Abstract Title Page
Not included in page count.

Title: How do we match instructional effectiveness with learning curves?

Authors and Affiliations:

Lee Branum-Martin, Georgia State University, BranumMartin@gsu.edu
Paras D. Mehta, University of Houston, Paras.Mehta@times.uh.edu
W. Patrick Taylor, University of Houston, Pat.Taylor@times.uh.edu
Coleen D. Carlson, University of Houston, Coleen.Carlson@times.uh.edu
Xiaoxuan Lei, Georgia State University, xlei1@student.gsu.edu
C. Vincent Hunter, Georgia State University, chunter1@student.gsu.edu
David J. Francis, University of Houston, dfrancis@uh.edu

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

In order to examine the effectiveness of instruction, we confront formidable statistical problems, including multivariate structure of classroom observations, longitudinal dependence of both classroom observations and student outcomes, and complex nesting as children change classrooms across years. As we begin to examine instruction, classroom observations involve multiple variables for which we need valid measurement models. These classrooms, however, involve students who are not only growing, but typically switching classrooms each year. While measurement models for multiple variables are commonplace as confirmatory factor analysis (CFA), and individual student growth can be modeled under switching classrooms with multilevel analysis software, connecting these two types of models is currently challenging and limited.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

Consequently, it becomes difficult to jointly examine two types of important substantive questions. First, we are interested in the nature of instruction: to what extent can we fit a measurement model which is consistent over time and what might that say about teachers and classrooms with respect to the stability of instructional quality? Second, how might instructional quality relate to student growth, given changing classrooms essentially every year?

In order to answer these two substantive questions, we can fit a model of instructional observations, a model of student growth, and then join these two models. We illustrate the joining of these two models using a new modeling framework, xxM (Mehta, 2013), a package for multilevel structural equation models (SEMs) in the R programming language (R Core Team, 2014). Overall, we intend to show how the core concepts of SEM for measurement invariance of classroom observation data can be combined with a model of individual student growth in an example of a six-level model.

Setting:

Description of the research location.
(May not be applicable for Methods submissions)

The current data were drawn from an evaluation of the implementation of Reading First in a state in the southern US.

Population / Participants / Subjects:

Description of the participants in the study: who, how many, key features, or characteristics.
(May not be applicable for Methods submissions)

The selected 146 schools had been identified as low-performing in reading scores by the state education agency. The students were 50% female, with 64% Hispanic, 20% African American, and 15% White (1% were other classifications). The teachers were 94% female, 40%

Hispanic, 15% African American, and 42% White (3% other). Overall, these schools can be seen as typical of low-performing schools in the Southwestern US.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

Current investigations of the effects of instruction are typically limited in several ways. First, most multilevel models require only a single outcome, and then multiple levels and predictors are allowed, even with changing classifications, as in mixed-effects regression (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). Second, latent variables as predictors or at higher levels are either not possible or are cumbersome to program in most mixed-effects software. Third, while multilevel SEM is becoming more practical (Kaplan & Elliot, 1997; Mehta & Neale, 2005; Muthén, 1991, 1994), the number of levels is frequently limited on an a priori basis to two or three levels, usually for computational reasons.

The software framework illustrated here, xxM, uses efficient estimation algorithms of mixed-effects models with the specification of SEM for observed and latent variables (Mehta, 2013). The current illustration shows how two commonplace models, a classroom CFA and a student growth model, can be combined to examine crucial questions of instructional effectiveness.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

The model used employs the concepts of standard SEM, with observed and latent variables for CFA. As implemented in xxM, a typical CFA is specified for each level with observed and latent variables, and then across-level links are specified as factor loadings as in a typical CFA. An across-level link (or random slope) in typical multilevel or mixed-effects models is mathematically equivalent to a factor loading in SEM {Mehta, 2000; Mehta, 2005; Mehta, 2013}. xxM uses this equivalence for random slopes so that SEMs can be specified as a typical SEM for each level involved, and then across-level links specified as factor loadings or regressions. There are two major pieces to each model: within-level models and across-level links. A schematic specification is given in Figure 2.

Usefulness / Applicability of Method:
Demonstration of the usefulness of the proposed methods using hypothetical or real data.

We present an empirical example involving a cohort sequential design of 13,236 students over three years (grades 1-3), nested in 974 classrooms, 762 teachers, in 146 schools. These data were taken from a quasi-experimental cohort sequential design for the statewide evaluation of Reading First in a southern US state. For the current analysis, students and teachers were selected to comprise a longitudinal sample, from grades one to three.

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.
(May not be applicable for Methods submissions)

At the child level, test scores were reported by each district (n = 71) at the end of the year for each child. For the classroom observations, a school-level sampling strategy was used in which twice per year (fall and spring), two teachers for each grade level for each school were randomly selected to be observed. Observers with experience in teaching reading were trained in the use of a modified version of the Early Language & Literacy Classroom Observation instrument (Smith, Dickinson, Sangeorge, & Anastasopoulos, 2002).

Measures: Classroom sessions were rated for instruction quality by trained reading teachers using a modified version of the Early Language & Literacy Classroom Observation (Smith et al., 2002). Factor scores for instructional quality in reading, oral language, and writing were used for the present analysis. The student outcome was reading comprehension Lexile score in the spring of each year (divided by 100 for estimation convenience, abbreviated cL).

Analysis: Prior to the current analysis, the modified ELLCO was fit as an item-level CFA to estimate factor scores for the quality of instruction in reading, oral language, and writing. Loadings and thresholds were held equal for all time points and years. These three factor scores were estimated for fall and spring for each classroom observed (total = 974). This preliminary CFA was fit in Mplus with good fit (e.g., CFI > .90 and RMSEA < .08).

Across grades one to three, the three measures of instructional quality in fall and spring make outcomes. A general factor of instructional quality for fall and spring was fit for each grade, holding factor loadings and regression intercepts equal over semesters and years.

For student growth in reading, preliminary multilevel models were fit in SAS PROC HPMIXED. Time was centered at grade one. The individual growth model provided estimates of student initial reading comprehension in grade one and yearly rate of growth, school deviations in initial status and growth, as well as yearly deviations per classroom (as is typical in mixed-effects growth)

Built up from smaller models of classroom observations structure and student growth, we present an overall, integrated model (Figure 1), including a confirmatory model of classroom instructional quality in reading, oral language, and writing in each grade (right hand side of Figure 1). This instructional model is linked to a linear model of student growth (left side of Figure 1), nested within schools and switching classrooms, grades 1-3. The cross-level links use definition variables (diamonds in Figure 1; Mehta & Neale, 2005; Mehta & West, 2000), as in the random slopes of mixed-effects regression.

Findings / Results:
Description of the main findings with specific details.
(May not be applicable for Methods submissions)

This six-level design highlights substantive issues of (1) student-level growth (2) school-level differences in growth (3) classroom deviations in growth akin to "value-added" estimates, (4) instructional quality measured in an equivalent metric across grades, and (5) the relations across these levels.

Results: (1) Student performance appeared typical of grade level expectations, with 3.35cL in first grade and 1.32cL yearly change. Students had substantial variability in intercept (SD = 1.14) and in annual growth (SD = .85). (2) Schools had substantial variability in intercept

(SD = .41) and in annual growth (SD = .31). (3) Classrooms had deviations for each grade which were substantial (SD by grades 1-3 = .45, .50, .68). The standardized loadings for the observation measures ranged from .72 to .96, showing good measurement quality (not in the Figure). (4) Across grades, there was an increase in instructional quality, up to half an SD over fall of grade 1, and a decrease in variability in instructional quality. (5) Instructional quality was related to classroom achievement in grades 1-2 (r = .10 to .23), but unrelated in grade 3 (r = .02 to .06).

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

Limitations. Three time points are not sufficient to test for other types of trajectories or longer trends. The current illustration, however, could be extended to more time points and curvilinear growth. Clustering within teacher was not modeled for space considerations, but could be added. The current design used school-level random sampling of teachers, and teacher variability could not be estimated dependably (this is an area for further testing). Student and school-level predictors can be added. Data were assumed missing at random, and deserve further testing for sensitivity and bias.

Instructional Quality. The measurement properties of the classroom observations were strong, with good fit and high validity coefficients (loadings). The instructional factors suggest an increase in quality and homogeneity across grades. However, instructional quality was not strongly related within year, from fall to spring. A lack of stability in classroom estimates has been found in value-added scores (McCaffrey, Sass, Lockwood, & Mihaly, 2009; National Research Council and National Academy of Education, 2010), and opens deep conceptual problems in classroom level measurement: either our estimates are too poor or teacher quality itself is elusive (Wainer, 2011). The current modeling framework helps to make such modeling issues explicit, and perhaps lays bare a conceptual conundrum.

The relation of instructional quality to classroom achievement appeared weaker in third grade. The decrease in relations of instruction to classroom achievement could suggest that perhaps reading comprehension may be a more complex skill which does not have a simple relation with a global factor of instructional quality in literacy.

Student growth. Classroom deviations were substantial, apparently increasing in variability. Classroom deviations by grade were similar in magnitude to school level differences, implying the possibility of important context effects within schools. Issues for future investigation include missing data, clustering due to teachers, and instructional carryover effects across years.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Kaplan, D., & Elliot, P. R. (1997). A model-based approach to validating education indicators using multilevel structural equation modeling. *Journal of Educational and Behavioral Statistics, 22*(3), 323-347.

Littell, R. D., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, NC: SAS Institute.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572-606.

Mehta, P. D. (2013). n-level structural equation modeling. In Y. Petscher, C. Schatschneider & D. L. Compton (Eds.), *Applied quantitative analysis in the social sciences* (pp. 329-362). New York: Routledge.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations models. *Psychological Methods, 10*(3), 259–284.

Mehta, P. D., & West, S. G. (2000). Putting the individual back in individual growth curves. *Psychological Methods, 5*(1), 23-43.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*(4), 338-354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22*(3), 376-398.

National Research Council and National Academy of Education. (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: The National Academies Press.

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

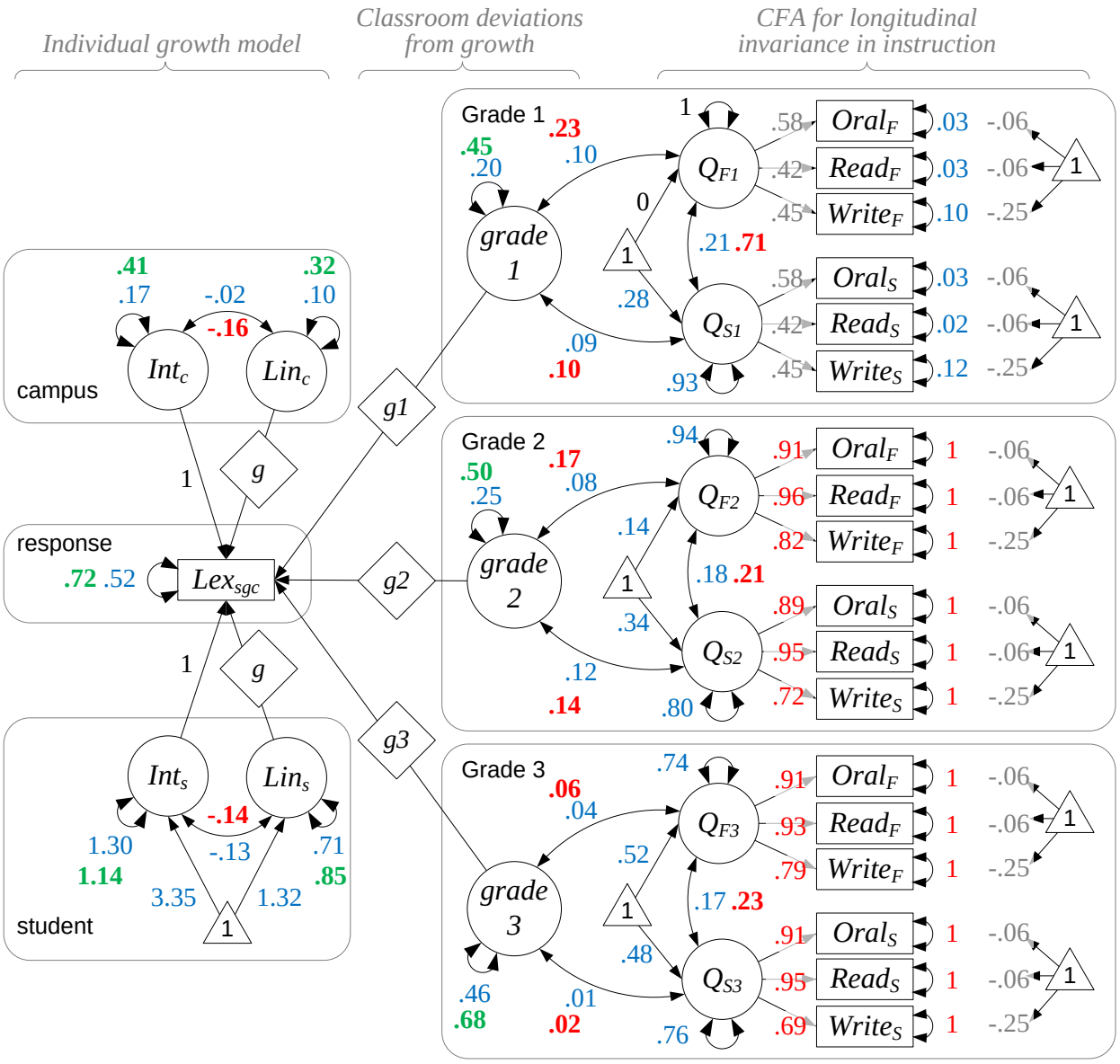
Smith, M. W., Dickinson, D. K., Sangeorge, A., & Anastasopoulos, L. (2002). *Early Language & Literacy Classroom Observation (ELLCO) Toolkit, Research Edition*. Baltimore, MD: Paul H. Brookes.

Wainer, H. (2011). Value-added models to evaluate teachers: A cry for help. *Chance, 24*(1), 11.

Appendix B. Tables and Figures

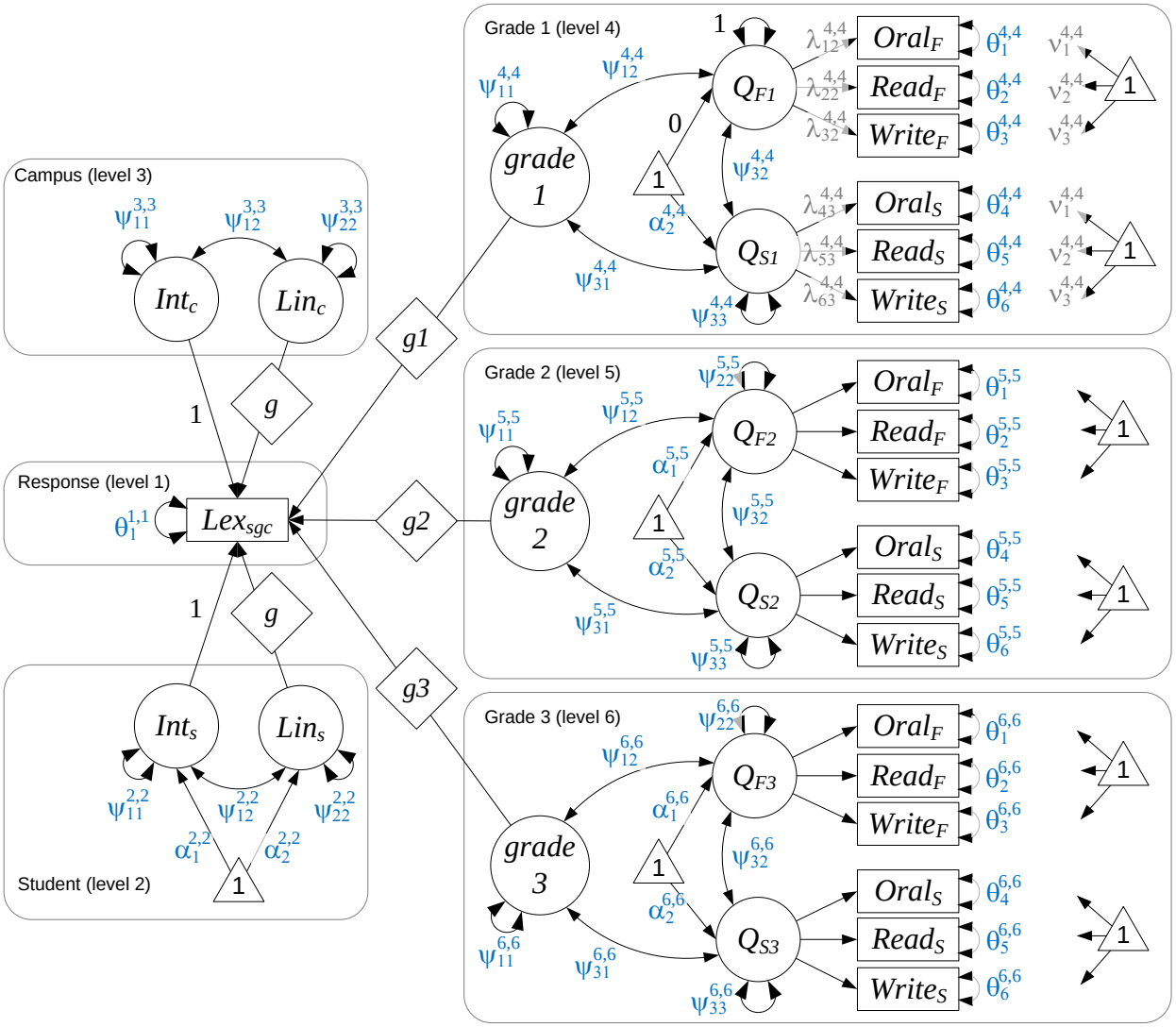
Not included in page count.

Figure 1. Substantive results of the 6-level SEM



Note: The left side of the model represents a linear growth model for students and schools. The right side represents fall and spring measures of instructional quality, with measurement held equivalent across semester and grades. Black values were fixed for model definition. Gray estimates were held equivalent over time. Blue values were freely estimated and represent unstandardized variances, covariances, and regression paths as in typical SEM. Red values show fully standardized estimates (i.e., correlations). Green values show SD for variance estimates.

Figure 2. Technical specification of parameters for xxM



Note: See Figure 1 for model estimates. Freely estimated parameters are shown in blue, fixed parameters in black, and parameters held equivalent over time are shown in gray (grade 1 only, redundant parameters omitted for grades 2-3). Each parameter is identified with a double-superscript indicating the levels it crosses (from, to), such that "2,2" represents a parameter solely within level 2. The subscripts for each parameter indicate the row and column of that matrix (latent variances and loadings), or the list number of that parameter (means, intercepts, and residual variances). The diamonds indicate individually specific values of time ("g" for grade, centered at grade 1), and dummy indicators for grade-specific classrooms. These are random slopes as is typical in mixed effects regression growth models (Mehta & Neale, 2005; Mehta & West, 2000).