.70	−0.25	p =.803
TREAT * HONOR	γ_{02}	1.89	0.90	2.11	p =.044
Slope	γ_{10}	0.67	0.06	10.91	p <.001
TREAT * HONOR	γ_{11}	−0.06	0.19	−0.32	p =.748
Random Effect		**Variance Component**	**df**	χ^2	**p**
Level 2 variance	τ_{00}	2.75	26	139.338	p <.001
Level 1 variance	σ^2	9.67			

Deviance = 2353.259

students within a class and classes when controlling for the pre-test, as indicated by γ_{11},. Neither is the interaction moderating a relationship from pre to post test scores within classrooms, i.e., the pre to post test relationships within honors level classrooms is not found to be significant.

Checking model assumptions

As in any statistical modelling, there are assumptions underlying the theoretical model, which should be verified in order for all subsequent claims to be ratified. In HLM, these focus on checking whether the level 1 and 2 residuals are normal and have constant variance, i.e., $U_{0j} \sim (0, \tau_{00})$ and $R_{ij} \sim (0, \sigma^2)$. We have verified four assumptions that need to be met for HLM. The first is to check for linearity at all levels. We built 2-level models and therefore needed to check for linearity at level 1 – the student level – and at level 2 – the class level. Using statistical software, we ran a linear regression to check the value of R^2 for our predictors onto outcome measures. The R^2 value with post-test as our dependent variable was greater than 0.5 and less than 0.7.

The second assumption pertains to normality of residuals at each level. We checked the residuals for normality using the Kolmogorov-Smirnov test and the Shapiro-Wilk test. We also plotted a Q-Q plot of the level 1 residuals; the plot was approximately linear, suggesting there is not a significant departure from a normal distribution. For the level 2 residuals, we plotted a Q-Q plot of the Mahalanobis distances; the plot resembled a 45-degree line, indicating that random effects are normally distributed.

The third assumption is homogeneity of level 1 variance assumption. This assumption asserts that the residual variance of the outcome – after adjusting for the level 1 covariate – is the same for all groups, in our case, classes. The HLM6 software can output a Chi-square test for homogeneity of variance. We also computed a skewness z-score and a kurtosis z-score for our level 1 residuals.

The fourth assumption pertains to independence, and can be met through an analysis of the correlation between the level 1 and level 2 residuals in which there was no correlation. Additionally, independence assumes observations in the highest level are independent of each other. Using the residual files saved from the HLM6 software, we plotted a scatter plot of the level 1 residuals against the fitted value for each level 1 unit, which are the values predicted based on the level 2 model. There was no strong relation here.

Conclusions

In this paper, we have demonstrated the impact of a SimCalc intervention under a cluster-randomised trial in algebra classrooms in the US, and have illustrated how multi-level analytical methods can be used to disaggregate datasets and examine the effect of contextual variables. The various models outlined in this paper have exemplified techniques to examine such approaches. We have also attempted to describe under what conditions certain performance measures are associated with individual and group-wise or contextual variables related to a SimCalc intervention.

We hope that such a delicate treatment, through progressively and methodically adding contextual and explanatory variables to nested models, can help to describe both the variation we observe in students' learning of important algebraic concepts and the analytical methods used to assess such claims. We propose that such models can help researchers, examining impact at varying degrees of scale, unpack their results and look more carefully at several contextual variables within school settings.

Sponsorship

This material is based upon work supported by the U.S. Department of Education, Institute of Education Sciences (IES) under Grant # R305B070430. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s), and do not necessarily reflect the views of the U.S. Department of Education or the National Science Foundation.

Notes

1. Download SimCalc curriculum and software at http://www.kaputcenter.umassd.edu/products/software/smwcomp/download/
2. In our models, all individual student level data were centred at the grand mean of all pre-tests vs. the group (class) level. Care should be taken in deciding which form of centring (if any) should be applied, and how to interpret the results. The reason for choosing grand versus group mean centring relates to research design and sample (see Lüdtke et al. 2009; Sloane 2008).

References

Beatty, R., and V. Geiger. 2010. Technology, communication, and collaboration: Re-thinking communities of inquiry, learning and practice. In *Mathematics education and technology – rethinking the terrain*, ed. C. Hoyes and J-B. Lagrange, 251–84. New York: Springer.

Cuban, L. 2001. *Oversold and underused: Computers in the classroom*. Cambridge, MA: Harvard University Press.

Hegedus, S., and J. Roschelle, eds. 2013. *Democratizing access to important mathematics through dynamic representations: Contributions and visions from the SimCalc research program*. Berlin: Springer.

Kaput, J. 1994. Democratizing access to calculus: New routes using old routes. In *Mathematical thinking and problem solving*, ed. A. Schoenfeld, 77–156. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Kaput, J.J., D.W.Carraher, and M.L. Blanton, eds. 2007. *Algebra in the early grades.* Mahwah, NJ: Lawrence Erlbaum Associates.

Lüdtke, O., H.W. Marsh, A Robitzsch, U. Trautwein, T. Asparouhov, and B. Muthén. 2008. The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods* 13: 203–29. doi:10.1037/a0012869.

Mislevy, R.J., L.S. Steinberg, R.G. Almond, G.D. Haertel, and W. Penuel. 2003. Improving educational assessment. In *Evaluating educational technology: Effective research designs for improving learning*, ed. B. Means and G.D. Haertel, 149–80. New York: Teachers College Press.

National Research Council. 2003. *Engaging schools: Fostering high school students' motivation to learn*. Washington, DC: National Academy Press.

Roschelle, J., N. Shechtman, D. Tatar, S.J. Hegedus, B. Hopkins, S. Empson, J. Knudsen, and L. Gallagher. 2010. Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal* 47, no. 4: 833–78. doi:10.3102/0002831210367426.

Sloane, F. 2008. Multilevel models in design research: A case from mathematics education. In *Handbook of design research methods in education*, ed. A.E. Kelly, R.A. Lesh, and J.Y. Baek, 459–76. New York: Routledge.

Snjiders, T.A.B., and R.J. Bosker. 1999. *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications.