ACT Research Report Series 2016 (2)

Size and Consistency of Relative Performance Measures of School Districts Across Models, Subjects, and Grade Levels

Chrys Dougherty
Teresa Shaw

Chrys Dougherty is a principal research scientist in Statistical and Applied Research at ACT.

Teresa Shaw is a statistical project consultant with Pastor, Behling & Wheeler, an environmental consulting firm. From 2007 to 2015, she worked as a senior software engineer and statistical analyst at ACT and researched performance growth and value-added modeling for schools and districts.

Acknowledgments

The authors would like to thank Linda Hiserote for creating SAS programs that were modified to construct the study datasets, and Jeff Allen, Dick Buddin, Michelle Croft, Dan Wright, Joanne Moore, Ty Cruce, and Richard Sawyer for their helpful comments on drafts of the report. In addition, we would like to thank the staff of the Arkansas Department of Education for making available the data that made this report possible.

 5293

Contents

Abstract

One purpose of statistical models in education is to identify schools and districts that perform above or below average, given their students' characteristics. This information can be used in further inquiry about the relationship between school- or district-wide student performance and educator practices. This report looks at the extent to which the relative performance of school districts is similar across different statistical models (status and value-added models, with and without district-level predictors), different subjects (English, reading, mathematics, and science), and different grade levels (grades 4, 8, and 11–12).

We find that results differ substantially across statistical models, subjects, and grade levels. We conclude that educators, researchers, and policymakers must take care in interpreting the statistical models that are used, and that "overall district performance" is a less useful idea than district performance in specific subjects and grade levels.

Introduction[1]

A substantial body of literature focuses on the role of public school districts in establishing conditions that support improvement in classroom instruction (ACT, 2012a; Supovitz, 2006; Hightower, Knapp, Marsh, & McLaughlin, 2002; Marsh et al., 2005). School districts have been a significant focus of education reform efforts (Whitehurst, Chingos, & Gallaher, 2013; Chingos, Whitehurst, & Gallaher, 2013) and public recognition (Broad Foundations, 2014).

School districts are important because they are often the smallest administrative unit that addresses the needs of students from preschool through grade 12 (P–12). School district leaders are in a position to create systems to improve teaching and learning that span all of those grades (ACT, 2012a). For example, they can work to ensure students receive a content-rich curriculum aligned across elementary, middle, and high school levels; establish assessment and data systems to monitor student progress; promote educators' use of those systems (Dougherty, 2015); ensure that time is set aside for teachers to collaborate; develop a coaching system for teachers; and lead efforts to involve parents and community leaders. These district-led efforts to improve teaching and learning throughout P–12 are likely to be particularly important for economically disadvantaged students, who are more likely to start out far behind academically and have more trouble catching up when they are behind (ACT, 2012b; Dougherty, 2014).

To guide their improvement efforts, school district and community leaders can benefit from using indicators that give them feedback on their students' progress and their success in creating conditions in which further student progress is likely. These indicators might fit into four categories:

- *Academic performance indicators*: ways of monitoring students' academic performance and progress in different subject areas as they move through the grades, and whether students are ready for college and other postsecondary learning opportunities by the time they graduate from high school. Performance indicators can include information about student learning, such as test scores, and information about student completion of academic requirements, such as course completion, high school graduation, and acquisition of postsecondary degrees and certificates.[2]
- *Behavior, attitude, and climate indicators*: ways of monitoring students' ongoing development of desired academic and social behaviors and their desire for further learning, as well as the quality of the learning environment in each school. Sample indicators of this type may include student attendance records, discipline reports, and results from surveys of students, teachers, or parents.
- *Practice indicators*: ways of monitoring whether educators across the district are making progress in developing systems to improve teaching and learning. Sample indicators may include results from classroom observations, curriculum audits, school quality reviews, and educator surveys.
- *Community environment indicators*: ways of monitoring the extent to which interventions in students' out-of-school lives are creating a more favorable environment for student learning. Sample community environment indicators might include results from parent and community

[1] This study uses data maintained by the Arkansas Department of Education and is published with its permission.

[2] Grades are often a mixed indicator representing student learning, on-time completion of assignments, and sometimes student behavior (e.g., class participation).

surveys about neighborhood safety, housing, and health care, and surveys of students about their use of learning opportunities outside of school. Examples of out-of-school interventions might be summer programs for disadvantaged students to reduce summer learning loss, and housing assistance, family counseling, and greater availability of health care, all of which might reduce student mobility and increase student attendance in school.[3]

Table 1 contains examples of different types of district and school performance indicators, with an emphasis on indicators based on student test scores.[4] Comparing the top and bottom rows of Table 1, the indicators in the top row are based on information about a group of students at a single moment in time, often referred to as status or snapshot data (Data Quality Campaign, 2010). Indicators in the bottom row are based on longitudinal data following individual students over time. Either status or longitudinal data may be aggregated across multiple cohorts (for example, fifth-grade test scores in a school averaged across three years), without changing the status or longitudinal character of the data.

Comparing the right and left sides of Table 1, the indicators in the right side are based on comparisons across different student cohorts, whereas the indicators on the left side do not compare cohorts.[5] The usual reason to compare student cohorts is to assess improvement over time. Thus, the upper and lower right quadrants of the table address improvement over time in snapshot and longitudinal indicators, respectively. For example, the lower right quadrant might contain an analysis of whether students in more recent cohorts experience more rapid growth than did students in earlier cohorts.

Table 1. Examples of District and School Academic Performance Indicators

	Current Period (1–3 years)		Trends Over Time	
	A. Our School or District	B. Our School or District Relative to Others	C. Our School or District	D. Our School or District Relative to Others
1. Status (does not use information on prior achievement)	1A. Average test scores; distribution of students across achievement categories	1B. Comparison of average scores or proficiency rates with those in comparable schools or districts; actual vs. predicted scores using statistical models with student demographics as predictors	1C. Trends over time in average scores; trends in distribution of students across achievement categories	1D. Trends over time in actual vs. predicted scores using statistical models with student demographics as predictors
2. Growth, value-added, or change in achievement level over time (uses information on prior achievement)	2A. Average growth; percentages of below-benchmark students reaching benchmarks; other results from transition tables[6]	2B. Comparison of growth with that in comparable schools or districts; actual vs. predicted scores using statistical models with prior achievement scores and student demographics as predictors	2C. Trends over time in average growth or in percentages of below-benchmark students reaching benchmarks	2D. Trends over time in actual vs. predicted scores using statistical models with prior achievement and student demographics as predictors

[3] If the summer program is operated by the school district rather than an outside entity, it should be classified as a district practice and addressed under practice indicators.

[4] Many of the indicators in Table 1 can be disaggregated by student demographic and prior achievement groups. For example, trends over time in average scores can be disaggregated by income and ethnic groups, and trends in student growth can be disaggregated for students who were on track, off track, or far off track in the prior year. See Dougherty (2014) and Dougherty, Hiserote, and Shaw (2014a, 2014b) for examples of disaggregating students by prior achievement level.

[5] One can aggregate cohorts without comparing them (left half of the table) or compare sets of aggregated cohorts (right half of the table). For example, student growth in the combined 2007–11 and 2008–12 cohorts might be compared with growth in the combined 2009–13 and 2010–14 cohorts. Such a comparison would go in the lower right quadrant of Table 1.

[6] Transition tables show the number and percentage of students at each prior-year achievement level who ended up at each current-year achievement level. We put these in the "growth" category (second row) although they can also be thought of as status measures disaggregated by students' prior achievement levels.

In addition, each of the four quadrants of the table is divided into two cells, based on whether the performance of educational institutions and jurisdictions is being compared. Thus, cell 1A contains indicators of snapshot student performance in a single school or district, whereas cell 1B entails comparisons of snapshot student performance across schools and districts.

In this paper, we focus on snapshot and longitudinal indicators of the academic performance of school districts relative to others (cells 1B and 2B in Table 1), which we refer to as *district relative performance statistics* or just *district performance statistics*. Educators and policy influencers should be interested in both the size and consistency of these statistics. Size addresses the question: If we rank districts based on the performance statistics, how much better do students in top-ranking districts do relative to similar students in middle- or bottom-ranking districts? These performance differences can reflect practices by educators in the districts and also unmeasured student, parental, and community influences that were not picked up as controls in the statistical analysis (Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010). Since the statistics do not in themselves tell the story of which behaviors are responsible for them in what proportions, they should be treated as the starting point of a cycle of inquiry about which educator practices and community interventions most affect student performance.

If differences in relative district performance are large, then the combination of educator practices and out-of-school influences would appear to be having a much more positive impact in some communities than in others. If the relative performance differences are small, then the effects of these practices and influences would appear to be similar across school districts. This does not rule out the possibility of improvement through changes in practices, but might focus attention on a different set of practices—for example, practices that are promising but rarely implemented.[7]

In this report, *consistency* addresses the question: Do district relative performance statistics vary markedly across subject areas, grade levels, or the choice of statistical models? District performance might vary across *subject areas* if, for example, a district's mathematics program is better than average but its reading program is not.[8] Relative performance might vary across *grade levels* if, for example, the district has unusually effective elementary schools but its high schools are average. When relative district performance varies across subjects and grade levels, educators may want to search for explanations that are likely to differ across subjects and levels—for example, classroom practices and the quality of support provided to teachers may differ across subjects, or high schools may face different challenges than elementary schools and therefore be either more or less responsive to district improvement initiatives. In general, educators should pay more attention to performance differences across districts in a given subject and grade level when those differences are relatively large, persistent over time, and consistent across statistical models. Differences in relative district performance across subjects and levels should also affect policymakers' ideas about how to rate district performance. If a district is above average in high school mathematics but below average in elementary school reading, assigning the district a single performance rating might hide too much information.

[7] Promising but rarely-implemented practices might also be the focus if performance of some student groups falls below desired levels in all districts, including those with the best performance of students from the group in question.

[8] In this report we use the terms "subject areas" or "subjects" interchangeably to refer to academic subjects such as English language arts, mathematics, or science.

Why relative performance might vary across *statistical models* for the same subject and grade level may be understood as follows: Educators and state officials typically recognize that comparing schools and districts based on average test scores and proficiency percentages gives an edge to districts that begin with more advantaged students. To level the playing field and create comparisons more likely to reflect differences in what educators are doing, statistical models may be used to adjust these comparisons for differences in student populations. To the extent that different models adjust for different population characteristics or make these adjustments in different ways, those models may show different results. Variation across statistical models would require us to pay close attention to what is being adjusted for when the districts are being compared.

In addition to varying across subjects, grade levels, and statistical models—the three kinds of variation that are the focus of this report—relative district performance may also vary across other dimensions. For example, performance may vary over time (student cohorts) and across different student demographic or prior achievement groups, so that today's higher-performing districts may be average or lower-performing tomorrow, and districts that are more effective than average with academically advanced students may be less effective than average with students who start out academically behind.[9] These kinds of variation will be the focus of future reports in this series.

Our research on the size of district relative performance differences is intended to complement prior research in the size of district effects (Whitehurst, Chingos, & Gallaher, 2013; Chingos, Whitehurst, & Gallaher, 2013). Our research on consistency of relative performance is intended to complement prior research on the consistency of performance statistics calculated at the school or teacher level. Several studies have examined the consistency of these statistics across subjects and years (Mandeville & Anderson, 1987; Mandeville, 1988; Luyten, 1994; Luyten, 1998; Ma, 2001; Mohiemang & Pretorius, 2012; Abe et al., 2015) and across statistical models (Tekwe et al., 2004; Abe et al., 2015). Other studies have looked at the consistency of teachers' value-added statistics over time (Sass, 2008; McCaffrey, Sass, Lockwood, & Mihaly 2009; Lissitz, 2012; Haertel, 2013), across statistical models (Newton, Darling-Hammond, Haertel, & Thomas, 2010; Lissitz, 2012; Goldhaber & Theobald, 2013), and across subject areas (Lissitz, 2012).

Data

This report used student-level enrollment and test data supplied by the Arkansas Department of Education for the 2006–07 through the 2012–13 school years. The test data we used included student scores on the state-administered Arkansas Benchmark Exams (ABE) in grade 4, the ACT Explore® tests in grade 8, and the ACT® in grades 11 and 12. All enrollment and test records had state-encrypted student IDs so that they could be linked anonymously at the student level to follow multiple cohorts of students from kindergarten through grade 4, from grade 4 through grade 8, and from grade 8 through grade 12.

Our dataset construction had three steps: (1) create cohorts of students to be included in the analysis, applying rules for coding students into demographic categories; (2) calculate district-level statistics and apply rules for including districts in the analysis; (3) merge district- and student-level data into files for hierarchical modeling.

[9] For example, Sawyer (2013) looked at changes in school performance over time (cell 1C of Table 1), while Whitehurst, Chingos, and Gallaher (2013) looked at changes in relative district performance (cell 1D).

Students in the Analysis

This section describes which students were included in the models and how students' demographic characteristics were derived from the Arkansas state enrollment and test data.

Creation of student cohorts. To determine what student cohorts to create, we decided to focus on academic achievement in the waypoint grades of 4, 8, and 12, using data from the ABE tests in grade 4, the ACT Explore tests in grade 8, and the ACT in grades 11 and 12. We were interested both in student achievement levels in those grades and in student growth between grades 4 and 8 and grades 8 and 11–12. In order to focus the analysis on students who have been in the district's instructional program over multiple years, we identified students who had been enrolled in the same district for periods of four or more years, e.g., since the fall of kindergarten for fourth-grade students, the fall of fourth grade for eighth graders, and the fall of eighth grade for twelfth graders.

An alternative approach would be to use an accountability subset—all students who were enrolled in the same district for at least the current school year, including those who were in a different district in the previous year. The single-year accountability subset would be more appropriate for a model of growth over one year, as opposed to one that assesses students who are educated in the district over multiple years. Some students in the accountability subset in grades 8 and 12 do not have prior score information from grades 4 and 8.

Our cohort creation process worked as follows: In the analyses using fourth-grade test results, we began with students enrolled in the district in kindergarten in the 2006–07, 2007–08, and 2008–09 school years (the three initial cohort years) and kept students who also took both the literacy and mathematics ABE tests in the fourth grade four years later, in the 2010–11, 2011–12, and 2012–13 school years (the three final cohort years), and were enrolled in the same district from kindergarten through fourth grade. This had the effect of excluding students who did not progress from kindergarten through fourth grade in four years, e.g., students retained in a grade during the cohort period and were not double-promoted later to catch up. However, a student retained in kindergarten but progressing normally after that could still be part of a cohort based on the student's second kindergarten year. The same logic applied for retained fourth graders in the grades 4–8 cohorts and retained eighth graders in the 8–12 cohorts.

For eighth graders, we began with students enrolled in fourth grade in each of the three initial cohort years, and kept students enrolled in the same district during the entire cohort period who took both fourth-grade ABE tests in the initial cohort year and all four ACT Explore tests (in English, mathematics, reading, and science) four years later in the final cohort year. Finally, for twelfth graders, we began with students enrolled in eighth grade in each of the three initial cohort years and kept students enrolled in the same district during the entire cohort period who took all four ACT Explore tests in the initial year and all four ACT tests (in English, mathematics, reading, and science) in grade 12 in the final cohort year or in grade 11 in the next-to-final year.

This process created three longitudinal cohorts at each level, referred to as the 2007–11, 2008–12, and 2009–13 cohorts based on the initial cohort year and the final cohort year four years later. At each level (K–4, 4–8, and 8–12), we concatenated the three student-level cohorts into a single dataset in order to avoid double-counting students who ended up in multiple cohorts because they were retained in the initial cohort grade and to create an indicator variable for those retained students.

Identification of students' demographic and program participation status. Students' characteristics may vary naturally over time: For example, a student's family may qualify for the free and reduced-price lunch program when the student is in fourth grade but not when the same student is in eighth grade. Likewise, a student's special education and English language learner (ELL) status or a student's self-identified ethnicity may change over time. Because an indicator of low-income, ELL, or special education status may signal a level of disadvantage even if the status is not consistent every year, we identified students as low-income, ELLs, or special education if they had that status in either the initial or final cohort grade level, e.g. either kindergarten or fourth grade for the students in the grade 4 analysis. Because no such logic applies to inconsistent reporting of student ethnic status, we used the student's reported ethnicity in the earliest cohort year as the determining factor for the student's overall ethnic status.[10] Students whose ethnicity, low-income status, special education status, or ELL status could not be ascertained using these criteria were dropped from the analysis. Only a small percentage of records were dropped based on incomplete demographic data (Appendix A, Tables A1–A3).

Calculation of attrition rates. Student attrition is an issue when following students over multiple years. If attrition is selective—for example, higher-poverty and lower-achieving students leave the sample in disproportionate numbers—then districts with high attrition may benefit from this selection bias.

Alternatively, districts with a lower-than-typical percentage of students included in the model might have an included student population that is *more* disadvantaged in both measured and unmeasured ways. High attrition could also result from the presence of a nearby military base or from the district's being less effective at retaining and educating students. In the last of these cases, the attrition variable picks up some of the district performance we are trying to measure. Further research may explore the variables that are associated with student attrition to identify when attrition should be treated as a district performance indicator (e.g., as in the case of high school dropout rates) and when it is simply an aspect of the environment in which the district operates.

Attrition is usually thought of in a "forward" sense, as in survival analysis: Students in the initial cohort year (e.g., enrolled eighth graders in 2008–09) are followed to the final cohort year (e.g. 2012–13, when students who progress through the grades at a normal rate would be high school seniors). In this analysis, the share of students included in the analysis (a measure of cohort selectivity) is calculated using eighth-grade enrollment as the denominator, and the attrition rate may be defined as one minus the share of included students. By contrast, a backward analysis might begin with a group of students in the cohort's final year (e.g., high school seniors in 2012–13) and look back to see what percentage of those students met the requirements to be included in the analysis.[11] In the latter case, the denominator is the group of enrolled students in the final year of each cohort. In grades 8 through 12 where many students drop out, it seems clear that a forward look from eighth grade is a better approach to assessing attrition; for consistency, we used a forward look based on the enrollment in the earliest cohort grade for grades K–4 and 4–8 as well.[12]

[10] The only exception was for a student with missing ethnic data for the earliest cohort grade (e.g., kindergarten in the fourth-grade analysis) but ethnic data present for the final cohort grade level (e.g., grade 4 in the fourth-grade analysis), in which case we used the data from the final grade level.

[11] Accountability data tends to focus on a backward look, as their starting point is a group of students tested in the current year. Accountability subsets are generally the group of current year tested students who were enrolled in the school for the full academic year.

[12] A backward- or forward-looking cohort selection process can create exactly the same cohort, with the only difference being the denominator to which the size of the cohort is compared.

In all, the 55,091 students in our K–4 cohorts, 56,188 students in our 4–8 cohorts, and 25,835 students in our 8–12 longitudinal cohorts comprised respectively 50%, 52%, and 24% of the students enrolled in the initial cohort years (Appendix A, Tables A1–A3).

School Districts in the Analysis

Our analysis began with 239 K–12 school districts in Arkansas that were in existence continuously from the 2006–07 through the 2012–13 school years.[13] Since our focus was on traditional K–12 school districts, charter schools that were not part of such a district were omitted from the analysis.[14]

School district demographics. We used each district's fall student-level enrollment data for kindergarten through twelfth grade for each year from 2006–07 through 2008–09, the starting years for the cohorts in this report, to derive annual statistics on the district-wide percentage of low-income, African American, Hispanic, Asian, White, Native American, ELL, and special education students in kindergarten through twelfth grade. These K–12 district-level demographic statistics for the 2006–07, 2007–08, and 2008–09 school years were used as district-level predictors in our two statistical models that used district-level data.[15]

Accuracy of low-income statistics. The use of students' low-income status as an important control in the statistical models made the accuracy of this classification an important consideration. To assess the accuracy of each district's low-income statistics in a given year, we regressed the district's overall percentage of low-income students in that year[16] on Census estimates of poverty rates of individuals aged 5–17 in the district to get a statewide relationship between the two variables, which in turn yielded a Census-predicted district low-income percentage for each year.[17] To have its students included in the analysis, a district's percentage of low-income students in kindergarten through grade 12 had to fall within 20 percentage points of its Census-predicted value in each school year from 2006–07 through 2008–09, the starting years for the cohorts in this report. Two hundred fifteen out of 239 Arkansas K–12 districts met this requirement.

Number of students in the analysis. To be included in the analysis for a given grade level (4, 8, or 12), districts were required to have at least 20 students in the three combined longitudinal cohorts for that grade level. In grades 4 and 8, all 215 districts that met the income data requirement also met this criterion. Because of the lower ACT Explore and ACT participation rates in the grades 8–12 cohorts used for the twelfth-grade analysis, 50 of the 215 districts meeting the low-income data criterion had fewer than 20 eligible students, leaving 165 districts eligible for the analysis.

[13] If District A consolidated into District B during that time period, then A's students were combined with B's for the years prior to the consolidation and everyone was treated as part of District B.

[14] In future research, it may be possible to analyze the performance of charter management organizations (CMOs), which manage more than one charter school. We did not have data on which charter schools were managed by CMOs.

[15] The district-wide low-income student percentages calculated this way differ from those that would be calculated by aggregating our cohort data, which do not cover all grades. As was the case in our student cohorts, we dropped students with missing demographic data when calculating the district-wide statistics. We also calculated district-wide demographic statistics for the 2009–10, 2010–11, 2011–12, and 2012–13 school years, but did not use those statistics in our analysis.

[16] Census-defined poverty uses a lower income threshold than the state definition of low-income, which is based on federal eligibility requirements for the free and reduced-price school lunch program. Thus we needed to derive a predicted low-income percentage from the Census data rather than just using the Census percentage. We hypothesized that a district with accurate low-income data would have a relationship between the two poverty measures that is not too different from the state average relationship between the two measures.

[17] The ordinary least squares model used for this regression included dummy variables for the year and interaction terms between the year and the district's Census-derived poverty rate, in order to allow the statewide relationship between Census-measured and lunch-program-measured poverty to differ by year.

At the end of this process, we had created three district-level files, one per grade level (K–4, 4–8, and 8–12), containing information on each district's eligibility by criterion, and the following information for each initial cohort year (2006–07 through 2007–09): district-level K–12 demographic statistics, the number and percentage of students in the analysis, and average prior scores for the grades 4–8 and 8–12 cohorts.

Combining Student- and District-Level Data

At each grade level (K–4, 4–8, and 8–12), we merged the concatenated file containing student-level data on the three cohorts with the district-level data created in the previous step. This process created a single dataset at each level with matched student- and district-level data.

Methods

Our analysis had three steps: (1) estimate status and value-added statistical models of students' test scores, generating the district-level random effect estimates that are used as district performance statistics; (2) compare the size of these district performance statistics across statistical models, subject areas, and grade levels; and (3) compare the consistency of these performance statistics across models, subjects, and grade levels. We describe these steps here.

Statistical Models Used to Estimate District Performance Statistics

Grade 4. We used two status models to predict student-level scores on each of the two fourth-grade ABE tests for students in our grades K–4 cohorts in the 2010–11, 2011–12, and 2012–13 school years. Table 2 shows the predictors used in these models. Model 1 contained mostly student-level predictors: information on students' income, ethnic, ELL, and special education status.[18] Model 2 added district-level averages of these predictors. District-level demographic predictors might be hypothesized to represent the impact of larger district-wide concentrations of students from different groups on school district academic culture, funding, and priorities.

Both fourth-grade status models contained predictors on the district's number and percentage of students in the analysis—in order to explore effects of district size and the selectivity of each district's group of students in the analysis, respectively.[19] Selectivity may be an issue whenever there is nonrandom attrition of students, either because a nonrandom group of students drop out or transfer from the local public schools or because the researchers choose to focus on a nonrandom subset of students. In addition, dummy variables for students who took the fourth-grade test in the spring of 2012 and 2013 were included to allow for shifts in average test scores across years.

[18] These demographic and program participation variables were coded as 1 for students in the category in question and 0 for students not in the category. The earlier section on identification of students' demographic and program participation status describes how a student's status was determined when multiple years of enrollment data were available.

[19] The selectivity variable appears in Tables 2 through 4 as "district percent of students in model."

Because Models 1 and 2 contained both student- and district-level predictors, we estimated them as hierarchical linear models with district random effects.[20] The random effect in the models modified the intercept, but not the slopes associated with the predictor variables; thus the effects of these predictor variables were assumed to be the same across districts. We used these estimated random effects as the relative district performance statistic for each model and subject. This process generated four sets of grade 4 district performance statistics based on two models and two fourth-grade subjects.

In this report we use the term "status model" to refer to any statistical model that does not use information on prior achievement. Our status models were unlike state accountability reports in two major ways. First, we used student scores, not proficiency status, as the dependent variable. Second, our analysis adjusted for students' demographic characteristics and (in the case of Model 2) district demographic characteristics as well.[21] Accountability reports often do not make these adjustments because policymakers do not want to convey the message that educators are free to aim for lower test results for disadvantaged students. However, the focus in this report is not on accountability targets, but on indicators that stand a better chance of reflecting relative district effectiveness. For that purpose, it is important to take the degree of difficulty in educating students into account (as reflected by the regression coefficients associated with student characteristics), just as the degree of difficulty is taken into account in scoring Olympic gymnastics or diving events.

We did not estimate value-added models for fourth grade. Such models could have been developed based on third-grade test data. However, in this report we were interested in indicators of student growth over multiple years, such as the periods of three and a half or four years represented by our grades 4–8 and 8–12 cohorts. No statewide assessment data were available for kindergarten or first grade to make possible a value-added model reflecting student growth over three or four years.

[20] SAS Proc Mixed was used for all of the statistical models in this report. The SAS code used for the models is available on request from the first author.

[21] Therefore, one might refer to Models 1 and 2 as "statistically adjusted status models." For brevity's sake, we refer to them simply as status models with the understanding that the reader will keep the models' use of statistical predictors in mind.

Table 2. Predictors in Grade 4 Statistical Models

Variable	Model 1	Model 2
Intercept	✓	✓
Low-Income Status	✓	✓
African American Status	✓	✓
Hispanic Status	✓	✓
Asian Status	✓	✓
Native American Status	✓	✓
ELL Status	✓	✓
Special Education Status	✓	✓
District Percent Low-Income		✓
District Percent African American		✓
District Percent Hispanic		✓
District Percent Asian		✓
District Percent Native American		✓
District Percent ELL		✓
District Percent Special Education		✓
District Number Students in Model	✓	✓
District Percent of Students in Model	✓	✓
Flag for Student's Earlier Record Deleted[22]	✓	✓
Flag for Taking Fourth-Grade Test in 2012	✓	✓
Flag for Taking Fourth-Grade Test in 2013	✓	✓

Note: Each model was estimated using fourth-grade ABE literacy and mathematics test scores as dependent variables, generating separate sets of fixed and random effect estimates for each model and dependent variable. Fixed effect estimates are provided in Appendix B.

Grade 8. We estimated two status and two value-added models to predict scores on the eighth-grade ACT Explore English, mathematics, reading, and science tests for students in our grades 4–8 cohorts. The status models were the counterparts to Models 1 and 2 in grade 4 (Table 3). Their value-added counterparts, Models 3 and 4, added information on individual students' prior fourth-grade test scores. Model 4 also used as predictors average scores for all fourth graders in the district in each subject in the cohort's initial (fourth-grade) year. The use of district average prior scores was intended to capture the effect that peer achievement had on the students in the cohort. Because students who did not meet all of the criteria for inclusion in the cohort still could have had such an effect, we used district-wide average fourth-grade scores rather than just averaging the prior scores of students in the cohort.

Both value-added models contained a variable for the length of time (in years) between the fourth- and eighth-grade tests.[23] We estimated both status and value-added models on the same student cohorts in order to isolate the difference made by changing the model while holding the student

[22] Since inclusion in a given cohort file was based on kindergarten enrollment, students retained in kindergarten in the 2007–08 or 2008–09 school years had two records in our database. We dropped the earlier record and kept the record from the student's second kindergarten year going forward, creating a flag for students for whom this was the case. The same applied to retained fourth- and eighth-grade students in 2007–08 and 2008–09 for the models in Tables 2 and 3, respectively.

[23] This was set equal to the number of days between the two test dates divided by 365. This variable was recorded at the student level even though in grades 4–8, all students in the same district generally took the two tests on the same dates. Most Arkansas districts offered ACT Explore in the fall, but a few gave ACT Explore in the spring. Students who took the test later would have had more learning time between tests, hence the need to control for the amount of time between tests.

population constant. This process generated sixteen sets of district performance statistics based on four models and four eighth-grade subjects.

Grade 12. We estimated two status and two value-added models to predict scores on the ACT English, mathematics, reading, and science tests in grade 11 or 12 for students in our grades 8–12 cohorts (Table 4). We used the most recent score for students who took the ACT more than once. As in grade 8, all statistical models were estimated on the same student cohorts. This process generated sixteen sets of district performance statistics based on four models and four ACT subjects.[24]

Table 3. Predictors in Grade 8 Statistical Models

Variable	Model 1	Model 2	Model 3	Model 4
Intercept	✓	✓	✓	✓
Low-Income Status	✓	✓	✓	✓
African American Status	✓	✓	✓	✓
Hispanic Status	✓	✓	✓	✓
Asian Status	✓	✓	✓	✓
Native American Status	✓	✓	✓	✓
ELL Status	✓	✓	✓	✓
Special Education Status	✓	✓	✓	✓
Fourth-Grade Literacy Score			✓	✓
Fourth-Grade Mathematics Score			✓	✓
Number Years between Tests			✓	✓
District Percent Low-Income		✓		✓
District Percent African American		✓		✓
District Percent Hispanic		✓		✓
District Percent Asian		✓		✓
District Percent Native American		✓		✓
District Percent ELL		✓		✓
District Percent Special Education		✓		✓
District Average Fourth-Grade Literacy Score				✓
District Average Fourth-Grade Mathematics Score				✓
District Number Students in Model	✓	✓	✓	✓
District Percent of Students in Model	✓	✓	✓	✓
Flag for Student's Earlier Record Deleted	✓	✓	✓	✓
Flag for Eighth-Grade Test in 2012	✓	✓	✓	✓
Flag for Eighth-Grade Test in 2013	✓	✓	✓	✓

Note: Each model is estimated using eighth-grade ACT Explore English, reading, mathematics, and science test scores as dependent variables, generating separate sets of fixed and random effect estimates for each model and dependent variable. Fixed effect estimates are provided in Appendix B.

[24] In grades 8–12, the length of time between the two tests depended on when the student took the ACT and thus differed by student. Students who took the test later would have had more time to learn, but other selection effects might offset this effect: The strongest students might take the test in grade 11 and, satisfied with their scores, skip the test in grade 12. Hence, we would be uncertain in advance what the coefficient of the "time between tests" variable would be, but would hope that this variable would adjust for some of these selection effects.

Table 4. Predictors in Grade 12 Statistical Models

Variable	Model 1	Model 2	Model 3	Model 4
Intercept	✓	✓	✓	✓
Low-Income Status	✓	✓	✓	✓
African American Status	✓	✓	✓	✓
Hispanic Status	✓	✓	✓	✓
Asian Status	✓	✓	✓	✓
Native American Status	✓	✓	✓	✓
ELL Status	✓	✓	✓	✓
Special Education Status	✓	✓	✓	✓
Eighth-Grade English Score			✓	✓
Eighth-Grade Mathematics Score			✓	✓
Eighth-Grade Reading Score			✓	✓
Eighth-Grade Science Score			✓	✓
Number Years between Tests			✓	✓
District Percent Low-Income		✓		✓
District Percent African American		✓		✓
District Percent Hispanic		✓		✓
District Percent Asian		✓		✓
District Percent Native American		✓		✓
District Percent ELL		✓		✓
District Percent Special Education		✓		✓
District Average Eighth-Grade English Score				✓
District Average Eighth-Grade Mathematics Score				✓
District Average Eighth-Grade Reading Score				✓
District Average Eighth-Grade Science Score				✓
District Number Students in Model	✓	✓	✓	✓
District Percent of Students in Model	✓	✓	✓	✓
Flag for Student's Earlier Record Deleted	✓	✓	✓	✓
Flag for Twelfth Grader in 2012	✓	✓	✓	✓
Flag for Twelfth Grader in 2013	✓	✓	✓	✓

Note: Each model is estimated using eleventh- or twelfth-grade ACT English, reading, mathematics, and science test scores as dependent variables, generating separate sets of fixed and random effect estimates for each model and dependent variable. Fixed effect estimates are provided in Appendix B.

Interpretation of Statistical Models. Status and value-added models are fundamentally different in the period of growth that they reflect. Status models, which assess performance in the current period without regard to performance in any prior period, may be thought of as reflecting students' cumulative growth over their entire lifetime. Value-added models, which condition on students' scores in a previous period, mostly reflect student growth since that period.[25] For example, measuring students' eighth-grade achievement conditional on their fourth-grade achievement may be thought

[25] We say that the value-added models "reflect" growth rather than "measuring" growth since there is no common unidimensional scale on the grades 4 and 8 tests to use as units for measuring growth. The knowledge and skills measured on the eighth-grade test may be very different from those measured on the fourth-grade test. Also, controlling for end-of-fourth-grade test scores does not completely control for what happened to the student prior to the end of grade 4, since those scores do not capture 100% of the effects of those prior experiences.

of as an indicator of the performance of the district's schools (and other unmeasured influences) between grades 4 and 8. Measuring students' eighth-grade achievement without conditioning on their fourth-grade scores, on the other hand, may be thought of as reflecting the cumulative performance of the district's schools prior to grade 8 (as well as other unmeasured influences throughout the student's lifetime). Educators and policymakers may be interested in the results from both status and value-added models.

Alternatively, the four models may be thought of as answering the following questions:

- Model 1: How did students in this district perform relative to what would have been predicted for students with the same demographics?
- Model 2: How did students in this district perform relative to what would have been predicted for students with the same demographics in districts with the same demographics?
- Model 3: How did students in this district perform relative to what would have been predicted for students with the same demographics and prior scores?
- Model 4: How did students in this district perform relative to what would have been predicted for students with the same demographics and prior scores in districts with the same demographics and average prior scores?

Size of District Relative Performance Statistics

To address the size of performance differences across districts for each grade level, subject, and model, we examined by how much the performance of above- and below-average districts differed from average performance.[26] We defined above-average districts in a given grade level, subject, and model as those whose random effect performance statistic a) fell in the top quintile for the grade, subject, and model in question and b) was statistically different from zero at the .05 confidence level.[27] Similarly, we identified below-average districts as those in the bottom quintile whose performance statistics were different from average at the .05 confidence level. (Districts not meeting these requirements—i.e., in the middle three quintiles and/or not statistically different from average at the .05 level—could be thought of as close to average.)

For each grade level, subject, and model, we used a weighted average of the statistics of the above-average districts—with the districts' number of students in the analysis as weights—to create a measure of the difference between those districts' performance and average performance. We calculated a similar weighted average for the below-average districts. We expressed these performance differences in standard deviation units, calculated by dividing the number of scale score points by the standard deviation of student-level scores for the test in question.[28] Performance differences based on standard deviations are more comparable across grades, subjects, and

[26] Other measures of the range of district performance could be used, e.g., the interquartile range or the difference between the 5th and 95th percentile district.

[27] Technically, the p-value is based on the hypothesis that the average of the district's predicted random effect over hypothetical repeated sampling is 0. This is like testing the hypothesis that the district is really of average performance. Districts with smaller numbers of students in the analysis could appear in the top quintile and yet fail to exceed average performance at the .05 confidence level.

[28] Standard deviations on the grade 4 ABE, calculated for all students tested in the 2006–07 and 2007–08 school years, were 186.37 scale score points in literacy and 100.93 in mathematics. Standard deviations for all students taking the grade 8 ABE in the 2010–11 and 2011–12 school years (used in the grades 4–8 growth analysis described in footnote 32) were 150.88 points in literacy and 89.27 in mathematics. Standard deviations on the ACT Explore were 4.2 points in English, 3.5 in mathematics, 3.9 in reading, and 3.3 in science (ACT 2013, Table 4.11). Standard deviations on the ACT were 6.5 in English, 5.3 in mathematics, 6.3 in reading, and 5.3 in science (ACT, 2014, Table 5.4). In previous work (e.g., Dougherty, 2014), we used the standard deviations on the ACT for students graduating in 2010.

tests (e.g. ABE versus ACT Explore) than are differences based on scale score points. In the rest of the report, we refer to the performance statistics denominated in standard deviation units as "standardized performance statistics."

Consistency of District Relative Performance Statistics

To analyze the consistency of districts' student performance between Indicators A and B (where A and B might be two different subjects, models, or grade levels), we used four different metrics:

1. The *correlation of the standardized performance statistics* between A and B.
2. The *mean absolute difference in the standardized performance statistics for the same district* using Indicators A and B. In other words, by how much does a typical district's measured performance change when different indicators are used? Standardized performance statistics are more useful than simple score point statistics for comparing results across grades and tests with different scales (e.g. the scales for the ABE and ACT Explore).
3. The *above-average agreement rate* between A and B: of the districts rated as above average under either A or B, the percentage of those districts that were above average under both A and B.[29]
4. The *below-average agreement rate* between A and B: of the districts with student performance rated below average under either A or B, the percentage of those districts below average under both A and B.

The matrix in Table 5 illustrates how the different types of agreement rates are calculated. In Table 5, the above-average agreement rate equals $100^*a_{11}/(a_{11} + a_{21} + a_{31} + a_{12} + a_{13})$, or number of districts in the upper left cell divided by the total number of districts in the L-shaped region that are above average on at least one indicator, expressed as a percentage. Likewise, the below-average agreement rate equals $100^*a_{33}/(a_{31} + a_{32} + a_{33} + a_{13} + a_{23})$, or the number of districts in the lower right cell divided by the total number of districts in the L-shaped region that are below average on at least one indicator, also expressed as a percentage.[30]

Table 5. Hypothetical Agreement Matrix between Indicators A and B

	Indicator B Performance		
Indicator A Performance	1. Above Average	2. Average	3. Below Average
1. Above Average	a_{11}	a_{12}	a_{13}
2. Average	a_{21}	a_{22}	a_{23}
3. Below Average	a_{31}	a_{32}	a_{33}

For example, cell a_{21} contains the number of districts that are average on Indicator A and above average on Indicator B.

[29] One can also calculate the *overall agreement rate* between A and B: the percentage of districts whose student performance rating of above average, average, or below average is the same for A and B. However, in a state with many small districts, overall agreement rates are likely to be high because small districts have difficulty meeting requirements for being statistically different from average, hence they tend to show up as average on any given pair of indicators. As expected, we found that overall agreement rates were higher than the above- and below-average agreement rates reported in this paper.

[30] The overall agreement rate consists of the percentage of all districts falling on the diagonal, $100^*(a_{11} + a_{22} + a_{33})$ divided by the sum of all cells in the matrix.

In general, we would expect consistency based on placing districts in discrete categories (such as above, below, or close to average) to be lower than many users might anticipate from consistency numbers based on comparisons of district statistics (such as correlation coefficients or absolute differences between performance statistics). That is because districts close to category boundaries can fall into different categories even if their performance statistics are similar.

In addition to calculating consistency statistics between pairs of models, subjects, or levels, we examined how many districts were above or below average across multiple subjects and levels. In particular, for Model 2 (a status model) and Model 4 (a value-added model), we addressed these questions:

a. How many districts were above (below) average across all subjects in a given level?
b. How many districts were above (below) average for all levels in a given subject?
c. How many districts were above (below) average across all subjects and levels for which data were available in the model?
d. How many districts were above average in at least one subject and level and below average in another?

Limitations

Though this report looks at district performance statistics, we make no attempt to differentiate "district effects" from "school effects." Thus, we do not make a distinction between "performance of the district in grades X to Y" and "performance of the district's school(s) in grades X to Y." The great majority of Arkansas school districts are small and rural, and many districts have only one school at a given level. For example, in 2013, 151 (92%) of the 165 districts in the grades 8–12 cohort analysis had only one high school. Likewise, 191 (89%) of the 215 districts in the grades 4–8 cohort analysis had only one school serving eighth grade, and 161 (75%) of the 215 districts in the K–4 cohort analysis had only one school serving fourth grade. Thus, for the great majority of Arkansas districts, the value-added performance statistic in grades 8–12 could also be thought of as an indicator of the performance of the district's single high school, and the comparable statistic for grades 4–8 as an indicator of the performance of the district's single middle or junior high school and its feeder elementary school(s).[31] The value of thinking of the district as the unit of analysis is to be able to compare the performance of the district's schools at different levels—even when districts have only one school at a given level—and to examine performance in grade spans such as 4–8 where students typically change schools.

Second, we do not attempt to compare the wide range of statistical models that could be used to generate district performance statistics. Our goal was to compare results from four relatively straightforward models that control for generally available student- and district-level demographic statistics that most educators would be likely to believe should be controlled for. We did not refine the models to eliminate variables that do not add much explanatory power to the models.

A third limitation is that the accuracy of agreement rates may be reduced when the number of above- or below-average districts is small.

[31] In a state with a number of larger districts, one could partition the variance in performance across schools in those districts into the variance across districts and the variance across schools within districts.

Results

Size of Performance Differences Across Districts

Grade 4. Based on the two status models estimated for fourth grade, students in above-average districts on average scored .18 or .21 standard deviations in literacy and .21 or .23 standard deviations in mathematics above their predicted achievement levels (Table 6). Similarly, students in below-average districts on average scored .21 or .22 standard deviations in literacy and .20 or .21 standard deviations in mathematics below predicted levels (Table 7). By comparison, average yearly growth between grades 4–8 on the ABE exam for students in the three study cohorts ranged from .24 to .30 standard deviations in literacy and from .27 to .35 standard deviations in mathematics.[32]

Table 6.[33] Performance Level of Above-Average Districts in Grade 4 Based on Students Enrolled in the District since Kindergarten (N = 215 Districts)[34]

Subject	Status Model	Standard Deviations Relative to Average	Number of Districts
Literacy	01	0.21	43
	02	0.18	39
Mathematics	01	0.21	36
	02	0.23	36

Table 7. Performance Level of Below-Average Districts in Grade 4 Based on Students Enrolled in the District since Kindergarten (N = 215 Districts)

Subject	Status Model	Standard Deviations Relative to Average	Number of Districts
Literacy	01	-0.22	39
	02	-0.21	37
Mathematics	01	-0.20	35
	02	-0.21	34

Grade 8. Prior fourth-grade ABE scores were available for students in our grades 4–8 cohorts, so we estimated all four models in grade 8. Not surprisingly, the standardized district performance statistics tended to be smaller in absolute value in value-added models (Models 3 and 4) that controlled for prior student scores, than in status models (Models 1 and 2) that did not. For example, those statistics for above-average districts ranged from .15 to .21 across model-subject combinations using Models 1 or 2, compared with .12 to .17 using Models 3 or 4 (Table 8). Likewise, for below-average districts, the statistics ranged from -.14 to -.23 for Models 1 and 2, compared with

[32] In general, students in earlier cohorts grew faster on this measure than students in later cohorts. To calculate average grades 4–8 growth for students in a given cohort and subject, we averaged the differences between fourth- and eighth-grade scale scores across the students in the cohorts for the subject in question, using the average of the standard deviations of the fourth- and eighth-grade scores to convert the resulting mean score point difference into standard deviation units. This process took advantage of the ABE's use of a common vertical scale in grades 3–8. No such common scale was available to measure growth between the fourth-grade ABE and eighth-grade ACT Explore. We used data from students in the study's three cohorts rather than all students in the state having both fourth- and eighth-grade ABE scale scores.

[33] Tables 6–11 include districts in the top (bottom) quintile whose performance was above (below) average at the .05 significance level. "Standard deviations above (below) average" are a weighted average of these districts' random- effect performance statistics.

[34] The above- and below-average districts were drawn from a total set of 215 districts. As discussed below, the above- or below-average districts differed across subjects and models.

-.11 to -.17 for Models 3 and 4 (Table 9). Grade 8 status models attempt to explain current-year test scores that result from differences in growth in the student's entire lifetime up through eighth grade, whereas value-added models that control for fourth-grade test scores leave differences in student growth from grades 4 to 8 to be explained.

Table 8. Performance Level of Above-Average Districts in Grade 8 Based on Students Enrolled in the District since Grade 4 (N = 215 Districts)

Subject	Status Model	Standard Deviations Relative to Average	Number of Districts	Value-Added Model	Standard Deviations Relative to Average	Number of Districts
English	01	0.16	30	03	0.13	31
	02	0.15	24	04	0.12	25
Mathematics	01	0.21	36	03	0.17	39
	02	0.18	36	04	0.15	34
Reading	01	0.16	35	03	0.15	30
	02	0.15	23	04	0.13	22
Science	01	0.19	31	03	0.15	37
	02	0.15	28	04	0.13	29

Table 9. Performance Level of Below-Average Districts in Grade 8 Based on Students Enrolled in the District since Grade 4 (N = 215 Districts)

Subject	Status Model	Standard Deviations Relative to Average	Number of Districts	Value-Added Model	Standard Deviations Relative to Average	Number of Districts
English	01	-0.15	24	03	-0.14	33
	02	-0.14	22	04	-0.11	27
Mathematics	01	-0.16	29	03	-0.16	34
	02	-0.17	30	04	-0.16	29
Reading	01	-0.18	27	03	-0.17	24
	02	-0.18	22	04	-0.15	23
Science	01	-0.23	29	03	-0.18	37
	02	-0.19	29	04	-0.16	28

Grades 11–12. Prior eighth-grade ACT Explore scores were available for students in our grades 8–12 cohorts, so we estimated all four models in grades 11–12. Standardized district performance statistics for above-average districts ranged from .16 to .24 using Models 1 and 2, versus .08 to .16 using Models 3 and 4 (Table 10). For below-average districts, these statistics ranged from –.16 to -.21 for Models 1 and 2, versus -.09 to -.13 for Models 3 and 4 (Table 11). By comparison, typical annual growth from grade 8 to grades 11–12 (using ACT Explore and the ACT as endpoints) is about .26 standard deviations in English, .28 in mathematics, .31 in reading, and .24 in science (ACT, 2012c), using the average of the ACT Explore and ACT standard deviations to convert typical growth in score points to standard deviations.

Table 10. Performance Level of Above-Average Districts in Grades 11 and 12 Based on Students Enrolled in the District since Grade 8 (N = 165 Districts)

Subject	Status Model	Standard Deviations Relative to Average	Number of Districts	Value-Added Model	Standard Deviations Relative to Average	Number of Districts
English	01	0.19	27	03	0.11	20
	02	0.17	17	04	0.09	18
Mathematics	01	0.24	29	03	0.16	24
	02	0.20	23	04	0.15	22
Reading	01	0.19	20	03	0.09	15
	02	0.16	13	04	0.09	12
Science	01	0.19	23	03	0.10	13
	02	0.18	15	04	0.08	11

Table 11. Performance Level of Below-Average Districts in Grades 11 and 12 Based on Students Enrolled in the District since Grade 8 (N = 165 Districts)

Subject	Status Model	Standard Deviations Relative to Average	Number of Districts	Value-Added Model	Standard Deviations Relative to Average	Number of Districts
English	01	-0.21	9	03	-0.11	9
	02	-0.18	9	04	-0.09	9
Mathematics	01	-0.20	24	03	-0.13	18
	02	-0.19	23	04	-0.12	18
Reading	01	-0.18	12	03	-0.11	13
	02	-0.16	10	04	-0.09	7
Science	01	-0.19	14	03	-0.10	9
	02	-0.16	10	04	-0.12	1

Consistency of District Performance Statistics across Models

Grade 4. Correlations between results for the same districts using the two models available for fourth grade were high: .94 for literacy and .98 for mathematics (Table 12). On average, performance statistics for the same district using different models differed by .04 of a standard deviation in literacy and .02 of a standard deviation in mathematics (the "mean absolute difference" column in Table 12). These differences are relatively small compared with the differences of .18–.23 standard deviations between above-average and average districts and .20–.22 between average and below-average districts (Tables 6 and 7).

However, when districts are placed in categories ("above average" or "below average"), there were a substantial percentage of above- or below-average districts for which it mattered which model was used. For example, 61% of districts that were identified as above average by at least one model in literacy and 85% of districts identified by at least one model in mathematics were identified as above average by both models, leaving 39% and 15% of districts identified as above average in one model but not the other (Table 12). Likewise, 73% of districts identified as below average in literacy and 82% as below average in mathematics by at least one model were identified as below average by both models, leaving 27% and 18% of those districts identified as below average in only one model.

Table 12. Correlations and Agreement Rates of District Performance Statistics between Models: Grade 4 (N = 215 Districts)

Subject	Model Comparison	Correlation of Performance Statistics	Mean Absolute Difference in Performance Statistics	Above-Average Agreement Rate	Below-Average Agreement Rate
Literacy	01 vs. 02	0.94	0.04	61%	73%
Mathematics	01 vs. 02	0.98	0.02	85%	82%

Note: Differences are measured in standard deviation units. The size of these numbers may be compared with the absolute size of the statistics in the "standard deviations relative to average" columns of Tables 6 and 7.

Grade 8. Correlations of eighth-grade performance statistics for the same districts across different models were high when comparing models that were similar in their use of prior score information (Table 13). For example, comparisons of district results from Models 1 and 2—both of which are status models that do not use prior scores—yielded correlations of performance statistics ranging from .91 to .94, depending on the subject. Likewise, correlations of district performance statistics between Models 3 and 4—both of which are value-added models based on their use of prior scores—also ranged from .91 to .94. For these two sets of comparisons, the mean absolute difference in performance statistics ranged from .03 to .04. For context, these mean absolute differences in Table 13 may be compared with the gaps of .12 to .21 between above-average and average district performance shown in Table 8.

Despite these similarities in general results, significant discrepancies in classifying districts as above or below average occurred when comparing results from these similar models. Above-average agreement rates for comparisons of Models 1 and 2 and Models 3 and 4 ranged from 53% to 74% with a median of 58%, and below-average agreement rates ranged from 58% to 81%, with a median of 60%. Thus, for some comparisons, almost half of the districts that were identified as above average in at least one model were not above average in both models.

Comparisons of status models with value-added models—Models 1 vs. 3, 2 vs. 4, or 1 vs. 4—produced greater discrepancies in district performance statistics.[35] Correlations between performance statistics generated by these pairs of models ranged from .51 to .85, with a median correlation of .75, and mean absolute differences ranged from .05 to .08, with a median of .06. These correlations are lower and differences are higher than those from the comparisons of pairs of status models or pairs of value-added models.

Comparisons of status with value-added models also produced higher discrepancies in the classification of districts as above or below average. Above-average agreement rates in those comparisons ranged from 22% to 55% with a median of 43%, and below-average agreement rates from 21% to 68% with a median of 42%.

[35] To simplify the table, we do not show the comparison between Models 2 and 3, which also differ in their use of prior score information.

Table 13. Correlations and Agreement Rates of District Performance Statistics between Models: Grade 8 (N = 215 Districts)

Subject	Model Comparison	Correlation of Performance Statistics	Mean Absolute Difference in Performance Statistics	Above-Average Agreement Rate	Below-Average Agreement Rate
English	01 vs. 02	0.94	0.03	69%	59%
	03 vs. 04	0.91	0.03	56%	58%
	01 vs. 03	0.51	0.08	22%	21%
	02 vs. 04	0.75	0.05	44%	36%
	01 vs. 04	0.70	0.06	34%	34%
Mathematics	01 vs. 02	0.94	0.03	71%	64%
	03 vs. 04	0.91	0.04	74%	62%
	01 vs. 03	0.68	0.08	44%	24%
	02 vs. 04	0.83	0.05	49%	48%
	01 vs. 04	0.78	0.07	46%	35%
Reading	01 vs. 02	0.92	0.03	53%	58%
	03 vs. 04	0.92	0.03	53%	74%
	01 vs. 03	0.65	0.07	35%	38%
	02 vs. 04	0.81	0.05	55%	50%
	01 vs. 04	0.75	0.06	43%	47%
Science	01 vs. 02	0.94	0.03	59%	81%
	03 vs. 04	0.94	0.03	53%	59%
	01 vs. 03	0.74	0.07	33%	50%
	02 vs. 04	0.85	0.05	54%	68%
	01 vs. 04	0.81	0.06	43%	63%

Grades 11–12. As was the case in eighth grade, correlations of performance statistics in grades 11–12 were higher when comparing models that were similar in their use of prior score information. For example, comparisons of Models 1 and 2 and of Models 3 and 4 (Table 14) yielded correlations between .88 and .92, with a median of .89, and mean absolute differences in performance statistics of .02 to .06, with a median of .03. For context, these mean absolute differences in Table 14 may be compared with the gaps of .08 to .24 between above-average and average district performance shown in Table 10.

However, significant discrepancies in classifying districts as above or below average in grades 11–12 occurred when comparing results from these similar models. Above-average agreement rates for comparisons of Models 1 and 2 and Models 3 and 4 ranged from 41% to 65% with a median of 55%, and below-average agreement rates ranged from 11% to 68% with a median of 41%. Thus, for many model comparisons in grades 11–12, the majority of districts that were identified as above (below) average in at least one model were not above (below) average in both models.

Switching to comparison of status with value-added models produced lower score correlations, higher mean absolute differences, and larger discrepancies in classifying districts as above or below average. For those comparisons, correlations of performance statistics ranged from .58 to .83, with a median of .69; mean absolute differences in those statistics ranged from .05 to .09 with a median of .07. Above-average agreement rates ranged from 31% to 53% with a median of 41%, and below-average agreement rates from 7% to 28% with a median of 20% (Table 14).

Table 14. Correlations and Agreement Rates of District Performance Statistics between Models: Grades 11–12 (N = 165 Districts)

Subject	Model Comparison	Correlation of Performance Statistics	Mean Absolute Difference in Performance Statistics	Above-Average Agreement Rate	Below-Average Agreement Rate
English	01 vs. 02	0.91	0.04	52%	64%
	03 vs. 04	0.88	0.03	65%	38%
	01 vs. 03	0.58	0.08	34%	20%
	02 vs. 04	0.71	0.06	46%	20%
	01 vs. 04	0.64	0.07	50%	20%
Mathematics	01 vs. 02	0.89	0.06	58%	68%
	03 vs. 04	0.90	0.03	64%	38%
	01 vs. 03	0.78	0.08	47%	17%
	02 vs. 04	0.83	0.06	50%	28%
	01 vs. 04	0.73	0.09	42%	24%
Reading	01 vs. 02	0.92	0.03	57%	38%
	03 vs. 04	0.91	0.02	50%	43%
	01 vs. 03	0.60	0.07	40%	25%
	02 vs. 04	0.69	0.05	39%	13%
	01 vs. 04	0.64	0.07	33%	19%
Science	01 vs. 02	0.89	0.04	52%	50%
	03 vs. 04	0.89	0.02	41%	11%
	01 vs. 03	0.70	0.07	38%	28%
	02 vs. 04	0.73	0.06	53%	10%
	01 vs. 04	0.63	0.08	31%	7%

Consistency of District Performance Statistics between Subjects

Grade 4. Overall, districts' performance in literacy was positively related to their performance in mathematics, but there were also substantial discrepancies. Correlations between results for the same districts between the subject areas of literacy and mathematics were .63 and .65 when Model 1 and Model 2 were used, respectively (Table 15). Mean absolute differences in standardized district performance statistics across subjects were .10 for both models. For context, these may be compared with the gaps in performance of .18 to .23 between above-average and average districts shown in Table 12. In addition, less than half of districts that were identified as above (below) average in at least one of the two subjects were above (below) average in the other subject (Table 15).

These discrepancies in districts' performance statistics across subjects were substantially greater than discrepancies from the use of different models in the same subject. The cross-subject correlations of .63 and .65 in Table 15 may be compared with the cross-model correlations of .98 and .94 shown in Table 12. Likewise, the mean absolute difference in district performance across subjects of .10 in Table 15 may be compared the performance differences across models of .04 and .02 in Table 12.

Table 15. Correlations and Agreement Rates of District Performance Statistics between Subjects: Grade 4 (N = 215 Districts)

Model	Subject Comparison	Correlation of Performance Statistics	Mean Absolute Difference in Performance Statistics	Above-Average Agreement Rate	Below-Average Agreement Rate
01	Literacy, Mathematics	0.63	0.10	41%	30%
02	Literacy, Mathematics	0.65	0.10	36%	25%

Grade 8. As in fourth grade, districts' performance statistics in different subject areas were positively related, but there were also substantial discrepancies (Table 16). For example, correlations of district performance between subjects ranged from .48 to .86 with a median of .68; mean absolute differences across subjects ranged from .05 to .08 with a median of .07. For context, these mean absolute differences in performance across subjects may be compared with gaps between above-average and average performance in specific subjects ranging from .12 to .21, as shown in Table 8.[36]

In the majority of comparisons, more than half of districts identified as above (below) average in one subject were not above (below) average in the other subject (Table 16). Above-average agreement rates ranged from 20% to 61% with a median of 40%, and below-average agreement rates from 24% to 50% with a median of 34%. Of those districts that were above average in any subject in eighth grade, the percentages that were above average in all four subjects were 22%, 17%, 11%, and 11% in Models 1 to 4, respectively (not shown in Table 16).[37] Of those districts that were below average in any subject in eighth grade, the percentages that were below average in all four subjects were 13%, 7%, 12%, and 11% in Models 1 to 4.[38]

Correlations and agreement rates between subjects were generally lower for value-added models (Models 3 and 4) than for status models (Models 1 and 2). For Models 1 and 2, cross-subject correlations ranged from .61 to .86, with a median of .72. For Models 3 and 4, cross-subject correlations ranged from .48 to .83 with a median of .64. Likewise, above-average agreement rates for Models 1 and 2 ranged from 28% to 61% with a median of 42%, whereas for Models 3 and 4 those rates ranged from 20% to 49% with a median of 33%. Below-average agreement rates were similar for status and value-added models, ranging from 24% to 50% with a median of 34% for Models 1 and 2; and from 24% to 46% with a median of 34% for Models 3 and 4.

Mean absolute differences in performance across subjects were similar for status and value-added models: those differences ranged from .05 to .08 with a median of .07 for both sets of models. However, the gap between above-average and average performance, shown in Table 8, was smaller for value-added models: a range of .12 to .17 for Models 3 and 4, compared with .15 to .21 for Models 1 and 2. So for the value-added models, cross-subject performance differences were a larger share of the (smaller) total variation in performance.

[36] For Models 3 and 4, the gaps between above-average and average performance ranged from .12 to .17 (Table 8).

[37] From Figure 2, the reader can see that in Model 2, nine districts were above average in all four subjects, representing 17% of the 52 districts above average in at least one subject. Likewise, in Model 4, seven districts were above average in all four subjects, representing 11% of the 61 districts above average in at least one subject.

[38] From Figure C2 in Appendix C, the reader can see that in Model 2, four districts were below average in all four subjects, representing 7% of the 54 districts below average in at least one subject. Likewise, in Model 4, six districts were below average in all four subjects, representing 11% of the 56 districts below average in at least one subject.

The highest cross-subject correlations in grade 8 were between reading and science, while the lowest were between English and mathematics and between reading and mathematics. The correlations between reading and science performance were higher than for mathematics and science.[39] Interestingly, the correlation between district reading and science performance was also higher than that between reading and English.

Table 16. Correlations and Agreement Rates of District Performance Statistics between Subjects: Grade 8 (N = 215 Districts)

Model	Subject Comparison	Correlation of Performance Statistics	Mean Absolute Difference in Performance Statistics	Above-Average Agreement Rate	Below-Average Agreement Rate
01 Student Demographics	English, Mathematics	0.66	0.08	38%	36%
	English, Reading	0.83	0.05	51%	42%
	English, Science	0.78	0.06	53%	33%
	Mathematics, Reading	0.65	0.08	34%	40%
	Mathematics, Science	0.69	0.08	40%	35%
	Reading, Science	0.86	0.05	61%	47%
02 Student and District Demographics	English, Mathematics	0.62	0.08	40%	24%
	English, Reading	0.80	0.05	42%	29%
	English, Science	0.75	0.06	44%	31%
	Mathematics, Reading	0.61	0.08	28%	30%
	Mathematics, Science	0.65	0.08	42%	28%
	Reading, Science	0.85	0.05	50%	50%
03 Student Demographics and Prior Scores	English, Mathematics	0.57	0.08	21%	37%
	English, Reading	0.75	0.05	42%	36%
	English, Science	0.71	0.06	36%	32%
	Mathematics, Reading	0.57	0.08	30%	29%
	Mathematics, Science	0.61	0.08	29%	31%
	Reading, Science	0.83	0.05	49%	42%
04 Student and District Demographics and Prior Scores	English, Mathematics	0.48	0.08	20%	27%
	English, Reading	0.71	0.05	47%	35%
	English, Science	0.67	0.06	35%	38%
	Mathematics, Reading	0.49	0.08	22%	24%
	Mathematics, Science	0.55	0.08	31%	30%
	Reading, Science	0.80	0.05	42%	46%

[39] Reading ability is based both on the fluent decoding needed to know what the text says and on the background knowledge and vocabulary needed to understand what the text means. Both decoding and background knowledge are likely to be important in helping students learn science.

Grades 11–12. As was the case in fourth and eighth grades, districts' performance statistics in different subject areas were positively related, but there were also substantial discrepancies (Table 17). Correlations in district performance across pairs of subjects ranged from .39 to .87 with a median of .67; mean absolute differences across subjects ranged from .03 to .09 with a median of .05. For context, the mean absolute performance differences across subjects may be compared with gaps between above-average and average performance in specific subjects ranging from .08 to .24, as shown in Table 10.

In the majority of comparisons, more than half of districts identified as above (below) average in one subject were not above (below) average in the other subject (Table 17). Above-average agreement rates ranged from 21% to 65% with a median of 40%, and below-average agreement rates from 0% to 58% with a median of 27%. Of those districts that were above average in any subject in grades 11–12, the percentages that were above average in all four subjects were 34%, 18%, 14%, and 8% in Models 1 to 4, respectively (not shown in Table 17).[40] Of those districts that were below average in any subject in grades 11–12, the percentages that were below average in all four subjects were 16%, 11%, 3%, and 0% in Models 1 to 4.[41]

Correlations and agreement rates between pairs of subjects were lower for value-added models (Models 3 and 4) than for status models (Models 1 and 2). For Models 1 and 2, cross-subject correlations ranged from .67 to .87, with a median of .82. For Models 3 and 4, cross-subject correlations ranged from .39 to .68 with a median of .60. Likewise, above-average agreement rates for Models 1 and 2 ranged from 24% to 65% with a median of 48%, whereas for Models 3 and 4 those rates ranged from 21% to 44% with a median of 31%. Below-average agreement rates for Models 1 and 2 ranged from 25% to 58% with a median of 34%, whereas for Models 3 and 4 those rates ranged from 0% to 38% with a median of 14%.

Mean absolute differences in performance across subjects were similar for status and value-added models: those differences ranged from .04 to .09 with a median of .06 for Models 1 and 2 and from .03 to .07 with a median of .05 for Models 3 and 4. However, to put these differences in context, the gap between above-average and average performance for Models 3 and 4 ranged from .08 to .16, whereas the same gap for Models 1 and 2 ranged from .16 to .24 (Table 10). So for the value-added models, cross-subject performance differences were a larger share of the (smaller) total variation in performance.

The highest cross-subject correlations in Grades 11 and 12 were between reading and English, while the lowest were between reading and mathematics. As in Grade 8, the correlations between reading and science performance were higher than for mathematics and science, particularly in Models 3 and 4.

[40] From Figure 3, the reader can see that in Model 2, six districts were above average in all four subjects, representing 18% of the 34 districts above average in at least one subject. Likewise, in Model 4, three districts were above average in all four subjects, representing 8% of the 37 districts above average in at least one subject.

[41] From Figure C3 in Appendix C, the reader can see that in Model 2, three districts were below average in all four subjects, representing 11% of the 28 districts below average in at least one subject. Likewise, in Model 4, no district was below average in all four subjects, representing 0% of the 29 districts below average in at least one subject.

Table 17. Correlations and Agreement Rates of District Performance Statistics between Subjects: Grades 11–12 (N = 165 Districts)

Model	Subject Comparison	Correlation of Performance Statistics	Mean Absolute Difference in Performance Statistics	Above-Average Agreement Rate	Below-Average Agreement Rate
01 Student Demographics	English, Mathematics	0.77	0.07	51%	38%
	English, Reading	0.87	0.05	57%	40%
	English, Science	0.85	0.05	61%	35%
	Mathematics, Reading	0.72	0.09	44%	29%
	Mathematics, Science	0.83	0.07	58%	31%
	Reading, Science	0.86	0.05	65%	44%
02 Student and District Demographics	English, Mathematics	0.73	0.07	43%	33%
	English, Reading	0.86	0.04	36%	58%
	English, Science	0.83	0.04	52%	36%
	Mathematics, Reading	0.67	0.08	24%	27%
	Mathematics, Science	0.79	0.06	41%	27%
	Reading, Science	0.82	0.04	40%	25%
03 Student Demographics and Prior Scores	English, Mathematics	0.50	0.07	38%	17%
	English, Reading	0.68	0.04	35%	38%
	English, Science	0.62	0.04	32%	29%
	Mathematics, Reading	0.45	0.07	30%	15%
	Mathematics, Science	0.59	0.06	42%	8%
	Reading, Science	0.67	0.04	40%	22%
04 Student and District Demographics and Prior Scores	English, Mathematics	0.49	0.06	25%	4%
	English, Reading	0.65	0.04	30%	14%
	English, Science	0.60	0.04	26%	11%
	Mathematics, Reading	0.39	0.07	21%	9%
	Mathematics, Science	0.54	0.06	22%	0%
	Reading, Science	0.62	0.03	44%	14%

Consistency of District Performance Statistics across Grade Levels

Grade 4 vs. Grade 8. Substantial differences existed in district performance statistics in the same or similar subjects[42] between grades 4 and 8. Correlations in district performance ranged from .30 to .39 (Table 18), and differences in performance of the same district in the two different grades averaged .12. These cross-grade-level correlations were generally lower and the discrepancies across grade levels generally higher than was the case for comparisons across subject areas at the same grade level (Tables 15 and 16). The majority of districts that were above (below) average in one of the two grade levels were not above (below) average in both levels. Above-average agreement rates ranged from 15% to 24%, and below-average agreement rates from 13% to 20%. These agreement rates were generally lower than the same-grade agreement rates across subjects shown in Tables 15 and 16. Only status models could be compared across these two grade levels as value-added models were not used in fourth grade.

Table 18. Correlations and Agreement Rates of District Performance Statistics across Grade Levels: Grade 4 vs. Grade 8 (N = 215 Districts)

Model	Subject	Correlation of Performance Statistics	Mean Absolute Difference in Performance Statistics	Above-Average Agreement Rate	Below-Average Agreement Rate
01	Literacy/Reading	0.38	0.12	24%	20%
	Mathematics	0.39	0.12	24%	16%
02	Literacy/Reading	0.30	0.12	15%	13%
	Mathematics	0.38	0.12	22%	16%

Grade 4 vs. Grades 11–12. As might be expected, correlations and agreement rates between district performance statistics in grades 4 and 11–12 were generally lower than between grades 4 and 8, and average differences in these statistics were for the most part higher. Correlations across the two levels were .21 or .24 in literacy/reading and .33 or .34 in mathematics (Table 19). Average performance differences ranged from .12 to .14, above-average agreement rates from 10% to 21%, and below-average agreement rates from 8% to 20%. These cross-grade-level correlations were lower and the discrepancies across grade levels generally higher than was the case for status-model comparisons across subject areas at the same grade level (Tables 15 and 17).

Table 19. Correlations and Agreement Rates of District Performance Statistics across Grade Levels: Grade 4 vs. Grades 11–12 (N = 165 Districts)

Model	Subject	Correlation of Performance Statistics	Mean Absolute Difference in Performance Statistics	Above-Average Agreement Rate	Below-Average Agreement Rate
01	Literacy/Reading	0.24	0.13	18%	8%
	Mathematics	0.33	0.14	21%	18%
02	Literacy/Reading	0.21	0.12	10%	8%
	Mathematics	0.34	0.13	11%	20%

[42] The fourth-grade ABE literacy test assessed both reading and writing. This may have reduced the correlations between district performance on that test with performance on the ACT Explore reading test in grade 8 (Table 18) and the ACT reading test in grades 11–12 (Table 19).

Grade 8 vs. Grades 11–12. Comparisons of both status and value-added models were available between grades 8 and 11–12. For the two status models (Models 1 and 2), correlations of district performance statistics between grades 8 and 11–12 ranged from .38 to .49, and differences in performance between .08 and .12. Above-average agreement rates ranged from 17% to 43%, and below-average agreement rates from 0% to 17% (Table 20).

The introduction of prior score information into the analysis (Models 3 and 4) virtually eliminated the correlation of district performance statistics across the two grade levels. Five of the correlations in value-added performance statistics were not statistically different from 0 at the .05 confidence level, and the remaining correlations ranged from .19 to .23. Above-average agreement rates ranged from 9% to 26%, and below-average agreement rates from 0% to 8%. Average performance differences ranged from .07 to .11, which were larger than most of the performance differences across subjects at the same grade level shown in Tables 16 and 17. The low correlations in these value-added measures indicate that there is not much of a relationship between student growth in late elementary and middle school and student growth in high school in the same district.

Table 20. Correlations and Agreement Rates of District Performance Statistics across Grade Levels: Grade 8 vs. Grades 11–12 (N = 165 Districts)

Model	Subject	Correlation of Performance Statistics	Mean Absolute Difference in Performance Statistics	Above-Average Agreement Rate	Below-Average Agreement Rate
01 Student Demographics	English	0.49	0.09	43%	4%
	Mathematics	0.47	0.12	35%	14%
	Reading	0.41	0.10	26%	16%
	Science	0.47	0.10	35%	13%
02 Student and District Demographics	English	0.48	0.08	23%	0%
	Mathematics	0.42	0.11	38%	17%
	Reading	0.38	0.09	17%	8%
	Science	0.43	0.09	28%	7%
03 Student Demographics and Prior Scores	English	*0.05*	0.09	10%	6%
	Mathematics	0.19	0.11	22%	3%
	Reading	*-0.08*	0.09	9%	7%
	Science	*0.08*	0.10	12%	3%
04 Student and District Demographics and Prior Scores	English	0.19	0.07	18%	7%
	Mathematics	0.23	0.10	26%	8%
	Reading	*-0.02*	0.08	15%	0%
	Science	*0.11*	0.09	15%	0%

Note: Correlations shown in *italics* were not statistically different from 0 at the .05 confidence level.

District Performance across Multiple Subjects and Levels

In this section we provide information on the number and percentage of districts that were above average across multiple subjects and grade levels.[43]

a. *How many districts were above average across all subjects in a given level?*

In grade 4, 20 (9% of 215 eligible districts) were above-average in both literacy and mathematics in Model 2, compared with 35 districts that were above average in one subject but not the other (Figure 1). In grade 8, nine districts (4% of eligible districts) were above average in all four subjects in Model 2 and seven districts (3%) in all four subjects in Model 4 (Figure 2). In grades 11–12, six districts (4% of 165 eligible districts) were above average in all four subjects in Model 2 and three districts (2%) in Model 4 (Figure 3).

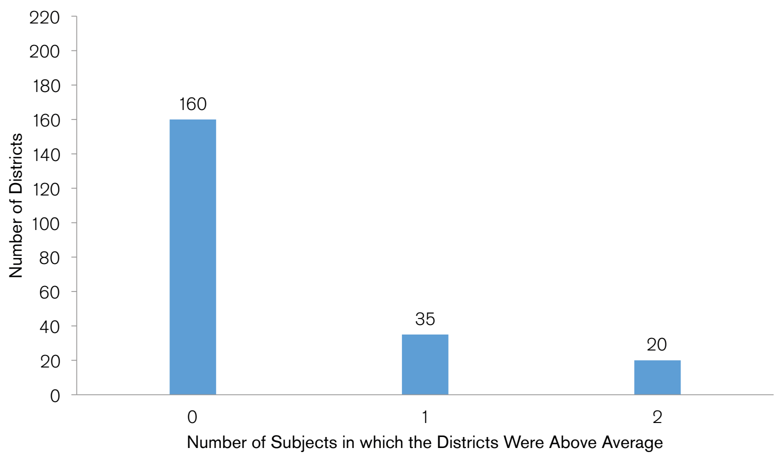

Figure 1. Number of districts in grade 4 by the number of subjects (out of two) in which they were above average: Model 2 (N = 215 districts)

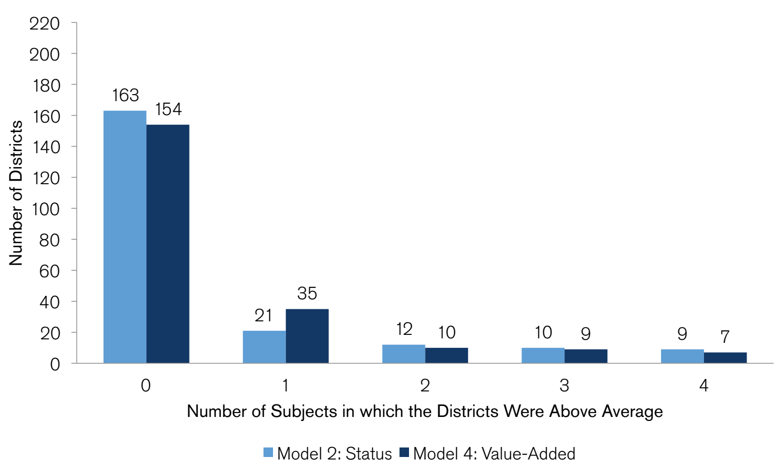

Figure 2. Number of districts in grade 8 by the number of subjects (out of four) in which they were above average: Models 2 and 4 (N = 215 districts)

[43] In Appendix C, we provide similar information on the number and percentage of districts that were below average in multiple grade levels and subjects.

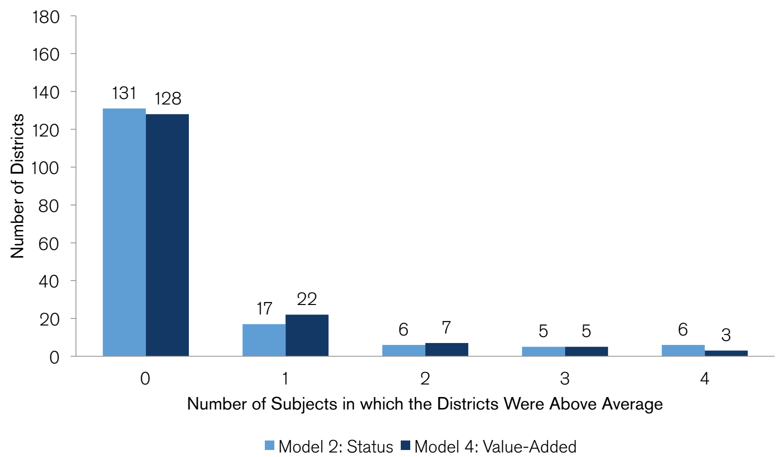

Figure 3. Number of districts in grades 11–12 by the number of subjects (out of four) in which they were above average: Models 2 and 4 (N = 165 districts).

b. *How many districts were above average for all levels in a given subject?*

Using Model 2, four districts (2% of 165) were above average in mathematics in all three levels (grades 4, 8, and 11–12), while only one district was above average in all three levels in reading/literacy (Figure 4). English and science were not tested in grade 4 so a three-level comparison was not available for those two subjects. Looking at a two-level comparison based on Model 4, six districts were above average in English in both grade 8 and grades 11–12, 10 districts in mathematics, four in reading, and five in science (Figure 5).[44]

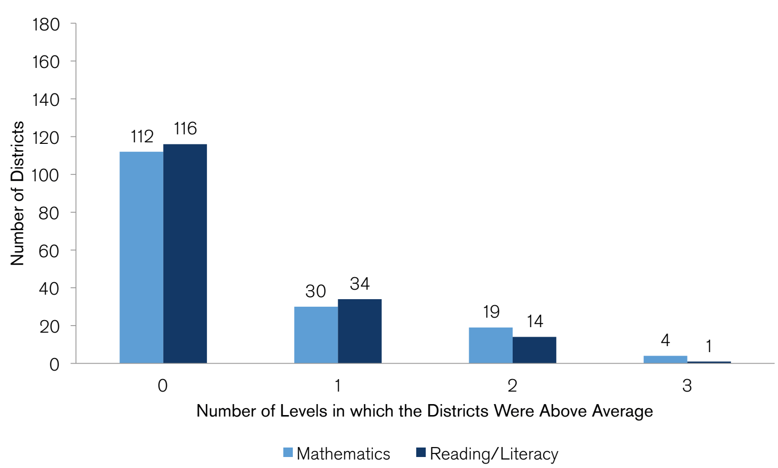

Figure 4. Number of districts by the number of levels (out of three) in which they were above average in each subject: Model 2 (N = 165 districts)

[44] These were not necessarily the same districts, as is evident from the answer to the next question.

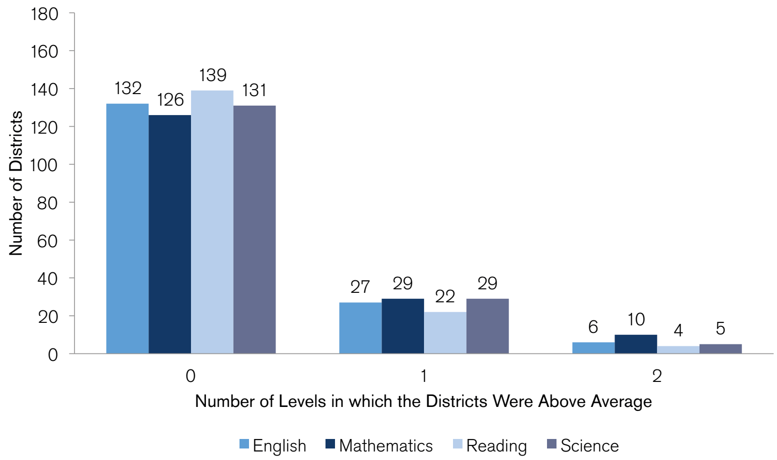

Figure 5. Number of districts by the number of levels (out of two) in which they were above average in each subject: Model 4 (N = 165 districts)

c. *How many districts were above average across all subjects and levels for which data were available in the model?*

Using Model 2, there were ten possible combinations of subjects and levels: literacy and mathematics in grade 4 and English, mathematics, reading, and science in each of grades 8 and 12. No district was above average across all ten of these combinations, and only one district was above average for nine of the ten combinations (Figure 6). Twelve districts (7% of 165 districts eligible in all three levels) were above average in more than half of the subject-level combinations.

Using Model 4, there were eight combinations of subjects and levels: four subjects each in grades 8 and 11–12. One district was above average for all eight combinations, and one additional district was above average for seven of the eight combinations (Figure 7). Seven districts (4% of 165 eligible districts) were above average for the majority of subject-level combinations.

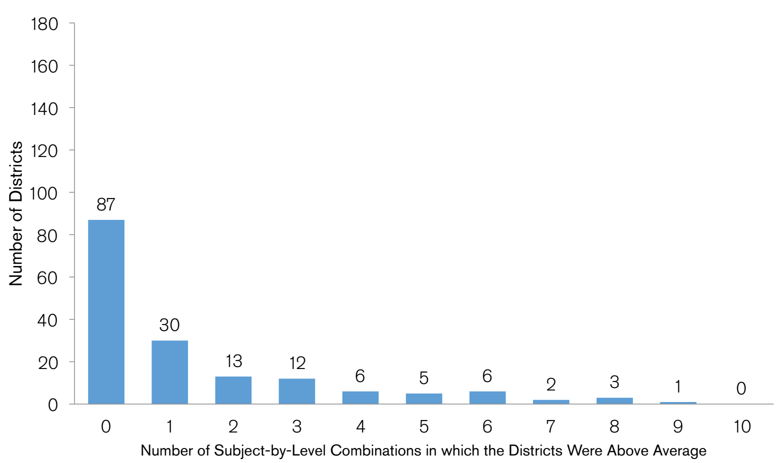

Figure 6. Number of districts by the total number of subject-level combinations (out of ten) in which they were above average: Model 2 (N = 165 districts)

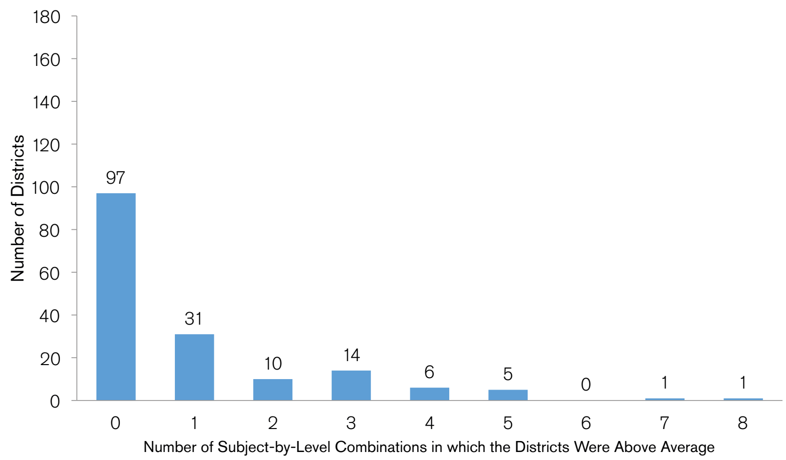

Figure 7. Number of districts by the total number of subject-level combinations (out of eight) in which they were above average: Model 4 (N = 165 districts)

d. *How many districts had opposite results in different subjects and grade levels: that is, they were above average in at least one subject and/or level and below average in another?*

Districts could have opposite results across subjects and levels in the following ways: 1) between different subjects in the same grade level (e.g., above average in fourth-grade literacy and below average in fourth-grade mathematics); 2) between different grade levels in the same subject (e.g., above average in fourth-grade mathematics but below average in mathematics in grades 11–12); 3) in cases where both the subject and level are different (e.g., above average in fourth-grade reading and below average in eighth-grade mathematics).

Using Model 2, we found the following combinations of opposite results:

- One district had opposite results in literacy and mathematics in fourth grade. No other districts had opposite results across subjects at the same grade level.
- Seven districts had opposite results in mathematics in different levels, five districts had opposite results in reading, and one had opposite results in science.
- Nineteen districts had opposite results where both the level and subject area were different.

Model 4 yielded the following combinations of opposite results:

- Two districts had opposite results in different subjects in eighth grade, and two districts had opposite results in different subjects in grades 11–12. (One of the two districts was on both lists.)
- Two districts had opposite results in English between grades 8 and 11–12; four districts had opposite results in mathematics; five in reading; and one in science.
- Fifteen districts had opposite results where both the level and subject area were different.

Conclusion

Our research found sizeable differences in student performance between above- and below-average districts. This is broadly similar to the results found by Chingos, Whitehurst, and Gallaher (2013).[45] However, we also found that comparisons of district performance often yielded different results by subject and grade level—for example, a district that was higher-performing than others in high school mathematics might not be higher-performing in elementary school English language arts. A relatively low percentage of districts were above or below average in all subjects in a given grade level; similarly, relatively few districts in a given subject were consistent across all grade levels. Much larger numbers and percentages of districts were above average in some subjects and levels but not others. District performance comparisons were also affected by the choice between value-added and status models, which respectively assess student growth and cumulative student achievement.

Our findings on consistency of district performance by subject and grade level were broadly comparable with the research literature on consistency of school performance across subjects (Mandeville & Anderson, 1987; Luyten, 1994; Luyten, 1998; Ma, 2001; Mohiemang & Pretorius, 2012) and grade levels (Mandeville & Anderson, 1987). Thus, this report reinforced the results of prior research on consistency across subject areas while adding new findings on the consistency of district performance across the elementary, middle, and high school levels and between status and value-added models.[46]

The inconsistency of measured district performance across grade levels and subject areas implies that educators and policymakers should pay close attention to performance in specific subjects and levels. Overall district performance is a less meaningful measure than, say, district performance in high school mathematics or elementary school English language arts. In addition, differences in district performance across models (e.g., between status and value-added models) imply that educators and policymakers should pay close attention to the exact statistical comparison being made and the type of performance being compared. For example, an eighth-grade value-added model that mainly measures growth since fourth grade may yield different results from a status model that measures eighth-grade performance without taking fourth-grade scores into account. All of this is consistent with the finding of Abe et al. (2015) that systems of rating schools may be highly sensitive to the method used to construct the ratings.[47]

Although statistical models can create fairer comparisons across districts than do rankings based on proficiency rates, the statistics calculated in these models nonetheless call for further exploration of what causes these differences in performance. Performance statistics should be treated not only as feedback on the results of past efforts, but also as the starting point for further inquiry that seeks to understand how current practices and environmental factors influence student outcomes. These statistics should be used to stimulate conversations among educators and community leaders about current progress, areas of need, and ways to improve educator practices and community support (ACT, 2012a). ■

[45] Chingos et al. conducted their research with data from two states, Florida and North Carolina, with large county-based school districts. Thus, they were able to break out school from district effects, whereas in this study, those effects were combined in the great majority of districts with only one school at a given level. Thus, our results are not directly comparable to theirs.

[46] Mandeville and Anderson (1987) compared school performance across grades 1–4. Since all of these studies used schools and not districts as the unit of analysis, they did not compare performance of schools in the same district across elementary, middle, and high school levels as is done in this report. Tekwe et al. (2004) compared results across various types of value-added models, but they did not compare results across status and value-added models, all using student covariates, as was done in this report.

[47] The rating systems compared by Abe et al. used different outcome measures as well as different ways of aggregating and comparing the measures. Thus the disparities in ratings they found were greater than the disparities we found in classifying districts as above average, average, or below average using the four models.

References

Abe, Y., Weinstock, P., Chan, V., Meyers, C., Gerdeman, R. D., & Brandt, W. C. (2015). *How methodology decisions affect the variability of schools identified as beating the odds* (REL 2015-071). Washington, DC: US Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. Retrieved from ies.ed.gov/ncee/edlabs.

ACT. (2008). *The forgotten middle: Ensuring that all students are on target for college and career readiness before high school.* Iowa City, IA: ACT. Retrieved from www.act.org/research/policymakers/reports/ForgottenMiddle.html.

ACT. (2012a). *Rising to the challenge of college and career readiness: A framework for effective practices.* Iowa City, IA: ACT. Retrieved from www.act.org/research/policymakers/pdf/RisingToChallenge.pdf.

ACT. (2012b). *Catching up to college and career readiness.* Iowa City, IA: ACT. Retrieved from www.act.org/research/policymakers/reports/catchingup.html.

ACT. (2012c). *Principles for measuring growth towards college and career readiness.* Iowa City, IA: ACT. Retrieved from www.act.org/research/policymakers/pdf/GrowthModelingReport.pdf.

ACT. (2013). *2013–2014 Technical Manual: ACT Explore®.* Iowa City, IA: ACT. Retrieved from www.act.org/explore/pdf/TechManual.pdf.

ACT. (2014). *Technical Manual: The ACT®.* Iowa City, IA: ACT. Retrieved from www.act.org/aap/pdf/ACT_Technical_Manual.pdf.

Bassiri, D. (2014). Research study: The forgotten middle. Iowa City, IA: ACT. Retrieved from www.act.org/research/policymakers/pdf/ForgottenMiddle-ResearchStudy2014.pdf.

Broad Foundations. (2014). The Broad prize for urban education. Web page. www.broadprize.org.

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement: Lessons from Chicago.* Chicago, IL: University of Chicago Press.

Chingos, M. M., Whitehurst, G. J., & Gallaher, M. R. (2013, March). *School districts and student achievement.* Washington, D.C.: The Brookings Institution, Brown Center on Education Policy.

Data Quality Campaign. (2010, November). *Creating reports using longitudinal data: How states can present information to support student learning and school system improvement.* Washington, D.C.: Data Quality Campaign. Retrieved from www.dataqualitycampaign.org/wp-content/uploads/files/1065_DQC%20Reports%20Issue%20Brief%20Nov8.pdf.

Dougherty, C. (2014). *Catching up to college and career readiness: The challenge is greater for at-risk students.* Iowa City, IA: ACT. Retrieved from www.act.org/research/policymakers/pdf/CatchingUp-Part3.pdf.

Dougherty, C. (2015). *Use of data to support teaching and learning: A case study of two school districts* (ACT Research Report 2015-1). Iowa City, IA: ACT.

Dougherty, C., Hiserote, L., & Shaw, T. (2014a). *Catching up to college and career readiness in Arkansas* (ACT Research Report 2014-3). Iowa City, IA: ACT.

Dougherty, C., Hiserote, L., & Shaw, T. (2014b). *Catching up to college and career readiness in Kentucky* (ACT Research Report 2014-4). Iowa City, IA: ACT.

Goldhaber, D., & Theobald, R. (2013). *Do different value-added models tell us the same things?* Carnegie Knowledge Network, What We Know Series: Value-Added Methods and Applications.

Haertel, E. H. (2013, March). Reliability and validity of inferences about teachers based on student test scores. Princeton, NJ: Educational Testing Service. Fourteenth William H. Angoff Memorial Lecture.

Hightower, A., Knapp, M., Marsh, J., & McLaughlin, M. (Eds.). (2002). *School districts and instructional renewal.* New York, NY: Teachers College Press.

Lissitz, R. W. (2012). *The evaluation of teacher and school effectiveness using growth models and value added modeling: Hope versus reality.* Paper presented at the annual meeting of the American Educational Research Association.

Luyten, H. (1994). Stability of school effects in Dutch secondary education: The impact of variance across subjects and years. *International Journal of Educational Research, 21*(2), 197–216.

Luyten, H. (1998). School effectiveness and student achievement, consistent across subjects?: Evidence from Dutch elementary and secondary education. *Educational Research and Evaluation: An International Journal on Theory and Practice, 4*(4), 281–306.

Ma, X. (2001). Stability of school academic performance across subject areas. *Journal of Educational Measurement, 38*(1), 1–18.

Mandeville, G. K. (1988). School effectiveness indices revisited: Cross-year stability. *Journal of Educational Measurement, 25*(4), 349–356.

Mandeville, G. K., & Anderson, L. W. (1987). The stability of school effectiveness indices across grade levels and subject areas. *Journal of Educational Measurement, 24*(3), 203–216.

Marsh, J. A., Kerr, K. A., Ikemoto, G. S., Darilek, H., Suttorp, M., Zimmer, R. W., & Barney, H. (2005). *The role of districts in fostering instructional improvement: Lessons from three urban districts partnered with the Institute for Learning.* Santa Monica, CA: Rand Corporation.

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

Mohiemang, I. L. & Pretorius, S. G. (2012). Stability and consistency of school effects and the implications for school improvement interventions: The case of Botswana. *Journal of Social Sciences, 30*(3), 273–281.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). V*alue-added modeling of teacher effectiveness: An exploration of stability across models and contexts*. Education Policy Analysis Archives, 18:23.

Sass, T. R. (2008, November). T*he stability of value-added measures of teacher quality and implications for teacher compensation policy* (CALDER issue brief 4). Washington, DC: Urban Institute.

Sawyer, R. (2013). *Interpreting changes over time in high school average ACT® college readiness assessment composite scores and ACT college readiness benchmark attainment rates* (ACT Research Report 2013-9). Iowa City, IA: ACT.

Supovitz, J. A. (2006). T*he case for district-based reform: Leading, building, and sustaining school improvement.* Cambridge, MA: Harvard University Press.

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., and Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11–35.

Whitehurst, G. J., Chingos, M. M., & Gallaher, M. R. (2013, March). *Do school districts matter*? Washington, DC: The Brookings Institution, Brown Center on Education Policy.

Appendix A

Tables A1 through A9
Numbers and Percentages of Students in the Analysis

Tables A1, A2, and A3 show the number and percentage of students from each cohort who were included in the statistical analysis. The highest attrition was in grades 8–12, owing both to a higher rate of enrollment attrition (27,852 in grades 8–12 versus 16,336 in grades 4–8 on a similar initial year enrollment base), and to a large percentage of students not taking the ACT Explore test in eighth grade or the ACT in grades 11–12. Higher enrollment attrition in the high school grades likely reflects students dropping out in those grades. The percentage of students taking ACT Explore rose substantially in 2010–11 when the State of Arkansas began paying the districts' costs of giving students the ACT Explore test in eighth grade. This policy increased the number and percentage of students in the 4–8 cohorts (Table A2) but not in the 8–12 cohorts (Table A3) in this study.

Table A4 illustrates how student attrition affected the percentages of students in various at-risk groups in the study cohorts. As would be expected, the students in the longitudinal cohorts—who were continuously enrolled in the same district, progressed by four grades in four years, and took all tests—were less at-risk than the general enrolled population in the initial cohort years. The percentages of low-income, African American, and special education students were lower in the study cohorts in all three levels (K–4, 4–8, and 8–12) than in the population from which the cohorts were drawn. Likewise, the percentages of Hispanic students and ELLs were lower in the high school study cohorts than in the population from which these cohorts were drawn. In addition, the attrition of low-income and special education students was greater in grades 8–12 than the other two levels.[48]

Tables A5 and A6 provide statistics on the demographics of the districts in the study. In general, fewer than half of the districts had substantial percentages of African American students, and Asian and Native American students were scarce across the state: The 95th percentile districts in the grades K–4 and 4–8 cohort analyses had 4% and 3% Asian and Native American students, respectively (Table A5). Likewise, districts across the state cope with significant amounts of student poverty: the 5th percentile district had 35% of students eligible for the free and reduced-price lunch program (Table A5). (As noted in the tables, the district that is at the Kth percentile on one variable is not necessarily the district at the same percentile on any other variable.) The "number of years between tests" reflects variation in testing dates: Students generally took the ABE on the same date in grade 4 and took ACT Explore on one of two different testing dates, while there was greater variation in when students took the ACT.

In Tables A7–A9, we show correlations of district-level statistics used in Models 2 and 4. Not surprisingly, districts' percentages of ELLs are strongly related to their percentages of Hispanic students (with a correlation of .94 in the grades 4 and 8 analyses and of .93 in the grades 11–12 analysis).[49] Other variables with moderately high correlations were the percentages of ELL and Asian students and the percentages of low-income and African American students. Districts with more low-income students had lower percentages of cohort students in the analysis, as shown by the negative correlations between those two variables in the bottom row of Tables A7–A9.

[48] If performance across indicators is less consistent for at-risk students, we would expect that losing (keeping) more of those students increases (reduces) the consistency of district performance statistics.

[49] This might indicate the possibility of paring down the model by dropping one of those two variables.

Table A1. Percentage of Arkansas Kindergarten Students in Grade 4 Analysis

Student Cohort[1]	Total Kindergarten Enrollment in K–12 Districts	Students Tested in Fourth Grade[2]	Students Eligible for Statistical Analysis[3]	Eligible Students in Eligible Districts	Percent of Students in Statistical Analysis
2007–2011	33,064	25,779	19,809	18,007	54%
2008–2012	35,944	27,494	21,029	18,501	51%
2009–2013	40,184	28,183	21,206	18,583	46%
Total	109,192	81,456	62,044	55,091	50%

[1] For example, the 2007–2011 cohort consists of students who were enrolled in kindergarten in the 2006–07 school year and who took the ABE tests in fourth grade in the 2010–11 school year.

[2] The attrition of 27,736 students between the first two data columns of this chart includes 262 kindergarten students with incomplete demographic data, 24,691 who were not enrolled in fourth grade four years later, and 2,783 students who were enrolled in fourth grade but did not take both state tests.

[3] The attrition of 19,412 students between the second and third columns of this chart consists of students who were enrolled in kindergarten and tested in both subjects four years later in grade 4, but who were not enrolled throughout grades K–4 and tested in grade 4 in the same district.

Table A2. Percentage of Arkansas Fourth-Grade Students in Grade 8 Analysis

Student Cohort	Total Fourth Grade Enrollment in K–12 Districts	Students Tested in Fourth and Eighth Grade[4]	Students Eligible for Statistical Analysis[5]	Eligible Students in Eligible Districts	Percent of Students in Statistical Analysis
2007–2011	34,567	25,509	20,281	18,013	52%
2008–2012	35,412	26,498	21,111	18,817	53%
2009–2013	38,188	28,018	21,883	19,358	51%
Total	108,167	80,025	63,275	56,188	52%

[4] The attrition of 28,142 students between the first two data columns of this chart includes eight students with incomplete demographic data; 16,336 students who were not enrolled in eighth grade four years later; 2,698 students who were enrolled four years later but had not taken both fourth-grade state tests; and 9,100 students enrolled and tested in grade 4 and enrolled in grade 8, but who did not take the ACT Explore test in eighth grade.

[5] The attrition of 16,750 students between the second and third columns of this chart consists of students enrolled and tested in both grades 4 and 8 but who were not enrolled throughout grades 4–8 and tested in grades 4 and 8 in the same district.

Table A3. Percentage of Arkansas Eighth-Grade Students in Grades 11–12 Analysis

Student Cohort	Total Eighth Grade Enrollment in K–12 Districts	Students Tested in Grade 8 and Grade 11 or 12[6]	Students Eligible for Statistical Analysis[7]	Eligible Students in Eligible Districts	Percent of Students in Statistical Analysis
2007–2011	34,774	10,016	8,466	7,294	21%
2008–2012	35,443	10,770	9,180	8,398	24%
2009–2013	37,208	14,078	11,455	10,143	27%
Total	107,425	34,864	29,101	25,835	24%

[6] The attrition of 72,561 students between the first two data columns of this chart includes 36 students with incomplete demographic data, 27,852 students who were enrolled in eighth grade but not in twelfth grade four years later; 31,520 students enrolled in grade 12 four years later but who had not taken ACT Explore in eighth grade; and 13,153 students enrolled and taking ACT Explore in grade 8 and following a normal grade progression between grades 8 and 12 but not taking the ACT.

[7] The attrition of 5,763 students between the second and third data columns of this chart consists of students who met the requirements for inclusion in the second column but who were not enrolled throughout grades 8–12 and tested in grades 8 and 11 or 12 in the same district.

Table A4. Demographics of Arkansas Student Cohorts

	Grades K–4		Grades 4–8		Grades 8–12	
Demographic Category	All Students in Initial Grade	Students in Cohort	All Students in Initial Grade	Students in Cohort	All Students in Initial Grade	Students in Cohort
Percent Low-Income	70%	63%	66%	59%	60%	46%
Percent African American	21%	18%	22%	18%	22%	19%
Percent Hispanic	11%	11%	9%	9%	8%	6%
Percent Asian	2%	2%	2%	1%	1%	1%
Percent Native American	1%	1%	1%	1%	1%	1%
Percent ELL	9%	10%	7%	7%	4%	2%
Percent Special Education	16%	13%	14%	11%	13%	5%

Table A5. Descriptive Statistics for Arkansas School Districts in the Grades 4 and 8 Analysis (N = 215)

	District Percentile						
District Statistic	5th	10th	25th	50th	75th	90th	95th
Percent Low-Income	35%	41%	52%	60%	69%	75%	84%
Percent African American	0%	0%	1%	2%	25%	54%	72%
Percent Hispanic	1%	1%	1%	2%	5%	11%	18%
Percent Asian	0%	0%	0%	0%	1%	2%	4%
Percent Native American	0%	0%	0%	0%	1%	1%	3%
Percent ELL	0%	0%	0%	1%	3%	7%	11%
Percent Special Education	9%	9%	10%	12%	14%	16%	17%
Number Students in Analysis—Grade 4	15	18	24	43	90	173	234
Number Students in Analysis—Grade 8	15	19	27	48	95	183	237
Percent Students in Analysis—Grade 4	35%	40%	46%	53%	58%	62%	65%
Percent Students in Analysis—Grade 8	42%	47%	54%	60%	66%	69%	73%
Number Years between Tests—Grade 4 to 8	3.59	3.59	3.59	3.59	3.67	3.67	3.67

Note: Because two districts did not have any students in the 2009–13 cohort, there were 213 districts in the distribution of the "number years between tests" variable. In general, reading down the columns, the percentiles refer to different districts: for example, the district with the median percentage of low-income students is not necessarily the district that with the median percentage of African American students.

Table A6. Descriptive Statistics for Arkansas School Districts in the Grades 11–12 Analysis (N = 165)

	District Percentile						
District Statistic	5th	10th	25th	50th	75th	90th	95th
Percent Low-Income	35%	38%	50%	60%	70%	77%	92%
Percent African American	0%	0%	1%	3%	24%	54%	72%
Percent Hispanic	1%	1%	1%	2%	5%	11%	18%
Percent Asian	0%	0%	0%	0%	1%	2%	4%
Percent Native American	0%	0%	0%	0%	1%	2%	3%
Percent ELL	0%	0%	0%	1%	2%	6%	11%
Percent Special Education	9%	10%	11%	12%	14%	16%	17%

Table A6. (continued)

District Statistic	District Percentile 5th	10th	25th	50th	75th	90th	95th
Number Students in Analysis—Grade 12	0	10	18	35	68	136	194
Percent Students in Analysis—Grade 12	0%	17%	35%	44%	50%	58%	60%
Number Years between Tests—Grade 8 to 12	3.65	3.73	3.83	3.94	4.04	4.12	4.16

Note: Because 14 districts did not have any students in the 2009–13 cohort, there were 151 districts in the distribution of the "number years between tests" variable. In general, reading down the columns, the percentiles refer to different districts: for example, the district with the median percentage of low-income students is not necessarily the district that with the median percentage of African American students.

Table A7. Pairwise Correlations between District-Level Statistics for Arkansas School Districts: Grade 4 Analysis (N = 215)

	Percent Low-Income	Percent African American	Percent Hispanic	Percent Asian	Percent Native American	Percent ELL	Percent Special Education	Number in Analysis
Percent African American	0.59							
	0.0000							
Percent Hispanic	0.06	-0.08						
	0.3925	*0.2725*						
Percent Asian	-0.16	-0.20	0.35					
	0.0191	*0.0033*	*0.0000*					
Percent Native American	-0.09	-0.25	0.15	0.36				
	0.1817	*0.0002*	*0.0284*	*0.0000*				
Percent ELL	0.06	-0.07	0.94	0.46	0.18			
	0.3849	*0.3146*	*0.0000*	*0.0000*	*0.0082*			
Percent Special Education	0.20	-0.16	-0.14	-0.10	0.00	-0.13		
	0.0037	*0.0210*	*0.0394*	*0.1367*	*0.9755*	*0.0557*		
Number Students in Analysis	-0.20	0.08	0.42	0.39	0.01	0.51	-0.20	
	0.0036	*0.2351*	*0.0000*	*0.0000*	*0.8845*	*0.0000*	*0.0039*	
Percent Students in Analysis	-0.41	-0.28	0.12	0.13	0.12	0.12	-0.22	0.22
	0.0000	*0.0000*	*0.0862*	*0.0560*	*0.0679*	*0.0779*	*0.0014*	*0.0013*

Note: p-values are in *italics*. These unweighted correlations and those in the next two tables were calculated using district-wide demographic data from the 2008–09 school year and the number and percentages of students in the analysis in the 2009–13 student cohort.

Table A8. Pairwise Correlations between District-Level Statistics for Arkansas School Districts: Grade 8 Analysis (N = 215)

	Percent Low-Income	Percent African American	Percent Hispanic	Percent Asian	Percent Native American	Percent ELL	Percent Special Education	Number in Analysis
Number Students in Analysis	-0.24	0.05	0.42	0.40	0.02	0.50	-0.20	
	0.0003	*0.4428*	*0.0000*	*0.0000*	*0.7971*	*0.0000*	*0.0029*	
Percent Students in Analysis	-0.43	-0.36	0.10	0.12	0.02	0.11	-0.12	0.13
	0.0000	*0.0000*	*0.1517*	*0.0694*	*0.7206*	*0.1236*	*0.0856*	*0.0482*

Note: p-values are in *italics*. The correlations of district-wide demographic percentages (not shown in this table) are the same as in Table A7 because the same district-wide demographic variables were used for all grade levels, and the set of districts included in the analysis was the same.

Table A9. Pairwise Correlations between District-Level Statistics for Arkansas School Districts: Grades 11–12 Analysis (N = 165)

	Percent Low-Income	Percent African American	Percent Hispanic	Percent Asian	Percent Native American	Percent ELL	Percent Special Education	Number in Analysis
Percent African American	0.61							
	0.0000							
Percent Hispanic	0.07	-0.08						
	0.3788	*0.3374*						
Percent Asian	-0.19	-0.20	0.19					
	0.0123	*0.0107*	*0.0145*					
Percent Native American	-0.18	-0.31	0.27	0.55				
	0.0225	0.0001	0.0004	0.0000				
Percent ELL	0.07	-0.05	0.93	0.29	0.32			
	0.3506	*0.5266*	*0.0000*	*0.0002*	*0.0000*			
Percent Special Education	0.14	-0.19	-0.17	-0.12	-0.12	-0.17		
	0.0719	*0.0167*	*0.0281*	*0.1134*	*0.1237*	*0.0318*		
Number Students in Analysis	-0.29	0.09	0.21	0.29	0.07	0.26	-0.19	
	0.0001	*0.2370*	*0.0063*	*0.0002*	*0.3826*	*0.0009*	*0.0140*	
Percent Students in Analysis	-0.22	-0.09	0.06	0.09	0.02	0.02	-0.21	0.26
	0.0040	*0.2365*	*0.4595*	*0.2738*	*0.7694*	*0.8049*	*0.0080*	*0.0008*

Appendix B

Tables B1 through B10
Fixed-Effect Coefficients from Statistical Models

Tables B1–B10 show the fixed-effect coefficients from the hierarchical models, measured in units of scale score points on the test used as the dependent variable. These are partial effects: For example, in Table B1, the fixed-effect coefficient for "low-income status" of -62.61 for Model 2 indicates that the predicted score of a low-income student is about 63 points lower on the grade 4 literacy ABE than the predicted score of a non-low-income student who has the same values of the other variables in the model. No interaction effects were modeled, i.e., we did not model how a student's ethnicity or prior score might affect the differences in predicted scores between low- and non-low-income students.

The tables also show the standard deviation of the district random effects in each model, labeled as "SD of random effects (pts)." This statistic estimates the variation across districts in the true random effect. For example, in Table B1, the "SD of random effects (pts)" is 29.1 for Model 2. Given the assumed normal distribution of the random effect in the model, for approximately two-thirds of the districts, the random effect is 29.1 score points or less; for approximately 95% of the districts, the random effect is 58.2 score points or less.

We also converted these standard deviations, measured in score points, into standardized form by dividing them by the standard deviation of student scores on the test in question. This facilitates comparisons with the sizes of the district effects of above- and below-average districts shown in Tables 6–11. For example, for ABE grade 4 literacy in Model 2, the standard deviation of 29.1 score points translates into a standardized standard deviation of .16. Thus, the random effect is .16 of a test score standard deviation or less in approximately two-thirds of the districts and .32 standard deviations or less in approximately 95% of the districts.

The sizes of the fixed effects and the district random effects in the tables can also be roughly compared with each other, with the caveat that these comparisons might change if we explored a wider variety of models. For example, based on the coefficient of the eighth-grade reading score in Table B9, Model 4, increasing the student's ACT Explore reading score by three points, or about three-quarters of a standard deviation, increases the student's predicted ACT score by over one and a half points (.54 x 3 = 1.62 points). This is larger than the difference in predicted score made by the student's poverty status (-.34 points) or African American minority status (-1.33 points). This reflects the large importance of prior student academic preparation found in prior ACT research (e.g., ACT, 2008; Dougherty, 2014; Bassiri, 2014).

Table B1. Fixed-Effect Coefficients in Value-Added Regressions Predicting Grade 4 Literacy Scores

Variable	Model 1		Model 2	
Intercept	777.12	***	794.22	***
Low-Income Status	-62.70	***	-62.61	***
African American Status	-62.01	***	-60.42	***
Hispanic Status	-1.50		-1.48	
Asian Status	28.98	***	29.31	***
Native American Status	1.21		1.80	
ELL Status	-40.07	***	-40.23	***
Special Education Status	-170.44	***	-170.56	***
District Percent Low-Income			-0.28	
District Percent African American			-0.31	**
District Percent Hispanic			1.00	
District Percent Asian			-1.29	
District Percent Native American			-4.50	***
District Percent ELL			-0.08	
District Percent Special Education			0.80	
District Number Students in Model	-0.02		-0.02	
District Percent Students in Model	0.20		0.08	
Earlier Record Deleted†	-97.55	***	-97.76	***
Took Fourth-Grade Test in 2012	41.28	***	41.98	***
Took Fourth-Grade Test in 2013	39.17	***	39.53	***
SD of Random Effect (pts)	30.9		29.1	
SD of Random Effect (std) ††	0.17		0.16	

*** Significant at the .01 level. ** Significant at the .05 level. * Significant at the .10 level.

† These students were retained in kindergarten in the 2007–08 or 2008–09 school years and thus had more than one record once data from different years were merged. When that occurred, the student's most recent record was retained and the record was flagged.

†† Obtained by dividing the random effect standard deviation in score points by 186.37, the standard deviation of students' fourth-grade literacy scores in the 2006–07 and 2007–08 school years.

Table B2. Fixed-Effect Coefficients in Value-Added Regressions Predicting Grade 4 Mathematics Scores

Variable	Model 1		Model 2	
Intercept	684.13	***	685.32	***
Low-Income Status	-37.21	***	-37.17	***
African American Status	-49.77	***	-49.58	***
Hispanic Status	-7.40	***	-7.34	***
Asian Status	12.89	***	12.96	***
Native American Status	-5.92		-5.67	
ELL Status	-16.75	***	-16.92	***
Special Education Status	-71.05	***	-71.10	***
District Percent Low-Income			-0.15	
District Percent African American			0.02	
District Percent Hispanic			-0.38	
District Percent Asian			0.07	
District Percent Native American			-1.87	**
District Percent ELL			1.08	*
District Percent Special Education			0.72	*
District Number Students in Model	0.01	*	0.00	
District Percent Students in Model	0.12		0.11	
Earlier Record Deleted†	-46.70	***	-46.74	***
Took Fourth-Grade Test in 2012	-1.16		-0.55	
Took Fourth-Grade Test in 2013	-0.92		-0.33	
SD of Random Effect (pts)	16.7		17.0	
SD of Random Effect (std) ††	0.17		0.17	

*** Significant at the .01 level. ** Significant at the .05 level. * Significant at the .10 level.

† These students were retained in kindergarten in the 2007–08 or 2008–09 school years and thus had more than one record once data from different years were merged. When that occurred, the student's most recent record was retained and the record was flagged.

†† Obtained by dividing the random effect standard deviation in score points by 100.93, the standard deviation of students' fourth-grade mathematics scores in the 2006–07 and 2007–08 school years.

Table B3. Fixed-Effect Coefficients in Value-Added Regressions Predicting Grade 8 English Scores

Variable	Model 1		Model 2		Model 3		Model 4	
Intercept	15.99	***	17.02	***	-0.21		6.24	***
Low-Income Status	-1.65	***	-1.64	***	-0.45	***	-0.44	***
African American Status	-1.91	***	-1.84	***	-0.35	***	-0.34	***
Hispanic Status	-0.09		-0.09		-0.11	*	-0.12	*
Asian Status	0.74	***	0.74	***	0.32	***	0.32	***
Native American Status	-0.41	**	-0.40	**	-0.24	*	-0.23	
ELL Status	-1.99	***	-1.99	***	0.01		0.03	
Special Education Status	-3.51	***	-3.51	***	0.20	***	0.21	***
Fourth-Grade Literacy Score					0.01	***	0.01	***
Fourth-Grade Mathematics Score					0.01	***	0.01	***
Number Years between Tests					0.28		0.21	
District Percent Low-Income			-0.01	***			-0.01	***
District Percent African American			0.00	*			0.00	**
District Percent Hispanic			0.01				0.02	*
District Percent Asian			0.01				0.03	
District Percent Native American			-0.05				-0.03	
District Percent ELL			0.00				-0.02	
District Percent Special Education			-0.01				-0.02	
District Avg. Grade 4 Literacy Score							0.00	***
District Avg. Grade 4 Mathematics Score							-0.01	***
District Number Students in Model	0.00	***	0.00		0.00		0.00	
District Percent Students in Model	0.00		-0.01	***	0.00		0.00	
Earlier Record Deleted†	-1.99	***	-1.98	***	-0.94	***	-0.94	***
Took Eighth-Grade Test in 2012	0.20	***	0.21	***	-0.56	***	-0.34	***
Took Eighth-Grade Test in 2013	0.38	***	0.40	***	-0.63	***	-0.26	***
SD of Random Effect (pts)	0.5		0.5		0.4		0.4	
SD of Random Effect (std) ††	0.12		0.11		0.11		0.09	

*** Significant at the .01 level. ** Significant at the .05 level. * Significant at the .10 level.

† These students were retained in fourth grade in the 2007–08 or 2008–09 school years and thus had more than one record once data from different years were merged. When that occurred, the student's most recent record was retained and the record was flagged.

†† Obtained by dividing the random effect standard deviation in score points by 4.2, the standard deviation of students' ACT Explore English scores.

Table B4. Fixed-Effect Coefficients in Value-Added Regressions Predicting Grade 8 Mathematics Scores

Variable	Model 1		Model 2		Model 3		Model 4	
Intercept	16.93	***	17.94	***	0.18		6.16	***
Low-Income Status	-1.30	***	-1.29	***	-0.43	***	-0.43	***
African American Status	-1.45	***	-1.42	***	-0.21	***	-0.20	***
Hispanic Status	0.08		0.07		0.05		0.04	
Asian Status	0.86	***	0.86	***	0.58	***	0.58	***
Native American Status	0.04		0.05		0.14		0.16	
ELL Status	-1.25	***	-1.25	***	0.09		0.10	
Special Education Status	-2.91	***	-2.91	***	-0.38	***	-0.37	***
Fourth-Grade Literacy Score					0.00	***	0.00	***
Fourth-Grade Mathematics Score					0.02	***	0.02	***
Number Years between Tests					0.86	***	0.72	**
District Percent Low-Income			-0.01	***			-0.01	***
District Percent African American			0.00				0.00	
District Percent Hispanic			0.02				0.03	***
District Percent Asian			-0.01				0.00	
District Percent Native American			-0.05	*			-0.05	**
District Percent ELL			0.00				-0.03	**
District Percent Special Education			-0.01				-0.02	*
District Avg. Grade 4 Literacy Score							0.00	
District Avg. Grade 4 Mathematics Score							-0.01	***
District Number Students in Model	0.00		0.00		0.00		0.00	
District Percent Students in Model	0.00		-0.01	**	0.00		0.00	
Earlier Record Deleted†	-1.78	***	-1.78	***	-0.98	***	-0.98	***
Took Eighth-Grade Test in 2012	0.18	***	0.19	***	-0.35	***	-0.22	***
Took Eighth-Grade Test in 2013	0.13	***	0.16	***	-0.68	***	-0.41	***
SD of Random Effect (pts)	0.5		0.5		0.4		0.4	
SD of Random Effect (std) ††	0.15		0.14		0.12		0.12	

*** Significant at the .01 level. ** Significant at the .05 level. * Significant at the .10 level.

† These students were retained in fourth grade in the 2007–08 or 2008–09 school years and thus had more than one record once data from different years were merged. When that occurred, the student's most recent record was retained and the record was flagged.

†† Obtained by dividing the random effect standard deviation in score points by 3.5, the standard deviation of students' ACT Explore mathematics scores.

Table B5. Fixed-Effect Coefficients in Value-Added Regressions Predicting Grade 8 Reading Scores

Variable	Model 1		Model 2		Model 3		Model 4	
Intercept	15.51	***	16.65	***	2.56	**	7.80	***
Low-Income Status	-1.39	***	-1.39	***	-0.36	***	-0.36	***
African American Status	-1.65	***	-1.57	***	-0.34	***	-0.30	***
Hispanic Status	0.00		0.00		-0.01		-0.02	
Asian Status	0.74	***	0.74	***	0.39	***	0.38	***
Native American Status	-0.16		-0.16		-0.01		-0.01	
ELL Status	-1.66	***	-1.66	***	0.08		0.09	
Special Education Status	-2.69	***	-2.69	***	0.52	***	0.53	***
Fourth-Grade Literacy Score					0.01	***	0.01	***
Fourth-Grade Mathematics Score					0.01	***	0.01	***
Number Years between Tests					0.11		0.04	
District Percent Low-Income			-0.01	**			-0.01	***
District Percent African American			-0.01	***			-0.01	***
District Percent Hispanic			0.01				0.01	
District Percent Asian			-0.02				0.00	
District Percent Native American			-0.03				-0.01	
District Percent ELL			0.01				-0.01	
District Percent Special Education			-0.02	*			-0.03	***
District Avg. Grade 4 Literacy Score							0.00	***
District Avg. Grade 4 Mathematics Score							0.00	***
District Number Students in Model	0.00	***	0.00	*	0.00		0.00	
District Percent Students in Model	0.00		-0.01	*	0.00		0.00	
Earlier Record Deleted†	-1.70	***	-1.68	***	-0.81	***	-0.80	***
Took Eighth-Grade Test in 2012	0.30	***	0.30	***	-0.35	***	-0.16	***
Took Eighth-Grade Test in 2013	0.45	***	0.45	***	-0.41	***	-0.11	**
SD of Random Effect (pts)	0.5		0.5		0.4		0.4	
SD of Random Effect (std) ††	0.13		0.12		0.11		0.10	

*** Significant at the .01 level. ** Significant at the .05 level. * Significant at the .10 level.

† These students were retained in fourth grade in the 2007–08 or 2008–09 school years and thus had more than one record once data from different years were merged. When that occurred, the student's most recent record was retained and the record was flagged.

†† Obtained by dividing the random effect standard deviation in score points by 3.9, the standard deviation of students' ACT Explore reading scores.

Table B6. Fixed-Effect Coefficients in Value-Added Regressions Predicting Grade 8 Science Scores

Variable	Model 1		Model 2		Model 3		Model 4	
Intercept	17.43	***	18.34	***	6.13	***	11.09	***
Low-Income Status	-1.15	***	-1.14	***	-0.35	***	-0.35	***
African American Status	-1.28	***	-1.23	***	-0.20	***	-0.18	***
Hispanic Status	0.08		0.07		0.06		0.05	
Asian Status	0.81	***	0.81	***	0.55	***	0.54	***
Native American Status	0.00		0.00		0.10		0.11	
ELL Status	-1.24	***	-1.24	***	0.05		0.06	
Special Education Status	-2.24	***	-2.24	***	0.16	***	0.16	***
Fourth-Grade Literacy Score					0.01	***	0.01	***
Fourth-Grade Mathematics Score					0.01	***	0.01	***
Number Years between Tests					0.06		-0.05	
District Percent Low-Income			-0.01	**			-0.01	***
District Percent African American			0.00				0.00	
District Percent Hispanic			0.01				0.02	*
District Percent Asian			0.02				0.04	
District Percent Native American			-0.02				-0.01	
District Percent ELL			0.00				-0.02	
District Percent Special Education			-0.02	*			-0.02	**
District Avg. Grade 4 Literacy Score							0.00	
District Avg. Grade 4 Mathematics Score							-0.01	***
District Number Students in Model	0.00	***	0.00	**	0.00	***	0.00	
District Percent Students in Model	0.00		-0.01	**	0.00		0.00	
Earlier Record Deleted†	-1.49	***	-1.48	***	-0.78	***	-0.78	***
Took Eighth-Grade Test in 2012	-0.05		-0.04		-0.53	***	-0.43	***
Took Eighth-Grade Test in 2013	0.25	***	0.26	***	-0.44	***	-0.22	***
SD of Random Effect (pts)	0.5		0.4		0.4		0.4	
SD of Random Effect (std) ††	0.14		0.13		0.13		0.12	

*** Significant at the .01 level. ** Significant at the .05 level. * Significant at the .10 level.

† These students were retained in fourth grade in the 2007–08 or 2008–09 school years and thus had more than one record once data from different years were merged. When that occurred, the student's most recent record was retained and the record was flagged.

†† Obtained by dividing the random effect standard deviation in score points by 3.3, the standard deviation of students' ACT Explore science scores.

Table B7. Fixed-Effect Coefficients in Value-Added Regressions Predicting Grades 11–12 English Scores

Variable	Model 1		Model 2		Model 3		Model 4	
Intercept	23.27	***	24.91	***	-4.83	***	0.93	
Low-Income Status	-2.37	***	-2.33	***	-0.91	***	-0.88	***
African American Status	-4.16	***	-4.20	***	-1.24	***	-1.32	***
Hispanic Status	-1.55	***	-1.59	***	-0.77	***	-0.78	***
Asian Status	1.09	***	1.05	***	0.84	***	0.81	***
Native American Status	-0.53		-0.55		-0.12		-0.11	
ELL Status	-3.84	***	-3.84	***	-0.84	***	-0.84	***
Special Education Status	-6.22	***	-6.21	***	-1.17	***	-1.15	***
Eighth-Grade English Score					0.62	***	0.62	***
Eighth-Grade Mathematics Score					0.28	***	0.28	***
Eighth-Grade Reading Score					0.29	***	0.29	***
Eighth-Grade Science Score					0.22	***	0.23	***
Number Years between Tests					1.15	***	1.14	***
District Percent Low-Income			-0.03	***			-0.02	***
District Percent African American			0.02	***			0.01	***
District Percent Hispanic			0.02				-0.01	
District Percent Asian			0.17	**			0.09	*
District Percent Native American			-0.09				-0.03	
District Percent ELL			0.00				0.02	
District Percent Special Education			-0.02				-0.02	
District Avg. Grade 8 English Score							-0.16	**
District Avg. Grade 8 Mathematics Score							-0.03	
District Avg. Grade 8 Reading Score							0.07	
District Avg. Grade 8 Science Score							-0.18	**
District Number Students in Model	0.00	***	0.00	***	0.00	***	0.00	**
District Percent Students in Model	-0.03	***	-0.03	***	-0.01	***	-0.01	***
Earlier Record Deleted†	-3.70	***	-3.79	***	0.49		0.43	
Twelfth Grader in 2012	-0.28	***	-0.22	**	-0.26	***	-0.18	***
Twelfth Grader in 2013	-0.35	***	-0.20	*	-0.15	**	-0.01	
SD of Random Effect (pts)	0.9		0.8		0.5		0.5	
SD of Random Effect (std) ††	0.14		0.12		0.08		0.08	

*** Significant at the .01 level. ** Significant at the .05 level. * Significant at the .10 level.

† These students were retained in eighth grade in the 2007–08 or 2008–09 school years and thus had more than one record once data from different years were merged. When that occurred, the student's most recent record was retained and the record was flagged.

†† Obtained by dividing the random effect standard deviation in score points by 6.5, the standard deviation of students' ACT English scores.

Table B8. Fixed-Effect Coefficients in Value-Added Regressions Predicting Grades 11–12 Mathematics Scores

Variable	Model 1		Model 2		Model 3		Model 4	
Intercept	21.71	***	22.98	***	2.88	***	6.32	***
Low-Income Status	-1.52	***	-1.50	***	-0.58	***	-0.56	***
African American Status	-2.78	***	-2.77	***	-0.71	***	-0.72	***
Hispanic Status	-0.83	***	-0.85	***	-0.29	***	-0.31	***
Asian Status	1.85	***	1.82	***	1.49	***	1.46	***
Native American Status	-0.53	*	-0.56	*	-0.07		-0.09	
ELL Status	-1.81	***	-1.81	***	0.10		0.11	
Special Education Status	-3.62	***	-3.62	***	0.08		0.09	
Eighth-Grade English Score					0.18	***	0.19	***
Eighth-Grade Mathematics Score					0.53	***	0.53	***
Eighth-Grade Reading Score					0.07	***	0.07	***
Eighth-Grade Science Score					0.29	***	0.29	***
Number Years between Tests					0.21	***	0.20	***
District Percent Low-Income			-0.03	***			-0.02	***
District Percent African American			0.01	***			0.01	***
District Percent Hispanic			0.04				0.01	
District Percent Asian			0.23	***			0.14	***
District Percent Native American			-0.03				0.02	
District Percent ELL			-0.02				0.00	
District Percent Special Education			0.00				0.00	
District Avg. Grade 8 English Score							-0.19	***
District Avg. Grade 8 Mathematics Score							-0.05	
District Avg. Grade 8 Reading Score							0.14	
District Avg. Grade 8 Science Score							-0.07	
District Number Students in Model	0.00	***	0.00	*	0.00	***	0.00	
District Percent Students in Model	-0.01	***	-0.01	***	-0.01	**	0.00	
Earlier Record Deleted†	-1.63	**	-1.69	**	1.32	**	1.26	**
Twelfth Grader in 2012	-0.20	***	-0.13	*	-0.22	***	-0.13	**
Twelfth Grader in 2013	-0.26	***	-0.14	*	-0.08		0.08	
SD of Random Effect (pts)	0.9		0.8		0.6		0.6	
SD of Random Effect (std) ††	0.17		0.15		0.11		0.11	

*** Significant at the .01 level. ** Significant at the .05 level. * Significant at the .10 level.

† These students were retained in eighth grade in the 2007–08 or 2008–09 school years and thus had more than one record once data from different years were merged. When that occurred, the student's most recent record was retained and the record was flagged.

†† Obtained by dividing the random effect standard deviation in score points by 5.3, the standard deviation of students' ACT mathematics scores.

Table B9. Fixed-Effect Coefficients in Value-Added Regressions Predicting Grades 11–12 Reading Scores

Variable	Model 1		Model 2		Model 3		Model 4	
Intercept	23.24	***	24.52	***	-1.12	***	3.01	***
Low-Income Status	-1.67	***	-1.64	***	-0.36	***	-0.34	***
African American Status	-4.07	***	-4.04	***	-1.30	***	-1.33	***
Hispanic Status	-1.47	***	-1.49	***	-0.78	***	-0.79	***
Asian Status	0.52	*	0.47		0.35	*	0.31	
Native American Status	-0.03		-0.08		0.34		0.30	
ELL Status	-3.32	***	-3.34	***	-0.53	***	-0.55	***
Special Education Status	-4.81	***	-4.80	***	-0.38	***	-0.37	***
Eighth-Grade English Score					0.38	***	0.38	***
Eighth-Grade Mathematics Score					0.16	***	0.16	***
Eighth-Grade Reading Score					0.54	***	0.54	***
Eighth-Grade Science Score					0.26	***	0.26	***
Number Years between Tests					0.58	***	0.57	***
District Percent Low-Income			-0.03	***			-0.02	***
District Percent African American			0.01	**			0.01	**
District Percent Hispanic			-0.01				-0.04	**
District Percent Asian			0.15	**			0.05	
District Percent Native American			0.01				0.06	
District Percent ELL			0.04				0.07	**
District Percent Special Education			0.00				0.00	
District Avg. Grade 8 English Score							-0.07	
District Avg. Grade 8 Mathematics Score							0.04	
District Avg. Grade 8 Reading Score							-0.02	
District Avg. Grade 8 Science Score							-0.17	*
District Number Students in Model	0.00	***	0.00	**	0.00	**	0.00	
District Percent Students in Model	-0.02	***	-0.03	***	-0.01	***	-0.01	***
Earlier Record Deleted†	-2.36	**	-2.42	**	1.30	*	1.26	*
Twelfth Grader in 2012	-0.15	*	-0.09		-0.11	*	-0.05	
Twelfth Grader in 2013	-0.25	***	-0.12		0.02		0.11	
SD of Random Effect (pts)	0.8		0.7		0.5		0.4	
SD of Random Effect (std) ††	0.13		0.11		0.08		0.07	

*** Significant at the .01 level. ** Significant at the .05 level. * Significant at the .10 level.

† These students were retained in eighth grade in the 2007–08 or 2008–09 school years and thus had more than one record once data from different years were merged. When that occurred, the student's most recent record was retained and the record was flagged.

†† Obtained by dividing the random effect standard deviation in score points by 6.3, the standard deviation of students' ACT reading scores.

Table B10. Fixed-Effect Coefficients in Value-Added Regressions Predicting Grades 11–12 Science Scores

Variable	Model 1		Model 2		Model 3		Model 4	
Intercept	22.18	***	23.59	***	2.54	***	5.08	***
Low-Income Status	-1.36	***	-1.32	***	-0.40	***	-0.37	***
African American Status	-3.19	***	-3.18	***	-1.11	***	-1.14	***
Hispanic Status	-1.12	***	-1.14	***	-0.62	***	-0.62	***
Asian Status	1.10	***	1.06	***	0.78	***	0.75	***
Native American Status	-0.35		-0.37		0.02		0.00	
ELL Status	-2.20	***	-2.20	***	-0.22		-0.24	
Special Education Status	-3.70	***	-3.69	***	-0.14		-0.14	
Eighth-Grade English Score					0.18	***	0.18	***
Eighth-Grade Mathematics Score					0.37	***	0.37	***
Eighth-Grade Reading Score					0.19	***	0.19	***
Eighth-Grade Science Score					0.34	***	0.34	***
Number Years between Tests					0.40	***	0.40	***
District Percent Low-Income			-0.03	***			-0.02	***
District Percent African American			0.01	***			0.01	***
District Percent Hispanic			0.01				-0.02	
District Percent Asian			0.21	***			0.10	***
District Percent Native American			-0.09				-0.01	
District Percent ELL			0.00				0.03	
District Percent Special Education			-0.01				0.00	
District Avg. Grade 8 English Score							0.10	
District Avg. Grade 8 Mathematics Score							-0.02	
District Avg. Grade 8 Reading Score							-0.10	
District Avg. Grade 8 Science Score							-0.09	
District Number Students in Model	0.00	***	0.00	***	0.00	***	0.00	***
District Percent Students in Model	-0.02	***	-0.02	***	-0.01	***	-0.01	***
Earlier Record Deleted†	-2.24	**	-2.31	***	0.67		0.64	
Twelfth Grader in 2012	-0.33	***	-0.27	***	-0.33	***	-0.27	***
Twelfth Grader in 2013	-0.35	***	-0.21	***	-0.14	**	-0.08	
SD of Random Effect (pts)	0.7		0.6		0.4		0.3	
SD of Random Effect (std)	0.14		0.12		0.08		0.06	

*** Significant at the .01 level. ** Significant at the .05 level. * Significant at the .10 level.

† These students were retained in eighth grade in the 2007–08 or 2008–09 school years and thus had more than one record once data from different years were merged. When that occurred, the student's most recent record was retained and the record was flagged.

†† Obtained by dividing the random effect standard deviation in score points by 5.3, the standard deviation of students' ACT science scores.

Appendix C

Number of Below-Average Districts across Multiple Subjects and Levels

In this appendix we provide information on the number and percentage of districts that were below average across multiple subjects and grade levels.

a. *How many districts were below average across all subjects in a given level?*

In grade 4, 14 (7% of 215 eligible districts) were below average in both literacy and mathematics in Model 2, compared with 43 districts that were below average in one subject but not the other (Figure C1). In grade 8, four districts (2% of eligible districts) were below average in all four subjects in Model 2 and six districts (3%) in all four subjects in Model 4 (Figure C2). In grades 11–12, three districts (2% of 165 eligible districts) were below average in all four subjects in Model 2 and no districts in Model 4 (Figure C3).

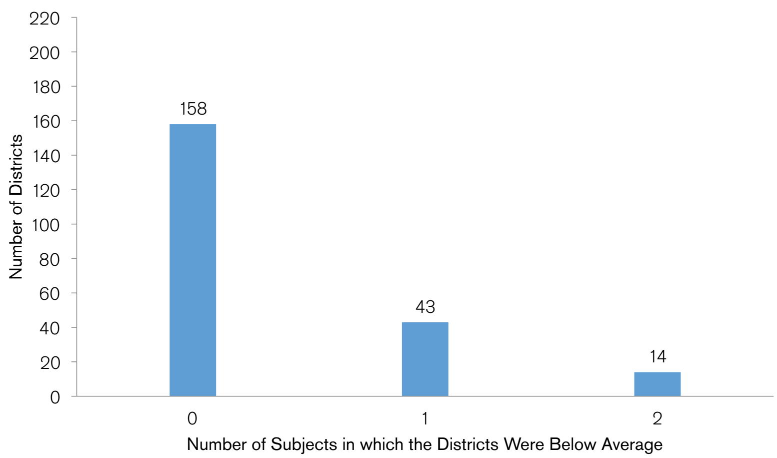

Figure C1. Number of districts in grade 4 by the number of subjects (out of two) in which they were below average: Model 2 (N = 215 districts)

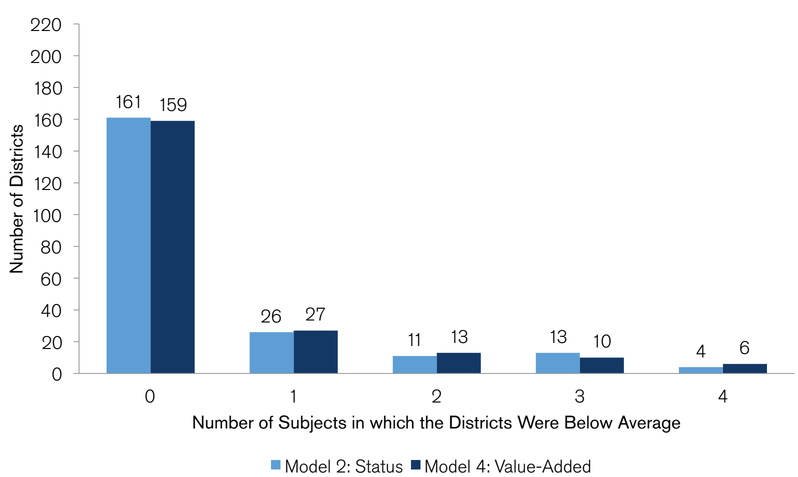

Figure C2. Number of districts in grade 8 by the number of subjects (out of four) in which they were below average: Models 2 and 4 (N = 215 districts)

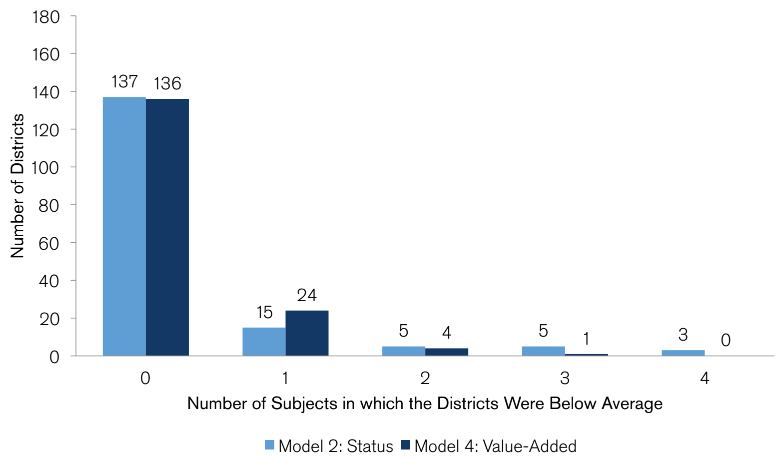

Figure C3. Number of districts in grades 11–12 by the number of subjects (out of four) in which they were below average: Models 2 and 4 (N = 165 districts)

b. *How many districts were below average for all levels in a given subject?*

Using Model 2, three districts (2% of 165) were below average in mathematics in all three levels (grades 4, 8, and 11–12), while two districts (1% of 165) were below average in all three levels in reading/literacy (Figure C4). English and science were not tested in grade 4, so a three-level comparison was not available for those two subjects. Looking at a two-level comparison based on Model 4, two districts were below average in English in both grades 8 and 11–12, three districts in mathematics, and no districts in reading or science (Figure C5).[50]

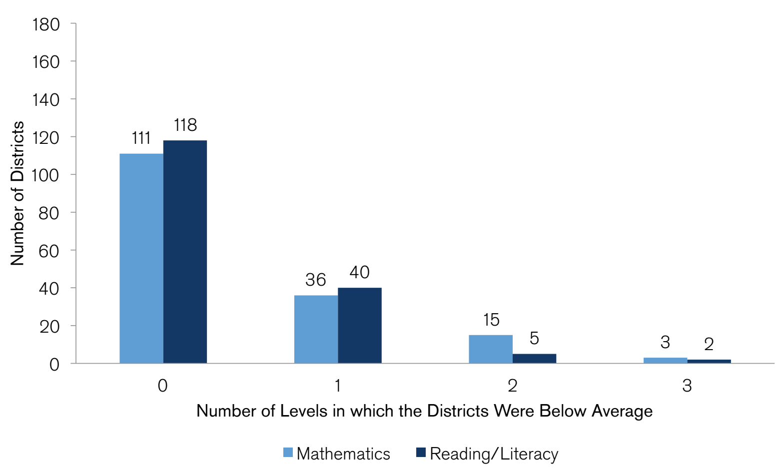

Figure C4. Number of districts by the number of levels (out of three) in which they were below average in each subject: Model 2 (N = 165 districts)

[50] These were not necessarily the same districts, as is evident from the answer to the next question.

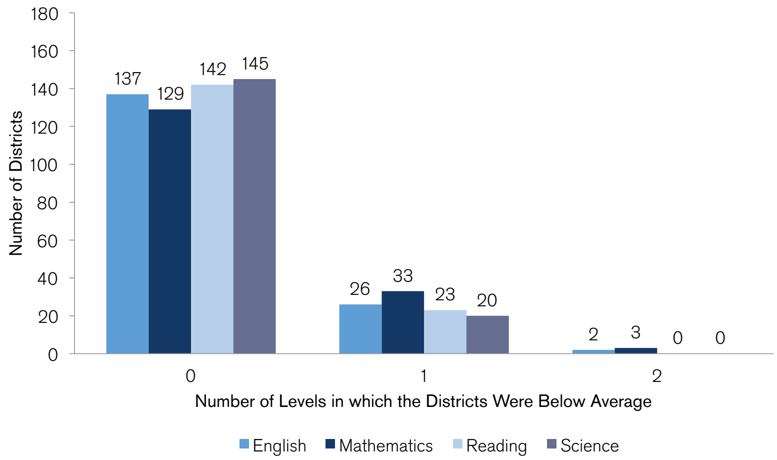

Figure C5. Number of districts by the number of levels (out of two) in which they were below average in each subject: Model 4 (N = 165 districts)

c. *How many districts were below average across all subjects and levels for which data were available in the model?*

Using Model 2, there were ten possible combinations of subjects and levels: literacy and mathematics in grade 4 and English, mathematics, reading, and science in each of grades 8 and 12. No district was below average across all ten of these combinations, and one district was below average for eight of the ten combinations (Figure C6). Four districts (2% of 165 districts eligible in all three levels) were below average in more than half of the subject-level combinations.

Using Model 4, there were eight combinations of subjects and levels: four subjects each in grades 8 and 11–12. No district was below average for all eight combinations, and no districts were below average for the majority of subject-level combinations (Figure C7).

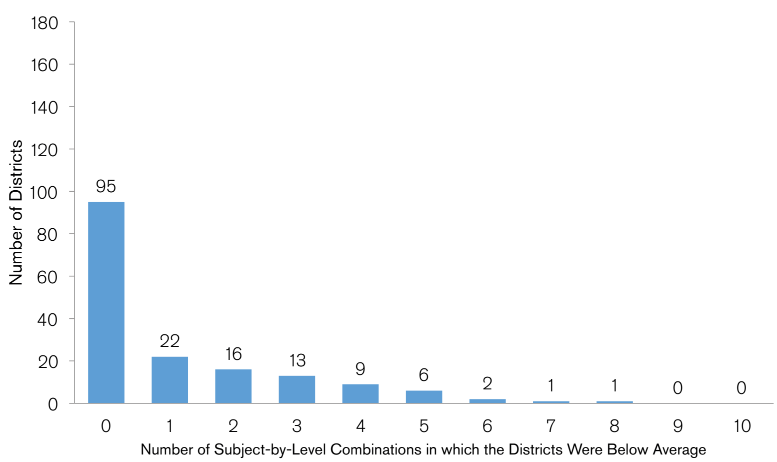

Figure C6. Number of districts by the total number of subject-level combinations (out of ten) in which they were below average: Model 2 (N = 165 districts)

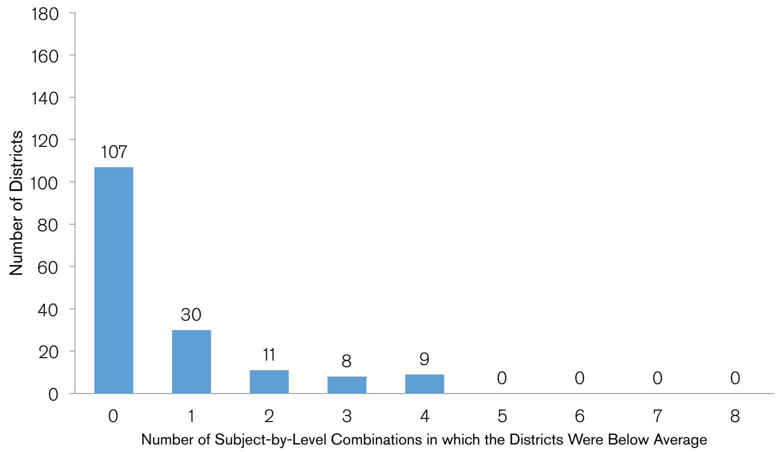

Figure C7. Number of districts by the total number of subject-level combinations (out of eight) in which they were below average: Model 4 (N = 165 districts)

070155160 Rev 1