Exploring the Impact of Cohort Variability on Teacher Effects

Daniel Anderson

Joseph Stevens

University of Oregon

Abstract

Value-added models (VAMs) attempt to isolate the contribution of teachers to students' achievement by conditioning current achievement on previous achievement. VAMs continue to gain popularity in large-scale accountability applications as one component of teacher evaluation systems. Generally, teacher effectiveness estimates are obtained from VAMs on an annual basis. Cohort variability, or year-to-year fluctuations in the student samples, therefore poses a threat to the validity of VAMs. The purpose of this paper was to explore the extent to which students' initial achievement and rate of growth during the school year varied as a function of the specific academic year. Overall, we found that students' within-year growth depended on *Academic Year*, independent of student demographics or classroom-level nesting. Implications for VAMs are discussed.

Exploring the Impact of Cohort Variability on Teacher Effects

Teachers impact on students' academic achievement is of considerable interest to a broad array of stakeholders, including parents, district administrators, and educational policy makers. Perhaps unsurprisingly, a wealth of previous research has attempted to isolate the "teacher effect" in the variance among student test scores (e.g., Chetty, Friedman, & Rockoff, 2014; Hanushek, Kain, O'Brien, & Rivkin, 2005; Heck, 2009; Kane, McCaffrey, Miller, & Staiger, 2013; Kane, Rockoff, & Staiger, 2008; Kane & Staiger, 2008; Koedel & Betts, 2007; Konstantopoulos & Chung, 2011; Nye, Konstantopoulos, & Hedges, 2004; Palardy & Rumberger, 2008; Rockoff, 2004). The statistical models used to estimate teacher effects, referred to generally as value-added models, or VAMs, vary considerably (see McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). In operational use, however, two models dominate: the student growth percentiles model (SGP; Betebenner, 2011) and the educational value-added assessment system (EVAAS; see Sanders & Horn, 1994; Wright, White, Sanders, & Rivers, 2010). Regardless of the specific model applied, VAMs provide estimates of teacher effectiveness based on the deviation of the mean classroom achievement from the overall sample mean achievement. "Achievement" is defined broadly, and can mean different things based on the specific study or model applied. Both the SGP and EVAAS models condition the current year's status (on the state test) on the previous years' status. The teacher effect is then defined as the average classroom-level deviation from the expected and observed achievement.

VAMs originated in the econometrics literature, with Hanushek (1971) applying models as early as the 1970s. The past decade, however, has seen an explosion in the quantity of research on teacher effects, with the past five years being particularly productive. At the time of this writing, a quick Google Scholar search revealed 192,000 records for the search terms “teacher effectiveness” since 2010. The enormous growth in research was prompted, at least in part, by federal legislation. In 2009, the United States Department of Education (USED) announced the Race to the Top (RTT) grant, a $4.35 billion dollar competitive grant that placed considerable emphasis on evaluating educator effectiveness (USED, 2009). The RTT grant was followed by a federal waiver, offering flexibility relative to the requirements of the *No Child Left Behind Act* (NCLB, 2002) “in exchange for rigorous and comprehensive State-developed plans designed to improve educational outcomes for all students” (USED, 2013, para 1). The state plans were required to “link teacher, principal, and student data and provide that information to educators to improve their practices”(USED, 2011, para 7). To date, 43 states, the District of Columbia and Puerto Rico have been approved for such a waiver (United States Department of Education, 2015), with many using VAMs as one component of educator evaluations.

It is important to note that VAMs can be used for many different purposes, including research purposes and, perhaps, as a formative diagnostic check for district personnel (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). Large-scale policy applications of VAMs, however, are in many ways restricted. For example, datasets are limited to include only state test scores, collected annually. Further, there is generally a desire to calculate estimates on an annual basis (Hull, 2013). This is problematic considering McCaffrey, Sass, Lockwood, and Mihaly (2009) found teacher effect estimates calculated during one year correlated with effects calculated the following year at only 0.2-0.5 for elementary school teachers, and 0.3-0.7 for middle school teachers. While some degree of year-to-year variability would be expected, these results suggest the effects may be highly unstable. Ballou (2005) found similar results, with roughly 30% of teachers estimated as being in the bottom quartile the first year moving to one of the top two quartiles the following year. The reverse was also true; approximately 25% of teachers who were in the top quartile during the first year fell to one of the bottom two quartiles in the second year. Both Ballou and McCaffrey et al. note that averaging performance over multiple years increased the stability of estimates.

Year-to-year variability in students' achievement may occur for multiple reasons outside of the specific teacher instructing the student, which may be part of the reason teacher effects have been found to be unstable. For example, district- or school-level policies, reform efforts, or leaders may have indirect effects on students' achievement beyond the individual teacher. Annual changes in the classroom or school student composition have also been shown to have a substantial impact on school effect estimates (Kane & Staiger, 2002; Linn & Haug, 2002), and the impact is likely to be greater at the classroom-level given the reduced sample size. Averaging effects across multiple years may result in more stable estimates in part because the estimate essentially becomes averaged over multiple years of external factors and samples of students. Classroom composition and interpersonal dynamics have also been shown to significantly relate to student achievement (i.e., peer effects; Hanushek, Kain, Markman, & Rivkin, 2003; Kang, 2007). Each school year, teachers are charged with instructing new groups of students with different dynamics and challenges, both academically and behaviorally. In some years, teachers may have to spend a substantial amount of time dedicated to classroom management issues, at the expense of academics, while in other years teachers may spend relatively little time on such issues.

We refer globally to the issue of year-to-year variability in students' achievement beyond the teacher effect as "cohort effects". This label is intended to be general, and not interpreted as purely "of the students" (i.e., cohort effects may be the result of a policy change). Because of the desire within large-scale policy applications to calculate estimates of teacher effects on an annual basis (see Hull, 2013), cohort effects cannot be accounted for, which introduces a potential source of bias. Many have questioned the validity of VAMs for producing unbiased, causal estimates of teacher effectiveness (American Statistical Association, 2014; Braun, 2005; McCaffrey, Lockwood, Mariano, & Setodji, 2005; Rothstein, 2010). Proponents argue that when prior achievement is used as a control variable in the model the classroom-level “intake” is essentially equated. The resulting models can then provide unbiased causal estimates of teachers’ impact on students’ achievement, it is argued, by evaluating student gains (Chetty et al., 2014). However, if peer effects, policy changes, or other factors influence student achievement, each annual cohort of students may exhibit different rates of growth during the school year. In this case, cohort variability would bias estimates of teacher effectiveness even if differences between classrooms in students’ initial achievement were statistically indistinguishable.

The purpose of this paper was to explore the extent to which differences in students' reading and math achievement were attributable to year-to-year changes in the sample of students beyond classroom-level nesting or student demographics. We fit within-year growth models to five years of data collected during the fall, winter, and spring of the school year in Grades 3-5 in reading and math in one large school district located in the southwestern United States. We explored the extent to which *Academic Year* (i.e., cohort) related to students’ initial achievement and rate of growth in each grade. We compared the fit of competing models representing three alternative hypotheses: (1) *Academic Year* does not relate to students’ achievement beyond classroom-level nesting and student demographic variables; (2) *Academic Year* relates to students’ initial achievement, but not their rate of growth; and (3) *Academic Year* relates to both students’ initial achievement and their subsequent rate of growth.

Method

Sample and Measures

The reading and mathematics portion of the Measures of Academic Progress (MAP) were used in this study. The MAP is an untimed computerized adaptive interim assessment that was administered seasonally (fall, winter, and spring), with students being presented different items conditional on their estimated ability level. Items are selected so the conditional probability of students correctly responding to items is approximately .5, maximizing information relative to the latent trait (Wang, McCall, Jiao, & Harris, 2013). The adaptive algorithm results in consistently higher test information and lower standard errors across a wide range of student abilities (NWEA, 2011). The tests include 50 multiple-choice items with 4 or 5 response options.

All items on the MAP were calibrated on a common, vertical scale, using a Rasch model (NWEA, 2011). Scores were reported on a transformed logit scale, called a Rasch unit, or RIT scale (RIT = $(\theta * 10) + 200)$. The developers report the equivalent of alternate form reliability for the MAP ranging from .705 to .914, while the equivalent of test-retest reliability ranged from .703 to 925 (NWEA, 2011). The marginal reliability (Samejima, 1977, 1994), which is interpreted similarly to coefficient alpha, ranged from .946 to .958. The bivariate correlation between MAP and state test performance when taken at approximately the same time (concurrent validity) ranged from 0.635 to 0.878 across states. The correlation between MAP taken at an earlier time and state test performance (predictive validity) ranged from 0.583 to 0.868 across states.

Five years of data were available in each of Grades 3-5, collected across the 2008-2009 to 2012-2013 school years. The total sample included ~7,000 students per grade, with approximately 1,350-1,400 students in each year. An outline of the data structure is displayed in Figure 1. Note that each grade was analyzed separately, and included five years of data. There was considerable overlap between grades in student cohorts. As displayed in the figure, three cohorts of students were represented in each analysis. Grades 3 and 4 and Grades 4 and 5 each had one additional cohort of students that overlapped - those who matriculated from Grades 3 to 4 from 2011-12 to 2012-13, and students who matriculated from Grades 4 to 5 from 2008-09 to 2009-10. Each of Grades 3 and 5 had one cohort of students who were not in common with any other analysis. In sum, Grades 3 and 4 and Grades 4 and 5 each shared 4 cohorts of students, while Grades 3 and 5 shared 3 cohorts of students.

Classrooms with fewer than 16 students at any point in the study were excluded to ensure adequate variance partitioning between student and classroom factors. These numbers are similar to those used in previous research (McCaffrey et al., 2009), and by states adopting teacher growth models for accountability purposes (American Institutes for Research, 2011). Means and standard deviations for each time point and cohort for reading and math are displayed in Table 1. A total of 131 teachers with at least 16 students were represented in Grade 3, with 111 in Grade 4, and 112 at Grade 5. The average class size was approximately 23 students in Grade 3 and 27 students at Grades 4 and 5.

Analyses

Our primary analyses were three-level linear growth models for each grade and subject across the seasonal time points. The assessment windows for each seasonal time point were quite broad (~6 weeks). We accounted for the variation by coding *Time* relative to the specific date the measures were administered. *Time* was coded in months, in decimal form, such that the coefficient could be interpreted as “monthly growth”, but accounted for the specific number of days between assessments. The intercept was centered on the first day of the school year for each cohort of students, and represented a backward projection of their initial achievement. An unconditional growth model was fit first as follows

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}Months_{tij} + e_{tij} \tag{1a}$$

$$\begin{aligned}\pi_{0ij} &= \beta_{00j} + r_{0ij}\\ \pi_{1ij} &= \beta_{10j} + r_{1ij}\end{aligned} \tag{1b}$$

$$\begin{aligned}\beta_{00j} &= \gamma_{000} + u_{00j}\\ \beta_{10j} &= \gamma_{100} + u_{10j}\end{aligned} \tag{1c}$$

where Y_{tij} represents the MAP assessment, either reading or math, at time *t* for student *i* nested in classroom j. The π_{0ij} and π_{1ij} terms represent the estimated intercept and slope for student *i*. The residual term, e_{tij}, represents variance not accounted for by the model and is assumed normally distributed with a mean of zero and variance σ^2. Equation 1a represents the within-student portion of the model (Level 1). Equations 1b and 1c represent the between-student and between-classrooms portion of the model (Levels 2 and 3, respectively). The β_{00j} and β_{10j} terms represent the average student-level initial achievement and growth slope, respectively, while r_{0ij} and r_{1ij} represent random student deviations around the average intercept and growth slope, respectively. Similarly, the γ_{000} and γ_{100} terms represent the average classroom initial achievement and growth, respectively, while the u_{00j} and u_{10j} terms represent individual classroom deviations around the averages. Both the student- and classroom-level random effects were assumed normally distributed with a mean of zero and an unstructured covariance matrix.

Following the unconditional model, a demographic-conditional model was estimated where student demographics were entered as fixed-effects predictors of students' initial achievement and rate of growth. Demographic variables included dummy-coded vectors of whether the student was male, actively enrolled in an English language learner program (ELL: Active), being monitored for their English language skills (ELL: Monitor), received special education services (SPED), were eligible for free or reduced price lunch (FRL), were of Hispanic/Latino ethnicity, or were of non-Hispanic/Latino and non-White ethnicity. The intercept thus represented White female students receiving general education instruction who were not enrolled in an ELL program and were not FRL eligible. Following the demographic-conditional model, *Academic Year* was entered as an effect-coded predictor of students' intercepts. Coefficients for specific cohorts of students therefore represented the difference between the average initial achievement for students within the specific academic year and the weighted grand mean (i.e., the mean of the group means). The weighted grand mean reasonably approximated the true grand mean for the reference group, given that the number of students within each academic year was roughly equivalent (see Table 1). Finally, the full model was estimated by including *Academic Year* as a predictor of students' rate of growth within the year.

We used multimodel inference when comparing competing models (Burnham & Anderson, 2004). Specifically, our model building strategy explicitly tested whether *Academic Year* contributed to the model as a predictor of students' initial achievement after controlling for student demographic variables and classroom level nesting (Hypothesis 1 versus Hypothesis 2). Following this model, we tested whether *Academic Year* contributed to the model as a predictor of students' rate of growth beyond the effect on the intercept or student demographic variables and classroom-level nesting (Hypothesis 2 versus Hypothesis 3). Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) were primarily used when selecting between competing models. We used general rules of thumb outlined by Burnham and Anderson to select a best-fitting model. Specifically, when the difference between competing models was less than two, we concluded there was little evidence to support one model over the other. Differences between four and seven indicated "considerably less support" (p. 271) for the model with the higher value, while differences greater than ten provided “essentially no support” (p. 271) for the model with the higher value. BIC tends to be the more conservative indicator (i.e., includes a greater penalty for the number of estimated parameters). We also used a χ^2 test to evaluate differences in the model deviance, which essentially became the moderator in the case of conflicting evidence between AIC/BIC.

Results

In the presentation of our results, we focus primarily on model selection. We briefly discuss the effect of all demographic variables, but focus primarily on the effect of including *Academic Year* in the model.

Model Selection

Fit indices for our competing models are displayed in Table 2 for math and in Table 3 for reading. The inclusion of student demographic variables universally resulted in a better fit to the data over the unconditional model, as indicated by AIC, BIC, and the χ^2 test of the change in the model deviance. Generally, AIC indicated that including *Academic Year* as a predictor of students’ intercepts increased model fit modestly (i.e., < 10). The exceptions were in Grade 5, where AIC stayed at the same estimated value in reading and decreased by 1 point in math. By contrast, the BIC increased with the inclusion of *Academic Year* as a predictor of the intercept across all models. The χ^2 test of the model deviance revealed similar evidence to AIC, with the deviance reducing significantly ($p < .05$) for all models outside of Grade 5 for both reading and math. Overall, the AIC decreased by 0 to 9 points, while BIC universally increased (range = 24 to 31 points). Because the χ^2 tests provided similar results to AIC, we concluded that the inclusion of *Academic Year* as a predictor of students’ intercepts moderately improved model fit after accounting for student demographics.

Following the evaluation of *Academic Year* as a predictor of students' intercepts, we estimated the full model, which included all student demographics and *Academic Year* as a predictor of students' intercepts and slopes. The AIC universally indicated that the full model displayed the best fit to the data, with all values being at least 10 points lower than any alternative model, with the exception of Grade 5 Reading, which was 8 points lower than the next best fitting model. The BIC was generally reduced, relative to the model with *Academic Year* as a predictor of students' intercepts only, with the exceptions being Grade 3, where the estimate stayed the same in math and decreased by 1 point in reading, and Grade 5 Reading, where the estimate increased by 24 points. All other models were at least 10 points lower than the model with *Academic Year* on the intercept along. However, the lowest BIC value remained the demographic-conditional model across all models except Grade 4 Math, for which the full model had the best fit to the data. The χ^2 tests of model deviance were universally significant, relative to the model with only *Academic Year* as a predictor of students' intercepts. Thus, while it seemed clear that the full model was a better fit to the data than the model with *Academic Year* as a predictor of intercepts only, there was conflicting evidence between AIC and BIC as to which model fit the data best overall, with AIC universally indicating the full model and BIC generally indicating the demographic-conditional model.

While not reported in Tables 2 and 3, additional χ^2 tests of the model deviance were conducted across all models in which the lowest AIC was estimated for the full model, but the lowest BIC value was estimated for the demographic-conditional model. Across all grades and models, χ^2 indicated the full model displayed the best fit to the data. Given the accumulation of evidence across model fit indices, we concluded that the full models displayed the best fit to the data across all grades and both subjects.

Parameter Estimates

Tables 4 and 5 display *Academic Year* parameter estimates for math and reading, respectively, as well as variance components for each of the selected models. To keep the tables of a manageable size, parameter estimates for demographic variables are not displayed[1], but rather are discussed briefly below. Tables 4 and 5 present only the final two models: *Academic Year* as a predictor of students' intercepts only, and *Academic Year* as a predictor of students' intercepts and slopes. When interpreting Tables 4 and 5, it is important to keep in mind the data structure (see Figure 1). For example, the coefficient for 2009 in Grade 3 represents largely the same group of students (outside of attrition) as the coefficient for 2010 in Grade 4.

Across all grades and subjects, students who were Active ELLs, FRL eligible, receiving SPED services, of Hispanic/Latino ethnicity, and/or of non-Hispanic/Latino and non-White ethnicity, had a significantly lower initial achievement than the reference group ($p < .05$). Male students were universally higher in initial math achievement, and lower in initial reading achievement (although the coefficient for reading at Grade 4 was not significant, $p = .08$). Students who had exited the district ELL program and were on monitor status had an initial achievement approximately 1 to 2 points higher in math and reading than the reference group, with the exception of Grade 5 Reading, which was not significantly different. Patterns for growth were less systematic. Active ELLs progressed significantly faster than the reference group in Reading in both Grades 4 and 5, while ELLs on monitor status progressed significantly faster in Grades 4 and 5 Math. FRL-eligible students progressed significantly faster in Grade 4 Reading, and significantly slower in Grade 5 math. Students receiving SPED services progressed significantly slower in Grades 4 and 5 Math. Male students progressed significantly faster in

[1] Please contact the lead author for complete parameter estimates.

Grade 3 Reading and Math, Grade 4 Math, and Grade 5 Reading. All other coefficients were not significantly different from the reference group.

In terms of cohort effects (see Tables 4 and 5), *Academic Year* was a sporadically significant predictor of students' intercepts, but was nearly a universally significant predictor of students' slopes in math across Grades 3-5 (with the coefficient for 2010 in Grade 4 and 2009 in Grade 5 being the exceptions). The coefficients all appeared relatively small; however, this was largely a function of the scale being monthly growth. For example, in Grade 4 the difference between the groups who displayed the lowest and highest growth within a given year (2009 and 2013) was 0.51 points per month, or roughly 4.9 points of growth over the course of the school year, independent of student demographics or classroom-level nesting. This difference corresponded to slightly more than 1/3 of a standard deviation on the Grade 4 spring measure when aggregated across cohorts. There were no discernable patterns to effects across grades within a given academic year.

In reading, the effect of *Academic Year* on the intercept was significant for between 1 and 3 student groups, depending on the specific model, while between 2 and 3 student groups progressed at significantly different rates. The results in reading were thus not as consistent as math, but students did still progress at different rates. For example, in Grade 3, the difference between students in 2009 and 2011 was 0.29 points of growth per month, or approximately 2.77 points over the course of the school year (0.22 standard deviations on the spring assessment).

Discussion

The purpose of this paper was to explore the extent to which students' initial achievement and rate of growth during the school year varied from one school year to the next, independent of classroom-level nesting or student demographics. Overall, models that included *Academic Year* as a predictor of students' intercepts generally displayed better fit to the data than models that included only student demographic variables. However, when *Academic Year* was entered as a predictor of students' growth within the school year, the model fit indices nearly universally improved. In math, all but two coefficients across grades for the effect of *Academic Year* on students' growth were significantly different than the weighted grand mean. In other words, each group of students, from one year to the next, progressed at significantly different rates, even after accounting for student demographics and the classroom in which the student was enrolled. In reading, the year-to-year variation in students' rate of growth was lower, with between two and three groups progressing at significantly different rates than the weighted grand mean.

There are many potential reasons that students' may progress at significantly different rates between school years, such as sampling variability (Kane & Staiger, 2002; Linn & Haug, 2002), changes in school- or district-wide leadership or policy (Heck & Hallinger, 2009), classroom dynamics (Hanushek et al., 2003), or changes in school- or district-wide reform efforts (e.g., changes to curriculum, implementation of response to intervention [RTI] and/or positive behavioral interventions and supports [PBIS], etc.). In our study, we did not find consistent effects across grades within a specific school year relative to growth. This lack of consistency implies that the differences in within-year growth were likely due primarily to sampling variability, rather than other factors. However, it is also possible that these external factors (e.g., leadership, reform efforts, etc.) had inconsistent effects across grades (e.g., large effects in Grade 3 versus small effects in Grade 5). In this case, the external factors may have influenced the observed differences in students' within-year growth between academic years, but attributing differences to specific factors would be difficult.

Regardless of the specific reasons that students within-year growth may differ between academic years, the results of this study suggest that students within a specific grade may progress at different rates depending on the specific school year. The variability in within-year growth rates likely impacts estimates of the value-added by teachers to students' achievement, and future research should examine this relation explicitly with more typical models used in high-stakes accountability policies. Controlling for classroom or school intake, however, as is the goal of many VAMs (Timmermans, Doolaard, & de Wolf, 2011), may be insufficient. If VAM estimates are calculated annually, then the estimated effects will depend, in part, upon the specific group of students within the specified grade during the specified year, given that within-year growth depends upon these specifics even after controlling for classroom-level nesting.

Limitations and Future Directions

This study utilized a large extant dataset, which inherently comes with certain limitations. For example, while we can speculate on possible reasons why students progressed at different rates during different academic years, we cannot state with any certainty why the differences occurred. While sampling variability appeared to be the primary mechanism, given the lack of consistent effects within an academic year across grades, it is likely that many unmodeled phenomena were also at work. School systems are complicated organizational units, with a myriad of factors potentially relating to students achievement (see Scheerens & Bosker, 1997). Yet, when viewed through the lens of VAMs and high-stakes accountability policies, it is perhaps less important to explain *why* students progress at different rates depending on the specific school, than it is to understand the potential repercussions on the accuracy of VAM estimates.

The model applied in this study is not generalizable to VAMs applied for high-stakes accountability purposes (e.g., SGP or EVAAS; Betebenner, 2011; Wright et al., 2010), because students growth within the academic school year was modeled, rather than between-year gains. Future research should explore how VAM estimates change when multiple cohorts of student gains are modeled simultaneously, with a cohort indicator, as opposed to being modeled annually. The former method would account for the year-to-year fluctuations in students' growth observed here, while the latter would not. If the VAM estimates were sufficiently similar, then annual estimates may be feasible. However, this seems unlikely given the instability of estimates observed between years in previous research (Ballou, 2005; McCaffrey et al., 2009). This method would also provide insight into the proportion of instability that is attributable to cohort effects, as opposed to other factors.

Conclusions

Attributing gains in student learning to teachers is a difficult proposition. Schools are not laboratory settings where all variables can be tightly controlled, with students randomly assigned to teachers and teachers randomly assigned to schools. Analysts employing VAMs attempt to isolate the effect of teachers on students' current year achievement by controlling for prior achievement. Yet, the results of this study suggest that even if prior achievement is controlled, students may progress at different rates during the school year depending on the specific academic year, independent of demographic variables and the classrooms in which students are enrolled. VAMs continue to grow in popularity for use within high-stakes statewide accountability systems. The practical repercussions of these models are substantial. It is critical that VAMs continue to be vetted so that a better understanding of what they can and cannot measure becomes understood. It is equally important that the results of these investigations be shared with educators and policy-makers who are using the models. If the limitations of the models are better understood then the validity of model-based inferences is likely to be increased.

References

American Institutes for Research. (2011). *2010-11 Beta Growth Model for Educator Evaluation Technical Report*: New York State Education Department.

American Statistical Association. (2014). ASA statement on using value-added models for educational assessment. Retrieved 2015, March 4, from https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Ballou, D. (2005). Value-added assessment: Lessons from Tennessee. In R. Lissetz (Ed.), *Value added models in education: Theory and applications* (pp. 272-303). Maple Grove, MN: JAM Press.

Betebenner, D. W. (2011). A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth trajectories/projections. The National Center for the Improvement of Educational Assessment. Retrieved March 4, 2015, from http://www.nj.gov/education/njsmart/performance/SGP_Technical_Overview.pdf

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods Research, 33*, 261-304. doi: 10.1177/0049124104268644

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *The American Economic Review, 104*, 2593-2632.

Hanushek, E. A. (1971). Teacher charactheristics and gains in student achievment: Estimation using micro data. *The American Economic Review, 61*, 280-288.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics, 18*, 527-544. doi: 10.1002/jae.741

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). The market for teacher quality. Working Paper 11154.

Heck, R. H. (2009). Teacher effectiveness and student achievement: Investigating a multilevel cross-classified model. *Journal of Educational Administration, 47*, 227-249. doi: 10.1108/09578230910941066

Heck, R. H., & Hallinger, P. (2009). Assessing the contribution of distributed leadership to school improvement and growth in math achievement. *American Educational Research Journal, 46*, 626-658. doi: 10.3102/0002831209340042

Hull, J. (2013). Trends in teacher evaluation: How states are measuring teacher performance. Retrieved July 29, 2014, from http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/Trends-in-Teacher-Evaluation-At-A-Glance/Trends-in-Teacher-Evaluation-Full-Report-PDF.pdf

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* MET Project Research Paper, Bill & Melinda Gates Foundation.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*, 615-631. doi: 10.1016/j.econedurev.2007.05.005

Kane, T. J., & Staiger, D. O. (2002). Volatility in school test scores: Implications for test-based accountability systems. *Brookings Papers on Education Policy, 5*, 235-283.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Working Paper 14607). http://www.nber.org/papers/w14607: National Bureau of Economic Research.

Kang, C. (2007). Classroom peer effects and academic achievement: Quasi-randomization evidence from South Korea. *Journal of Urban Economics, 61*, 458-495. doi: 10.1016/j.jue.2006.07.006

Koedel, C., & Betts, J. R. (2007). Re-examining the role of teacher quality in the educational production function. Working Paper No. 708, University of Missori-Columbia.

Konstantopoulos, S., & Chung, V. (2011). The persistence of teacher effects in elementary grades. *American Educational Research Journal, 48*, 361-386. doi: 10.3102/0002831210382888

Linn, R. L., & Haug, C. (2002). Stability of school building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24*, 29-36. doi: 10.3102/01623737024001029

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*, 67-101. doi: 10.3102/10769986029001067

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability. Santa Monica, CA: RAND Corporation.

McCaffrey, D. F., Lockwood, J. R., Mariano, L. T., & Setodji, C. (2005). Challenges for value-added assessment of teacher effects. In R. Lissitz (Ed.), *Value Added Models in Education: Theory and Application*. Maple Grove, Minnesota: JAM Press.

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. doi: 10.1162/edfp.2009.4.4.572

No Child Left Behind (NCLB) Act of 2001. Pub. L. No. 107-10 § 115, Stat. 1425 (2002).

Northwest Evaluation Association. (2011). Technical Manual For Measures of Academic Progress (MAP) and Measures of Academic Progress fro Primary Grades (MPG). Portland, OR.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237-257. doi: 10.3102/01623737026003237

Palardy, G. J., & Rumberger, R. W. (2008). Teacher effectiveness in first grade: The importance of background qualifications, attitudes, and instructional practices for student learning. *Educational Evaluation and Policy Analysis, 30*, 111-140. doi: 10.3102/0162373708317680

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*, 247-252.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics, 125*, 175-214. doi: 10.1162/qjec.2010.125.1.175

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*, 233-247.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18*, 229-244.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299-311. doi: 10.1007/BF00973726

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.

Timmermans, A. C., Doolaard, S., & de Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. *School Effectiveness and School Improvement, 22*, 393-413. doi: 10.1080/09243453.2011.590704

United States Department of Education. (2009). President Obama, U.S. secretary of education Duncan announce national competition to advance school reform. Retrieved March 4, 2015, from http://www2.ed.gov/news/pressreleases/2009/07/07242009.html

United States Department of Education. (2011). Letters from the Education Secretary or Deputy Secretary. Retrieved February 16, 2014, from http://www2.ed.gov/policy/gen/guid/secletter/110923.html

United States Department of Education. (2013). Elementary & Secondary Education: ESEA Flexibility. Retrieved February 16, 2014, from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

United States Department of Education. (2015). ESEA Flexibility. Retrieved March 19, 2015, from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

Wang, S., McCall, M., Jiao, H., & Harris, G. (2013). Construct validity and measurement invariance of computerized adaptive testing: Application to Measures of Academic Progress (MAP) using confirmatory factor analysis. *Journal of Educational and Developmental Psychology, 3*, 88-100. doi: 10.5539/jedp.v3n1p88

Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010). SAS EVAAS Statistical Models. Retrieved March 19, 2015, from http://www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf

Table 1

Means and Standard Deviations

Time point	Cohort									
	2008-09		2009-10		2010-11		2011-12		2012-13	
	(*n* = 1,433)		(*n* = 1,469)		(*n* = 1,331)		(*n* = 1,323)		(*n* = 1,354)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Math										
Grade 3										
Fall	189.03	11.26	189.58	11.23	189.47	11.14	189.49	11.53	189.55	11.27
Winter	194.71	11.03	195.97	11.2	195.62	11.41	196.42	11.49	197.38	11.16
Spring	201.46	11.79	202.77	11.53	202.9	12.09	203.93	12.17	204.94	11.48
Grade 4										
Fall	200.55	11.80	201.05	12.47	200.62	12.07	201.16	11.82	200.24	11.79
Winter	204.77	12.38	205.36	12.72	205.65	12.73	206.73	12.81	204.92	11.94
Spring	210.32	13.40	212.33	13.65	212.44	13.68	215.36	13.9	209.84	13.27
Grade 5										
Fall	209.87	12.92	210.05	13.25	209.82	13.35	209.43	12.78	209.51	13.28
Winter	215.13	14.18	215.76	13.79	216.78	14.72	214.54	13.59	213.29	13.82
Spring	221.52	15.39	222.01	15.36	222.76	15.26	220.38	15.06	219.02	14.67
Reading										
Grade 3										
Fall	187.46	14.15	187.42	14.35	187.67	13.33	187.74	14.65	186.42	15.13
Winter	192.84	13.47	193.62	14.2	193.48	13.25	193.96	14.17	194.39	13.64
Spring	197.13	13.57	199.14	13.61	198.73	13.12	198.99	14.11	199.41	13.57
Grade 4										
Fall	197.03	13.98	197.63	14.55	196.87	14.78	196.85	14.5	197.74	14.4
Winter	201.62	13.6	201.43	13.82	201.12	13.66	201.94	13.7	201.05	13.72
Spring	205.03	14.22	206.04	13.54	204.96	14.18	207.01	13.64	204.97	13.78
Grade 5										
Fall	204.99	13.58	205.46	14.25	203.07	14.12	204.39	14.28	204.14	14.03
Winter	209.00	13.38	208.81	13.9	208.11	13.29	208.49	13.71	207.63	13.71
Spring	212.86	13.43	212.51	13.99	211.7	13.16	211.99	13.84	211.38	13.24

Table 2

Model Fit Indices: Math

Grade	Model	*DF*	AIC	BIC	Deviance	χ^2	*p*
3	Unconditional	9	126025	126095	126007		
	Demos	23	122035	**122215**	121989	4017.80	0.00
	Intercept on cohort	27	122028	122239	121974	15.02	0.00
	Full	31	**121997**	122239	**121935**	39.23	0.00
4	Unconditional	9	128423	128494	128405		
	Demos	23	122476	122656	122430	5975.00	0.00
	Intercept on cohort	27	122472	122682	122418	12.36	0.01
	Full	31	**122327**	**122569**	**122265**	152.80	0.00
5	Unconditional	9	128887	128958	128869		
	Demos	23	123721	**123900**	123675	5194.00	0.00
	Intercept on cohort	27	123720	123930	123666	9.19	0.06
	Full	31	**123673**	123915	**123611**	54.68	0.00

Note. Values for the χ^2 test relate to the difference between the model deviance between the corresponding model and the preceding model (i.e., the model one row above).

Table 3

Model Fit Indices: Reading

Grade	Model	*DF*	AIC	BIC	Deviance	χ^2	*p*
3	Unconditional	9	134701	134771	134683		
	Demos	23	130026	**130205**	129980	4703.00	0.00
	Intercept on cohort	27	130019	130230	129965	14.37	0.01
	Full	31	**129987**	130229	**129925**	40.12	0.00
4	Unconditional	9	134526	134596	134508		
	Demos	23	127598	**127777**	127552	6956.20	0.00
	Intercept on cohort	27	127589	127800	127535	16.25	0.00
	Full	31	**127545**	127787	**127483**	52.28	0.00
5	Unconditional	9	131789	131859	131771		
	Demos	23	125702	**125881**	125656	6114.90	0.00
	Intercept on cohort	27	125702	125912	125648	8.15	0.09
	Full	31	**125694**	125936	**125632**	15.74	0.00

Note. Values for the χ^2 test relate to the difference between the model deviance between the corresponding model and the preceding model (i.e., the model one row above).

Table 4

Multilevel Growth Model Results, Cohort Parameter Estimates: Math

Fixed effects / year	Grade 3				Grade 4				Grade 5			
	Intercept Only		Intercept/Slope		Intercept Only		Intercept/Slope		Intercept Only		Intercept/Slope	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercept, γ_{000}	194.68*	0.39	194.71*	0.39	205.64*	0.44	205.66*	0.44	215.05*	0.45	215.05*	0.45
2009, γ_{010}	-0.70*	0.25	-0.46	0.27	-0.79*	0.27	-0.25	0.28	-0.33	0.30	-0.37	0.31
2010, γ_{020}	0.37	0.25	0.72*	0.27	0.65*	0.27	0.74*	0.28	0.37	0.31	0.26	0.32
2011, γ_{030}	0.68*	0.27	0.28	0.29	-0.17	0.28	0.18	0.29	-0.25	0.32	0.00	0.32
2012, γ_{040}	-0.40	0.25	-0.21	0.27	-0.02	0.29	-0.26	0.29	0.80*	0.33	0.62	0.33
2013, γ_{050}	0.05	0.26	-0.31	0.27	0.32	0.31	-0.41	0.31	-0.59	0.30	-0.51	0.31
Slope, β_{100}	1.69*	0.04	1.68*	0.04	1.44*	0.05	1.44*	0.04	1.52*	0.04	1.53*	0.04
2009, γ_{110}			-0.07*	0.03			-0.21*	0.03			0.02	0.03
2010, γ_{120}			-0.09*	0.02			-0.04	0.03			0.08*	0.03
2011, γ_{130}			0.12*	0.03			-0.15*	0.03			-0.18*	0.03
2012, γ_{140}			-0.05*	0.02			0.09*	0.03			0.15*	0.03
2013, γ_{150}			0.09*	0.02			0.30*	0.03			-0.07*	0.03
Variance components	Var	*SD*	Var	*SD*	Var	*SD*	Var	*SD*	Var	*SD*	Var	*SD*
Stu int, r_{0ij}	74.00	8.60	74.02	8.60	77.04	8.78	76.98	8.77	98.08	9.90	98.10	9.90
Stu slope, r_{1ij}	0.13	0.36	0.13	0.35	0.16	0.40	0.14	0.38	0.15	0.39	0.14	0.38
Class int, u_{00j}	4.23	2.06	4.03	2.01	7.22	2.69	6.81	2.61	5.80	2.41	5.79	2.41
Class slope, u_{10j}	0.08	0.29	0.08	0.28	0.12	0.34	0.09	0.31	0.10	0.31	0.09	0.31
Residual, e_{tij}	18.09	4.25	18.08	4.25	19.67	4.43	19.67	4.43	21.09	4.59	21.09	4.59

Note. Only the coefficients for *Cohort* are presented. All models also included demographic controls. See Methods section.

**p* < .05

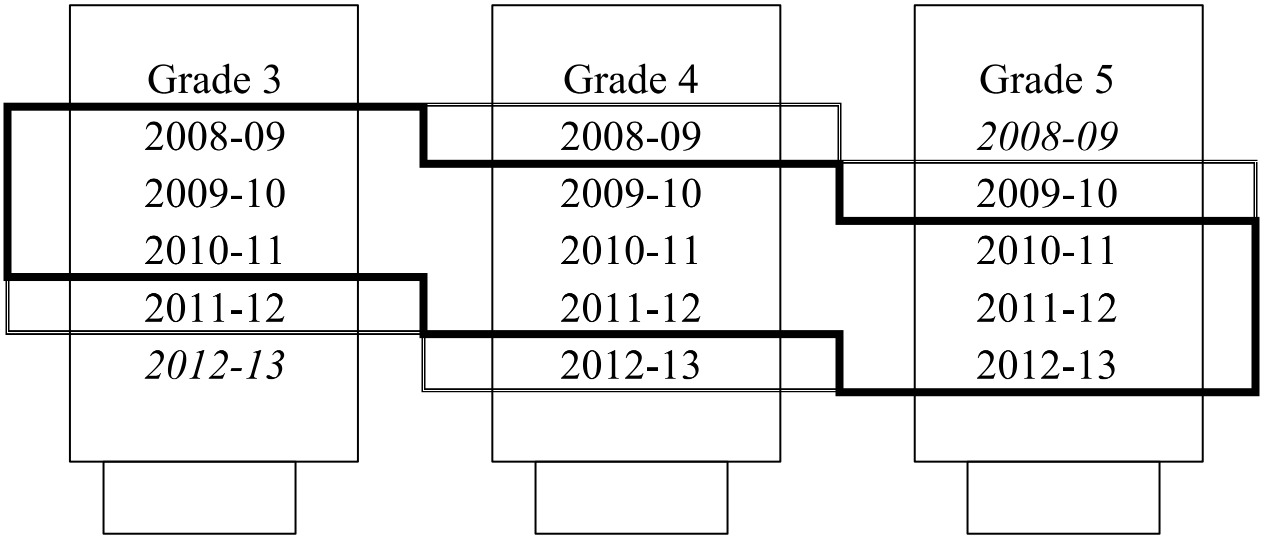
Table 5

Multilevel Growth Model Results, Cohort Parameter Estimates: Reading

Fixed effects / year	Grade 3				Grade 4				Grade 5			
	Intercept Only		Intercept/Slope		Intercept Only		Intercept/Slope		Intercept Only		Intercept/Slope	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Intercept, γ_{000}	196.89*	0.47	196.91*	0.46	206.34*	0.47	206.29*	0.46	213.18*	0.44	213.18*	0.44
2009, γ_{010}	-0.29	0.29	0.47	0.33	-0.39	0.28	-0.12	0.31	0.23	0.28	0.18	0.31
2010, γ_{020}	1.05*	0.28	0.80*	0.33	0.86*	0.28	0.96*	0.32	0.53	0.29	0.91*	0.32
2011, γ_{030}	-0.20	0.31	-0.97*	0.35	0.50	0.29	1.02*	0.33	0.11	0.29	0.22	0.32
2012, γ_{040}	-0.32	0.29	-0.13	0.33	-0.62*	0.30	-0.55	0.33	-0.71*	0.31	-1.17*	0.33
2013, γ_{050}	-0.24	0.29	-0.17	0.34	-0.35	0.32	-1.31*	0.35	-0.17	0.28	-0.13	0.31
Slope, β_{100}	1.36*	0.04	1.35*	0.04	0.93*	0.04	0.93*	0.04	0.85*	0.04	0.85*	0.04
2009, γ_{110}			-0.14*	0.03			-0.06*	0.03			0.01	0.03
2010, γ_{120}			0.04	0.03			-0.02	0.03			-0.07*	0.03
2011, γ_{130}			0.15*	0.03			-0.12*	0.03			-0.03	0.03
2012, γ_{140}			-0.04	0.03			-0.01	0.03			0.10*	0.03
2013, γ_{150}			-0.01	0.03			0.21*	0.03			-0.01	0.03
Variance components	Var	*SD*	Var	*SD*	Var	*SD*	Var	*SD*	Var	*SD*	Var	*SD*
Stu int, r_{0ij}	114.88	10.72	114.81	10.71	101.31	10.07	101.19	10.06	99.68	9.98	99.64	9.98
Stu slope, r_{1ij}	0.12	0.35	0.12	0.34	0.10	0.31	0.09	0.30	0.08	0.28	0.08	0.28
Class int, u_{00j}	5.04	2.24	4.57	2.14	5.65	2.38	5.57	2.36	4.12	2.03	4.08	2.02
Class slope, u_{10j}	0.04	0.19	0.03	0.18	0.04	0.21	0.04	0.21	0.04	0.20	0.04	0.20
Residual, e_{tij}	31.01	5.57	31.02	5.57	30.40	5.51	30.41	5.51	28.41	5.33	28.41	5.33

Note. Only the coefficients for *Cohort* (year) are presented. All models also included demographic controls. See Methods section.

**p* < .05



Grade 3	Grade 4	Grade 5
2008-09	2008-09	*2008-09*
2009-10	2009-10	2009-10
2010-11	2010-11	2010-11
2011-12	2011-12	2011-12
2012-13	2012-13	2012-13

Figure 1. Outline of Cohorts in each analysis. Each grade was analyzed separately, with five academic years in each analysis. Three cohorts of students were represented in each analysis (displayed with thick black border). Grades 3 and 4 and Grades 4 and 5 each shared one additional cohort (displayed with small double black line borders). Grades 3 and 5 each had one group of students who were not represented in any other analysis (displayed in italic font).