Methodological Summary: Analysis of Kentucky School Performance on Grade 3 Mathematics and Reading State Assessments

Background

In 2018, the Kentucky Department of Education (KDE) released a new strategic plan (KDE, 2018) prioritizing improved outcomes for students in mathematics and reading. As described in the plan, KDE's retrospective analyses of Kentucky students' data demonstrated that a majority of students in the 2018/19 grade 9 cohort who scored proficient in mathematics did so initially in grade 3—the first year they were tested; the same was true for reading. Of those grade 9 students who had ever scored proficient in math, 63 percent did so initially in grade 3; the corresponding statistic for reading was 61 percent. Given these results, KDE concluded that having strong foundational literacy and numeracy skills set these students up for success. As a result, KDE is pursuing efforts to get more students on track academically in their early years so that by grade 3 they are scoring at or above proficient in mathematics and reading.

To further this objective, Regional Educational Laboratory Appalachia (REL AP) supports KDE staff with training, coaching, and technical support to execute quantitative analyses aimed at identifying schools that are doing better, worse, or about the same as statistically predicted on outcomes of interest, given certain non-malleable factors, such as demographic characteristics.[1] Two research analysts in the Kentucky Commissioner's Office codesigned the analysis and are in the process of replicating the quantitative analyses with REL coaching support. REL AP staff have worked with these analysts to enhance their capacity to design and execute relevant quantitative analyses and share results with their leadership and other stakeholders. REL AP plans to support KDE's continued learning about schools performing better, worse, or about the same as predicted through ongoing coaching with the two research analysts. Specifically, we will provide coaching on their analysis of extant survey data and their collection of qualitative data to identify practices associated with success in schools that outperform predictions. Part of this endeavor may be to identify whether practices identified as evidence-based in federal clearinghouses, including the What Works Clearinghouse, are more prevalent

[1] The REL program has several publications using similar analyses: see Abe and colleagues (2015); Culbertson and Billig (2016); Koon, Petscher, and Foorman (2014); Meyers and Wan (2016); Partridge and Koon (2017); and Partridge, Rudo, and Herrera (2017).

in schools that outperform predictions than in other schools.

Several individuals are currently involved in this project from KDE. Two research analysts who work in the Commissioner's office codeveloped this project with REL AP staff, with one taking the lead role and the other collaborating substantively throughout the project. With coaching and technical support from REL AP staff, these analysts make all final design decisions, replicate quantitative analysis, and will conduct the extant data analysis and any additional data collection in the follow-on activities. The project also involves the state's chief performance officer and the associate commissioner, Office of Teaching and Learning, who provide strategic guidance and oversight. KDE invites additional staff to meetings with REL AP as needed. For example, the director of the division of program standards and an academy program consultant guide and advise REL AP and the core KDE staff on the development of follow-on activities to ensure the results can inform KDE-supported professional development efforts.

This document is a *methodological summary* of quantitative analyses performed by REL AP and KDE analysts. It is coupled with a PowerPoint slide deck describing results from a subset of quantitative analyses completed as of winter 2020.

- The primary audience for the methodological summary is the KDE analysts who have worked with REL AP to design and execute the analyses. The summary will provide a reference for the KDE analysts moving forward as they perform similar work in the future. The summary will also provide reference information to any broader research audiences that REL AP may engage with in cooperation with KDE.
- The primary intended audience for the PowerPoint presentation is KDE leadership. As such, the presentation has a sharper focus. Per KDE analysts' request, after providing background information on the full set of quantitative analyses, the presentation focuses on results for the second of two research questions described below. REL AP may also repurpose slides for additional presentations delivered with KDE staff to broader audiences (for example, REL AP webinar, National Center for Education Statistics STATS-DC conference).

The *methodological summary* serves two purposes. First, it describes how REL AP and KDE staff generated statistical models to predict school performance and changes in school performance over time based on the demographic makeup of schools and shifts in these student populations over time. Second, it describes how REL AP and KDE staff compared these predictions with actual school performance and change over time. The quantitative analyses addressed four school-level outcomes of interest: grade 3 mathematics scale scores (math status), grade 3 reading scale scores (reading status), growth in grade 3 mathematics scale scores over time (math growth), and growth in grade 3 reading scale scores over time (reading growth). Schools in which observed status or growth was greater than predicted were classified as outperforming predictions with respect to status or growth, respectively.

Primary research questions

This investigation is based on two primary research questions that jointly address the status and growth over time of school performance in grade 3 students' mathematics and reading achievement. The status research question (RQ1) investigates schools' grade 3 mathematics and reading performance in the most recent two school years after accounting for student and school demographic characteristics. The growth research question (RQ2) examines schools' adjusted school-level gains in grade 3 mathematics and reading performance over five school years regardless of their starting point with respect to student performance.[2]

The status research question (RQ1) focuses on identifying high-performing schools. Some of these schools may not have shown substantial school-level gains in recent years, but they may have been consistently high-performing, with long-standing, well-developed strategies for supporting students' performance in early-grade mathematics and reading. RQ2 involves the identification of high-growth schools, which may have adopted new interventions, policies, or practices in recent years to boost student performance. Staff at low-performing schools may be more amenable to drawing lessons from high-growth schools that were similarly situated just five years ago than they would be from persistently high-performing schools. Over time, KDE can investigate both high-performing and high-growth schools in comparison to other schools to determine what is driving their success and, ultimately, to inform school improvement efforts in Kentucky.

The two research questions are as follows:

1. Status: Which schools performed better, worse, or about the same as predicted with respect to grade 3 students' (a) mathematics performance and (b) reading performance in 2017 and 2018, given student and school demographic characteristics?

[2] KDE data analysts decided to focus on the growth research question (RQ2) in their presentation of findings to KDE leadership. As a result, the accompanying PowerPoint slide deck focuses on RQ2 results. Although not prioritized in their presentation of findings to KDE leadership, KDE data analysts remain interested in the status research question (RQ1) results, as well.

2. Growth: Which schools have shown larger, smaller, or about the same as predicted average annual growth in grade 3 student (a) mathematics performance and (b) reading performance during the five years from 2014 to 2018, given student and school demographic characteristics and their changes over time?

Data

The quantitative analyses used deidentified student-level administrative data supplied by the Kentucky Center for Statistics (KSTATS), which collects and links data from KDE and other sources to evaluate education and workforce efforts in the commonwealth.

Analytic sample

The analytic sample comprised all first-time grade 3 students who had grade 3 mathematics and reading scale scores on the Kentucky Performance Rating for Educational Progress (K-PREP) assessment and who attended A1 schools. A1 schools, which serve 99.9 percent of public elementary students in the commonwealth,[3] are traditional public schools "under administrative control of a principal and eligible to establish a school-based decisionmaking council" and "not a program operated by, or as a part of, another school" (KDE, 2019). A1 schools serve the vast majority of Kentucky's students who receive special education services (more than 9 in 10) and all students in magnet schools. Education programs not included in the analysis, which jointly serve 0.1 percent of public elementary students in Kentucky, are district-operated alternative programs, special education programs where all enrollments are students in special education (for example, schools for the blind and schools for the deaf), and programs for children committed to or in the custody of Kentucky funded by the Kentucky Educational Collaborative for State Agency Children. The primary status analyses included student observations from the two most recent years available: the 2016/17 and 2017/18 school years.[4]

The growth analyses included observations from each school year from 2013/14 through 2017/18. The two-year analytic sample included 91,337 first-time grade 3 students enrolled in 700 elementary schools, and the five-year analytic sample included 233,343 first-time grade 3 students enrolled in 727 elementary schools.[5] Because only first-time grade 3 students were included in the sample, each

[3] Personal communication with A. Butler (July 11, 2019) from the Office of the Commissioner in the Kentucky Department of Education.
[4] As described in the supplemental analyses section, we also performed status analyses using five years of data.
[5] The discrepancy in the number of schools is because of schools opening and closing over time.

student contributes only a single record to the analyses.

Sample exclusions

In addition to excluding students enrolled in non-A1 schools, we excluded first-time grade 3 students enrolled in their school for less than 100 days because of the limited time the schools had to affect these students' academic performance.

Method

To identify which schools are performing better, worse, or about the same as predicted in mathematics and reading status based on student and school demographics, we fit two-level multilevel models to predict student scale scores and school-level effects on those scale scores.[6] As described below, we captured school effects by allowing the level-1 intercepts to vary randomly at the school level. The level-2 residuals associated with these parameters represent "school effects" after accounting for individual- and school-level demographics. As recommended in the literature (for example, Bowers, 2010; Trujillo, 2013), to reduce the possibility that findings from these status analyses are driven by chance differences across schools in student cohorts, we used the two most recent years of student data available (2016/17 and 2017/18) as opposed to basing status estimates off of a single year of data.

Building upon the status analyses, we investigated average annual growth over time in schools' average mathematics and reading scale scores. The growth analyses incorporated five years of data (2013/14, 2014/15, 2015/16, 2016/17, and 2017/18) so we could identify the schools that made the greatest improvements in grade 3 student mathematics and reading performance over the five school years.[7] As shown below, incorporating a year count variable in the first level of the model and allowing the coefficient on this variable to vary randomly at the school level enabled us to estimate the average annual growth in the outcomes of interest from 2014 to 2018 by school, accounting for the influence of changes in school demographics over time.

[6] Historically, multilevel modeling has been a relatively rare approach in the school and district effectiveness literature (Trujillo, 2013). Recent REL and other studies have used the approach (for example, Bowers, 2015; Partridge, Rudo, & Herrera 2017).

[7] KDE and REL AP chose to examine five years of growth data because it is a reasonable time frame for identifying schools that show sustained growth in student outcomes over time and allows KDE and REL AP to focus on relatively recent school performance.

Benefits of a multilevel model using student-level data

Multilevel models, like the hierarchical linear models (HLMs) used in the present study, are preferable to a more traditional approach, such as ordinary least squares, for several reasons. First, they generate standard errors that account for the nesting of data (in our case, observations of first-time grade 3 students and observations of schools from different years are nested within schools). Second, they allow investigations into the extent of variation in outcomes (and in changes over time in outcomes) at the student and school levels.[8] This provides a sense of the extent of variation in the overall outcomes that student- and school-level variables may be able to predict, along with information researchers can use when planning future studies. Third, multilevel modeling enables us to use the same analytical framework to investigate which schools have shown the most improvement in grade 3 student mathematics and reading performance (growth) and which schools have demonstrated the best relative performance in recent years (status), conditional on student and school demographics.

Potential benefits to using student-level data to estimate a multilevel model, as opposed to aggregating data to the school level and running a single-level model, also exist. Aggregating to a group level suppresses within-group variation, and this can lead to misleading results (for example, Aitkin & Longford, 1986). In contrast, multilevel models based on individual data nested within groups with individual- and group-level predictor variables can increase efficiency, reduce aggregation bias, and enable investigations into the extent of variation that lies at the student and school levels (Raudenbush & Bryk, 2002). Including student-level data in the multilevel model allows the researcher to account for both individual- and school-level influences on outcomes. For example, we know that there is both an individual effect on student achievement of living in a poor family and an effect of attending a school serving a high concentration of poor students (for example, Caldas & Bankston, 1999). Models based on student-level data can help disentangle individual-level and contextual effects in a way that aggregate school-level models cannot.

Variables

The analyses drew on an array of variables from KDE administrative data. Table 1 describes each variable included in the analyses: outcomes of interest; student-level covariates; school-level covariates; time variables; sample inclusion and exclusion variables; and reporting variables, such as school name or

[8] We report intraclass correlation coefficients when presenting findings to describe the extent of variation that exists at different levels of the analyses.

magnet status, which identify schools and provide context when presenting results. The outcomes of interest are grade 3 mathematics and reading scale scores. The student-level covariates are student age (in years), as well as indicator (dummy) variables for English learner status, free and reduced-price lunch eligibility, individualized education program (IEP) status, male, and race and Hispanic origin (variables for Black alone, non-Hispanic; Hispanic; and Other race, non-Hispanic; with White alone, non-Hispanic as the reference category). The school-level covariates are school means of the student-level covariates, such as mean student age. Note that taking the mean of a student-level indicator variable at the school level generates a proportion ranging from 0 to 1. Time variables include an indicator variable for the 2017/18 school year in the status analyses and a year count variable in the school-level growth analyses. The sample inclusion and exclusion variables align with the concepts discussed above in the analytic sample and sample exclusion sections. The reporting variables are school and district name, magnet status, and variables describing receipt of support under the Every Student Succeeds Act (ESSA) via Comprehensive Support and Improvement (CSI) or Targeted Support and Improvement (TSI) efforts.

Magnet schools. These are public schools with specialized schoolwide curricula that typically draw students from across a school district via an application process. The school district may provide transportation to magnet schools for participating students.

CSI schools. Identified by Kentucky for the first time in the 2018/19 school year, these schools are the lowest-performing 5 percent of schools in the commonwealth, according to its accountability system.

TSI schools. Any school with at least one ESSA student subgroup (such as economically disadvantaged students) whose performance was at or below that of all students in any of the lowest 5 percent of all schools (Kentucky Revised Statutes Title XIII. Education § 160.346).

KDE works with local education agencies to help improve CSI and TSI schools by providing interventions, allocating resources, and delivering technical assistance.

Table 1. Variables in the analyses

Variable	Description
Outcomes of interest	
Grade 3 mathematics scale score	Student scale score on the grade 3 Kentucky Performance Rating for Educational Progress (K-PREP) mathematics assessment, a mandatory criterion-referenced test to measure student performance on Kentucky's mathematics standards and to provide data for the state accountability system.
Grade 3 reading scale score	Student scale score on the grade 3 K-PREP reading assessment, a mandatory criterion-referenced test to measure student performance on Kentucky's reading standards and to provide data for the state accountability system.
Student-level covariates	
Age	Student age estimated by subtracting the student's year of birth from the year of the spring when the student first participated in the grade 3 K-PREP in mathematics or reading.
English learner status	Indicator variable for whether the student was identified as an English learner in the current school year. English learners are students whose primary language is a language other than English whose difficulties in English may undermine their ability to meet state proficiency standards, achieve in classes taught in English, or participate fully in society.[a] Kentucky is part of the World-Class Instructional Design and Assessment Consortium.[b] As such, students are identified as English learners if they score below a cut point on a placement test or screener and if they have not later scored above a cut point on an annual assessment of English proficiency.[a]
Free and reduced-price lunch eligibility	Indicator variable for whether a student is eligible to participate in the National School Lunch Program.
Individualized education program (IEP) status	Indicator variable for whether a student is receiving special education services via an IEP.
Male	Indicator variable for whether a student reported gender as male (female is the reference category). Students not reporting gender as male or female are counted as missing for this variable.
Black	Student is Black alone, non-Hispanic.
Hispanic	Indicator variable for whether the student traces his or her origin or descent to Mexico, Puerto Rico, Cuba, Central and South America, or other Spanish cultures, regardless of race.
Other race	Student is non-Hispanic and either American Indian or Alaska Native, Asian, Hawaiian or Other Pacific Islander, two or more races, or of unknown race and ethnicity.
School-level covariates	
Mean age	School average student age among students in the analytic sample by year.
Proportion English learners	School proportion of English learners among students in the analytic sample by year.
Proportion eligible for free and reduced-price lunch	School proportion eligible for free and reduced-price lunch among students in the analytic sample by year.
Proportion with an IEP	School proportion with an IEP among students in the analytic sample by year.
Proportion male	School proportion male among students in the analytic sample by year.

Variable	Description
Proportion Black	School proportion Black alone, non-Hispanic among students in the analytic sample by year.
Proportion Hispanic	School proportion Hispanic among students in the analytic sample by year.
Proportion Other race	School proportion Other race (not White or Black only or Hispanic) among students in the analytic sample by year
Time variables	
Year 2018	Indicator variable in the status analyses identifying observations from the 2017/18 school year.
Year count	School year count, centered at the 2013/14 school year, so that 2013/14 is 0, 2014/15 is 1, 2015/16 is 2, 2016/17 is 3, and 2017/18 is 4. This variable is used in the growth analyses.
Sample inclusion and exclusion variables	
First-time grade 3 student status	Using data from student enrollment over time, we include students who are first-time grade 3 enrollees in the school district. Students enrolled in grade 3 in the school district for the second time (or beyond) will be excluded from the analyses.
A1 school	Indicator variable for traditional public school, including magnet schools. Excludes district-operated special education programs, alternative programs, and programs for children committed to or in the custody of Kentucky funded by the Kentucky Educational Collaborative for State Agency Children. No charter schools exist in Kentucky.
Enrolled 100 days or more	Indicator variable for whether students were enrolled in their school for at least 100 days in their first-time grade 3 school year. We excluded from the analyses students who did not meet this criterion.
Reporting variables	
Comprehensive Support and Improvement (CSI) school	Indicator variable showing whether the school is receiving CSI under the Every Student Succeeds Act (ESSA).
Targeted Support and Improvement (TSI) school	Indicator variable showing whether the school is receiving TSI under ESSA.
District name	Name of the school district.
Magnet status	Indicator variable for whether the school is a magnet school.
School name	Name of the school.

[a] **https://education.ky.gov/districts/tech/sis/Documents/Standard-LEP.pdf**
[b] **https://education.ky.gov/AA/Assessments/Pages/EL-Testing.aspx**

Approach to missing data

In accord with KDE's typical approach to missing data, we used complete case analysis. Any individual students with data missing on any of the outcomes of interest or covariates were excluded from the analyses. Because the analyses relied on variables that typically have little missing data, such as student assessment scores or demographic characteristics, the level of missingness in the data was limited. Just 5.57 percent of first-time grade 3 students were excluded from the analyses, mainly due to missing assessment data. As a result of low levels of missingness, complete case analysis was warranted.

That being said, it is important to note that results of the present analysis only pertain to students who participated in state assessments, and some students are less likely to participate in state assessments than others (table 2). For example, compared with those students who participated in assessments, more non-participants received special education services via an IEP (34 versus 15 percent), were English learners (7 versus 4 percent), and were eligible for free or reduced-price lunch (76 versus 63 percent).

Table 2. Descriptive statistics of analytic sample students and those excluded due to missing assessment or other data.

Student characteristics	Analytic sample average	Analytic sample SD	Excluded student average	Excluded student SD	Effect size of average difference
Age	9.41	0.536	9.65	0.654	-0.44
English learner	0.04	0.189	0.07	0.261	-0.43
Free or reduced-price lunch eligible	0.63	0.484	0.76	0.428	-0.38
Male	0.51	0.500	0.55	0.498	-0.09
Race and Hispanic origin (reference category is white, non-Hispanic)					
Black	0.11	0.313	0.15	0.355	-0.21
Hispanic	0.07	0.260	0.08	0.272	-0.06
Other race	0.04	0.189	0.04	0.207	-0.12
Receiving special education services via IEP	0.15	0.360	0.34	0.474	-0.64

NOTE: There were 233,341 cases in the analytic sample, and 13,764 cases were excluded due to missing data. All excluded cases had information on English learner status, gender, and race and Hispanic origin, 13,762 had information on eligibility for free or reduced-price lunch and receipt of special education services via an IEP, and 2,458 had age data. Effect size of average difference is Hedges' *g* for continuous variables and Cox index for dichotomous variables.

Status models

For the status models, using data from 2016/17 and 2017/18, we fitted two-level models separately for each of two different student outcomes of interest: grade 3 mathematics scale score and grade 3 reading scale score. These two outcomes are represented by the subscript *k* in the following two-level model:

Level 1

$$Y_{ijk} = \beta_{0j} + \beta_{1j}AGE_{ij} + \beta_{2j}ELL_{ij} + \beta_{3j}FRPL_{ij} + \beta_{4j}IEP_{ij} + \beta_{5j}MALE_{ij} + \beta_{6j}BLACK_{ij} + \beta_{7j}HISP_{ij} + \beta_{8j}OTHRACE_{ij} + \beta_{9j}\overline{AGE}_{jt} + \beta_{10j}\overline{ELL}_{jt} + \beta_{11j}\overline{FRPL}_{jt} + \beta_{12j}\overline{IEP}_{jt} + \beta_{13j}\overline{MALE}_{jt} + \beta_{14j}\overline{BLACK}_{jt} + \beta_{15j}\overline{HISP}_{jt} + \beta_{16j}\overline{OTHRACE}_{jt} + \beta_{17j}Y2018_t + r_{ijt} \quad (1)$$

Level 2

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (2)$$

$$\beta_{1j} = \gamma_{10} \quad (3)$$

$$\vdots$$

$$\beta_{17j} = \gamma_{17,0} \quad (4)$$

where each outcome of interest for individual *i* in school *j* is a function of student demographic characteristics and school-level averages of the same demographic characteristics at time *t*, along with a year effect ($Y2018_t$) representing the effect of being in the 2017/18 school year as opposed to the 2016/17 school year. Student-level demographic variables include age in years (AGE_{ij}) and dummy variables (which take the value of 0 for no and 1 for yes) for whether in grade 3 the student was:

- An English learner (ELL_{ij}).
- Eligible for free and reduced-price lunch ($FRPL_{ij}$).
- An IEP holder (IEP_{ij}).
- Male ($MALE_{ij}$).
- Black ($BLACK_{ij}$).
- Hispanic ($HISP_{ij}$).
- Other race ($OTHRACE_{ij}$).

School-level means of these student demographic characteristics are represented by variable names with single bars over their tops, with subscripts *j* and *t*, as the variables vary across *j* schools and over *t* years. For example, the school mean age of first-time grade 3 students in school *j* at time *t* is represented by $\overline{AGE}_{jt}$. All school-level means of dummy variables are proportions that can range from 0 to 1. For example, if no students in a school in a given year were eligible for free and reduced-price lunch, the variable $\overline{FRPL}_{jt}$ would be 0; if 100 percent were eligible, the variable would be 1; and if 50 percent of students were eligible, $\overline{FRPL}_{jt}$ would take on the value 0.5. School-level means of demographic characteristics are included at level 1 of the model because they vary over time. Variable coefficients are represented by the vector β′, with β_{0j} representing the model intercept. For the status model, all coefficients are held fixed at level 2 (the school level), except for the level-1 intercept, which we allow to vary randomly around a cross-school mean (γ_{00}).

We assume that the level-1 error term (r_{ijt}) and the error term associated with the random intercept at level 2 (u_{0j}) are normally distributed with means of zero. The level-2 error term associated with the random intercept (u_{0j}) represents the deviation of school *j* from the cross-school mean (γ_{00}) (see equation 2). As such, it represents the extent to which a school is over- or underperforming predictions with respect to the outcome of interest after accounting for student and school demographic factors and a year fixed effect. Some of this deviation from predicted performance may be due to chance, and some may be due to systemic factors not accounted for in the model. Some of these systemic factors may be school-caused and others may be the result of non-school factors. To the extent that these systemic factors represent factors within the purview of the school (for example, school policies, practices, procedures, climate, curricula, instruction, staffing, and decisions and efforts of teachers and leaders), they jointly represent school influences on student performance. For each school, we reported the level-2 error term associated with the random intercept (u_{0j}) and tested whether the empirical Bayes residual was statistically significantly different from zero ($p < .05$) using a two-tailed *t*-test. We then categorized each school as:

- Overperforming relative to predictions based on its students' demographic characteristics (those schools with u_{0j}'s that are positive and statistically significant).
- Underperforming relative to predictions based on its students' demographic characteristics (u_{0j}'s that are negative and statistically significant).
- Performing in accordance with predictions based on its students' demographic characteristics (schools with u_{0j}'s that are not statistically significantly different from zero).

To facilitate interpretation, we presented the status school effects both on the assessment scale and a standard deviation scale (based on the standard deviation of the relevant assessment among the two-year status model analytic sample). At KDE's request, to ease interpretation, we also grouped schools with statistically significant effects according to the size of their effects on the assessment scale: less than 5 points, 5 to 9.99 points, or 10 points or higher than predicted. Five points is roughly a quarter, and 10 points is roughly one half, of a standard deviation for both tests.

Growth models

As with the status models, for the growth models we fit two-level models separately for each of two different student outcomes of interest: grade 3 mathematics scale score and grade 3 reading scale score. These two outcomes are represented by the subscript *k* in the following two-level model:

Level 1

$$Y_{ijk} = \beta_{0j} + \beta_{1j}AGE_{ij} + \beta_{2j}ELL_{ij} + \beta_{3j}FRPL_{ij} + \beta_{4j}IEP_{ij} + \beta_{5j}MALE_{ij} + \beta_{6j}BLACK_{ij} + \beta_{7j}HISP_{ij} + \beta_{8j}OTHRACE_{ij} + \beta_{9j}\overline{AGE}_{jt} + \beta_{10j}\overline{ELL}_{jt} + \beta_{11j}\overline{FRPL}_{jt} + \beta_{12j}\overline{IEP}_{jt} + \beta_{13j}\overline{MALE}_{jt} + \beta_{14j}\overline{BLACK}_{jt} + \beta_{15j}\overline{HISP}_{jt} + \beta_{16j}\overline{OTHRACE}_{jt} + \beta_{17j}YEAR_t + r_{ijt} \quad (5)$$

Level 2

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (6)$$

$$\beta_{1j} = \gamma_{10} \quad (7)$$

$$\vdots$$

$$\beta_{17j} = \gamma_{17,0} + u_{17j} \quad (8)$$

where, as in the status models described above, each outcome of interest for individual *i* in school *j* is a function of student demographic characteristics and school-level averages of the same demographic characteristics at time *t*. The only differences between the specification of the status and growth models are that time is no longer accounted for with a single year dummy. Rather, because the growth models are drawing on data from five years (2013/14 through 2017/18), we have replaced the year dummy with a year count variable ($YEAR_t$), centered at the 2017/18 school year so that it ranges from –4 in 2013/14 to 0 in 2017/18. By including this year count variable, we have specified a linear growth model, where the coefficient on year (β_{17j}) represents the average annual change in our outcomes of interest from 2013/14 to 2017/18, and the intercept (β_{0j}) represents the status of those outcomes in 2017/18.[9]

Furthermore, we have allowed the coefficient, or slope parameter, on the year count variable to vary randomly at the school level (equation 8). The error term for this slope parameter (u_{17j}), which we assume to have a normal distribution and mean of zero, represents the deviation of each school, *j*, from the cross-school average annual change in the outcome of interest over time ($\gamma_{17,0}$). For each school, we tested whether the error term (u_{17j}) is statistically significantly different from zero. We reported the magnitude of the empirical Bayes residuals for each school, and those schools with residuals that are positive and statistically significant at the $p < .05$ level are classified as overperforming statistical predictions based on their students' demographic characteristics with respect to change over time. We categorized those schools with u_{17j}'s that are negative and statistically significant as underperforming with respect to change over time in the outcome of interest. Finally, we categorized those schools with u_{17j}'s that are not statistically significantly different from zero as performing roughly as statistically predicted with respect to the average annual change in the outcome of interest over time.

In addition to testing the significance of these estimates, we set cut points to ease interpretation at KDE's request. Per KDE's guidance, we grouped schools into categories according to whether their cumulative average annual gains were less than 5 points, 5 to 9.99 points, or 10 points or higher than predicted over the five-year period. Ten points is roughly equal to a half a standard deviation, and the 5 points is about a quarter of a standard deviation of first-time grade 3 students' scale scores on the mathematics and reading assessments. Unlike the random intercept estimate results from the status model, few random slope estimates under 5 points were statistically significantly different from zero due to relatively larger confidence intervals associated with the slope estimates.

[9] This intercept varies randomly at level 2; thus, the empirical Bayes residuals associated with u_{0j} provide alternate status estimates of the extent to which schools are over- or underperforming predicted performance in 2017/18.

Supplemental analyses

School-readiness analyses

Quantitative analyses aimed at understanding whether schools are performing in ways that differ from statistical predictions often include students' prior achievement in their models to identify schools that are doing better than predicted in improving student performance, given baseline student performance. That is, to measure school performance more accurately, these analyses often model school effects on growth in individual student achievement over time. Because grade 3 is the first year in which students participate in mandatory state assessments, comparable baseline student performance data were not readily available statewide.

Kentucky collects school-readiness data on students from teacher observations during kindergarten using the BRIGANCE Early Childhood Kindergarten Screen III. These screener data, however, are not directly comparable to grade 3 state assessment data. Unlike the summative grade 3 state assessment data, kindergarten screener data are designed to help teachers identify students with potential delays, support referrals for special education services, and inform personalized instruction. Furthermore, comparable and appropriately lagged data on school readiness are available in Kentucky only for 2016/17 and 2017/18 grade 3 students (who received the kindergarten screener in 2013/14 and 2014/15, respectively), meaning that school-readiness data could not be used for the five-year school growth analyses. Finally, in any potential cases where large numbers of students transferred into a school district after kindergarten, any complete case analyses including measures of school readiness could substantially reduce the analytic sample size, potentially undermining generalizability of results.

To investigate how the inclusion of school-readiness data in the status analyses might affect results, REL AP and KDE investigated which schools were performing better, worse, or about the same as predicted on grade 3 students' mathematics and reading scale scores in 2017 and 2018, given student and school demographic characteristics and school readiness as measured in kindergarten for the subsample of students who had kindergarten screening data and grade 3 test scores. For the same subsample, we also ran our original status models without information on student school readiness as measured in kindergarten, as described in equations 1–4, and compared the school categorizations. When we ran our original status models on both the overall sample and the subsample, we found similar results, leading us to determine that estimating school effects based on the subsample (limited to students with kindergarten-readiness information) was a reasonable approach.

Drawing on additional years of data for status estimates

To investigate the stability of status estimates, REL AP and KDE ran the status models on five years of data using two approaches. The first generated status estimates by incorporating all five years of data in a modified version of the model that included dummy variables for four of the years in level 1, holding the year effects fixed at level 2. We then compared each school's estimated effects from the two-year and the five-year models. The second approach measured status using the level-2 empirical Bayes residuals associated with the randomly varying intercept of the growth model, providing alternate status estimates. These status estimates indicated the extent to which schools were over- or underperforming predictions in the 2017/18 school year. We compared these estimates with our previously described status model estimates to determine whether the growth models provided status estimates consistent with our preferred status models.

Summary of supplemental analysis results

Tables 3 and 4 offer Pearson correlation coefficients among school performance status model estimates for math and reading for the two-year status model, and the supplemental status models. These supplemental models include the:

- Five-year status model,
- Two-year status model based on the restricted sample,
- Two-year status model based on the restricted sample including school-readiness predictor variables, and
- Supplemental status estimates based on the intercept of the five-year growth model.

The two-year status model estimates were very highly positively correlated (0.97 or above) with all supplemental model estimates aside from those associated with the five-year status model, with which they had a correlation of 0.86 for both math and reading.

Table 3. Pearson correlation coefficients among school math performance status model estimates

	School math performance status model estimates				
			Two-year restricted sample[a]		
School math performance status model estimates	Two-year	Five-year	Without school readiness[b]	With school readiness[b]	Five-year growth intercept[c]
Two-year	1.00	0.86	0.99	0.97	0.97

School math performance status model estimates	Two-year	Five-year	Two-year restricted sample[a] Without school readiness[b]	With school readiness[b]	Five-year growth intercept[c]
Five-year	0.86	1.00	0.85	0.82	0.86
Two-year restricted sample[a]					
Without school readiness[b]	0.99	0.85	1.00	0.98	0.96
With school readiness[b]	0.97	0.82	0.98	1.00	0.94
Five-year growth intercept[c]	0.97	0.86	0.96	0.94	1.00

[a]The restricted sample includes only those first-time grade 3 students who had school-readiness data collected in kindergarten.

[b]School-readiness variables included (1) whether the student scored "ready," (2) whether the student scored "ready with enrichments," (3) the proportion of sample students in the school who scored "ready," and (4) the proportion of students in the school who scored "ready with enrichments" on the BRIGANCE Early Childhood Kindergarten Screen III.

[c]This is a 2017/18 status estimate based on the intercept of the five-year growth model with random intercept and random slope on year, with year centered at 2017/18.

Table 4. Pearson correlation coefficients among school reading performance status model estimates

	School reading performance status model estimates				
			Two-year restricted sample[a]		
School reading performance status model estimates	Two-year	Five-year	Without school readiness[b]	With school readiness[b]	Five-year growth intercept[c]
Two-year	1.00	0.86	0.99	0.97	0.97
Five-year	0.86	1.00	0.84	0.82	0.90
Two-year restricted sample[a]					
Without school readiness[b]	0.99	0.84	1.00	0.98	0.95
With school readiness[b]	0.97	0.82	0.98	1.00	0.93
Five-year growth intercept[c]	0.97	0.90	0.95	0.93	1.00

[a]The restricted sample includes only those first-time grade 3 students who had school-readiness data collected in kindergarten.

[b]School-readiness variables included (1) whether the student scored "ready," (2) whether the student scored "ready with enrichments," (3) the proportion of sample students in the school who scored "ready," and (4) the proportion of students in the school who scored "ready with enrichments" on the BRIGANCE Early Childhood Kindergarten Screen III.

[c]This is a 2017/18 status estimate based on the intercept of the five-year growth model with random intercept and random slope on year, with year centered at 2017/18.

Limitations

The primary limitation of our analyses is that while they identified schools that were performing better or worse than statistically predicted or showing larger or smaller school-level gains than statistically predicted, they cannot, in and of themselves, explain why schools were doing so. Attributing school performance and changes in school performance solely to the effectiveness of the schools themselves or to changes in the effectiveness of schools would be naïve. In fact, any factors omitted from the initial models could be driving the school effects we estimated from these analyses, even factors outside the realm of a school's direct influence. For example, due solely to the luck of the draw, a school may have ended up with grade 3 cohorts that have, on average, greater cognitive abilities, more perseverance, or parents with higher educational expectations for their children than is the norm. Furthermore, some schools may be in communities with increasing levels of drug abuse, declining access to health care, or decreasing availability of social services.

This is not to say that factors within schools' purviews do not play a role in whether a school is over- or underperforming predictions. In fact, a wide array of literature on school effects suggests that numerous school factors, including principal and teacher effectiveness, educator expectations for student performance, data use, school climate, enacted curriculum, and instructional practices, can drive school performance (for example, Bryk, Sebring, Allensworth, Easton, & Luppescu, 2010; Edmunds, 1979; Teddlie & Reynolds, 2000). However, to successfully investigate the effect of malleable school-related factors on the results requires additional research. The results of the present analyses should be considered the launching point for a more thorough investigation.

A related limitation, unique to the present investigation, is the lack of baseline measures clearly aligned to the outcomes of interest. The absence of student mathematics and reading achievement measures prior to grade 3 may increase the likelihood that student cohort effects, and not school performance, are driving results. Incorporating demographic variables associated with the outcomes of interest helps mitigate this problem but does not eliminate it.[10]

[10] Similarly, using two cohorts of student data may mitigate this concern somewhat, but the results of the status models focused on the two most recent cohorts of student data are not necessarily generalizable to prior cohorts.

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* (Applied Research Methods No. REL 2015-071). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. https://eric.ed.gov/?id=ED558161

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society: Series A (General), 149*(1), 1–26.

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

Bowers, A. J. (2010). Toward addressing the issues of site selection in district effectiveness research: A two-level hierarchical linear growth model. *Educational Administration Quarterly, 46*(3), 395–425. https://doi.org/10.1177/0013161X10375271

Bowers, A. J. (2015). Site selection in school district research: A measure of effectiveness using hierarchical longitudinal growth models of performance. *School Leadership & Management, 35*(1), 39–61. https://doi.org/10.1080/13632434.2014.962500

Caldas, S. J,. & Bankston, C. L. (1999). Multilevel examination of student, school, and district-level effects on academic achievement. *The Journal of Educational Research*, *93*(2), 91–100.

Culbertson, M. J., & Billig, S. H. (2016). *Decision points and considerations for identifying rural districts that have closed student achievement gaps* (Applied Research Methods No. REL 2016-130). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Central. https://eric.ed.gov/?id=ED565611

Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership, 37*(1), 15–24.

Kentucky Department of Education. (2018). *Our kids, our future: 2018–2023 strategic plan*. Frankfort, KY: Author. https://web.archive.org/web/20190701161652/https://education.ky.gov/CommOfEd/CDU/Documents/Strategic%20Plan%20Final%202018.pdf

Kentucky Department of Education (KDE). (2019). *School classification*. Frankfort, KY: Author. https://applications.education.ky.gov/sdci/Classification.aspx

Koon, S., Petscher, Y., & Foorman, B. R. (2014). *Beating the odds: Finding schools exceeding achievement expectations with high-risk students* (REL 2014-032). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. https://eric.ed.gov/?id=ED544802

Meyers, C. V., & Wan, Y. (2016). *A comparison of two methods of identifying beating-the-odds high schools in Puerto Rico* (Applied Research Methods No. REL 2017-167). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. https://eric.ed.gov/?id=ED570966

Partridge, M. A., & Koon, S. (2017). *Beating the odds in Mississippi: Identifying schools exceeding achievement expectations* (What's Happening No. REL 2017-213). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. https://eric.ed.gov/?id=ED572606

Partridge, M. A., Rudo, Z., & Herrera, S. (2017). *Identifying South Carolina charter schools that are "beating the odds"* (What's Happening No. REL 2017-236). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. https://eric.ed.gov/?id=ED572602

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Teddlie, C., & Reynolds, D. (Eds.). (2000). *The international handbook of school effectiveness research*. London, ENG.: Falmer Press.

Trujillo, T. (2013). The reincarnation of the effective schools research: Rethinking the literature on district effectiveness. *Journal of Educational Administration, 51*(4), 426–452. https://doi.org/10.1108/09578231311325640