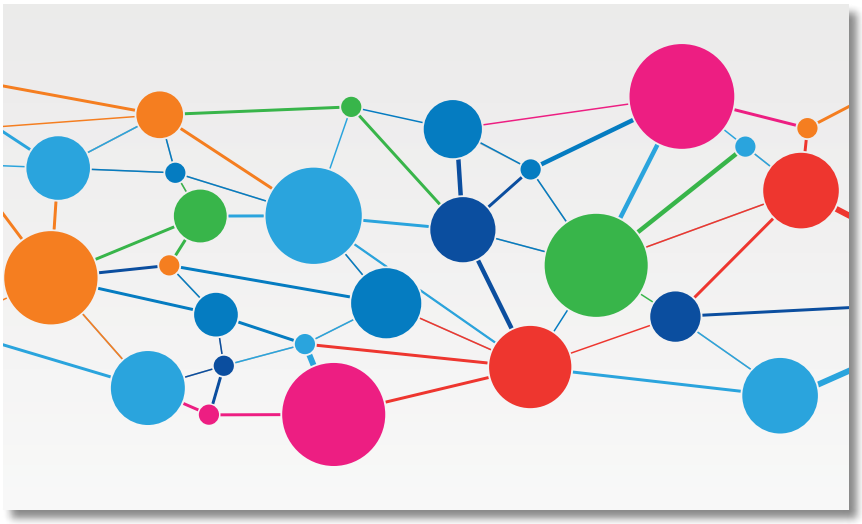
ACT Research Report Series 2015 (5)

Statistical Properties of School Value-Added Scores Based on Assessments of College Readiness



Dina Bassiri, PhD

ACT®

Dina Bassiri is a senior research scientist in the Statistical and Applied Research Department, specializing in educational outcomes research and student growth models.

Acknowledgments

Thanks to Joann Moore for her assistance in preparing the data for this study and to Jeff Allen, Jim Sconing, Richard Sawyer, Chrys Dougherty, Greg Bales, and Karen Zimmerman for their reviews and helpful comments and suggestions on an earlier draft of this report.

Contents

Abstract

The 2001 reauthorization of the Elementary and Secondary Education Act (ESEA) known as No Child Left Behind and more recent federal initiatives such as Race to the Top and the ESEA flexibility waiver have brought student growth to the forefront of education reform for assessing school and teacher effectiveness. This study examined growth projections and school value-added scores that are based on the ACT college and career readiness system for different growth periods (grades 8–10, 9–10, 10–11, 10–12, and 8–12). For this report, I investigated methodological questions related to value-added modeling. The analyses are based on large samples of students who took two or more of the tests in the ACT college and career readiness system in grades 8, 9, 10, and 11/12, respectively. For each growth period, I examined high schools' effects on student achievement in English, mathematics, reading, and science by controlling for prior test scores and varying time between tests.

Across the various growth periods, I found differences in the variance of school effects (value-added scores), as well as differences among types of schools in the proportion of their students classified as significantly above or below average. Most school effects were not significantly different from the average school effect and could not usually be distinguished from average with confidence. This was more often the case for small schools than for large schools.

I found evidence of consistency in value-added scores by examining same-year and same-cohort correlations for different growth periods and by examining correlations of value-added scores for adjacent cohorts (one, two, and three years apart) at the same schools. Value-added scores based on grades 8–12 data show greater reliability over time. The grades 8–12 data also generated value-added scores that perform better than scores based on shorter growth periods in terms of differentiating schools and achieving statistical significance with smaller sample size.

I also found positive associations between value-added scores and mean prior academic achievement and an inverse relationship with school poverty level and the proportion of racial/ethnic minority students. Generally, compared to other school characteristics, grade level enrollment and the proportion of students tested had weaker associations with value-added measures. Overall, the relationships of school characteristics and value-added scores suggest that different types of schools would be expected to have different mean value-added scores.

Introduction

The No Child Left Behind reauthorization of the Elementary and Secondary Education Act (ESEA) in 2001 (NCLB, 2002) brought great attention to the monitoring of educational progress and sparked questions on how best to measure the effectiveness of schools and teachers. More recent initiatives, including Race to the Top and the ESEA flexibility guidelines, continue to spur development of methods to accurately and fairly monitor school and teacher performance. Under NCLB, schools are accountable for increasing the percentage of students meeting proficiency cutoffs. Status measures, such as proficiency rates, use test scores from a single point in time, so they cannot directly measure how much schools or teachers affect learning throughout the year. There is considerable agreement among educators and researchers that status measures do not measure educator effectiveness well. Growth models, on the other hand, are less affected by the influence of student background and prior academic achievements. This is particularly important with non-random assignment of students to schools. As pointed out by Ballou, Sanders, and Wright (2004), the non-random distribution of student characteristics results in inequalities in the composition of student bodies from one school to another that may be mistakenly construed as disparities in school effectiveness. Furthermore, research indicates that in addition to effects associated with students' social class and prior achievement, aggregate school socioeconomic status has a significant effect on student outcomes (Willms, 1986). Research has also shown that academically related psychosocial behaviors such as motivation, social connectedness, school attendance, obedience of rules, and avoidance of drugs are important predictors of academic success in middle school and high school (Kaufman & Bradbury, 1992; Rumberger, 1995; Worrell & Hale, 2001; Jones & Byrnes, 2006).

Generally, growth models are used to measure how much growth occurred over a period of time for a student or group of students, project how much growth will occur, or indicate how much growth needs to occur to meet a certain achievement level (also known as a growth-to-standards model). Relative to status measures, measures of student growth are less vulnerable to the influence of student backgrounds (Riddle, 2008).

Projection and Value-Added Models

Projection and value-added models (VAMs) are types of growth models. As the name suggests, VAMs attempt to determine the value that is added to student learning by a teacher or school (i.e., producing more growth than expected, given specific characteristics of the student or school or teacher). Braun, Chudowsky, and Koenig (2010) refer to VAMs as "a variety of sophisticated statistical techniques that use one or more years of prior student test scores, as well as other data, to adjust for preexisting differences among students when calculating contributions to student test performance" (p. 1). According to Harris (2009), "the term is used to describe analyses using longitudinal student-level test score data to study the educational input-output relationship, including especially the effects of individual teachers (and schools) on student achievement" (p. 321).

Unlike some applications of growth modeling, VAMs are often specified with the intent of estimating the causal effects of schools (or teachers) on students (Briggs, 2011). VAMs purport to quantify

the effect on student achievement that, on average, a school or a teacher would have on similar students.[1]

Projection models are methods for forecasting (predicting) future test scores based on current and/or prior test scores and other variables. Typically, projection models are implemented by regressing later test scores on prior test scores (and other variables) for one cohort, and then applying the regression parameters to the current cohort to obtain projected scores.[2] VAMs and projection models can work hand-in-hand. VAMs can be based on *residual* scores, defined as the difference between actual and projected scores. For example, a school value-added score can be obtained by averaging the residual scores for students who attended the school for the growth period spanned by the tests.

School districts and state education agencies across the country are increasingly relying on VAMs to measure school and teacher performance, sometimes with high stakes attached. Using value-added measures to inform high-stakes decisions is certainly controversial, and there is not currently a consensus in the research community on the use of value-added measures for evaluation and decision making. Some of the disagreement is rooted in technical aspects and statistical properties of VAMs and their use in accountability (Harris, 2009; Braun et al., 2010). In addition, the American Statistical Association (2014) has issued an official statement on the use of VAMs. It urged states and school districts to exercise caution in the use of VAM scores for high stakes purposes and offered reasonable guidelines for practice. Some have expressed skepticism (Amrein-Beardsley, 2014; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; National Research Council, 2010), and others, including prominent foundations and some think tanks, have been more positive (Bill & Melinda Gates Foundation, 2012; Glazerman, Goldhaber, Loeb, Raudenbush, Staiger, & Whitehurst, 2010; Gordon, Kane, & Staiger, 2006; Hanushek & Rivkin, 2004). It appears that teachers and principals trust classroom observations more than VAMs. This conclusion emerges from both Goldring, Grissom, Rubin, Neumerski, Cannata, Drake, and Schuermann's (2015) analysis across eight districts and an in-depth case study of Chicago (Jiang, Sporte, & Luppescu, 2015).

The general consensus is that a set of VAM scores does contain some useful information that meaningfully differentiates among teachers, especially in the tails of the distribution. However, individual VAM scores do suffer from high variance and low year-to-year stability as well as an undetermined amount of bias (Goldhaber & Hansen, 2013; Kane, McCaffrey, Miller, & Staiger, 2013; McCaffrey, Sass, Lockwood, & Mihaly, 2009), leading some to suggest that it is not reliable enough to be used for high-stakes purposes (Darling-Hammond et al., 2012). Consequently, if VAM scores are to be used for evaluation, they should not be given inordinate weight and certainly not treated as the "gold standard" to which all other indicators must be compared (Braun, 2015).

In view of the importance of VAMs, a special issue of the *Journal of Educational and Behavioral Statistics* (Wainer, 2004) was devoted to the careful examination of this methodology; the issue

[1] Value-added analyses return norm-referenced effectiveness quantities indicating whether a teacher or school is significantly more or less effective than a group's average (e.g., district or state). Note that the effects are centered about zero (or in the case of the Student Growth Percentile, around the 50th percentile), indicating average effectiveness. Like all norm-referenced quantities, value-added effects do not implicitly include criteria stipulating what is enough effectiveness. The interest in incorporating growth measures into criterion-referenced assessment led to the development of criterion-referenced growth-to-standard models, stipulating adequate progress toward desired achievement outcomes (Betebenner, 2009).

[2] In this study, projected scores are not used to predict the score of different students in the future.

focused almost entirely on the statistical properties of the measures. Recently, a special issue of *Educational Researcher* (Harris & Herrington, 2015) was devoted to the careful examination of value-added–based teacher accountability. It focused on the effects on teaching and learning that come from embedding VAMs into policies like teacher evaluation, tenure, and compensation.

Value-added–based teacher accountability is likely to be different from school-based accountability. Although both use students test scores, there is much wider acceptance of school-level measures generally, and they have better statistical properties.

Research Questions

In this study, I address questions that may be of interest to users of the ACT college and career readiness system of assessments, but also to nonusers wishing to evaluate the utility of growth models based on college readiness assessments. The system included ACT Explore® (for grades 8 and 9), ACT Plan® (for grade 10), and the ACT® test (for grades 11 and 12).[3] I examine projection and value-added models for different growth periods, defined by student grade level at the two times of testing (8–10, 9–10, 10–11, 10–12, and 8–12).[4] I sought to answer the following questions pertaining to inferences about the growth of individual students (Question 1) and about the effectiveness of schools (Questions 2–6) across six different growth periods. I also evaluated and compared three types of growth models based on the accuracy with which they predict individual students' test scores in subsequent grades.[5]

1. How strongly related are students' projected scores to their actual scores?
   - Does the strength of relationship vary for different projection models?
   - Is measurement error in ACT college and career readiness system scores properly addressed by using multiple test scores?
   - Is the strength of relationship enhanced by using test scores from multiple prior years?
2. How often are school value-added scores classified as being significantly above or below average?
3. What is the variability across school cohorts in the proportion of students meeting or exceeding expected growth?
4. What school characteristics are related to school value-added scores?
5. To what extent do value-added scores vary across high- and low-poverty schools?
6. How consistent are a school's value-added scores?
   - across growth periods for the same school year?
   - across growth periods for the same cohort of students?
   - across cohorts of students (for each growth period)?

[3] The ACT Explore and ACT Plan assessments are only available to existing customers through 2016 and will be replaced by the ACT Aspire™ system afterward.

[4] For student grade levels 9–10, two different growth periods (10–14 and 9–18 months) are considered.

[5] The extent to which the assumptions of these models are met will affect answers to these questions, but in this study I do not investigate how well these assumptions are met or fit the data. Castellano and Ho (2015) investigated assumptions in different types of growth models, including the "*meanResid*" model (referred to as Std model in this study) by classifying them by aggregation functions (mean or median), regression approach (linear mean and nonlinear quantile), and the scales that support interpretations (percentile rank and score scale), among other factors using both simulated and empirical data.

Methodology

This study used longitudinal data from a large sample of students who took ACT Explore in eighth or ninth grade, ACT Plan in tenth grade, and the ACT in eleventh or twelfth grade to investigate certain statistical properties of school value-added scores. Clearly, to be relevant for studying school value-added scores, tests should measure the knowledge and skills specified in schools' curricula. All three components of the ACT college and career readiness system measure academic achievement in four subject areas (English, mathematics, reading, and science), with respect to the curriculum of the grade level for which it is intended.[6] The tests use a common score scale, making it easier for users to monitor progress (students' strengths and weaknesses).

Data collected as part of the assessment programs were used to create longitudinal datasets for different growth periods: grades 8–10, 9–10, 10–11, 10–11/12 (10–12 for brevity), and 8–11/12 (8–12 for brevity). Using value-added models, I examine high schools' effects on student achievement, controlling for prior academic achievement level as measured by the prior scores in all four subject areas. I also include a covariate in the models that accounts for varying lengths of time between Test 1 and Test 2. By using the number of months between tests as a covariate, the projected scores accommodate varying time spans between tests, as more growth is expected as more time passes (Allen, Bassiri, & Noble, 2009). I then fit this model to each of the four subject areas to obtain the estimated school effects on students' academic performance in English, mathematics, reading, and science.

Models and Growth Projection Accuracy

There are several options for modeling school (or teacher) performance. In a study comparing growth models, Goldschmidt, Choi, and Beaudoin (2012) classify growth models into five categories (categorical, gain score, regression, value-added, and normative). In another study, Ehlert, Koedel, Parsons, and Podgursky (2012) classify growth models into three broad classes (Student Growth Percentiles, one-step VAMs, and two-step VAMs). According to Cody, McFarland, Moore, and Preston (2010), most growth models tend to fall into three general categories of simple growth, growth-to-standard, and projection models.

In this study, I consider three types of growth models: a Standard VAM, a Standard-Classified VAM, and the Quantile Regression model. I evaluate these models based on the accuracy with which they predict test scores at Time 2. As we will see later, due to small differences observed between the standard VAM and the other two models with respect to the growth projection accuracy, I will then adopt the standard VAM model and base the remaining analyses only on this model. I briefly describe each of the models studied in the next few sections.

Standard VAM (Std). I modeled Test 2 (ACT Plan or the ACT test) scores for each student using as predictors the student's Test 1 (ACT Explore or ACT Plan) scores in all four subject areas and the number of months that passed between the two tests. To account for unobserved high school effects, a random intercept model was used. The random intercept model consists of fixed effects

[6] Every three to four years, ACT conducts the ACT National Curriculum Survey®, in which more than 20,000 educators nationwide in grades 7–14 are asked to identify the knowledge and skills that are important for students to know to be ready for college-level work. ACT also reviews textbooks on state-approved lists for courses in grades 7–12 and examines objectives for instruction for all states that have published such objectives. ACT then analyzes the information to refine the scope and sequence for each subject area test (ACT, 2011a; ACT, 2011b; ACT, 2007).

(the intercept and regression coefficients associated with the predictors) and a random effect that adjusts the intercept for individual schools. The model for projected scores is given in Equation 1:

$$Y_{ij} = \beta_0 + \sum_{p=1}^{4} \beta_p X_{ijp} + \beta_5 T_{ij} + U_j + R_{ij} \qquad (1)$$

In Equation 1, Y_{ij} is the projected Time 2 score for the ith student from the jth high school, β_0 is the overall intercept term, X_{ijp} (p = 1, 2, 3, 4) are the prior test scores in the four subject areas (English, mathematics, reading, and science) for the ith student from the jth high school, β_p (p = 1, 2, 3, 4) are the regression coefficients for the four Test 1 scores, T_{ij} is the number of months between Test 1 and Test 2 for the ith student from the jth high school, β_5 is the regression coefficient for the testing span, and U_j is the school-specific intercept and can be interpreted as the school value-added score assumed to be normally distributed with mean 0 and variance τ^2. Student-level error term R_{ij} is assumed to be normally distributed with mean 0 and variance σ^2. I then fit this model to each of the four subject areas to obtain student projections and school value-added scores. The model is a special case of a hierarchical linear model (Raudenbush & Bryk, 2002) and can be fit using statistical software packages such as HLM or SAS.

In this study, I calculated the value-added score for a particular school as the average of the residual scores of its students, where the residual score for a particular individual student is defined as the difference between the student's projected score, based on the estimated fixed effects in Equation 1 and the student's actual score. School value-added scores can also be based on both the fixed effects and random effects in Equation 1; see Snidjers and Bosker (1999).

Standard-Classified VAM (Std-cls). In this model, Y_{ij} is defined as in Standard VAM. But prior test scores for the four subject areas in the right-hand side of Equation 1 are coded differently. That is, the subject-specific prior test score (sspts) is classified into 10 levels,[7] whereas the other three nonspecific prior test scores (nspts) are treated as linear effects, as in Equation 1. For example, when projecting the Test 2 English score, the Test 1 English score is classified into 10 levels while the Test 1 scores in the other three subject areas (mathematics, reading, and science) remain continuous. Comparing results for Std and Std-cls models allows us to examine the adequacy of modeling only the linear relationship between Test 1 and Test 2 score.

Quantile Regression (QR). Quantile regression can be conceptualized as an elaboration of ordinary least squares linear regression, i.e., from parameterizing trends in conditional means of the response variable Y to parameterizing trends in many different conditional quantiles of Y. The QR model demonstrates growth not in terms of conditional mean test score, but in terms of conditional percentile of test score. The main advantage of QR over least-squares regression is its flexibility for modeling data with heterogeneous conditional distributions (i.e., it makes no distributional assumption about the error term, and so it offers model robustness). QR is especially useful with data that are heterogeneous in the sense that the tails and the central location of the conditional distributions vary differently with the covariates. See Chen (2005) for more formal definition of QR.

[7] In theory, since ACT Explore scale scores range from 1 to 25, I could have defined up to 25 classified levels, but because of data sparsity, especially in the lower tail of the distribution, I limited the classification of subject-specific test score to 10 levels.

For a student with Grade 3 and Grade 4 test scores, for example, QR could be used to identify the percentile of the Grade 4 score conditional on the Grade 3 score. A student that has a percentile of 85 would indicate that the student's grade 4 score is in the 85th percentile among her peers with the same Grade 3 score. The student's growth can then be described as being as good/better than 85% of his/her academic peer group, as defined by initial test score. These growth percentiles are a normative measure of students' academic progress, comparing a student's progress over time to his/her academic peers with a similar academic (test score) history (Betebenner, 2009).

Given this normative foundation, the QR model can be used to identify "typical growth" (e.g., the predicted score from a QR model of the 0.50 quantile). Similar to residuals from a linear regression model, residuals can be obtained from a QR model as the difference between actual and projected scores, where projected scores are the 50th percentile scores from the QR. School mean QR residuals can then be used to support value-added inferences.

In this study, the linear quantile regression model (using the SAS QUANTREG procedure) was used to fit the data. In contrast, the Student Growth Percentile (SGP) model fits the data by using a nonparametric spline model using the R software package (Betebenner, 2011). This model has been used in several states (e.g., Colorado, Massachusetts, Arizona, and Indiana).

Due to the different modeling assumptions used in the SAS QUANTREG procedure and the SGP package, the estimated quantiles from the two programs can be quite different when initial test scores are at the tails of the score distribution but almost identical for middle-range scores. In the SAS QUANTREG procedure, a linear function of the independent variables is used to model the quantiles of the response variable, while the SGP package models the quantiles of the response variable as a linear combination of cubic B-splines of the independent variables. The B-spline quantile regression is a nonlinear (or nonparametric) extension of the linear quantile regression and can also handle heteroscedasticity and skewness of the conditional distributions of the response variable associated with values of the independent variables. Unlike the linear quantile regression, the B-spline quantile regression does not assume that the conditional densities of the response variable have the same shape across different values of the independent variables. Thus, it is a more flexible method than the linear QR model.

School Value-Added Score. The models discussed above can be used to produce projected scores (also known as predicted scores) for individual students. Two of the models (Std and Std-cls) are mixed-effects models, and so projected scores for individual students can either be based on the fixed effects parameters or based on both the fixed effects and random intercepts. The fixed effect parameters produce projected scores based on the population-averaged model, and the random intercept is set to 0. The projections are thus interpreted as expected Time 2 scores, assuming the typical school effect. Using projections based only on the fixed-effects model, residual scores can be calculated for each student as the difference between the actual Time 2 score and projected Time 2 score. This residual describes a student's growth in terms of how much she performed above expected, given her prior scores and assuming the typical school effect. This approach is known as the residual gain model (Castellano & Ho, 2013). A school's mean residual score then describes how much higher or lower than expected students score, on average, and can be used to support school value-added interpretations. The one-sample *t*-test can be used to test whether the mean residual is different from 0 for each school.

For the Std and Std-cls models, an alternative to the fixed-effects based residual approach described above would be to use the estimates of school intercepts as the school value-added score. This approach has produced school effect estimates that are highly correlated (e.g., $r > 0.995$) with those produced by the residual gain model (Castellano et al., 2015). The residual gain model applies to all three models (Std, Std-cls, and QR) because all three produce projected scores assuming typical school effects. The residual gain model is of special interest in this study because it is consistent with growth modeling resources offered by ACT that enable ACT clients to implement specific types of growth models, including the residual gain model.[8] For these reasons, school value-added scores were based on school mean residuals rather than the random-effects solutions.

Sample and Data

Six datasets, corresponding to six different growth periods, defined by initial and subsequent grade levels, were created from longitudinal test score data from the academic years 2006–2007 through 2011–2012 (see Table 1). Note that Dataset 2 (Grade 9–10, 10–14-month interval) is a subset of Dataset 3 (Grade 9–10, 9–18-month interval).[9]

Student records included in the analyses were required to have valid school identification. The following additional requirements were imposed for data from particular schools to be included in the analyses:

- The proportion of students tested on both the initial and follow-up tests (designated "Test 1" and "Test 2") must have been at least 0.50 for a given high school cohort.
- The cohort size must have been at least 10.
- The school sample size combined across years (2006–2007 through 2011–2012) must have been at least 30.
- The school had valid data on grade-level enrollment (hereafter referred to as "class size"), poverty and minority levels, and prior mean academic achievement. In this study, poverty level was defined as the school's proportion of students eligible for free or reduced lunch, and proportion minority was defined as the proportion of students who are African American, American Indian, or Hispanic.

Proportion tested in the first requirement above was defined as the number of students who took both assessments divided by the school's enrollment for the higher grade level. For example, proportion tested for Dataset 4 (Grade 8 ACT Explore to Grade 11 or 12 for the ACT test) was defined as $N \div (Enroll_{11} + Enroll_{12})/2$, where N is the number of students who took both assessments (ACT Explore and the ACT), $Enroll_{11}$ is the high school cohort's eleventh-grade enrollment, and $Enroll_{12}$ is the high school cohort's twelfth-grade enrollment. With this inclusion criterion, the sample was restricted to high school cohorts where the majority of students were represented.

[8] ACT growth modeling resources are described at www.act.org/growthmodeling/. Each model allows clients to determine projected scores, and residual scores can be aggregated to support value-added interpretations. Unlike random-effects approaches, the growth modeling resources allow users to construct aggregate growth scores (e.g., mean residual scores) without the use of the entire multischool dataset needed to fit random effects models.

[9] Because students could choose to take ACT Explore in fall, spring, or summer of Grade 9 and also choose to take ACT Plan in fall, spring, or summer of Grade 10, to account for varying time spans both 10–14- and 9–18-month intervals between assessments were considered.

Analysis

In this section, I briefly describe the statistical methods used for addressing each of the five research questions. The accuracy of growth projections made by the selected value-added models for different growth periods is measured by the correlation of projected and actual Test 2 scores. Additional analyses were run on the three selected growth models to assess the proportion of Test 2 projected scores that are within 1, 2, or 3 points of the actual score. With respect to these two criteria, the three growth models generally performed similarly. I therefore decided to base the remaining analyses on the Standard VAM.

To determine the extent that using multiple prior test scores enhances projection accuracy, I compared the accuracy of growth projections of two different models for predicting ACT scores (each a variation of the model represented in Equation 1). In the first model, the four ACT Plan test scores were used as predictor variables. In the second model, the four ACT Explore scores and the four ACT Plan scores were used as predictors.

Next, I investigated how often school averages are classified as being statistically significantly above or below average using a *p*-value of 0.05. (For a description of the statistical procedure used, see Appendix B.) Value-added scores were classified as below average (estimated effect < 0, *p*-value $< .05$), above average (estimated effect > 0, *p*-value $< .05$), or uncertain (*p*-value $\geq .05$). Additionally, I examined the variability across school cohorts in the proportion of students meeting or exceeding expected growth.

For each growth period, I assessed the association of value-added measures with school characteristics (prior mean academic achievement, school poverty level, proportion of racial/ethnic minority students, class size, and proportion tested) through simple correlations. I also regressed each value-added measure on the set of school characteristics using a multiple linear regression model and report standardized regression coefficients (beta weights). Next, I compared the distributions of the value-added measures for high- and low-poverty schools.

Lastly, correlations were used to evaluate the consistency of value-added scores; three different types of consistency were evaluated: (1) same-year school effects (correlating scores from different growth periods for the same school year), (2) same-cohort school effects (correlating scores from different growth periods for the same cohort), and (3) cross-year school effects (correlating scores from different cohorts for the same growth period).

Results

Sample sizes and demographic breakdowns of each of the six datasets are presented in Table 1. It is worth noting that annually more students take the ACT test than take ACT Plan and ACT Explore. Also, more students take ACT Explore in Grade 8 than in Grade 9. In 2012, for example, 1,666,017 high school graduates took the ACT test in eleventh or twelfth grade, 1,245,990 tenth graders took ACT Plan, and only 925,130 and 357,072 eighth and ninth graders took ACT Explore, respectively. As expected, the total number of students (1,895,825), schools (4,457), and cohorts (15,924) are the largest in Dataset 6. In contrast, in Dataset 5 these counts are smallest—partly because of the difference in time span in the two datasets (10–14 vs. 9–30) and because of the grade level of students who took the ACT test (eleventh vs. eleventh or twelfth). Next to Dataset 6, Dataset 1 has the largest sample sizes. As previously mentioned, Dataset 2 is a subset of Dataset 3.

To access how well these datasets represent some larger population, the gender and racial/ethnic group breakdowns for the total population were derived from tenth- and eleventh-grade totals of public high school students nationally in the 2010–2011 Common Core of Data (Keaton, 2012).[10] Females are slightly overrepresented in the sample (50–53% vs. 49%). White students are overrepresented in the sample (61–78% vs. 53–55%), while African American (7–16% vs. 16%), Hispanic (7–18% vs. 21–22%), and Asian American students (3–4% vs. 5%) are underrepresented.

Generally, more female than male students were represented in these datasets (50–53% vs. 47–50%). While percentages of Asian and Native American students were stable across the six datasets (3–4% vs. 1–2%), White students had higher percentages for Datasets 4–6 (78%), whereas African American and Hispanic students had higher percentages for Datasets 1–3 (15–16% and 9–18%, respectively). This may be due to higher high school dropout rates among African American and Hispanic students (American Psychological Association, 2012) or more likely due to the fact that Hispanic and African American students tend not to be college bound and do not take the ACT (Buddin, 2012).

[10] For the first three datasets, the gender and racial/ethnic group breakdowns for the total population were derived from tenth-grade totals with the same order as presented in Table 1 (49%, 51%, 16%, 5%, 53%, 22%, and 1%, respectively), and for the last three datasets, the same group breakdowns were derived from eleventh-grade totals (49%, 51%, 16%, 5%, 55%, 21%, and 1%, respectively).

Table 1. Sample Sizes and Student Demographics

Dataset	Growth period	Range of month span	*N*	$N_{schools}$ $N_{cohorts}$	Female	Male	African American	Asian	Caucasian	Hispanic	Native American
1	8–10	18–30	1,564,550	3,155 11,185	51%	49%	15%	3%	71%	9%	2%
2	9–10	10–14	505,717	976 2,181	51%	49%	16%	4%	61%	18%	1%
3	9–10	9–18	532,915	1,003 2,269	51%	49%	16%	4%	62%	17%	1%
4	8–12	30–54	525,194	1,836 4,361	53%	47%	10%	4%	78%	7%	1%
5	10–11	10–14	146,690	495 1,176	50%	50%	7%	3%	78%	11%	1%
6	10–12	9–30	1,895,825	4,457 15,924	53%	47%	10%	3%	78%	7%	1%

Note. Each dataset represents the total number of high schools ($N_{schools}$) and the total number of high school cohorts ($N_{cohorts}$). For each high school, there are up to six cohorts of available data (assessment results from multiple years on the same cohort of students).

In Table 2, the six datasets are described with respect to Test 1, Test 2, and gain scores. To assess how well these datasets represent ACT national normative data, I compared the Test 1 means and *SD*s for all four subject areas from the samples with the nationally representative norms for grades 8, 9, and 10.[11] For Datasets 1–3, the Test 1 sample means were very similar to the national norms means—they were either identical (Dataset 1) or slightly larger (by 0.1 to 0.3). However, the sample means for Datasets 4–6 (in which Test 2 is the ACT) were larger than national norms means (by 0.9 to 1.8). This may be due to the fact that dropout rates between grades 10 and 12 for academically low-achieving students are higher than those for academically high-achieving students (American Psychological Association, 2012; Chapman, Laird, & KewalRamani, 2011), or more likely due to the fact that Hispanic and African American students do not take the ACT (Buddin, 2012). The *SD*s for all six samples were similar (but not identical) to the national norms *SD*s.

Table 2. Summary Statistics on Test Scores and Gains

Dataset	Growth period	Range of month span	Subject	Test 1		Test 2		Gain	
				M	*SD*	*M*	*SD*	*M*	*SD*
1	8–10	18–30	English	14.7	4.1	16.7	4.4	2.0	2.9
			Mathematics	15.5	3.7	17.7	4.6	2.2	3.2
			Reading	14.6	3.7	16.9	4.6	2.3	3.4
2	9–10	10–14	English	15.7	4.4	16.7	4.5	1.0	2.8
			Mathematics	16.3	4.0	17.8	4.8	1.5	3.2
			Reading	15.4	4.1	17.0	4.7	1.6	3.3
3	9–10	9–18	English	15.7	4.4	16.7	4.5	1.1	2.8
			Mathematics	16.4	4.0	17.9	4.8	1.5	3.2
			Reading	15.4	4.1	17.1	4.7	1.6	3.3
4	8–12	30–54	English	15.9	4.0	21.3	6.2	5.4	4.0
			Mathematics	16.6	3.4	21.3	5.3	4.8	3.6
			Reading	15.6	3.8	21.7	6.0	6.0	4.3
5	10–11	10–14	English	18.0	4.4	20.0	6.1	2.0	3.4
			Mathematics	19.0	4.7	20.5	5.0	1.6	2.8
			Reading	18.2	4.6	20.6	6.0	2.4	3.9
6	10–12	9–30	English	18.0	4.3	21.0	6.1	3.0	3.7
			Mathematics	18.9	4.6	21.1	5.2	2.2	3.0
			Reading	18.1	4.5	21.5	6.0	3.4	4.1
			Science	19.2	3.6	21.3	4.9	2.1	3.5

[11] National normative data are based on results for students who took all four academic tests within standard time limits as part of a national study conducted in fall 2010. See the technical manuals for ACT Explore and ACT Plan at www.act.org for information about the 2010 national norming sample.

How Strongly Related Are Students' Projected Scores to Their Actual Scores?

Does the strength of relationship vary for different projection models? I begin by evaluating the three selected growth models (two types of Standard VAMs and the QR model) based on two criteria:

- the correlation of actual and projected scores
- the proportion of Test 2 projected scores that are within 1, 2, or 3 points of the actual score. Note that for the QR models the median score is used as the projected score (i.e., single quantile value of 0.50).

Table 3 presents the projection accuracy of the different models across different growth periods for the four subject areas. Across all growth periods, English has the highest and reading and science have the lowest projection accuracies. The projection accuracy generally remains the same across the three different models, suggesting that projection accuracy is not improved by choosing one model over the other.

Table 3. Projection Accuracy

Dataset	Growth period	Range of month span	Subject	R_{Std}	$R_{Std\text{-}cls}$	R_{QR}
1	8–10	18–30	English	0.81	0.81	0.81
			Mathematics	0.76	0.77	0.76
			Reading	0.74	0.74	0.74
			Science	0.73	0.73	0.73
2	9–10	10–14	English	0.84	0.84	0.84
			Mathematics	0.80	0.80	0.80
			Reading	0.77	0.77	0.77
			Science	0.77	0.78	0.77
3	9–10	9–18	English	0.84	0.84	0.84
			Mathematics	0.80	0.80	0.80
			Reading	0.77	0.77	0.77
			Science	0.77	0.77	0.77
4	8–12	30–54	English	0.82	0.79	0.82
			Mathematics	0.78	0.77	0.78
			Reading	0.78	0.74	0.78
			Science	0.75	0.73	0.75
5	10–11	10–14	English	0.86	0.86	0.86
			Mathematics	0.85	0.86	0.85
			Reading	0.82	0.82	0.82
			Science	0.80	0.80	0.80
6	10–12	9–30	English	0.85	0.85	0.85
			Mathematics	0.84	0.84	0.84
			Reading	0.81	0.81	0.81
			Science	0.79	0.79	0.79

Note. R_{Std} is the correlation of projected Test 2 score and actual Test 2 score, using the standard value-added model. $R_{Std\text{-}cls}$ is the correlation of projected Test 2 and actual Test 2, using the Std model (where subject-specific Test 1 is used as a categorical covariate). R_{QR} is the correlation of projected Test 2 and actual Test 2, under the QR method (where the median score (0.50) is used as the projected score).

In Table 4, the three growth models are compared with respect to the proportion of Test 2 projected scores that are within 1, 2, or 3 points of the actual scores in absolute values. Again, with respect to this criterion, the three selected growth models generally revealed the same performance. That is, across the six datasets, the proportion of projected scores that are within 1, 2, or 3 points of the actual Test 2 score generally remained the same or differed negligibly (by .01 for Std vs. QR or by .01 to .06 for Std-cls vs. the other two models). Note that even though, across all datasets, English has the highest projection accuracy and science has the lowest (R_{Std} in Table 3), W_1, W_2, and W_3 for science are all larger than those for English (see Table 4). This is due to the fact that science has the smallest standard deviation of the four subject areas (see Table 2), and that causes the proportion of absolute distances between projected and actual scores (the *W*s) for science to have relatively inflated values.

Table 4. Comparison of Projection Accuracy

Dataset	Growth period	Range of month span	Subject	Std			Std-cls			QR		
				W_1	W_2	W_3	W_1	W_2	W_3	W_1	W_2	W_3
1	8–10	18–30	English	0.32	0.59	0.77	0.32	0.58	0.77	0.32	0.59	0.77
			Mathematics	0.31	0.56	0.73	0.32	0.58	0.74	0.32	0.57	0.74
			Reading	0.27	0.50	0.68	0.27	0.50	0.68	0.27	0.50	0.68
			Science	0.35	0.62	0.80	0.35	0.63	0.80	0.35	0.63	0.80
2	9–10	10–14	English	0.34	0.61	0.80	0.33	0.61	0.79	0.34	0.61	0.80
			Mathematics	0.30	0.56	0.73	0.32	0.57	0.74	0.31	0.56	0.73
			Reading	0.28	0.52	0.70	0.28	0.52	0.70	0.28	0.52	0.70
			Science	0.36	0.63	0.81	0.36	0.64	0.81	0.36	0.64	0.81
3	9–10	9–18	English	0.34	0.61	0.79	0.33	0.61	0.79	0.34	0.61	0.80
			Mathematics	0.30	0.55	0.73	0.32	0.57	0.73	0.31	0.56	0.73
			Reading	0.28	0.52	0.70	0.28	0.52	0.70	0.28	0.52	0.70
			Science	0.36	0.63	0.81	0.36	0.64	0.81	0.36	0.64	0.81
4	8–12	30–54	English	0.23	0.43	0.61	0.20	0.40	0.56	0.23	0.44	0.62
			Mathematics	0.23	0.45	0.64	0.24	0.46	0.63	0.24	0.46	0.64
			Reading	0.21	0.41	0.58	0.18	0.37	0.53	0.21	0.41	0.58
			Science	0.26	0.48	0.66	0.24	0.46	0.64	0.26	0.49	0.67
5	10–11	10–14	English	0.27	0.50	0.69	0.26	0.49	0.68	0.27	0.50	0.69
			Mathematics	0.31	0.57	0.76	0.33	0.59	0.77	0.31	0.58	0.77
			Reading	0.23	0.45	0.63	0.23	0.45	0.63	0.24	0.45	0.63
			Science	0.29	0.53	0.71	0.28	0.53	0.71	0.29	0.53	0.72
6	10–12	9–30	English	0.25	0.48	0.66	0.24	0.47	0.65	0.26	0.48	0.67
			Mathematics	0.28	0.53	0.72	0.29	0.55	0.73	0.28	0.54	0.73
			Reading	0.23	0.43	0.61	0.22	0.43	0.61	0.23	0.44	0.61
			Science	0.28	0.52	0.70	0.27	0.51	0.70	0.28	0.52	0.70

Note. W_1 is the proportion of Test 2 projected scores that are within 1 point of the actual score. W_2 is the proportion of Test 2 projected scores that are within 2 points of the actual score. W_3 is the proportion of Test 2 projected scores that are within 3 points of the actual score.

Next, I look at the linearity assumption by comparing results for Std Model and Std-cls Model to examine the adequacy of modeling the linear relationship between the Test 1 and Test 2 scores. For illustration purposes, I show the results of fitting the Std-cls Model to Dataset 4 only. In Figures A1–A4 (see Appendix A), I present scatter plots showing the relationship between classified subject-specific ACT Explore scores (classified into 10 levels, as shown in the figures)[12] and least-square means of the projected ACT scores based on the four subject areas (English, mathematics, reading, and science), respectively. In Table A1, projection parameters generated by Std-cls Model using Dataset 4 are presented for the four subject areas. Similar results are obtained for the other five datasets. It appears that the estimated trajectories—except for the lower-classified ACT Explore scores—tend to show a mostly linear relationship between Test 1 (classified ACT Explore) and Test 2 (continuous ACT test) scores.[13] This is more likely due to the fact that students with very low scores are not making full effort on Test 1. It is worth noting that linear trend in English, reading, and science begins at chance scores (see Figures A1, A3, and A4).[14] In mathematics, however, for unknown reasons the linear trend begins above chance scores (see Figure A2). There does appear to be some nonlinearity, but perhaps not enough to bias VAM estimates.

Standard Value-Added Model (Std). As noted above, due to small differences in projection accuracy observed between the Standard VAM and the other two selected growth models in this study, the remaining analyses will be based on the Standard VAM. In Table 5, projection parameters generated by Equation 1 are presented for different growth periods (Datasets 1–6). As expected, it is apparent that the estimated regression coefficients for the four prior test scores ($\hat{\beta}_p$) tend to be higher for the same-subject regression coefficients. There are some exceptions: for example, for predicting ACT Plan science scores the estimated regression coefficients in Dataset 3 are comparable for the ACT Explore mathematics score (0.309) and the ACT Explore science score (0.311). The estimated regression coefficients for the testing span ($\hat{\beta}_5$) are all positive (ranging from 0.029 to 0.120), indicating that more growth is expected as more time passes between the prior and later test (Test 1 and Test 2). These coefficients are slightly higher for mathematics projections in Datasets 1–3 and for English projections in Datasets 4–6. It is worth noting that projected scores are generally less heavily influenced by testing span relative to prior academic achievement (particularly in the same subject).[15] Also, the estimated regression coefficients (particularly in English and mathematics) for the same-subject are higher in Datasets 4–6 when the ACT is Test 2 (relative to Datasets 1–3 when ACT Plan is Test 2).

[12] Note that except for the lowest classified ACT Explore level (i.e., marked as 6.5) that corresponds to ACT Explore scores of 1–7, the other nine levels correspond to the midpoint of classified ACT Explore levels.

[13] Of the total sample of 525,194, the percentages of ACT Explore scores falling within the first two classifications is about 3% for English and less than 2% for math and reading; for science, less than 2% of ACT Explore scores fall within the first three classifications.

[14] Chance scores are scores that a student might have received if the student guessed randomly on every test question.

[15] As an example, for predicting ACT English scores in Dataset 5, I calculated the standardized regression coefficients for ACT Plan English (0.61) and time span (.015). By extension, across all datasets and for all subject areas, similar results are expected to be obtained.

Table 5. Growth Model Projection Parameters

Dataset	Growth period	Range of month span	Subject	Intercept	English	Math	Reading	Science	Month span	Level-2 *SD*
1	8–10	18–30	English	−0.505	0.474	0.180	0.212	0.143	0.081	2.546
			Mathematics	−1.457	0.169	0.507	0.091	0.273	0.120	2.864
			Reading	−0.633	0.299	0.088	0.421	0.181	0.110	3.093
			Science	3.264	0.149	0.226	0.179	0.276	0.084	2.500
2	9–10	10–14	English	0.908	0.469	0.148	0.206	0.144	0.029	2.412
			Mathematics	−0.314	0.133	0.512	0.085	0.307	0.084	2.832
			Reading	1.019	0.279	0.066	0.438	0.178	0.056	2.974
			Science	3.871	0.126	0.209	0.186	0.310	0.062	2.456
3	9–10	9–18	English	0.572	0.469	0.148	0.206	0.145	0.056	2.417
			Mathematics	−0.693	0.134	0.515	0.084	0.309	0.111	2.843
			Reading	0.696	0.279	0.066	0.437	0.180	0.081	2.981
			Science	3.827	0.126	0.211	0.186	0.311	0.061	2.466
4	8–12	30–54	English	−8.367	0.623	0.303	0.305	0.269	0.108	3.397
			Mathematics	−2.615	0.168	0.680	0.071	0.344	0.051	3.099
			Reading	−3.951	0.396	0.182	0.519	0.322	0.050	3.705
			Science	−0.556	0.182	0.426	0.182	0.393	0.044	3.181
5	10–11	10–14	English	−5.324	0.706	0.203	0.257	0.130	0.120	2.998
			Mathematics	0.196	0.156	0.615	0.004	0.234	0.090	2.545
			Reading	−3.431	0.403	0.134	0.495	0.196	0.115	3.412
			Science	1.138	0.189	0.324	0.140	0.352	**0.048**	2.897
6	10–12	9–30	English	−5.748	0.697	0.225	0.252	0.144	0.110	3.155
			Mathematics	−0.066	0.166	0.625	0.003	0.253	0.054	2.723
			Reading	−2.893	0.418	0.150	0.478	0.208	0.059	3.517
			Science	0.937	0.191	0.352	0.125	0.348	0.051	2.975

Note. The regression coefficient in bold is statistically significant at $p < .05$. The coefficient in red is not statistically significant. All others are statistically significant at $p < .001$. In the last column, the standard deviation of random effects is presented.

Table 6 summarizes the distributions of the value-added measures (i.e., mean residual scores[16]) generated by Equation 1 for each subject area and growth period. The largest variation in value-added scores was obtained in Dataset 4 (*SD* ranging from 0.99 to 1.18) and in Dataset 6 (*SD* ranging from 0.78 to 0.91), with each having time spans of 30–54 and 9–30 months, respectively.

Next, I calculated the number of students needed for scores at the 75th and 90th percentiles to be statistically significant. The statistical significance (e.g., *p*-value) of a value-added score is a function of the size of the score and the number of students included. I derived a formula for sample size needed in order for the value-added score at 75th (and 25th) or 90th (and 10th) percentiles to be statistically significant at $p < .05$ (see Appendix B).

The value-added score estimates (at the 75th and 90th percentiles) obtained from Dataset 4 not only are higher than those obtained from the other five datasets, but also require relatively fewer students in order for the estimates to be statistically significant. The results suggest that a small school (e.g., with 50 tested students) would need to have a value-added score above the 75th percentile in order to reach statistical significance.

Table 6. School Cohort Value-Added Score Distributions

Dataset	Growth period	Range of month span	Subject	*SD*	P_{25}	P_{50}	P_{75}	P_{90}	N_{75}	N_{90}
1	8–10	18–30	English	0.64	−0.37	0.03	0.43	0.81	134	37
			Mathematics	0.81	−0.58	−0.09	0.46	1.02	140	28
			Reading	0.76	−0.48	−0.01	0.47	0.94	168	41
			Science	0.65	−0.43	−0.02	0.40	0.80	142	35
2	9–10	10–14	English	0.65	−0.31	0.06	0.42	0.80	129	36
			Mathematics	0.77	−0.54	−0.09	0.39	0.83	197	44
			Reading	0.78	−0.40	−0.01	0.38	0.75	251	65
			Science	0.64	−0.36	0.02	0.38	0.71	154	46
3	9–10	9–18	English	0.59	−0.31	0.06	0.42	0.78	128	35
			Mathematics	0.70	−0.53	−0.08	0.39	0.82	197	43
			Reading	0.64	−0.39	0.00	0.37	0.73	233	62
			Science	0.56	−0.34	0.03	0.38	0.70	152	44
4	8–12	30–54	English	1.14	−0.70	0.01	0.76	1.48	78	20
			Mathematics	1.18	−0.71	−0.01	0.83	1.70	50	12
			Reading	1.00	−0.64	0.06	0.70	1.25	109	35
			Science	0.99	−0.59	0.06	0.69	1.27	83	24
5	10–11	10–14	English	0.89	−0.60	0.04	0.56	1.07	110	30
			Mathematics	0.73	−0.50	−0.07	0.43	0.93	127	28
			Reading	0.73	−0.45	0.02	0.48	0.89	198	56
			Science	0.70	−0.47	0.01	0.48	0.82	144	48
6	10–12	9–30	English	0.91	−0.57	0.02	0.62	1.17	98	27
			Mathematics	0.88	−0.54	0.00	0.63	1.23	68	18
			Reading	0.81	−0.52	0.01	0.53	0.98	170	49
			Science	0.78	−0.47	0.06	0.56	0.98	108	35

Note. N_{75} is the number of students needed in order for the value-added score at the 75th (and 25th) percentiles to be statistically significant at $p < .05$. N_{90} is the number of students needed in order for the value-added score at the 90th (and 10th) percentiles to be statistically significant at $p < .05$.

[16] Residual scores are defined as the difference between students' actual and projected scores based on the fixed effects in Equation 1.

Is measurement error in the ACT college and career readiness system scores properly addressed by using multiple test scores? It is well known that standardized tests measure student achievement with a certain degree of error, but less understood are the implications of test measurement error (TME) for VAMs. Recent studies that have examined the implications of TME for VAMs, include Koedel, Leatherman, and Parsons (2012); Boyd, Grossman, Lankford, Loeb, and Wyckoff (2008); and Lockwood and McCaffrey (2012).

If measurement error is not properly addressed (at the student level and/or at classroom or school levels) in VAMs, instability in value-added estimates from year to year can occur. This instability can affect inferences made about schools' or teachers' effectiveness (particularly because teacher-level sample sizes are inherently small). It has been argued that, when teachers or schools are the unit of analysis to which inferences are being made, some portion of the observed variability in value-added estimates can be explained by chance (McCaffrey et al., 2009). According to Briggs (2011):

> Measurement error at the student level is assumed to have a functional relationship with the number of test items that are administered to students; at the teacher or school level, measurement error is assumed to have a functional relationship with the number of students. The analogy here is essentially that students are to schools what items are to students. Taken together, both of these sources of measurement error could explain the instability in value-added estimates from year to year—even if the true value were actually constant over time. (p. 31)

There are ways to lessen the impact of measurement error at the student and school levels. At the classroom or school level, three steps are typically taken, sometimes in tandem, to address measurement error. One is to increase the number of years of data over which value-added is being estimated from one to three years (McCaffrey et al., 2009). Two other steps are to use estimates of reliability to "shrink" value-added estimates back to the grand mean (i.e., the average value-added of all teachers in the system), and/or to compute a confidence interval around each value-added estimate (Briggs, 2011).

In the Education Value-Added Assessment System (EVAAS),[17] the value-added analyses are multivariate and longitudinal in that all prior scores in all subjects are analyzed simultaneously. The inclusion of all available scores dampens the measurement error in any single score. The EVAAS Multivariate Response Model (MRM) process addresses the measurement error problem by using multiple predictors (at least three and as many as six for each student), showing that with a sufficient number of test scores, no adjustment is necessary at the student level (Sanders et al., 1997; Ballou et al., 2004). It has been shown that three prior scores are necessary to eliminate nontrivial bias in the school estimates due to the errors of measurement in the previous test scores. This conclusion has been reached by using both simulation and empirical data results (Sanders, 2006). In this study, I used four subject-specific initial test scores (English, mathematics, reading, and science) in the model. However, because they all emanate from a single test administration (as opposed to multiple years of test administrations), it is not yet clear whether further adjustment is necessary at the student level in order to dampen the effect of measurement error (particularly for schools with small sample size).

[17] The SAS Education Value-Added Assessment System (SAS EVAAS) is the most widely used value-added system in the country. It is also self-proclaimed as "the most robust and reliable" system available, with its greatest benefit to help educators improve their teaching practices (Sanders, Saxton, & Horn, 1997).

Another source of error is specification error (not including other covariates in the model) that potentially could bias the results and thereby affect inferences made about school effectiveness. One way to assess the effect of potential bias is through sensitivity analysis that is beyond the scope of this study.

Is the strength of relationship enhanced by using test scores from multiple prior years? To answer this question, using longitudinal data from a large sample of students who took ACT Explore in eighth grade, ACT Plan in tenth grade, and the ACT in eleventh or twelfth grade, I looked at two different models (where the time span of 9–30 is the same in both models):

- The first model is the same as Equation 1. However, Y_{ij} in this case is the ACT score for the ith student from the jth high school and X_{ijp} (p = 1, 2, 3, 4) are the ACT Plan test scores (prior test scores) in the four subject areas for the ith student from the jth high school.
- The second model is an extension of the first, where both ACT Plan and ACT Explore test scores in the four subject areas are used as covariates in the model.

Next, I look at the accuracy of projected ACT scores by evaluating these two models based on two criteria (see Table 7): (1) the correlation between projected and actual ACT scores, and (2) the proportion of projected ACT scores that are within 1, 2, or 3 points of the actual ACT scores. Table 7 presents the projection accuracy of these two models across the four subject areas with respect to the two criteria set forth above.

As shown in Table 7, across the four subject areas the correlation between projected and actual ACT scores (R), and the proportion of projection accuracy (W_1, W_2, and W_3) are only slightly higher in the second model than in the first. This reveals that projection accuracy is slightly enhanced by using test scores from multiple prior years—that is, by using both ACT Explore and ACT Plan scores as covariates (in the second model) rather than by using only ACT Plan scores as covariates (in the first model). Arguably, the small increase in accuracy with using multiple years of test scores does not warrant the use of that model. The model using multiple years of prior test scores requires more data and in practice would only be available for a subset of students that are included in the simpler model.

Table 7. Comparison of Projection Accuracy (ACT Plan vs. ACT Explore and ACT Plan Prior Test Scores)

Growth period	Range of month span	Subject	R	W_1	W_2	W_3
10–12	9–30	English	0.85	0.25	0.48	0.66
		Mathematics	0.84	0.27	0.52	0.71
		Reading	0.81	0.23	0.43	0.61
		Science	0.79	0.27	0.51	0.70
10–12 (with EXP)	9–30	English	0.87	0.27	0.50	0.69
		Mathematics	0.86	0.29	0.54	0.73
		Reading	0.83	0.24	0.45	0.63
		Science	0.80	0.28	0.53	0.71

Note. n = 525,194 students; R is the correlation between projected and actual ACT scores. W_1 is the proportion of ACT projected scores that are within 1 point of the actual score. W_2 is the proportion of ACT projected scores that are within 2 points of the actual score. W_3 is the proportion of ACT projected scores that are within 3 points of the actual score.

How Often Are School Value-Added Scores Classified as Significantly Above or Below Average?

As I have demonstrated, value-added models can be used to produce estimates of school effects and to produce measures of the uncertainty of the estimated school effects. Typically, standard errors and/or confidence intervals are used to quantify the uncertainty of estimates. If the standard error is larger, the confidence interval is wider, and there is greater uncertainty about the estimate. Estimated school effects for large cohorts tend to have less sampling error (i.e., smaller standard errors); larger schools are therefore more likely to be classified with certainty. Statisticians sometimes use *p*-values as a measure of uncertainty. In this case, the *p*-value represents the probability that the observed value or more extreme value would have resulted if the "true" value were actually 0. Therefore, smaller *p*-values imply greater certainty that an estimated school effect is different from that for the average school. *P*-values of .01 and .05 are commonly used thresholds for certainty.

For each of the estimated school effects, a *p*-value reflects the degree of certainty that the estimated school effect is greater or less than that for the average school. In Table 8, I present classifications of high school cohorts in the sample for the value-added measures (i.e., mean residual scores) generated from the six datasets.[18] Each high school cohort is classified as below average (estimated effect < 0, *p*-value < .05, *z*-score < −1.96), above average (estimated effect > 0, *p*-value < .05, *z*-score > +1.96), or uncertain (*p*-value ≥ .05, −1.96 ≤ *z*-score ≤ +1.96). Note that across all six datasets, the highest percentages of the estimated school effects that are classified as below average with certainty are in mathematics, and the lowest percentages are in reading. Results for percentages of the estimated school effects that are classified as above average with certainty are mixed, but generally the lowest percentages are in reading. In all, however, between 46% (in mathematics) to 70% (in reading) of the estimated school effects are classified as uncertain. This finding has important implications. It suggests that most estimated school effects should not be classified as "below average" or "above average" with high confidence. It reflects the reality that most school effects are not significantly different from the average school effect and cannot usually be distinguished from average with a high degree of confidence.

[18] In order for a value-added measure (VA-m) to be classified as "very certain" or to be included under "Above average (%)," the ratio of VA-m/$SE_{VA\text{-}m}$ (where $SE_{VA\text{-}m}$ is the standard error of VA-m) has to be greater than +1.96. Similarly, the ratio has to be between −1.96 to +1.96 for it to be classified under "Uncertain (%)," and less than -1.96 for "Below average (%)" classification. This approach was used as the school value-added scores in this study are based on the fixed effects parameters. For school value-added scores that are based on both the fixed effects and random intercepts see Snidjers et al. (1999, pp. 60–63) for recommendations on how to construct confidence intervals for school effects.

Table 8. Classifications of School Cohort Effects

Dataset	Growth period	Range of month span	Subject	Uncertain (%)	Below average (%)	Above average (%)
1	8–10	18–30	English	60	18	22
			Mathematics	53	27	21
			Reading	61	20	19
			Science	57	22	21
2	9–10	10–14	English	52	22	26
			Mathematics	50	29	21
			Reading	57	21	22
			Science	52	23	25
3	9–10	9–18	English	52	21	27
			Mathematics	50	29	21
			Reading	57	21	22
			Science	52	23	25
4	8–12	30–54	English	52	21	27
			Mathematics	46	25	29
			Reading	59	18	23
			Science	55	19	26
5	10–11	10–14	English	56	21	23
			Mathematics	54	24	22
			Reading	70	15	15
			Science	61	17	22
6	10–12	9–30	English	58	17	25
			Mathematics	52	22	26
			Reading	68	14	18
			Science	61	15	23

Note. Below Average implies school effect < 0 with a *p*-value < .05; *Above Average* implies school effect > 0 with a *p*-value < .05.

What Is the Variability Across School Cohorts in the Proportion of Students Meeting or Exceeding Expected Growth?

Table 9 summarizes, across school cohorts, the proportion of students meeting expected/average growth (i.e., those with positive residual scores) for each of the four ACT subject tests and different growth periods. (Note that across schools there is variability in the proportion of students meeting expected growth.) For example, at a school that is at the 75th percentile of all schools in Dataset 1, only 58% of students were at or above expected growth in English; 42% of students performed below the expected growth. Moreover, in Dataset 4, even at a school that is at the 90th percentile of

all schools, only 68% of students were at or above expected growth in English; one-third of students performed below the expected growth. Clearly, even at the high-ranking schools (at the 75th and 90th percentile), high percentages of students perform below expected growth in all subject areas and across all datasets. However, compared to average-ranking schools (at 50th percentile), expected growth for students is higher at 75th-percentile schools by 5–12% and by 10–22% at 90th-percentile schools. (The lower and upper ranges of expected growth correspond to reading and mathematics, respectively.) In all, at 75th and 90th percentiles, the proportion of students meeting or exceeding expected growth are higher for Dataset 4 than for the other five datasets—ranging from 0.57 in reading to 0.61 in science at P_{75} and from 0.63 in reading to 0.69 in mathematics at P_{90}.

Table 9. School Cohort Distributions of Proportion Meeting Expected Growth

Dataset	Growth period	Range of month span	Subject	*SD*	P_{25}	P_{50}	P_{75}	P_{90}
1	8–10	18–30	English	0.11	0.44	0.51	0.58	0.64
			Mathematics	0.12	0.38	0.46	0.54	0.62
			Reading	0.11	0.43	0.50	0.56	0.63
			Science	0.11	0.41	0.49	0.56	0.63
2	9–10	10–14	English	0.11	0.45	0.51	0.58	0.64
			Mathematics	0.11	0.39	0.45	0.52	0.59
			Reading	0.10	0.44	0.50	0.55	0.61
			Science	0.11	0.42	0.49	0.55	0.60
3	9–10	9–18	English	0.10	0.45	0.51	0.58	0.64
			Mathematics	0.10	0.39	0.46	0.52	0.59
			Reading	0.09	0.44	0.50	0.55	0.60
			Science	0.09	0.43	0.49	0.55	0.60
4	8–12	30–54	English	0.14	0.40	0.50	0.59	0.68
			Mathematics	0.16	0.36	0.47	0.59	0.69
			Reading	0.12	0.41	0.49	0.57	0.63
			Science	0.13	0.43	0.52	0.61	0.68
5	10–11	10–14	English	0.12	0.41	0.50	0.58	0.65
			Mathematics	0.13	0.38	0.45	0.54	0.63
			Reading	0.10	0.43	0.49	0.55	0.60
			Science	0.11	0.45	0.52	0.59	0.64
6	10–12	9–30	English	0.12	0.42	0.50	0.59	0.66
			Mathematics	0.14	0.37	0.47	0.57	0.66
			Reading	0.10	0.42	0.49	0.56	0.61
			Science	0.11	0.45	0.53	0.60	0.66

What School Characteristics Are Related to School Value-Added Scores?

In this section, in addition to controlling for student factors, I also control for school factors (school size, proportion of students tested, poverty level, and proportion of racial/ethnic minority students).[19] To account for possible peer effects that may introduce bias (Haertel, 2013), the projected scores are also adjusted for the prior subject-specific mean ACT Explore (or ACT Plan) scores observed in each high school.

Table 10 contains simple correlations between value-added measures (i.e., mean residual scores) and school characteristics, including prior mean academic achievement, school poverty level, proportion of racial/ethnic minority students, class size (i.e., grade level enrollment), and proportion tested. As expected, due to peer effect (Dahl, Loken & Mogstad, 2013), prior mean academic achievement is positively related to the value-added measures ($r = 0.13$ to $r = 0.52$), whereas school poverty level and proportion of racial/ethnic minority students have inverse relationships with the school effects ($r = -0.29$ to $r = -0.56$ and $r = -0.08$ to $r = -0.45$, respectively). Generally, compared to other school characteristics, class size and proportion of students tested had weaker associations with value-added measures.[20] The only exceptions for class size were in Datasets 4–6, where later assessment is the ACT test in grades 11 or 12. Further research is needed to understand why there is little or essentially no association between class size and value-added measures in Datasets 1–3, where later assessment is ACT Plan in grade 10 (note that Dataset 2 is a subset of Dataset 3).

In two earlier influential articles, Hanushek (1986; 1997) has surveyed much of the early research on class size as well as other educational inputs such as per-pupil spending. Based on these surveys, he concluded that "there is not a strong or consistent relationship between student performance and school resources" such as class size or spending. In a thorough re-analysis of Hanushek's literature summary, Krueger (2003) demonstrates that this conclusion relies on a faulty summary of the data and that there is indeed a systematic positive relationship between school resources and student performance in the literature surveyed by Hanushek. Nonetheless, in *David and Goliath*, Gladwell (2013) uncritically cites the Hanushek literature summary and its argument that the class size literature is inconclusive. In a policy brief, however, Schanzenbach (2014) summarizes the academic literature on the impact of class size and finds that class size is an important determinant of a variety of student outcomes, ranging from test scores to later-life outcomes such as college completion. Smaller classes are particularly effective at raising achievement levels of low-income and minority children, and an effective strategy to reduce the black-white achievement gap. The best evidence on the impact of reducing class sizes comes from Tennessee's Student Teacher Achievement Ratio (STAR) experiment (Mosteller, 1995). Results of this experiment suggest that the internal rate of return from reducing class size from 22 to 15 students is around 6%.

[19] School factors are derived from 2010–11 NCES Common Core of Data (Keaton, 2012).

[20] Given the large sample sizes, most of the correlations are statistically different from zero. However, one must pay attention to the strength of the correlation to determine if the relationship has any practical significance.

Table 10. Correlations of School Characteristics and Value-Added Scores

Dataset	Growth period	Range of month span	Subject	Class size	Proportion tested	Poverty level	Proportion minority	Prior mean achievement (Test 1)
1	8–10	18–30	English	**0.04**	0.12	−0.32	−0.31	0.37
			Mathematics	0.13	0.08	−0.47	−0.25	0.52
			Reading	0.06	0.14	−0.43	−0.38	0.45
			Science	0.09	0.15	−0.47	−0.35	0.47
2	9–10	10–14	English	0.00	0.03	−0.36	−0.34	0.42
			Mathematics	0.11	0.09	−0.35	−0.24	0.38
			Reading	0.01	0.15	−0.39	−0.41	0.45
			Science	−0.01	0.12	−0.43	−0.42	0.47
3	9–10	9–18	English	−0.01	0.03	−0.39	−0.38	0.43
			Mathematics	0.11	0.12	−0.40	−0.28	0.43
			Reading	−0.01	0.17	−0.42	−0.45	0.48
			Science	0.00	0.12	−0.46	−0.44	0.49
4	8–12	30–54	English	0.38	**0.05**	−0.44	−0.14	0.24
			Mathematics	0.38	0.07	−0.51	−0.11	0.34
			Reading	0.29	0.01	−0.44	−0.23	0.26
			Science	0.35	0.04	−0.56	−0.22	0.35
5	10–11	10–14	English	0.16	−0.03	−0.36	−0.08	0.23
			Mathematics	0.28	0.05	−0.42	**−0.10**	0.42
			Reading	0.04	−0.05	−0.29	−0.18	0.13
			Science	0.12	0.01	−0.35	−0.19	0.26
6	10–12	9–30	English	0.41	0.06	−0.43	−0.17	0.35
			Mathematics	0.39	0.10	−0.55	−0.19	0.54
			Reading	0.28	−0.05	−0.43	−0.25	0.35
			Science	0.31	0.04	−0.56	−0.32	0.47

Note. Correlations in bold are statistically significant at $p<.05$. Correlations in red are not statistically significant. All others are statistically significant at $p<.01$.

Multiple linear regression was also used to assess the relationships of the school characteristics and value-added scores. This analysis goes beyond the correlational analysis because the effect of each school characteristic is simultaneously controlled for. This analysis lets us see which school characteristics predict value-added measures above and beyond what is predicted by the other school characteristics.

Table 11 contains standardized regression coefficients (beta weights) by regressing each value-added measure on these characteristics using a multiple regression model. Generally, prior mean academic achievement level was positively related to the value-added measures, whereas poverty level was inversely related to the value-added measures (Table 11). Thus, cohorts with higher entering student achievement levels, as well as cohorts from wealthier schools, had significantly higher value-added scores. In Datasets 1–3, prior mean academic achievement level has larger positive effect (ranging from $b = 0.24$ to $b = 0.40$) in predicting growth than in Datasets 4–6, where later assessment is ACT Plan and the ACT, respectively. In contrast, in Datasets 4–6, school poverty level has larger negative effect (ranging from $b = -0.54$ to $b = -0.24$) in predicting growth than in Datasets 1–3. This suggests the effect of school poverty level is larger on ACT scores than on ACT Plan scores; perhaps the effect becomes larger with longer passage of time. School poverty level was inversely related to the mathematics value-added measures in Datasets 1–3, suggesting that wealthier schools have slightly greater effects on mathematics achievement. Generally, compared to other school characteristics, class size and proportion of students tested had weaker associations with value-added measures. The only exceptions were for class size in Datasets 4 and 6.

It is worth comparing the results obtained from Dataset 4 to those reported in an earlier study (Allen et al., 2009), that were also based on the same growth period (grades 8–12). With respect to proportion of racial/ethnic minority students in the school, proportion tested, and class size, results were similar. However, with respect to school poverty level and prior mean academic achievement, results were somewhat dissimilar: in Dataset 4 school poverty level has larger negative effect on the value-added measures (ranging from $b = -0.54$ to $b = -0.36$ vs. $b = -0.23$ to $b = -0.09$), but prior mean academic achievement level and the value-added measures are not related. A negative association was reported in the earlier study (ranging from $b = -0.37$ to $b = -0.22$).

Overall, the relationships of school characteristics and value-added scores suggest that different types of schools would be expected to have different mean value-added scores. Allen et al. (2009) demonstrated that school characteristics can be used as additional covariates in the value-added model (e.g., by including them as predictors in Equation 1), which then force the school characteristics to be uncorrelated with the value-added scores.

Table 11. Beta Weights for Predicting School Value-Added Scores

Dataset	Growth period	Range of month span	Subject	Class size	Proportion tested	Poverty level	Proportion minority	Prior mean achievement (Test 1)
1	8–10	18–30	English	0.01	0.03	−0.04	−0.13	0.27
			Mathematics	0.00	0.02	−0.24	0.10	0.40
			Reading	0.01	0.03	−0.18	−0.14	0.24
			Science	0.01	0.05	−0.24	−0.06	0.27
2	9–10	10–14	English	0.02	−0.04	−0.01	−0.14	0.32
			Mathematics	0.11	0.05	−0.16	0.05	0.28
			Reading	0.10	0.08	0.04	−0.23	0.31
			Science	0.03	0.04	−0.04	−0.18	0.31
3	9–10	9–18	English	0.02	−0.05	−0.05	−0.17	0.29
			Mathematics	0.10	0.08	−0.20	0.06	0.30
			Reading	0.08	0.09	0.03	−0.25	0.32
			Science	0.03	0.03	**−0.10**	−0.18	0.30
4	8–12	30–54	English	0.26	0.04	−0.39	0.02	−0.05
			Mathematics	0.19	0.06	−0.51	0.14	0.01
			Reading	0.19	−0.03	−0.36	−0.10	**−0.06**
			Science	0.17	0.01	−0.54	0.03	−0.04
5	10–11	10–14	English	0.05	**−0.10**	−0.46	**0.11**	−0.03
			Mathematics	0.16	0.01	−0.29	**0.11**	0.24
			Reading	0.02	−0.14	−0.38	−0.09	−0.15
			Science	0.07	−0.07	−0.31	−0.04	0.03
6	10–12	9–30	English	0.33	0.05	−0.24	0.00	0.16
			Mathematics	0.24	0.08	−0.30	0.15	0.37
			Reading	0.21	−0.09	−0.27	−0.08	0.11
			Science	0.20	−0.01	−0.35	−0.06	0.18

Note. Beta weights in bold are statistically significant at p<.05. Beta weights in red are not statistically significant. All others are statistically significant at p<.01.

To What Extent Do Value-Added Scores Vary Across High- and Low-Poverty Schools?

From this point forward, I focus the analysis on Datasets 1, 4, and 6 (I refer to these as the 8–10, 8–12, and 10–12 datasets). These capture grades 8 through 12 and are very common among users of the ACT college and career readiness system.

In the previous section, I examined the correlations between several school characteristics and value-added scores and, as expected, obtained inverse relationships between school poverty level and these measures (see Table 10). I now summarize the distributions of the value-added scores for high school cohorts with high poverty rates (at least 50% of the students are eligible for free or reduced lunch) and for its complement—high school cohorts with lower poverty rates (i.e., less than 50% of the students are eligible for free or reduced lunch). Tables 12 and 13 show the distributions of the value-added measures for high school cohorts with high poverty rate and for low poverty rate, respectively. Contrasting the two tables, we can see that across all three datasets average value-added scores at the 50th percentile for low-poverty schools are all above the total group average (0), while the medians for high-poverty schools are all below the total group average. At the 75th percentile, however, value-added scores for high-poverty schools are modestly above average in all four subject areas, with higher scores in English and reading (Table 12). As we can see in Table 13, the value-added scores in low-poverty schools are higher than those obtained from high-poverty schools in all subject areas. The 75th and 90th percentiles are particularly high in mathematics compared to other subject areas. Notice that in Dataset 6, even though the standard deviation in mathematics is not the largest compared to those for the other subject areas, the value-added scores at the 75th and 90th percentiles are the largest. Differences in shape of the value-added distributions across growth periods and subject areas may have caused the discrepancy between standard deviations and percentiles.

Table 12. School Value-Added Score Distributions, High Poverty

Dataset	Growth period	Range of month span	Subject	*SD*	P_{25}	P_{50}	P_{75}	P_{90}
1	8–10	18–30	English	0.66	−0.50	−0.09	0.34	0.73
			Mathematics	0.72	−0.77	−0.33	0.14	0.62
			Reading	0.78	−0.66	−0.22	0.25	0.75
			Science	0.62	−0.60	−0.20	0.17	0.56
4	8–12	30–54	English	1.00	−1.05	−0.47	0.20	0.88
			Mathematics	0.89	−1.13	−0.53	0.08	0.61
			Reading	0.93	−1.00	−0.39	0.24	0.78
			Science	0.84	−1.06	−0.48	0.08	0.52
6	10–12	9–30	English	0.85	−0.86	−0.33	0.20	0.72
			Mathematics	0.73	−0.89	−0.45	0.01	0.48
			Reading	0.81	−0.82	−0.34	0.16	0.66
			Science	0.76	−0.87	−0.40	0.07	0.49

Note. n = 5,095 high school cohorts for Dataset 1, n = 1,484 for Dataset 4, and n = 4,739 for Dataset 6.

Table 13. School Value-Added Score Distributions, Low Poverty

Dataset	Growth period	Range of month span	Subject	*SD*	P_{25}	P_{50}	P_{75}	P_{90}
1	8–10	18–30	English	0.61	−0.24	0.11	0.49	0.86
			Mathematics	0.81	−0.37	0.14	0.69	1.24
			Reading	0.72	−0.30	0.15	0.60	1.04
			Science	0.62	−0.25	0.14	0.54	0.93
4	8–12	30–54	English	1.13	−0.43	0.26	0.98	1.70
			Mathematics	1.18	−0.42	0.32	1.19	2.00
			Reading	0.98	−0.36	0.28	0.85	1.39
			Science	0.94	−0.26	0.32	0.93	1.48
6	10–12	9–30	English	0.89	−0.40	0.19	0.76	1.28
			Mathematics	0.85	−0.31	0.23	0.84	1.39
			Reading	0.77	−0.35	0.16	0.64	1.07
			Science	0.72	−0.23	0.25	0.70	1.09

Note. n = 6,090 high school cohorts for Dataset 1, n = 2,877 for Dataset 4, and n = 11,185 for Dataset 6.

How Consistent Are a School's Value-Added Scores?

I now assess across growth periods the association of the value-added measures for the same school year (Table 14) and for the same cohort of students (Table 15). Next, I examine the reliability of value-added measures by looking at the correlations of these measures for adjacent cohorts for the same growth period (Table 16). The correlations are weighted according to the average sample size (across cohorts) for each high school.

Correlations across growth periods for the same school year. Table 14 contains same-school-year correlations of school value-added scores obtained by correlating value-added scores of pairs of datasets where Test 2 was taken in the same academic year. The correlations involving the 8–12 dataset are highest, ranging from 0.49 in English to 0.77 in mathematics (with the 8–10 dataset) and ranging from 0.74 in reading to 0.89 in mathematics (with the 10–12 dataset). It is not surprising that correlations are strongest between the 8–12 and 10–12 datasets because of the large degree of overlap of student data (the 10–12 dataset is a subset of the 8–12 dataset, and so we would expect a strong correlation of school effect estimates based on the two datasets). The correlations between the 8–10 and 10–12 datasets (and the 8–10 and 8–12 datasets) are perhaps more interesting because there is no data overlap. For the 8–10 and 10–12 datasets, we see that the correlations range from 0.34 in English to 0.66 in mathematics. This suggests that school effects in mathematics, relative to the other subject areas, are more likely to permeate across different grade levels. This trend also holds when correlating the 8–10 and 8–12 datasets.

Table 14. Across Growth Periods, Same School-Year Correlation of School Value-Added Scores

Dataset	Growth period	Range of month span	Subject	$R_{8,10}$	$R_{8,12}$	$R_{10,12}$
1	8–10	18–30	English	1.00	0.49	0.34
			Mathematics		0.77	0.66
			Reading		0.53	0.35
			Science		0.66	0.51
4	8–12	30–54	English	3,505	1.00	0.82
			Mathematics			0.89
			Reading			0.74
			Science			0.85
6	10–12	9–30	English	6,564	4,243	1.00
			Mathematics			
			Reading			
			Science			

Note. Correlations are presented in the upper triangle and the average sample sizes (across cohorts) are presented in the lower triangle.

Correlations across growth periods for the same cohort of students. In Table 15, I correlate value-added scores of the two datasets with the same cohort of students. By design, the 8–12/10–12 correlations in Tables 14 and 15 are identical because the school-year and students' cohorts are the same for the two growth periods; as a result, both have the same average sample size (n = 4,243). The comparisons of interest for Table 15 are the 8–10/10–12 correlations versus the 8–10/8–12 correlations, where we observe substantially stronger associations between 8–10/8–12 growth periods. There are two plausible explanations for this: (a) projected scores are heavily influenced by prior academic achievement (Test 1), whereby Test 1 is the same in 8–10/8–12 growth periods (ACT Explore), but not the same in 8–10/10–12 (ACT Explore vs. ACT Plan); and (b) there is an overlap in school performance between 8–10/8–12 growth periods (as the former is a subset of the latter).

Table 15. Across Growth Periods, Same Cohort of Students Correlation of School Value-Added Scores

Dataset	Growth period	Range of month span	Subject	$R_{8,10}$	$R_{8,12}$	$R_{10,12}$
1	8–10	18–30	English	1.00	0.61	0.18
			Mathematics		0.81	0.54
			Reading		0.62	0.14
			Science		0.70	0.34
4	8–12	30–54	English	3,311	1.00	0.82
			Mathematics			0.89
			Reading			0.74
			Science			0.85
6	10–12	9–30	English	4,291	4,243	1.00
			Mathematics			
			Reading			
			Science			

Note. Correlations are presented in the upper triangle and the average sample sizes (across cohorts) are presented in the lower triangle.

Correlations across cohorts of students (for each growth period). Table 16 summarizes the cross-cohort correlations of the value-added measures for adjacent cohorts (one year apart), as well as for cohorts that are two and three years apart. The correlations are weighted according to the average sample size (across cohorts) for each high school. Kane and Staiger (2008) have argued that such year-to-year correlation can be interpreted as an estimate of reliability, in which case any intertemporal correlation less than 0.5 would imply that at least half of the variability in value-added scores are unrelated to school quality that persists over time. Generally, as shown in Table 16, the correlations are larger for adjacent cohorts and decrease as time between cohort increases. It appears that the 8–12 dataset generates value-added measures that have substantially greater consistency over time. In all, the correlations in Table 16 suggest that across the four subject areas, schools that have high value-added scores for one cohort are likely to have high value-added scores for future (and past) cohorts. Using the analogy from Briggs (2011) that "students are to schools what items are to students" (p. 31), I can put these correlations into context by comparing school score consistency (measured with longitudinal correlations of school value-added scores) to student score consistency (measured with longitudinal test score correlations). For example, correlations of ACT Plan and ACT scores range from 0.68 in science to 0.81 in mathematics (ACT, 2011b). The one-year correlations in school value-added scores are generally smaller than this for the 8–10 growth period, but are comparable to the longitudinal student correlations for the 8–12 and 10–12 growth periods. Similar to the student longitudinal test correlations, the school value-added correlations are affected by measurement error and by true change in performance over time. The reliability of school value-added scores is also affected by the number of tested students; thus larger schools are likely to see greater consistency in scores over time.

Table 16. Across Cohorts of Students Correlations of School Value-Added Scores

Dataset	Growth period	Range of month span	Subject	Correlation for cohorts (years apart)		
				1	2	3
1	8–10	18–30	English	0.54	0.48	0.49
			Mathematics	0.75	0.71	0.70
			Reading	0.56	0.42	0.48
			Science	0.63	0.58	0.55
4	8–12	30–54	English	0.84	0.80	0.73
			Mathematics	0.91	0.89	0.86
			Reading	0.78	0.72	0.67
			Science	0.84	0.82	0.77
6	10–12	9–30	English	0.73	0.65	0.62
			Mathematics	0.84	0.80	0.78
			Reading	0.63	0.54	0.54
			Science	0.73	0.66	0.63

Note. For Dataset 1, n = 6,538 high schools for one year between cohorts, 4,751 for two years, 3,141 for three years; for Dataset 4, n = 1,776 high schools for one year between cohorts, 1,252 for two years, 782 for three years; for Dataset 6, n = 9,067 high schools for one year between cohorts, 6,800 for two years, 4,608 for three years.

Summary

In this study, I considered three types of growth models (Standard VAM, a Standard-Classified VAM, and a Student Growth Percentile model) and evaluated them based on the accuracy with which they project growth between earlier and later assessments across six different growth periods. As the data suggested, the choice of model did not matter—different models produced very similar results when applied to the same growth period. Accordingly, I chose the Standard VAM model to address the research questions set forth in this study and based the remaining analyses on this model.

Comparing results for different growth periods, I found that using data from a growth period of grades 8–12 has relatively better statistical properties. This was expected given that it was the longest growth period, thus should be better able to differentiate schools. The findings supporting this include:

- Value-added score estimates obtained using 8–12 data have more variability across high school cohorts and require fewer students in order for the estimates to be statistically significant (see Table 6).
- At schools that are at the 75th and 90th percentiles of all schools, the proportion of students meeting or exceeding expected growth are larger for the 8–12 dataset than for those obtained from other datasets (see Table 9). Again, this is related to value-added scores based on 8–12 data having the most variability across schools.
- Cross-model same-year and cross-model same-cohort correlations of school value-added scores are strongest for the 8–12 data (see Tables 14 and 15).
- The 8–12 data generates value-added scores that have greater consistency over time compared to shorter growth periods (see Table 16).

Other important findings include:

- Score projections are heavily influenced by prior academic achievement, particularly in the same-subject, but including off-subject scores as well (see Table 5).
- Score projections increase as more time passes between the two tests. However, this effect is much smaller than the effect of prior academic achievement (see Table 5).
- Of school characteristics, prior mean academic achievement is positively related to the value-added measures, whereas school poverty level and proportion of racial/ethnic minority students have inverse relationships. Generally, compared to other school characteristics, class size and proportion of students tested had weaker associations with value-added measures (see Table 10).
- The importance of school characteristics varied by growth periods. When the ACT is the outcome variable, poverty level and class size tend to be more predictive of value-added scores. When ACT Plan is the outcome variable, prior mean academic achievement tend to be more predictive (see Table 11).
- Value-added scores in low-poverty schools are higher than those obtained from high-poverty schools in all subject areas (see Tables 15 and 16).

In each model, I used four subject-specific initial test scores (English, mathematics, reading, and science) to control for prior academic achievement. Some value-added models use multiple years of data from multiple subject areas to generate value-added scores. However, I found that projection accuracy was only slightly enhanced by using test scores from multiple prior years—using both ACT Explore and ACT Plan scores as covariates versus using only ACT Plan scores as covariates (see Table 7).

Conclusion

The point is made on several grounds that the grades 8–12 data produce more reliable results than those obtained from grades 10–11 and grades 10–12 data. But collecting grades 8–12 data for all students might not be feasible in many circumstances. Additionally, the 8–12 data are not ideal because they typically come from different schools; as a result, it is difficult to assess when and where growth actually occurs (e.g., if it is occurring in grade 9, 10, or 11). Thus, in considering a growth period to evaluate school effects (value-added measures) one must be cognizant of the availability of data for the assessment spans (prior and later test scores) and the timeline for assessing the growth.

Most school effects are not statistically significantly different from the "average" school effect and cannot usually be reliably distinguished from "average" (see Table 8). In other words, the confidence intervals for most schools' value-added scores overlap 0. This is more often the case for small than for large schools.

Overall, the relationships of school characteristics and value-added scores suggest that different types of schools would be expected to have different mean value-added scores. In other words, value-added scores depend on characteristics over which schools have no control (e.g., poverty), thus limiting the usefulness of a "one-size-fits-all" approach to VAM.

In this study, I examined high schools' effects on student achievement in four subject areas (English, mathematics, reading, and science), by controlling only for students' prior academic achievements.

Because academic achievement can be influenced by student's psychosocial (that is, psychological and social) development, additional research is needed to explore the predictive strength of school value-added measures by introducing covariates in the models that account for not only academic but also nonacademic factors. Research has shown that academically related psychosocial behaviors such as motivation, social connectedness, school attendance, obedience of rules, and avoidance of drugs are important predictors of academic success in middle school and high school (Kaufman, et al., 1992; Rumberger, 1995; Worrell et al., 2001; Jones et al., 2006). Other beneficial academic behaviors include academic discipline (i.e., good work and study habits, such as consistently completing homework), orderly conduct, and positive relationships with school personnel (Casillas, Robbins, & Schmeiser, 2007). ■

Appendix A

Table A1. Growth Model Projection Parameters for Predicting ACT from ACT Explore: Based on Standard-Classified Value-Added Model (for Dataset 4)

	Subject							
	English		Math		Reading		Science	
Effect	*N*	Estimate	*N*	Estimate	*N*	Estimate	*N*	Estimate
Intercept		6.44		14.92		7.97		9.19
exp_sspts1 (1–7)	1,672	−9.04	4,715	−8.94	699	−6.27	1,469	−4.02
exp_sspts2 (8–9)	14,919	−9.43	4,232	−9.16	9,177	−6.51	1,958	−4.31
exp_sspts3 (10–11)	60,514	−8.71	23,554	−9.14	49,966	−6.55	3,504	−4.29
exp_sspts4 (12–13)	94,500	−7.36	41,214	−8.93	100,178	−6.03	26,015	−4.47
exp_sspts5 (14–15)	87,184	−6.02	121,544	−8.14	133,053	−4.83	89,444	−4.26
exp_sspts6 (16–17)	93,038	−4.82	152,415	−6.29	85,830	−3.40	159,436	−3.54
exp_sspts7 (18–19)	63,592	−3.64	101,655	−3.98	56,584	−2.41	136,062	−2.48
exp_sspts8 (20–21)	48,331	−2.56	28,455	−2.36	44,063	−1.48	62,388	−1.47
exp_sspts9 (22–23)	45,086	−1.29	31,156	−1.32	23,740	−0.75	23,820	−0.74
exp_sspts10 (24–25)	16,358	0	16,254	0	21,904	0	21,098	0
exp_nspts1	exp_m = 0.31		exp_e = 0.16		exp_e = 0.40		exp_e = 0.18	
exp_nspts2	exp_r = 0.31		exp_r = 0.06		exp_m = 0.19		exp_m = 0.43	
exp_nspts3	exp_s = 0.28		exp_s = 0.33		exp_s = 0.33		exp_r = 0.18	
span	0.11		0.06		0.05		0.05	

Note. Subject-specific prior test score is denoted by sspts and non-specific prior test score by nspts. The ranges of the ACT Explore scores for each category are given in the parentheses.

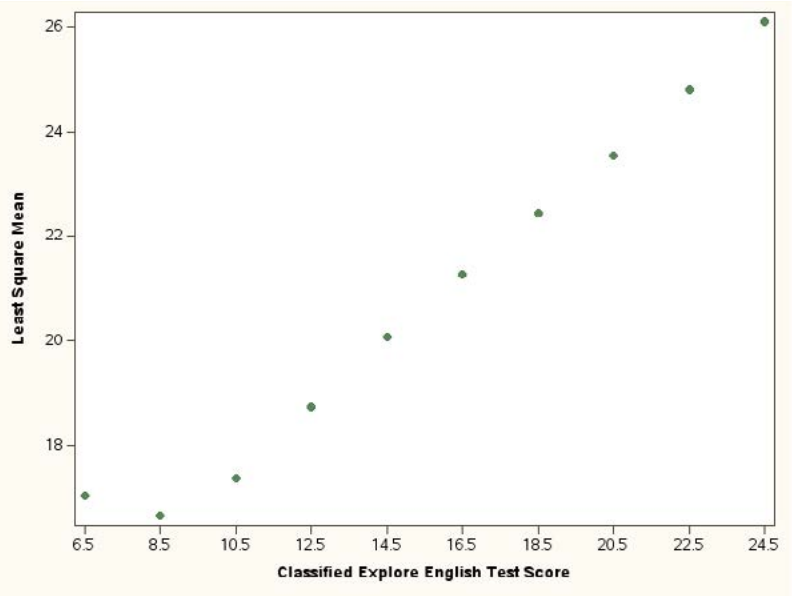


Figure A1. Estimated Least-Square Mean ACT English Score by ACT Explore English Score Level

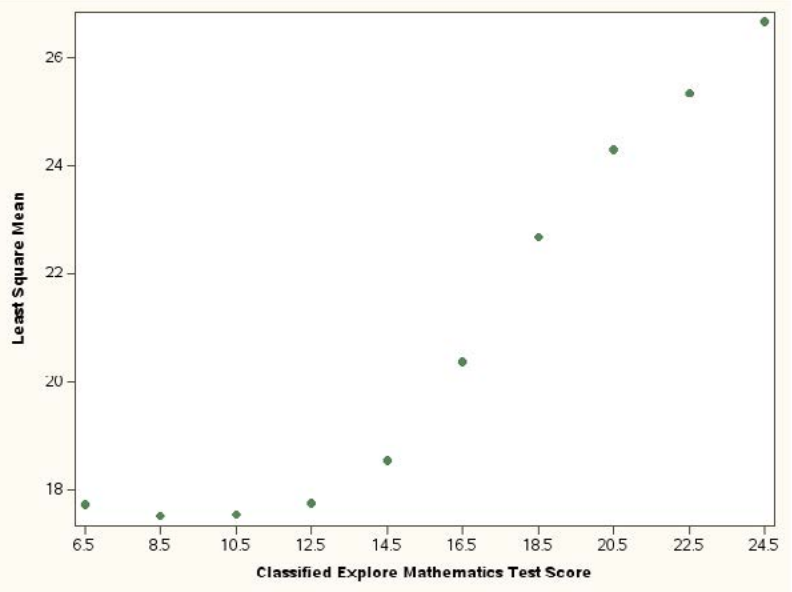


Figure A2. Estimated Least-Square Mean ACT Mathematics Score by ACT Explore Mathematics Score Level

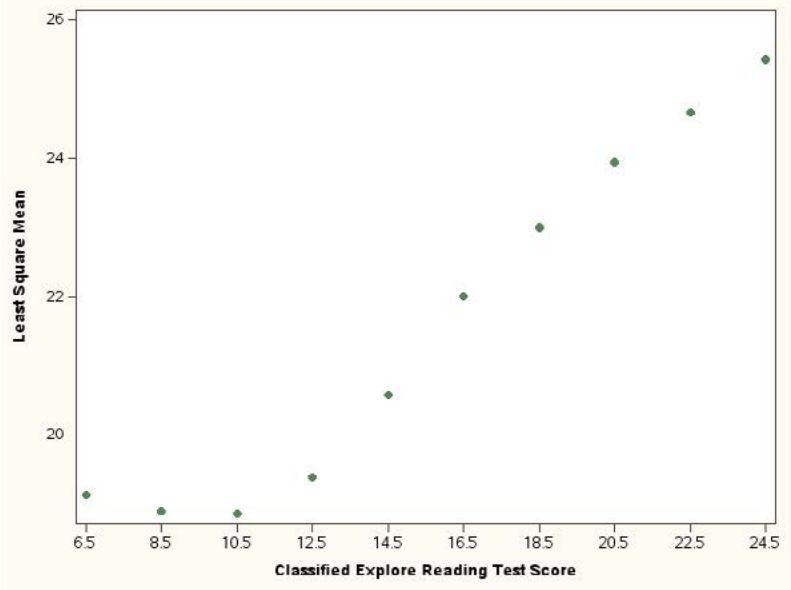


Figure A3. Estimated Least-Square Mean ACT Reading Score by ACT Explore Reading Score Level

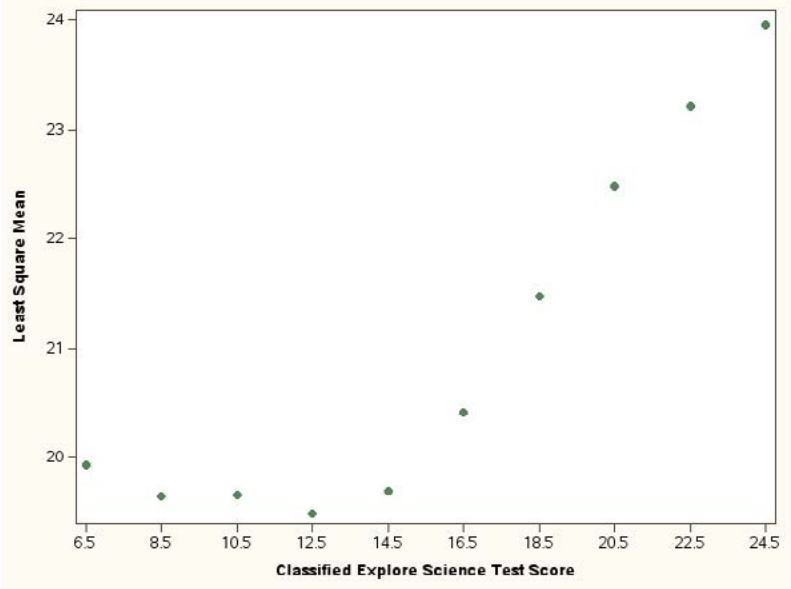


Figure A4. Estimated Least-Square Mean ACT Science Score by ACT Explore Science Score Level

Appendix B

- Generally, standard errors of means are proportional to $(1/\sqrt{n})$, where n is the sample size. For each high school cohort, the standard error of the value-added score (SE) is a function of the standard deviation of residual scores (s) and the sample size (n): $SE = s(1/\sqrt{n})$. The typical relationship between SE and n can be approximated by $SE = K(1/\sqrt{n})$, where K is the median standard deviation of residuals across high school cohorts.
- A z-ratio is established for a value-added score at the 75th (or 90th) percentile as $z = VA_{75\%or90\%}/SE$, where $z = \pm 1.96 \cong \pm 2.0$. I then replace z with 2 and solve for SE.
- The formula for the typical sample size needed to have a statistically significant value-added score is established replacing SE in the equation given in part 1 by $VA_{75\%or90\%}/2$ (or simply $VA/2$). Then, solving for n, I obtain $n = \left(\frac{K}{\frac{VA}{2}}\right)^2$.

References

ACT. (2011a). *EXPLORE technical manual*. Iowa City, IA: Author.

——. (2011b). *PLAN technical manual*. Iowa City, IA: Author.

——. (2007). *The ACT technical manual*. Iowa City, IA: Author.

Allen, J., Bassiri, D., & Noble, J. (2009). *Statistical properties of accountability measures based on ACT's educational planning and assessment* (ACT Research Report 2009-1). Iowa City, IA: ACT.

American Psychological Association. (2012). *Facing the school dropout dilemma*. Washington, DC: Author. Retrieved from http://www.apa.org/pi/families/resources/school-dropout-prevention.aspx

American Statistical Association. (2014). ASA statement on using value added models for educational assessment. Alexandria, VA: Author. Retrieved from http://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Amrein-Beardsley, A. (2014). Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability. New York, NY: Routledge.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*, 37–65.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42–51.

Betebenner, D. W. (2011). *A technical overview of the student growth percentile methodology: Student growth percentile and percentile growth projections/trajectories*. National Center for the Improvement of Educational Assessment, Dover, NH.

Bill & Melinda Gates Foundation. (2012). *Gathering feedback for teaching*. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). *Measuring effect sizes: The effect of measurement error*. National Center for Analysis of Longitudinal Data in Education Research (CALDER), Working Paper 19. Retrieved from http://www.wcer.wisc.edu/news/events/VAM%20Conference%20Final%20Papers/MeasuringEffectSizes_BoydEtAl.pdf

Braun, H., Chudowsky, N., & Koenig, J. (eds.). (2010). *Getting value out of value-added*. Report of a workshop. Washington, DC: National Research Council, National Academies Press.

Braun, H. (2015). The value in value added depends on the ecology. *Educational Researcher*, *44*(2), 105–16. doi: 10.3102/0013189X15576341

Briggs, D. C. (2011). *Making inferences about growth and value added: Design issues for the PARCC consortium*. Retrieved from http://parcconline.org/files/40/Technical%20Advisory%20Committee/49/Growth-and-Value-Added-Briggs.pdf

Buddin, R. (2012). *Implications of educational attainment trends for labor market outcomes.* (ACT Research Report 2012-7). Iowa City, IA: ACT.

Casillas, A., Robbins, S., & Schmeiser, C. (2007). *Developing academic and retention risk indicators for middle school students: The high school readiness inventory*. Unpublished manuscript.

Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models*. Washington, DC: Council of Chief State School Officers.

Castellano, K. E., & Ho, A. D. (2015). Practical differences among aggregate-level conditional status metrics: From median student growth percentiles to value-added models. *Journal of Educational and Behavioral Statistics*, *40*(1), 35–68.

Chapman, L., Laird, J., & KewalRamani, A. 2011. *Trends in high school dropout and completion rates in the United States: 1972–2009* (NCES 2012-006). US Department of Education, National Center for Education Statistics.

Chen, C. L. (2005). *An Introduction to Quantile Regression and the QUANTREG Procedure.* Retrieved from http://www2.sas.com/proceedings/sugi30/213-30.pdf

Cody, C. A., McFarland, J., Moore, E., & Preston, J. (2010). *The evolution and use of growth models.* Raleigh, NC: Financial and Business Services Internship Program, Public Schools of North Carolina. Retrieved from http://www.ncpublicschools.org/docs/intern-research/reports/growth.pdf

Dahl, G. B., Løken, K. V., & Mogstad, M. (2013). *Peer effects in program participation*. Retrieved from http://econweb.ucsd.edu/~gdahl/papers/peer-effects-in-program-participation.pdf

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, R. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, *93*(6), 8–15.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2012). *Selecting growth measures for school and teacher evaluations* (National Center for Analysis of Longitudinal Data in Education Research Working Paper 80). Retrieved from http://scee.groupsite.com/uploads/files/x/000/086/488/CALDERrpt.pdf

Gladwell, M. (2013). *David and Goliath: Underdogs, misfits, and the art of battling giants.* New York, NY: Little, Brown & Company.

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: Brookings Institution.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*, *80*(319), 589–612.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, *44*, 96–104.

Goldschmidt, P., Choi, K., & Beaudoin, J.P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance*. Washington, DC: Council of Chief State School Officers (CCSSO).

Gordon, R., Kane, T., & Staiger, D. (2006). *Identifying effective teachers using performance on the job*. Washington, DC: Brookings Institution.

Haertel, E. H. (2013). *Reliability and validity of inferences about teachers based on student test scores.* William H. Angoff Memorial Lecture Series presentation. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/PICANG14.pdf

Hanushek, E.A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, 24, 1141-77.

Hanushek, E.A. (1997). Assessing the effects of school resources on student performance: An Update. *Educational Evaluation and Policy Analysis*, 19(2), 141-64.

Hanushek, R., & Rivkin, S. (2004). How to improve the supply of high-quality teachers. In D. Ravitch (Ed.), *Brookings papers on education policy* (pp. 7–25). Washington, DC: Brookings Institution.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy*, *4*(4), 319–350.

Harris, D. N. & Herrington, C. D. (2015). The use of teacher value-added measures in schools: New evidence, unanswered questions, and future prospects. *Educational Researcher*, *44*(2), 271–276. doi:10.3102/0013189X15576142

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform: Chicago's REACH students. *Educational Researcher*, *44*(2), 105–116. doi: 10.3102/0013189X15575517

Jones, K. K., & Byrnes, J. P. (2006). Characteristics of students who benefit from high-quality mathematics instruction. *Contemporary Educational Psychology*, *31*, 328–343.

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

Kane, T., J. & Douglas O. Staiger, D., O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (National Bureau of Economic Research Working Paper 14607). Cambridge, MA: National Bureau of Economic Research.

Kaufman, P., & Bradbury, D. (1992*). Characteristics of at-risk students in NELS: 88* (NCES 92–042). Washington, DC: US Department of Education, National Center for Educational Statistics.

Keaton, P. (2012). *Documentation to the NCES common core of data public elementary/ secondary school universe survey: School year 2010–11* (NCES 2012-338rev). Washington, DC: US Department of Education, National Center for Education Statistics.

Koedel, C., Leatherman, R., & Parsons, E. (2012). *Test measurement error and inference from value-added models*. Columbia: University of Missouri, Department of Economics.

Krueger, A. B. (2003). Economic considerations and class size. *Economic Journal*, 113(485).

Lockwood, J. R., & McCaffrey, D. F. (2012). *Reducing bias in observational analyses of education data by accounting for test measurement error*. Unpublished manuscript. RAND Corporation.

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

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children*, 5(2), 113-127.

National Research Council. (2010). *Getting value out of value-added*. Washington, DC: National Academies Press.

No Child Left Behind Act of 2001. (2002). 107th Congress: Public Law 107-110, 115 Stat. 1425. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Riddle, W. C. (2008). *Adequate yearly progress (AYP): Growth models under the No Child Left Behind Act.* Retrieved from http://congressionalresearch.com/RL33032/document.php?study=Adequate+Yearly+Progress+AYP+Growth+Models+Under+the+No+Child+Left+Behind+Act

Rumberger, R. W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal*, *32*(3), 583–625.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid educational measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

Sanders, W. L. (2006). *Comparisons among various educational assessment value-added models*. Columbus, OH: SAS Institute.

Schanzenbach, D. W. (2014). *Does class size matter?* Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/publication/does-class-size-matter.

Snidjers, T. A. B., & Bosker, R. J. (1999). *Multilevel modeling: An introduction to basic and advanced multilevel modeling.* London: Sage Publications.

Wainer, H. (Ed.) (2004). Special issue of the *Journal of Educational and Behavioral Statistics* on value-added assessment. *Journal of Educational and Behavioral Statistics*, *29*(1).

Willms, J. D. (1986). Social class segregation and its relationship to student's examination results in Scotland. *American Sociological Review*, *51*, 224–241.

Worrell, F. C., & Hale, R. L. (2001). The relationship of hope in the future and perceived school climate to school completion. *School Psychology Quarterly*, *16*, 370–388.

ACT is an independent, nonprofit organization that provides assessment, research, information, and program management services in the broad areas of education and workforce development. Each year, we serve millions of people in high schools, colleges, professional associations, businesses, and government agencies, nationally and internationally. Though designed to meet a wide array of needs, all ACT programs and services have one guiding purpose—helping people achieve education and workplace success.

For more information, visit www.act.org.

* 0 5 0 2 0 5 1 5 0 * Rev 1