IES Institute of Education Sciences

APPENDIXES
September 2021

Regional Educational Laboratory Mid-Atlantic
At Mathematica

# Using Promotion Power to Identify the Effectiveness of Public High Schools in the District of Columbia

Appendix A. Methods

Appendix B. Supporting analysis

See https://go.usa.gov/xFVxQ for the full report.

## Appendix A. Methods

This section describes the outcomes used to measure high schools' promotion power, the student background characteristics used as control variables in the models, and the analytic sample. It also provides an overview of promotion power models and a detailed description of the analytic approach used to estimate the models.

### *Outcomes analyzed in the promotion power models*

This study examined public District of Columbia (DC) high schools' promotion power for three outcomes: meeting the College Board's college and career readiness benchmarks for both subjects on the SAT (college-ready SAT scores), high school graduation, and college enrollment.

*College-ready SAT scores.* This outcome is based on each students' highest math and reading scores across all the times they took the SAT. The study team created a binary indicator that equaled 1 if a student scored at or above the College Board's college and career readiness benchmarks and 0 otherwise. The DC Office of the State Superintendent of Education (OSSE) encourages all students in grades 11 and 12 to take the SAT and offers it for free. Students missing SAT scores were counted as not meeting the benchmarks; a discussion of the sensitivity of the analysis to this decision is in appendix B. Most students take the SAT during the spring of grade 11 or the fall of grade 12, so if students transfer to a new school after grade 11, it is unlikely that the new school contributes meaningfully to SAT scores.[1] Thus, in the main specification student attendance in grade 12 was not counted toward this outcome (see the section below "Attributing student outcomes to schools" for further details).

The study team used the binary indicator for achieving the SAT college and career readiness benchmarks rather than the SAT score itself to avoid introducing bias. If the team had instead used the SAT score as the outcome and excluded the 21 percent of students with missing SAT scores, a school that performed poorly in preventing dropouts might have only a small, unrepresentative sample of students who made it to grade 11 or 12 and took the SAT. Such a school might appear to perform better than another school serving similar students that had greater success in preventing dropouts and thus had lower-achieving students take the SAT. By including students who dropped out of high school (or exited public DC high schools and enrolled elsewhere) and counting them as not achieving the benchmarks, the measure avoids bias possibly introduced by differential SAT taking that could be a result of school effectiveness. Even so, supplemental analysis in appendix B shows that among students who

[1] The study team had data only on students' maximum score in each subject, not on the number of times or the dates they took the SAT.

were not missing SAT scores, the promotion power scores using the binary indicator are highly correlated with estimates using the continuous SAT score.

*High school graduation*. The study team also analyzed high schools' contribution to the likelihood that a student in the analytic sample would graduate from high school four years after starting grade 9 in 2015/16. For this grade 9 cohort, data were available for students' four-year graduation outcome through the 2018/19 school year. This outcome took the value of 1 for students who earned a regular high school diploma from a public DC high school and 0 otherwise. The administrative data did not distinguish students who dropped out of high school from those who exited public DC high schools and enrolled elsewhere. The analysis treats both groups as not graduating from high school.

*College enrollment.* The study team estimated a promotion power model for whether students were enrolled in college (two-year or four-year) in the fall four years after they began grade 9. These data were obtained from OSSE through a partnership with the National Student Clearinghouse (NSC). The NSC data cover 97 percent of all college students in the United States and 99 percent of all public and private postsecondary institutions (National Student Clearinghouse, 2019). For this study, NSC data were available through winter 2020. OSSE sent data for all students who graduated from a public DC high school within four years to the NSC for matching, including data for students who graduated early, received a GED credential, or received a Certificate of an Individualized Education Program Completion. Data on students who received a DC Tuition Assistance Grant were also sent to the NSC for matching regardless of where the student earned a high school credential. Data on students who did not receive one of these diplomas/certificates or grants were not sent to the NSC for matching, and these students were counted as not enrolling in college.

### *Background characteristics used as control variables in the promotion power models*

A key feature of promotion power models is that they account for the fact that high schools serve different populations of students, some of whom might be more likely to achieve relevant outcomes and others of whom might be less likely, regardless of which high school they attend. The promotion power models in this study controlled for a student's prior achievement (grade 8 standardized math and English language arts test scores), demographic characteristics (gender and race/ethnicity), and prior receipt of targeted academic services (grade 8 special education and language services). Baseline student-level information on discipline was unavailable, so the study instead controlled for the suspension rate among peers of each student's same race/ethnicity in the middle school that the student attended to proxy for the student's own middle school discipline rate.

The promotion power models included various background characteristics for students (table A1). All time-varying student-level characteristics are for school year 2014/15, when students in the 2015/16 grade 9 cohort were in grade 8. All student-level characteristics are from OSSE administrative data, and middle school–level discipline data are from the U.S. Department of Education's Civil Rights Data Collection (http://ocrdata.ed.gov).

**Table A1. Background characteristics included in promotion power models**

| Covariate | Description |
|---|---|
| **Student level** | |
| Partnership for Assessment of Readiness for College and Careers (PARCC) math score in grade 8 | Student's score on a districtwide assessment in math[a] |
| PARCC English language arts score in grade 8 | Student's score on a districtwide assessment in English language arts[a] |
| Asian | An indicator for whether the student identified as Asian |
| Black | An indicator for whether the student identified as Black |
| Hispanic | An indicator for whether the student identified as Hispanic |
| White | An indicator for whether the student identified as White |
| Other race/ethnicity | An indicator for whether the student identified as multiracial, Pacific Islander, or American Indian |
| Female | An indicator for whether the student identified as a female |
| Received special education services in grade 8 | An indicator for whether the student received special education services in grade 8 |
| English learner student in grade 8 | An indicator for whether the student was an English learner student in grade 8 |
| **Middle school level** | |
| Suspension rate, same race/ethnicity[b] | The share of students in the student's middle school who identified as the same race/ethnicity who were suspended |

Note: *n* = 3,568 first-time grade 9 students in 36 District of Columbia public high schools during the 2015/16 school year. All student-level data are for the grade 8 year 2014/15. Middle school–level data are for 2013/14 (when most students in the analysis cohort were in grade 7).
a. PARCC scores were standardized at the year-grade-subject level to have a mean of 0 and standard deviation of 1.
b. The Civil Rights Data Collection does not include discipline information for the 2014/15 school year.
Source: District of Columbia Office of the State Superintendent of Education for student-level characteristics and the U.S. Department of Education's Civil Rights Data Collection for data on the suspension rate in middle school.

### *Description of the analytic sample*

The analytic sample for all three research questions included students in public DC high schools who first entered grade 9 during the 2015/16 school year. The sample does not include students who entered a public DC high school after grade 9, although it does include students who first entered in grade 9 but later dropped out of high school or transferred out of public DC high schools. Three criteria were applied to this group of 4,349 students to create the analytic sample. First, 776 students were excluded who did not have grade 8 test score data for both math and English language arts. Second, one student was excluded who was missing information on multiple background characteristics that were available for all other students (grade 8 status for special education services and English learner student services). Third, to enhance the precision of promotion power scores, four students were excluded who attended very small high schools (schools that enrolled fewer than 10 student equivalents, an adjusted enrollment total that accounts for the proportion of students' high school years that they attend at each school; see the section below "Attributing student outcomes to schools" for further details). With these exclusions the final analytic sample included 3,568 students across 36 public DC high schools.

*Characteristics*. Some 77 percent of students in the sample identified as Black, and 14 percent identified as Hispanic (table A2). In middle school 18 percent of students in the sample had an Individualized Education Program and received special education services and 5 percent were identified as English learner students. On average, students in the analytic sample attended a middle school in which 24 percent of students of their same race/ethnicity received at least one suspension.

*Outcomes*. Some 17 percent of students in the sample met or exceeded the College Board's college and career readiness SAT benchmarks; 76 percent graduated high school within four years, roughly 5 percentage points below the national average; and 47 percent enrolled in college (table A3). The data do not include outcome information for students who might have achieved the outcome but did so only after moving out of the district or transferring to a private school. In this sense the study slightly undercounts the proportion of students who achieved each outcome. Even so, all outcomes show substantial variation across high schools.

**Table A2. Summary statistics for background characteristics of the analytic sample**

| Characteristic | Average | Across-school standard deviation |
|---|---|---|
| **Grade 8 standardized test scores** | | |
| PARCC math score (standard deviations) | 0.02 | 0.57 |
| PARCC English language arts score (standard deviations) | 0.01 | 0.60 |
| Missing PARCC math score (percent) | 2 | 0.02 |
| Missing PARCC English language arts score (percent) | 1 | 0.02 |
| **Sociodemographic characteristics (percent)** | | |
| Asian | 1 | 0.02 |
| Black | 77 | 0.22 |
| Hispanic | 14 | 0.17 |
| White | 7 | 0.11 |
| Other race/ethnicity | 1 | 0.02 |
| Female | 52 | 0.08 |
| Received special education services | 18 | 0.09 |
| English learner student | 5 | 0.08 |
| **Middle school level (percent)** | | |
| Middle school suspension rate, same race/ethnicity | 24 | 0.11 |
| Missing suspension rate, same race/ethnicity | 4 | 0.08 |

PARCC is Partnership for Assessment of Readiness for College and Careers.
Note: *n* = 3,568 first-time grade 9 students in 36 District of Columbia public high schools during the 2015/16 school year. All student-level data are for the grade 8 year 2014/15. Middle school–level data are for 2013/14 (when most students in the analysis cohort were in grade 7). Missing indicators are included as rows in the table.
Source: District of Columbia Office of the State Superintendent of Education for student-level characteristics and the U.S. Department of Education's Civil Rights Data Collection for data on the suspension rate in middle school.

**Table A3. Summary statistics for outcomes of the analytic sample**

| Outcome | Average (percent) | Across-school standard deviation |
|---|---|---|
| College-ready SAT scores | 17 | 0.23 |
| High school graduation | 76 | 0.16 |
| College enrollment | 47 | 0.22 |

Note: *n* = 3,568 first-time grade 9 students in 36 District of Columbia public high schools during the 2015/16 school year.
Source: Authors' analysis of data from the District of Columbia Office of the State Superintendent of Education.

### *Promotion power models*

The study used a statistical model to calculate promotion power for a given outcome. For high school graduation, for example, the model predicted the chances that each student would have graduated had they gone to the average school, based on a variety of background characteristics, including prior test scores and demographic

information. Next, the study compared the actual graduation status among the students in a high school with those predicted by the model. High schools with a higher proportion of graduates than predicted based on students' background characteristics are considered effective at promoting graduation. In contrast, high schools with a lower proportion of students graduating relative to the prediction can be considered to be falling short.

*Regression equation.* The study team used a linear probability model to estimate the following regression equation:

(A1) $$Y_i = \mathbf{X}_i\boldsymbol{\gamma} + D_i\boldsymbol{\delta} + e_i$$

where $Y_i$ is the outcome variable for student *i,* such as high school graduation or college enrollment; $\mathbf{X}_i$ is a vector of students' scores in math and English language arts (ELA) on the Partnership for Assessment of Readiness for College and Careers in grade 8 and the other background characteristics (see table A2), such as status for special education services and English learner services;[2] $\boldsymbol{\gamma}$ is a vector of coefficients corresponding to the student background characteristics in $\mathbf{X}_i$ (see table A5 later in the appendix); $D_i$ is a set of indicator variables, one for each high school, equal to 1 if student *i* attended the school and equal to 0 otherwise; $\boldsymbol{\delta}$ is a vector containing the estimated effect of attending each school, conditional on student background characteristics; and $e_i$ is an error term.

The team selected this approach because a comparison by Guarino et al. (2015) found that this model (referred to as dynamic ordinary least squares) was the most robust to a variety of assumptions. Also, this is the model that Deutsch et al. (2020) used in their study of the promotion power of Louisiana high schools and is comparable to the fixed effects specification that is commonly used to estimate teacher and school effectiveness (for example, Resch & Deutsch, 2015; Walsh et al., 2018).

*Linear probability model.* Some researchers use nonlinear models, such as logit or probit models, rather than linear probability models when examining binary outcomes. Although promotion power scores from a linear probability model highly correlate with those from logistic regression (see appendix B), the study team chose a linear probability model as the main specification. An advantage of the linear probability model is that it can estimate high school effects for a school when all students have the same value of the outcome. In contrast, logistic models are unable to estimate a school's power to promote high school graduation if, for example, all (or none) of the students at that school graduate. In the analytic sample there are three schools in which no students achieved a college-ready SAT score and one school in which all students graduated.

Prior research notes that when using binary outcomes, linear probability models and nonlinear models often lead to similar results (Wooldridge, 2013). To assess whether this is the case in the current study, the team re-estimated the promotion power models using logistic regression. The school effects from this alternative specification highly correlate with those from the linear probability model (see the discussion in appendix B).

*Attributing student outcomes to schools.* There are two key issues to consider in attributing student outcomes to schools. First, students can attend multiple schools throughout their high school careers, meaning more than one school can contribute to students' success. In the current study about half of students attended more than one high school. The team used the full roster method (Hock and Isenberg, 2017) to attribute student outcomes to multiple high schools. The full roster method was developed in the context of teacher value-added models to account for the fact that students can be taught by more than one teacher during a school year. Many studies

[2] Because the promotion power models in this study were estimated using a single grade 9 cohort, they could not control for peer effects (such as the average achievement of a student's high school classmates) because there was no variation in these characteristics within schools. Chetty et al. (2014) found that a teacher value-added model that includes average school- and classroom-level test scores generates value-added scores that are similar to those generated by a model that excludes these characteristics.

have used the method in settings such as this one, in which students can attend multiple schools (Deutsch et al., 2020; Resch & Deutsch, 2015; Walsh et al., 2018).

The full roster method attributes a student's outcome to schools based on the amount of time that student spent at that school. For example, if a student attended school A for grades 9 and 10 and school B for grades 11 and 12, the full roster method attributes the student's outcomes evenly to both schools. But if a student attended school A for grades 9–11 and school B for grade 12, the method would assign more weight to school A than to school B in attributing the outcomes to each school.

To implement the full roster method, the study team structured the data to contain unique observations for each student–school combination and weighted observations based on the amount of time that a student spent at a school. For example, students who attended only one high school had one observation in the data, which received a weight of 1. Students who attended more than one high school had multiple observations in the data, each of which received a weight of less than 1.

All the schools that a student attended each year were observable in the OSSE data, but there was no information on exact transfer dates. For example, the data might have shown that a student attended two schools for grade 9 but did not indicate whether the student transferred one month or six months into the school year. Therefore, students who attended multiple schools within a year were treated as spending equal time at each school. For example, if a student attended schools A and B for grade 9 and only school B for grades 10–12, the study team would assign a weight of .125 to school A (half of the grade 9 year) and a weight a .875 to school B (half of the grade 9 year plus all of the grade 10–12 years). Ultimately, weighted least squares was used to estimate the regression equation and cluster standard errors at the student level to account for the correlation between multiple observations of the same student.

Following Deutsch et al. (2020), the study team constructed the weights so that they summed to 1 across all of a student's observations in the data. This held regardless of the number of high schools that a student attended or the number of years that a student attended a public DC high school. For example, students who graduated from a public DC high school received the same weight in this analysis as students who dropped out after grade 9. This is because the data could not distinguish between students who dropped out and those who exited DC schools, and dropping out of high school is related to students' long-term outcomes. Therefore, allowing the weights to vary based on the number of years that a student attended a public DC high school would have introduced bias into the promotion power scores.

A second issue related to appropriately attributing student outcomes to schools is determining the set of schools that could have contributed to each outcome. Consider a student who transfers to a new school for grade 12. This school could not meaningfully contribute to the student's performance on the SAT because most students take the SAT in the spring of grade 11 or the fall of grade 12. Therefore, the main analysis of SAT scores was limited to the schools that students attended for grades 9–11. In contrast, the analysis of high school graduation and college enrollment included the schools that students attended for grades 9–12. Accordingly, the weights associated with each student–school combination could vary by outcome.

*Steps to enhance stability: Excluding small schools and applying empirical Bayes shrinkage*. The study team took two steps to enhance the stability of the promotion power scores. First, the analysis of each outcome was restricted to schools with at least 10 student equivalents. This means that the sum of the student weights within a school must be greater than or equal to 10 for that school to be included in the analysis. Many of the schools that did not meet this threshold focused exclusively on adult education. In the sample nearly all the students who attended schools that did not meet this threshold also attended other, larger schools. Therefore, although this restriction led to several schools being dropped from the analysis, it resulted in just four students being dropped.

The remaining high schools were larger; each high school in the analytic sample had at least 25 student equivalents and at least 45 unique students.

The second step to enhance the stability of promotion power scores was to apply empirical Bayes shrinkage, a common approach in value-added modeling (Angrist et al., 2017; Chetty et al., 2014; Kane & Staiger, 2008). This method addressed the fact that the promotion power scores were more likely to be farther from the average for smaller schools than for larger schools, even after the analysis was restricted to schools with at least 10 student equivalents. The study team followed the procedure described in Morris (1983), which moves school estimates toward the mean in proportion to their precision, or estimated standard error. Estimates of schools with fewer student equivalents, and thus larger standard errors, get "shrunk" toward the mean more than those of schools with greater student equivalents because they are less precise. The resulting estimate for each school is (approximately) a weighted average of the school's original estimate and the estimate for the average school—which is 0 by construction—where the weight placed on the school's original estimate is greater when that estimate is more precise.

The model included students with some missing background characteristics. In all, 2 percent of students in the analysis were missing data on grade 8 math test scores, and 1 percent were missing data on grade 8 ELA test scores (see table A2). Students in the sample who were missing data on both math and ELA test scores were excluded from the analysis. Also, 4 percent of students were missing data on the proportion of students of their race/ethnicity in their middle school who were suspended. The team set missing values to the average and included indicator variables for missing values as additional background characteristics in the model (Puma et al., 2009).

A single student was missing baseline data on status for special education services and English learner student services and was excluded from the analysis. No other students were missing either of these fields.

*Variation in student outcomes explained by the model*. The R-squared statistic for each outcome measures the amount of variation in the outcome that is explained by the school effects and student background characteristics. The R-squared from the promotion power model is .49 for college-ready SAT scores, .20 for high school graduation, and .26 for college enrollment (table A4). The R-squared for SAT promotion power is nearly twice as large as the R-squared for high school graduation and college enrollment. This is likely because baseline test scores are a stronger predictor of future test scores than of these other outcomes.

**Table A4. R-Squared statistics of promotion power models, by outcome**

| Measure | College-ready SAT scores | High school graduation | College enrollment |
|---|---|---|---|
| R-squared | .49 | .20 | .26 |

Note: *n* = 3,568 first-time grade 9 students in 36 District of Columbia public high schools during the 2015/16 school year.
Source: Authors' analysis of data from the District of Columbia Office of the State Superintendent of Education.

The R-squared values for high school graduation and college enrollment in this study are comparable to those from promotion power models estimated on high schools in Louisiana. Deutsch et al. (2020) find an R-squared of .27 for high school graduation models and .24 for college enrollment models. As noted in Deutsch et al., however, the R-squared values from promotion power models, including for SAT performance, are much lower than the R-squared values typically found in value-added studies with test score outcomes. Prior value-added work tended to report R-squared values of .60–.80. Deutsch et al. cite two potential reasons for this discrepancy. First, value-added studies usually account for a lagged version of the outcome, which is not possible for promotion power because there is no lagged version of graduation or college enrollment. Second, the outcomes in the current study are binary and therefore are expected to yield lower R-squared values than similar continuous outcomes because there is less variation to be explained (Cox & Wermuth, 1992). These lower R-squared values could indicate that

promotion power models might be less effective at removing bias from students selecting into schools than typical school value-added models.

*Promotion power regression output*

**Table A5. Regression coefficients, significance levels, and standard errors for promotion power models, by outcome**

| Variable | College-ready SAT scores | High school graduation | College enrollment |
|---|---|---|---|
| PARCC math score in grade 8 | 0.10** | 0.05** | 0.07** |
| | (0.01) | (0.01) | (0.01) |
| Missing PARCC math score in grade 8 | –0.03 | –0.26** | –0.10* |
| | (0.02) | (0.06) | (0.05) |
| PARCC English language arts score in grade 8 | 0.06** | 0.06** | 0.10** |
| | (0.01) | (0.01) | (0.01) |
| Missing PARCC English language arts score in grade 8 | 0.01 | –0.16* | –0.25** |
| | (0.03) | (0.07) | (0.04) |
| Asian | 0.23** | –0.03 | –0.04 |
| | (0.06) | (0.04) | (0.06) |
| Hispanic | 0.04 | –0.07** | –0.10** |
| | (0.02) | (0.02) | (0.03) |
| White | 0.26** | –0.02 | –0.12** |
| | (0.03) | (0.02) | (0.04) |
| Other race/ethnicity | 0.12 | –0.08 | –0.07 |
| | (0.07) | (0.05) | (0.07) |
| Female | –0.03** | 0.07** | 0.08** |
| | (0.01) | (0.01) | (0.02) |
| Received special education services | 0.02** | –0.03 | –0.05* |
| | (0.01) | (0.02) | (0.02) |
| English learner student | –0.01 | –0.06 | 0.01 |
| | (0.02) | (0.04) | (0.04) |
| Middle school suspension rate, same race/ethnicity | 0.10** | –0.11* | –0.08* |
| | (0.03) | (0.05) | (0.05) |
| Missing middle school suspension rate, same race/ethnicity | –0.01 | 0.09** | 0.08* |
| | (0.03) | (0.03) | (0.04) |

*Statistically significant at $p < .05$; **statistically significant at $p < .01$.
PARCC is Partnership for Assessment of Readiness for College and Careers.
Note: $n$ = 3,568 first-time grade 9 students in 36 District of Columbia public high schools during the 2015/16 school year.
Source: Authors' analysis of data from the District of Columbia Office of the State Superintendent of Education.

## *Research question 1: Additional details*

The study team used the standard deviation (in student standard deviation units) to assess the variation in promotion power across schools. For each outcome this statistic was calculated as the adjusted standard deviation after empirical Bayes shrinkage divided by the student-level standard deviation.

The benchmarks for effect sizes introduced in Kraft (2019) were used to determine whether promotion power meaningfully distinguished among high schools: effect sizes less than 0.05 are classified as small, 0.05 to less than 0.20 as medium, and 0.20 or greater as large. Differences in promotion power scores across schools were classified

as large if the standard deviation of promotion power scores, in student standard deviation units, was 0.20 or greater.

The study team adopted these thresholds because they are tailored to education interventions; they are based on the distribution of effect sizes from more than 700 randomized controlled trials of education interventions. One limitation of using these benchmarks for this study is that they are derived from effects on standardized test scores, whereas this study also examined high school graduation and college enrollment outcomes. However, Kraft (2019) notes that effect sizes tend to be smaller for longer-term outcomes. Therefore, if anything, a data-driven threshold for a large effect size on the outcomes in this study might be less than 0.20.

In addition to examining the standard deviation of promotion power, the study team estimated the proportion of high schools that were statistically different from the average. The team classified high schools as having promotion power scores statistically different from the average if the promotion power score divided by its standard error was –1.96 or less or 1.96 or greater. This is equivalent to using a 95-percent confidence interval to determine the statistical significance of the promotion power score. Furthermore, a school was classified as having a promotion power score that was statistically higher than the average if the score was statistically different from the average and the promotion power score was greater than 0 and statistically lower than the average if the promotion power score was statistically different from the average and the promotion power score was less than 0.

### *Research question 2: Additional details*

Research question 2 sought to determine the relationship between a school's power to promote different student outcomes. The study team estimated the correlation between a school's promotion power score for high school graduation and college enrollment, college-ready SAT scores and college enrollment, and college-ready SAT scores and high school graduation.

### *Research question 3: Additional details*

To determine how promotion power scores compare to status measures in their relationship with student background characteristics, the study team assessed whether grade 8 test scores were less strongly related to promotion power scores than to status measures. For each outcome the magnitude of two correlations was compared: the correlation between the average grade 8 achievement of a high school's students and the school's promotion power score, and the correlation between average grade 8 achievement and the school's average outcome. Average grade 8 achievement was defined as the average of grade 8 math and English language arts standardized test scores (*z*-scores). For each outcome the study team tested whether the difference between the two correlations was statistically significant.[3] If the first correlation was smaller than the second and the difference was statistically significant, the promotion power scores were determined to be less strongly related to average grade 8 achievement than status measures.

### *References*

Angrist, J. D., Hull, P. D., Pathak, P. A., & Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics, 132*(2), 871–919.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593–2632.

[3] A two-tailed test was used that accounted for the fact that the two correlations came from the same sample and employed a common variable (average grade 8 achievement).

Cox, D. R., & Wermuth, N. (1992). A comment on the coefficient of determination for binary responses. *The American Statistician*, *46*(1), 1–4.

Deutsch, J., Johnson, M., & Gill, B. (2020). *The promotion power impacts of Louisiana high schools* (No. 2c041387caf14e9eac49cd539408824f). Mathematica.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, *10*(1), 117–156.

Hock, H., & Isenberg, E. (2017). Methods for accounting for co-teaching in value-added models. *Statistics and Public Policy*, *4*(1), 1–11.

Kane, T. J., & Staiger, D. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (NBER Working Paper No. 14607). National Bureau of Economic Research.

Kraft, M. (2019). *Interpreting effect sizes of education interventions* (No. 19-10). Annenberg Institute at Brown University.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of American Statistical Association, 78*(381), 47–55.

National Student Clearinghouse. (2019). *More than just data…* https://studentclearinghouse.info/onestop/wp-content/uploads/NSCFactSheet.pdf.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE No. 2009-0049). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Resch, A., & Deutsch, J. (2015). *Measuring school and teacher value added in Charleston County School District, 2014-2015 school year*. Mathematica Policy Research.

Walsh, E., Dotter, D., & Liu, A. Y. (2018). *Can more teachers be covered? The accuracy, credibility, and precision of value-added estimates with proxy pre-tests* (Working Paper No. 64). Mathematica Policy Research.

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach*. South-Western Cengage Learning.

# Appendix B. Supporting analyses

## *Robustness to alternative models and samples*

The study team's main approach to estimate promotion power models was to use a linear probability model and control for background characteristics (see table A5 in appendix A). But there are other reasonable ways to estimate high school promotion power. To assess the robustness of the main results, the study team estimated alternative promotion power models and calculated the correlations between the main and alternative estimates. A high correlation indicates that the results would have been similar had the alternative model been used, and a low correlation indicates that the results are sensitive to the modeling decision. Research assessing the sensitivity of value-added and promotion power models typically finds correlations greater than or equal to .85, which offers a useful benchmark for how to define a high correlation (Deutsch et al., 2020; Goldhaber et al., 2013; Guarino et al., 2015; Johnson et al., 2015).

The rest of this section details alternative decisions that the study team could have made and reports the correlation between the main promotion power scores and the alternatives for each outcome (table B1).

**Table B1. Robustness of promotion power scores to alternative models and samples, by outcome (correlations)**

| Alternative model | College-ready SAT score | High school graduation | College enrollment |
|---|---|---|---|
| Logistic regression | .88 | .95 | >.99 |
| Errors-in-variables regression | >.99 | >.99 | >.99 |
| Polynomial grade 8 test scores | .99 | .99 | .99 |
| At risk in high school[a] | >.99 | .97 | .99 |
| No gender and race/ethnicity | .99 | >.99 | .99 |
| No middle school–level controls | >.99 | .99 | >.99 |
| No missing baseline covariates | >.99 | .98 | .99 |

Note: A high correlation indicates that the results would have been similar had the alternative model been used rather than the main model. Research assessing the sensitivity of value-added and promotion power models typically finds correlations greater than or equal to .85, which offers a useful benchmark for defining a high correlation; see text.
a. A student with any of the following conditions is considered at risk: student's family received Temporary Assistance for Needy Families benefits or Supplemental Nutrition Assistance Program benefits, student was homeless or in foster care, or student was over-age for the grade.
Source: Authors' analysis of data from the DC Office of the State Superintendent of Education.

*Logistic regression.* The main analysis used a linear probability model to estimate promotion power, but the study team also assessed robustness to using a logistic model.

The logistic model estimated school effects only for high schools that had variation in the outcome. In the analytic sample no students at three schools achieved a college-ready SAT score, and all students graduated at one school. The logistic model could not estimate promotion power scores for these schools, whereas the linear probability model could.

To compare results, the study team re-estimated linear probability models after excluding schools in which all or no students achieved the outcome. For example, the study team calculated correlations for high school graduation using the sample of schools in which some students graduated and others did not. The correlations range from .88 to greater than .99 across outcomes.

*Errors-in-variables regression.* The main analysis did not account for measurement error in grade 8 math and English language arts test scores, which could induce bias in promotion power scores. As a measure of student ability, test scores contain measurement error. This measurement error causes attenuation in the estimated coefficients associated with the test scores, and this bias in turn causes bias in the other regression coefficients, including promotion power scores (Buonaccorsi, 2010).

To account for this possibility, the study team assessed robustness to using errors-in-variables regression models described in Buonaccorsi (2010). This technique uses the published reliability of the test score to disattenuate the coefficient estimates associated with the test scores and thus avoids inducing bias in other coefficient estimates (Partnership for Assessment of Readiness for College and Careers, 2016).[4] The promotion power scores from errors-in-variables regressions highly correlate with those from the main model. All correlations are greater than .99.

*Flexibly controlling for prior test scores.* The main analysis included linear controls for grade 8 math and English language arts test scores. Other studies often include polynomial functions of baseline test scores (Chetty et al., 2014; Jackson, 2014; Mansfield, 2015). Although the true relationship between prior test scores and students' long-term outcomes might be nonlinear, the study team included only linear controls in the main model because higher order terms can exacerbate the problem of measurement error in test scores (Lockwood & McCaffrey, 2014). Using more flexible controls for grade 8 test scores made little difference in practice: the correlations for each outcome are at least .99.

*Controlling for an "at risk" indicator measured during high school.* DC Office of the State Superintendent of Education (OSSE) data do not include measures of background economic disadvantage, such as eligibility for the National School Lunch Program or public assistance, that are typically included as controls in value-added or promotion power models. However, the data did include an indicator variable for whether a student was at risk in high school. The variable was equal to 1 if any of the following conditions were met: the student's family received Temporary Assistance for Needy Families benefits or Supplemental Nutrition Assistance Program benefits, the student was homeless or in foster care, or the student was over-age for the grade. The at-risk indicator was measured in grade 12 for most students, but it was measured earlier for students who exited public DC high schools before grade 12.

The study team did not include the at-risk variable in the main analysis because it is measured after students enter high school. In this sense whether a student was deemed at risk could be influenced in part by a student's high school. For example, school retention decisions directly affect whether students are over-age for their grade.

The results are robust to whether an at-risk indicator measured in high school is controlled for: the correlations range from .97 to greater than .99.

*Excluding controls for gender and race/ethnicity.* Some states and districts prefer to exclude gender and race/ethnicity from teacher or school value-added models. For example, Deutsch et al. (2020) did not include these control variables in the promotion power models for Louisiana high schools. Models with and without controls for gender and race/ethnicity highly correlate with each other, and the correlations are all at least .99.

*Excluding middle school–level discipline variables.* Because the OSSE data did not include baseline student-level discipline information, the study team included middle school–level discipline information from the Civil Rights Data Collection in the main model as a proxy for student suspension rates. Specifically, the analysis controlled for the proportion of students in a student's middle school who were of the same race/ethnicity who were suspended. Controlling for prior school-level information is not standard in promotion power or value-added models,

[4] The error-in-variables regression, as implemented in Stata, does not allow standard errors to be adjusted for clustering. As a result, the team implemented the error-in-variables regression in two steps: first, the error-in-variables regression was estimated, and an adjusted outcome variable was generated that was equal to the original outcome less the effects of the grade 8 test scores—that is, $\tilde{Y} = Y - X^T\hat{\gamma}^T$, where $X^T$ represents the grade 8 test scores and $\hat{\gamma}^T$ represents the error-adjusted coefficients on those test scores. In the second step a standard regression is run, with standard errors adjusted for clustering, using $\tilde{Y}$ as the outcome and without test scores on the right-hand side. Promotion power scores were estimated from the second regression. The standard errors in this second regression, including those associated with the promotion power scores, are biased downward because the regression treated the coefficients on the grade 8 test scores as if they were known rather than parameters to be estimated.

however. Therefore, the study team assessed the sensitivity of the results to the exclusions of these variables. The correlations are all at least .99.

*Excluding students who were missing baseline covariates.* As discussed in appendix A, the analytic sample included students who were missing grade 8 test scores in either math or English language arts but not students who were missing scores in both subjects. It also included students missing middle school discipline information. The results would not have changed substantially had the team instead excluded students who were missing any baseline covariates; the correlations range from .98 to greater than .99 (see table B1).

### *Robustness to alternative SAT outcomes*

The main analysis of schools' promotion power for college-ready SAT scores used a binary outcome that equaled 1 if students scored at or above the College Board's college and career readiness benchmarks and 0 otherwise. Students who were missing SAT scores were coded as not meeting the benchmarks, regardless of whether they did not take the SAT or whether they took the SAT after exiting public DC high schools. There are three potential concerns with this outcome.

First, students who took the SAT after exiting public DC high schools might have scored at or above the benchmarks, yet they were coded as not meeting the benchmarks. To address this concern, the study team compared the background characteristics of students who were missing SAT scores with students who were not missing scores. The students who were missing SAT scores had lower baseline test scores, were more likely to receive special education services, had a lower grade 9 attendance rate, and had a higher grade 9 discipline rate (table B2). Therefore, it appears unlikely that many of these students would have scored at or above the benchmarks even if they had taken the SAT after exiting public DC high schools.

**Table B2. Summary statistics of outcomes and background characteristics for students who were and those who were not missing SAT scores**

| Background characteristic | Not missing SAT scores | Missing SAT scores |
|---|---|---|
| **PARCC grade 8 standardized test scores (standard deviations)[a]** | | |
| Math score | 0.11 | –0.33 |
| English language arts score | 0.13 | –0.43 |
| **Sociodemographic characteristics (percent)** | | |
| Asian | 20 | 0 |
| Black | 75 | 84 |
| Hispanic | 15 | 11 |
| White | 7 | 3 |
| Other race/ethnicity | 1 | 1 |
| Female | 53 | 44 |
| Received special education services | 15 | 28 |
| English learner student | 5 | 6 |
| At risk in high school[b] | 59 | 83 |
| Grade 9 attendance rate | 91 | 76 |
| Grade 9 discipline rate | 0 | 2 |
| **Sample** | | |
| Number of students | 2,806 | 762 |
| Percentage of students | 79 | 21 |

PARCC is Partnership for Assessment of Readiness for College and Careers.
a. PARCC grade 8 math and English language arts scores are *z*-scores: they are set to have a mean of 0 and a standard deviation of 1.
b. A student with any of the following conditions is considered at risk: student's family received Temporary Assistance for Needy Families benefits or Supplemental Nutrition Assistance Program benefits, student was homeless or in foster care, or student was over-age for the grade.
Source: Authors' analysis of data from the DC Office of the State Superintendent of Education.

A second potential concern is that the analysis might have missed important variation in the outcome by using a binary measure of SAT scores. That is, when a school improves a student's test score but does not move it from below the threshold to over it, that change would not be observable in the SAT outcome measure used in the analysis. To address this concern, the study team first compared the promotion power scores using the binary outcome with the full sample and with a sample that excluded students who were missing SAT scores. This correlation was quite high (.97; table B3). With this established, the study team compared estimates using the binary outcome and continuous SAT scores among students who were not missing SAT scores. The results were similar when using binary and continuous measures, with a correlation of .91.

Estimates might also have differed if a score threshold other than the College Board's college and career readiness benchmarks (at least 530 in math and 480 in reading) had been used. For example, schools that appeared to be effective at meeting the College Board benchmarks might have been less effective at promoting even higher performance. Therefore, the study team assessed the robustness of the main analysis to several alternative score thresholds: 800, 900, 1,000, and 1,100. For this check, students were defined as scoring at or above the threshold if the sum of their highest math and reading scores is greater than or equal to the threshold. The results were fairly robust to alternative thresholds that indicate much higher performance on the SAT: the correlations were .96 for a threshold of 1,000 and .97 for a threshold of 1,100.

The correlation with the measure that used a threshold of 800 was much lower, at .45. Many more students scored at least 800 on the SAT (58 percent) than met the higher College Board benchmarks (17 percent). Thus, a score threshold of 800 measures a much different level of competency than college and career readiness, and it might not be surprising that the correlation is so low.

**Table B3. Robustness of SAT promotion power to alternative SAT outcomes**

| Alternative model | Correlation |
|---|---|
| Not missing SAT scores[a] | .97 |
| SAT scale score[b] | .91 |
| SAT score of 800 or higher | .45 |
| SAT score of 900 or higher | .77 |
| SAT score of 1,000 or higher | .96 |
| SAT score of 1,100 or higher | .97 |

Note: The main analysis used a binary measure of SAT score for SAT promotion power (the College Board's college and career readiness benchmarks of scoring at least 530 in math and 480 in reading).
a. Reports the correlation between promotion power scores using the binary outcome with the full sample and with a sample that excludes students who were missing SAT scores.
b. Reports the correlation between estimates using the binary outcome (met or did not meet the College Board's college and career readiness benchmarks) and estimates using continuous SAT scores among students who were not missing SAT scores.
Source: Authors' analysis of data from the District of Columbia Office of the State Superintendent of Education.

### *Sensitivity to a broader set of student background characteristics*

The administrative data from OSSE included information on key background characteristics such as grade 8 math and English language arts test scores, gender, race/ethnicity, and grade 8 status for special education services and English learner student services. The OSSE data did not, however, contain other information commonly used to account for factors outside a school's control, such as eligibility for the National School Lunch Program, grade 8 attendance, and disciplinary incidents in grade 8. Without these characteristics, the promotion power models in this study might not isolate a school's contribution to students' outcomes from factors outside of a school's control.

The study team used administrative records from the Louisiana Department of Education (LDOE) to assess the sensitivity of promotion power scores to controlling for a broader set of background characteristics than were available in the OSSE data. The background characteristics that were included in the main promotion power

models using OSSE data were compared with those available in the LDOE data (table B4). Deutsch et al. (2020) provide further details about these variables and the LDOE data in their study of promotion power for Louisiana high schools.

**Table B4. Comparison of background characteristics available from the District of Columbia Office of the State Superintendent of Education and Louisiana Department of Education**

| Background characteristic | DC Office of the State Superintendent of Education data | Louisiana Department of Education data |
|---|---|---|
| Math test scores | X | X |
| English language arts test scores | X | X |
| Science test scores | | X |
| Social studies test scores | | X |
| Absences | | X |
| Attended fewer than 45 days | | X |
| Suspension rate | | X |
| Over-age for grade | | X |
| Disability: Emotional disturbance | | X |
| Disability: Learning disability | | X |
| Disability: Intellectual disability | | X |
| Disability: Other health impairment | | X |
| Disability: Speech impairment | | X |
| Received special education services | X | |
| Gifted student status | | X |
| English learner student | X | X |
| Eligible for free lunch | | X |
| Eligible for reduced price lunch | | X |
| American Indian/Alaska Native | X | X |
| Asian/Pacific Islander | X | X |
| Black | X | X |
| Hispanic | X | X |
| Female | X | X |

Note: All characteristics are measured in grade 8 except for over-age for grade which was measured in the fall of grade 9. Test scores from the DC Office of the State Superintendent of Education come from the Partnership for Assessment of Readiness for College and Careers. Test scores from the Louisiana Department of Education come from the Louisiana Educational Assessment Program. Race/ethnicity and gender characteristics were not used in the promotion power models developed for the Louisiana Department of Education but were available in Louisiana Department of Education data and were used in this analysis.
Source: Data from the District of Columbia Office of the State Superintendent of Education and the Louisiana Department of Education.

To assess the sensitivity of the estimates to the set of student background characteristics used, the study team first estimated promotion power models using the LDOE data and controlling for the limited set of background characteristics available in the OSSE data (see column 1 in table B4). Then, the team re-estimated promotion power, again using LDOE data, but this time controlling for the broader set of background characteristics in column 2. As before, a high correlation between estimates from these two models indicates that the results are robust to including only a limited set of student controls.

The set of controls in the main model were slightly different from the limited set of background characteristics shown in column 1 of table B4. Specifically, the main model also included middle school–level student discipline

information from the U.S. Department of Education's Civil Rights Data Collection (CRDC). The LDOE data did not include identifiers for students' middle schools that could enable students to be linked to CRDC data. Even so, the main promotion power scores highly correlate with those from a model that excluded middle school–level discipline information (see table B1).

The sensitivity analysis examined high school graduation and college enrollment outcomes but not SAT scores, which are not included in the LDOE data. The team focused on the cohort of first-time grade 9 students in Louisiana public schools during the 2013/14 school year, which was the most recent cohort with available data. This analysis is slightly different from that of Deutsch et al. (2020), which used information from multiple cohorts to estimate promotion power models.

The results of this replication exercise are reported in table B5. The first row of results shows the correlation between promotion power scores when a broad rather than a limited set of background characteristics are controlled for. Across both outcomes the promotion power scores highly correlate with each other: .95 for high school graduation and .93 for college enrollment. This suggests that accounting for the broader set of background characteristics did not considerably alter promotion power scores.

**Table B5. Comparison of promotion power models for high school graduation and college enrollment, controlling for broad and limited sets of student background characteristics**

| | High school graduation | | College enrollment | |
|---|---|---|---|---|
| Background characteristic | Broad | Limited | Broad | Limited |
| Correlation between measures that control for broad versus limited set of student background characteristics | .95 | | .93 | |
| R-squared | .27 | .17 | .25 | .20 |
| Correlation between promotion power scores and grade 8 standardized test scores | .34 | .49 | .27 | .48 |
| Correlation between school average outcome (status measure) and grade 8 standardized test scores | .70 | | .78 | |
| Correlation between promotion power scores and school average outcome (status measure) | .87 | .96 | .75 | .90 |

Source: Authors' analysis of data from Louisiana Department of Education.

Even though the promotion power scores highly correlate with each other, models that controlled for the limited set of background characteristics appear less successful at removing bias. The second row of results in table B5 reports the R-squared from each model, which signifies the proportion of variation in student outcomes explained by the covariates in the model. The broad set of background characteristics explains 10 percentage points more of the variation in high school graduation (.27 versus .17) and 5 percentage points more of the variation in college enrollment (.25 versus .20).

Promotion power scores from models that controlled for the limited set of characteristics were more strongly related to school-average background characteristics than measures from models that controlled for the full set. For example, the third row of results in table B5 reports the correlation between promotion power scores and average grade 8 standardized test scores (*z*-scores). The correlation between promotion power for high school graduation and grade 8 standardized test scores is .34 when the broad set of background characteristics in the LDOE model is used and .49 when the more limited set of characteristics in the main analysis of the study is used.

Even so, estimates from the model with the limited set of covariates reduced the association between the populations of students that schools serve and schools' outcome measures. In Louisiana the correlation between high school graduation rate and grade 8 standardized test scores was .70 (fourth row of results in table B5), and

the correlation between promotion power for high school graduation (using the broad set of covariates) and grade 8 standardized test scores was .34 (third row of results). The difference between these two correlations (.36) could be thought of as the amount of bias reduction associated with using the promotion power model with the broad set of covariates as compared with using a status measure. In contrast, when the limited set of covariates is used, the correlation between promotion power for high school graduation and grade 8 test scores is .49. As expected, the amount of bias reduction associated with using the promotion power model with the limited set of covariates (.70 – .49 = .21) was smaller than with the broad set of covariates (.36). However, promotion power models with the limited set of covariates still appeared to remove more than half the bias associated with traditional status measures. Specifically, for high school graduation, promotion power models with the limited set of covariates appear to remove 58 percent of the bias (.21 / .36 = .58), and for college enrollment they appear to remove 59 percent of the bias (.30 / .51 = .59). This analysis indicates that even promotion power models that account for relatively few student background characteristics, such as the main analysis estimates in this study for OSSE, are considerably less biased than traditional status measures.

A related concern is whether promotion power models are as successful in removing bias as teacher and school value-added measures, which prior research indicates have limited, if any, bias (Angrist et al. 2016; Chetty et al. 2014; Deming 2014; Deutsch, 2012; Kane et al. 2013). After the broad set of student background characteristics is controlled for, the correlation between promotion power and school average outcomes is .87 for high school graduation and .75 for college enrollment (see fifth row of results in table B5). In contrast, Chetty et al. (2014) found a correlation of .30 between teacher value-added and student average test scores. Thus, promotion power models might be less successful than value-added models in removing bias.

Overall, this exercise reveals that promotion power scores are robust to controlling for a more limited set of student background characteristics. And models that use fewer student background characteristics than the model in Deutsch et al. (2020) still provide a less-biased indication of a school's influence than status measures such as high school graduation and college enrollment rates.

### *Sensitivity to multiple student cohorts*

Some schools' estimates could be further from the schools' true promotion power because data were available for only a single cohort of first-time grade 9 students. The study team again used LDOE data to compare school estimates when using information from one cohort compared with two cohorts of first-time grade 9 students. To ensure that the analysis with LDOE data was comparable to what might have been found in the main analysis with OSSE data, this analysis accounted for only the limited set of student background characteristics detailed in column 1 of table B4.

The study team followed the method introduced in Deutsch et al. (2020) to combine the school estimates when using multiple cohorts. Specifically, the promotion power model was estimated separately by cohort. Then, before empirical Bayes shrinkage was applied, an average of the two school estimates, weighted by the share of student equivalents in each cohort, was generated. Similarly, a standard error was constructed for the combined estimate that treated the two cohort-specific estimates as independent because they were based on distinct sets of students. Finally, empirical Bayes shrinkage was applied using the combined school estimates and standard errors.

As expected, models with multiple cohorts were more precise than models with only a single cohort (table B6). Models that are more precise will identify a greater proportion of schools as statistically different from the average. For high school graduation 52 percent of schools are statistically different from the average when using one cohort compared with 62 percent when using two cohorts. The results were similar for college enrollment; 40 percent of schools were statistically different from the average with one cohort compared with 52 percent with two. These findings are consistent with those of Deutsch et al. (2020), who found precision gains to including a second cohort, yet less so for adding a third.

**Table B6. Comparison of promotion power models using one and two cohorts**

| | High school graduation | | College enrollment | |
|---|---|---|---|---|
| Statistic | One cohort | Two cohorts | One cohort | Two cohorts |
| Percentage of schools different from average | 52 | 62 | 40 | 52 |
| Correlation between measures based on one and two cohorts | .98 | | .95 | |

Source: Authors' analysis of data from Louisiana Department of Education.

School estimates using one cohort and estimates using two cohorts are highly correlated: .98 for high school graduation and .95 for college enrollment. This indicates that the main school effects for public DC high schools were unlikely to have changed by much had information on multiple student cohorts been available.

## *References*


Angrist, J., Hull, P., Pathak, P., & Walters, C. (2016). Interpreting tests of school VAM validity. *American Economic Review, 106*(5), 388–392.

Angrist, J. D., Hull, P. D., Pathak, P. A., & Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics, 132*(2), 871–919.

Buonaccorsi, J. P. (2010). *Measurement error: Models, methods, and applications.* Chapman and Hall/CRC.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593–2632.

Deming, D. J. (2014). Using school choice lotteries to test measures of school effectiveness. *American Economic Review, 104*(5), 406–411.

Deutsch, J. (2012). *Using school lotteries to evaluate the value-added model.* University of Chicago Working Paper.

Deutsch, J., Johnson, M., & Gill, B. (2020). *The promotion power impacts of Louisiana high schools* (No. 2c041387caf14e9eac49cd539408824f). Mathematica.

Goldhaber, D. D., Goldschmidt, P., & Tseng, F. (2013). Teacher value-added at the high-school level: Different models, different answers? *Educational Evaluation and Policy Analysis*, *35*(2), 220–236.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, *10*(1), 117–156.

Jackson, K. (2014). Teacher quality at the high school level: The importance of accounting for tracks. *Journal of Labor Economics*, *32*(4), 645–684.

Johnson, M. T., Lipscomb, S., & Gill, B. (2015). Sensitivity of teacher value-added estimates to student and peer control variables. *Journal of Research on Educational Effectiveness*, *8*(1), 60–83.

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

Lockwood, J. R., and McCaffrey, D. F. (2014). Should nonlinear functions of test scores be used as covariates in a regression model? In R. W. Lissitz & H. Jiao (Eds.), *Value-added modeling and growth modeling with particular application to teacher and school effectiveness*. Information Age Publishing.

Mansfield, R. (2015). Teacher quality and student inequality. *Journal of Labor Economics*, *33*(3), 751–788.

Partnership for Assessment of Readiness for College and Careers. (2016). *PARCC final technical report for 2015 administration*. https://eric.ed.gov/?id=ED599097.