A Method for Identifying Comprehensive Support and Improvement Schools and Holes in the Every Student Succeeds Act

Tony Moss

Kansas State Department of Education

30 August 2016, revised 20 July 2017

Author Note

Tony Moss is a researcher and data analyst at the Kansas State Department of Education. The interpretations and views expressed in this paper are those of the author, not KSDE nor the Kansas State Board of Education.

Correspondence concerning this article should be addressed to Tony Moss, Kansas State Department of Education, Landon State Office Building, 900 SW Jackson Street, Suite 653, Topeka, KS 66612-1212.

E-mail: tmoss@ksde.org

Abstract

The 2015 federal law, the Every Student Succeeds Act (ESSA), obligates states to identify the lowest-performing Title I schools based on four academic measures and at least one qualitative measure. A workgroup identified nine factors that they believed would accurately identify high-risk schools: chronic absence, student mobility, cumulative poverty, higher concentrations of Students with Disabilities, migrant students, or English learners, the rate of suspensions and expulsions, the demographic distance in gender and ethnicity between teachers and students, and the percentage of new teachers. Four of the nine, cumulative poverty, percentage of English learners, the rate of suspensions and expulsions, and chronic absences, were predictive of lower school-level performance as measured by an index derived from state assessments. One factor, the demographic distance between teachers and students, contrary to expectations, was strongly predictive of improved school academic performance. The four negative school risk factors were used to identify the highest-risk schools in a formula with a 60 percent weight on the state assessment index, and a 40 percent combined weight on the four factors predictive of lower school performance. This method almost exclusively identified high-poverty urban schools as the highest-need Title I schools. The nine factors were also used as control variables in a regression predicting a state assessment index. When actual school academic performance was compared to predicted results, a diverse group of rural as well as urban schools were identified as performing below expectations. ESSA, continuous improvement models, and policy implications are discussed.

Keywords: accountability; risk; state education agencies; federal state relationship; student mobility; attendance; ethnic groups; social bias; school turnaround; Every Student Succeeds Act; Total Quality Management; United States; 2016; Kansas public schools; quantitative/regression

A Method for Identifying Comprehensive Support and Improvement Schools and Holes in the Every Student Succeeds Act

The new federal law, the Every Child Succeeds Act (ESSA), requires state education agencies (SEAs) to identify the lowest performing Title I schools for Comprehensive Support and Improvement, or CSI status. The SEA, the Kansas State Department of Education (KSDE), formed a collaborative workgroup to develop a method to identify the CSI schools. The workgroup sought to use the flexibility of the new law to redefine school identification away from the public embarrassment of NCLB, to one based on validated school risk-factors. Factors predictive of lower school performance are more likely to direct attention toward the causes of lower-school performance.

In a parallel change, the Kansas Board of Education has moved away from an emphasis on assessment results to a broader set of developmental student outcomes, including the social-emotional skills that are predictive of labor market and health outcomes (Kautz, Heckman, Diris, ter Weel, & Borghans, 2014). KSDE invited representatives from some of the largest school districts in Kansas—Denise Seguine and Neil Guthrie from the Wichita school district, Juanita Erickson and Tammy Austin from USD 501 Topeka, David Rand and Kristen Scott from USD 500 Kansas City, and Renae Hickert from USD 480 Liberal—to participate in the workgroup. The SEA was represented by Brad Neuenswander, the Deputy Commissioner, Tammy Mitchell, the director of KSDE's school improvement project, Michele Hayes from the Kansas Learning Network (KLN), and Tony Moss, a researcher and the author of this report.

In a series of meetings, the workgroup identified the nine risk factors, tested the risk factors with data aggregated to the school level, and reviewed the schools each factor, if applied by itself, would identify as high-risk schools. Then the nine factors were used as predictors in a regression

model to see if, and how much, they predicted schools' academic performance, and how well they collectively identified the highest-risk CSI schools. This report shares the results of this experiment so that other states and districts attempting to identify their highest-need schools under ESSA may borrow or improve upon these methods. It also considers the policy issues influencing the experiment.

Literature Review

The school accountability movement, as expressed in the federal NCLB Act, strongly emphasized school assessment scores and a rigidly ascending staircase of improving proficiency. The schools which were identified as failing were expected to make exceptionally large improvements—to turnaround—or face increasingly severe consequences. A science for converting low-performing schools into continuously improving ones never developed (Klute, Cherasaro, & Apthorp, 2016; Kutash, Nico, Gorin, Rahmatullah, & Tallant, 2010). Case studies of apparently successful school turnarounds were sometimes put forward by advocacy organizations using poor methodology (Trujillo & Rivera, 2016), or by the U.S. Department of Education blogs and newsletters. Schools in decline were typically excluded from studies (Hochbein, 2012). NCLB was widely discredited—for narrowing the curriculum, for failing to reduce academic gaps, for high-stakes that sometimes led to gaming and cheating. To its credit, NCLB quantified academic gaps between student groups, between those with disabilities and those without, between those with subsidized family lunches and those with family-paid lunches, and between ethnic groups.

In December 2015, NCLB was replaced by the Every Student Succeeds Act (ESSA). While lacking the rigidity and threats of NCLB, ESSA continues to emphasize state assessments as the principal means for identifying low-performing schools and subgroups. Current assessments are now based on more demanding college-level standards, so without effective measures to counter the

causes of academic and skills gaps, states should expect to observe larger proficiency gaps than under NCLB.

There are three bodies of research that have undermined assumptions of assessment-driven school improvement but have been ignored by the law. First, human development research has managed to explain the mechanisms by which developmental stressors shape a child's brain and, to some extent, their later academic and social capacities. What is variously called the Barker hypothesis, fetal programming, life-course epidemiology, or the Developmental Origins of Health and Disease, with advances in epigenetics and endocrine influences on early development, can now explain the correlations of the 1966 Coleman Report. Research can now explain the causal mechanisms between family poverty and later suppressed academic capacities. It can even measure how developmental stressors physically change the brains of young children (Brooks-Gunn & Markman, 2005; Heckman, 2006; Lerner, 2006; Luby, et al., 2013; Noble, Houston, Kan, & Sowell, 2012). This growing body of research makes clear that the expansion of equal educational opportunity, and the goal of cultivating a globally competitive workforce, to some extent depend on early developmental interactions, the quality of child-rearing, family and early child-care environments, and chronic exposures to stress. These developmental discoveries undermine the accountability movement's assumption as expressed in ESSA, that testing and transparency in test scores between groups will signal to the public, administrators, teachers, and students, how well they are performing, and that all stakeholders, now informed by test scores, will do what is needed to improve academic achievement. The NCLB experiment demonstrated that transparent test scores will not remedy the interactions that suppress optimum child development.

Second, also absent from ESSA's design is any influence of the international comparative studies of teacher selection, training, and retention, and the large role they seem to play in successful educational reforms. In the United States, the accountability movement focused on evaluating

teacher performance by linking student assessments to individual teachers. This value-added approach suggested that the quality of teaching was the most important school-based influence on student achievement (Sawchuk, 2011) but it ignored teacher selection, training, and retention. Value-added approaches that blamed teachers were so compromised and controversial (Baker, et al., 2010), that ESSA removed provisions requiring teacher evaluations based on test scores, prohibited the U.S. Department of Education from interfering in teacher evaluations, and dropped the NCLB requirements that teachers have to have a bachelor's degree and be certified in the subjects they teach.

This federal retreat from teacher quality is contradicted by international country case studies. McKinsey's *Closing the talent gap: Attracting and retaining top-third graduates to careers in teaching* (Auguste, Kihn, & Miller, 2010) found that a crucial ingredient to reform was the careful selection and deep training of exceptionally capable individuals as teachers. Other factors identified by the international comparative study included the matching of teacher supply to teacher demand, better working conditions, lower teacher turnover, higher salaries relative to comparable professions, and government-paid higher education for teacher-candidates. Book-length international case studies by Mark Tucker (2011) and Peter Sahlberg (2011) also emphasized the primary role of careful teacher selection and deep training. The lack of teacher selection and training experiments in ESSA undermines the law's professed goal of improving the educational success of lower-income students and turning around lower performing schools.

A third influential body of research has demonstrated the importance of a set of social and personality skills as parallel and comparably weighty contributors to academic and life success (Kautz, Heckman, Diris, ter Weel, & Borghans, 2014). This body of research has had some influence on ESSA. ESSA requires a measure of school quality and suggests social-environmental factors like teacher or student engagement, or school climate, as qualitative measures in identifying the CSI

schools. But measuring social characteristics, which are both within a person, and drawn out and cultivated by social interactions and environments, is fraught with technical difficulties (Duckworth & Yeager, 2015). If these soft skills, motivations, or personality skills (there are a host of overlapping terms applied to them) are as important as academic skills in determining success in life, then SEAs and schools will need guidance identifying the causal, social, and developmental origins of these skills and how they can best be measured and cultivated within school environments. ESSA leaves these complex research and validation riddles to state educational agencies, most of which are ill equipped to solve them. One exception is a consortium of nine large California districts which has field tested measures of four social-emotional skills as reported by the students themselves. Preliminary evidence is positive (West, 2016).

The Research Problem

The federal education law, ESSA, requires states to identify schools at highest risk for poor academic performance, but does not point to causal factors suggested by current research. What school risk factors, as closely supported by research as possible, are available to states? Which are demonstrably predictive of school academic performance?

Theory

Like NCLB, the theory underlying ESSA, with its emphasis on assessments, reporting and transparency, is a signaling theory. By making academic performance and gaps clear, stakeholders—parents, teachers, staff, policymakers—are left to define causes at the state, district, or school levels.

In addition to this signaling theory, ESSA also inherited a continuous improvement model from NCLB. A consortium of foundations, the Council of Chief State School Officers, advocate organizations, and some business interests like Standard & Poor's, acting through the Data Quality

Campaign (DQC), successfully advocated for data-driven improvement models in education. The DQC rated states based on their steps in building longitudinal data systems that, at their core, required individual identification numbers for all students so that students' performance on tests, attendance, and graduation could be followed across time. Theoretically, these longitudinal data systems, if combined with administrative data, could be used to evaluate curricula, programs, and interactions between individual students, teaching methods, environments and more. Under NCLB, the U.S. Department of Education, and the DQC, the focus turned to evaluating teachers using state assessment scores. This project wasn't able to overcome confounding factors and other technical limitations (Baker, et al., 2010; Rothstein, 2008). Teacher evaluations based on student test scores were removed from ESSA.

Could continuous improvement models be adapted to help cultivate better human beings? Data-intensive, continuous improvement models were developed by Walter Shewhart early in the last century, first to identify and then to eliminate the sources of defects in manufacturing. Later, W. Edwards Deming and Japanese businesses enlarged statistical controls into a management system for manufacturing. Fetuses, children, and how social interactions shape genetic expression involve extremely complicated interactions between billions of genes, environmental influences, and time. Early developmental events are often foundational to later capacities and health. By comparison, manufacturing computers in a data-intensive continuous improvement system seems simple.

But the evidence is growing that optimizing development through risk identification and improved interactions is theoretically possible. Caspi and Moffitt, working with two groups of researchers in separate projects, have produced strong evidence of genetic-environmental interactions. In one prospective, longitudinal study, individuals with a particular polymorphous short allele, exposed to stressful life events, were much more vulnerable to depression than individuals with the long, homogenous version of the same allele who had also been exposed to stressful life

events (Caspi, et al., 2003). In a second study, the researchers produced evidence that individuals with low activity in a certain polymorphism, if exposed to maltreatment in childhood, were at much greater risk for developing antisocial behavior (Kim-Cohen, J., et al., 2006). Using evolutionary theory, the Caspi and Moffitt studies of human adaptation to environmental exposures, suggest that yes, continuous improvement models, combined with child development models, could optimize the development and the capacities, academic and social, of many, maybe even most, children. Theoretically, reducing exposures to early developmental stressors could also extend life expectancies and reduce health care costs for large segments of the population, especially among the lower classes, since the same stressors that shape brain architecture in early development, also reallocate physical and immune system resources in ways that predict later disease and shorter lives (Johnson & Schoeni, 2015). Admittedly, the technical and political obstacles to building a preventive and optimizing longitudinal system are politically unlikely and enormously complex. Though the potential for public and private benefits is great, both NCLB and ESSA failed to translate the continuous improvement model from tools developed for manufacturing and business management into a model designed to optimize child development.

Hypotheses

Nine factors were identified by school district and SEA administrators as predictors of poor school academic performance. The workgroup expected that all nine factors, especially cumulative poverty, would predict school academic performance.

Data and Methods

Included Schools and Students

All public schools within the State, Title I and non-Title I, were included in the analysis. Virtual schools and stand-alone alternative schools were excluded. Virtual schools often have

students with joint membership in non-virtual schools and records that are non-comparable to regular brick-and-mortar schools, for example, in attendance data. Stand-alone alternative schools, as opposed to alternative programs within regular schools, serve students who have either previously dropped out or have been at-risk of dropping out, so they too would have risk factors which would not be comparable to regular schools. All public school students within the State, and all their students with records in either the pre-audited enrollment files or the end-of-year files reported to the SEA, were included. At the individual student level, cumulative measures included data from 2007 through 2015. In some cases, apparent errors in the enrollment and exit dates were corrected.

Dependent Variable

Academic Performance Index (API). Used as a school and district accountability measure, this index is based on the four performance levels of the State assessment. On the general assessments, the four levels have been split to create eight performance levels. Except for the lowest level, which is awarded no points, each increment is worth 100 points more than the level below it. Students scoring in the top level win 700 points for the school or district the student attends. The total points of all students taking a state assessment are then divided by the total number of assessed students. The intention is to reward schools and districts for advancing each student to the highest performance category possible and avoid the biased incentives of a single proficiency line. Normally, for accountability reporting, the API is calculated separately for each subject. For identifying the highest-risk schools, reading and math were aggregated into a single index. In developing this measure, the workgroup had only one year of assessment data available, but in the final determination of the highest-need schools in the fall of 2016, KSDE used two years of assessment data and will add a third year in the fall of 2017.

Approximately one percent of students, all with Individual Education Plans and severe disabilities, take a different assessment, the Dynamic Learning Map (DLM). With fewer items, the

DLM assessments can't be split from four into eight levels. To avoid creating an incentive to not give the DLM to qualified students, the DLMs were scored generously: at the first level, they were awarded 100 points; at the second, 300 points; the third 500 points; and at the fourth, 700.

Independent Variables

Chronic Absence. For the workgroup's review, I prepared the following variables, aggregated to the school level: 1) the yearly attendance rate, calculated as the total student days attending school divided by the total student days enrolled (sometimes referred to as days of membership); 2) the cumulative attendance rate, or the total days attended divided by the total days enrolled across all student years enrolled; 3) the absence rate, the total days absent divided by the total days in membership converted to a percentage; 4) the cumulative absence rate, the total days absent over the total days in membership across all years attended, converted to a percentage; 5) the percentage of students attending less than 140 days in the school year (140 days was the approximate inflection point where the long left tail of abnormally low attendance appeared to break into a normal distribution of days attended); and 6) the percentage of students who were absent ten or more days in the year. The workgroup chose the percentage of students missing ten or more days in the year as a measure of chronic absence. This differs from the federal Civil Rights Data Collection's more restrictive definition of chronic absence, which is fifteen or more days' absence during the school year (Office for Civil Rights, 2016).

Student Mobility. To offer a flexible range of measures of mobility, I identified several mutually exclusive student subgroups: 1) the percentage of students who changed schools within the school year; 2) the percentage of students who were new to the school—in their first year of attendance at the school—and attended less than 140 days; 3) the students who were stable—that is, continuing in the same school—but attended less than 140 days; 4) students who were stable during the year, but left the school at the end of the year for reasons other than matriculation or graduation;

5) students who were in their first year within the school, and attended at least 140 days or more; and 6) students who were continuing in the school and attended at least 140 days or more. My intention was to create ordered subgroups, from the least socially stable—those who change schools within the school year—to the most socially stable—those who were continuing in the same school and within a normal distribution of attendance. The work group chose the percentage of students who changed schools within the year.

Cumulative Poverty. The group rejected the traditional measure of poverty, the percentage of students in a school who are qualified for free or reduced lunch in a single school year. Under the traditional measure, schools that have a high proportion of reduced lunch students in temporary poverty may be equated to schools with students who have lived in poverty and poor neighborhoods since birth. A better measure would include measures of the developmental stressors associated with poverty, especially in early childhood (Brooks-Gunn & Markman, 2005; Luby, et al., 2013; Noble, Houston, Kan, & Sowell, 2012). But these measures weren't available. The closest substitute was a measure of cumulative poverty, the total number of years a school's students were in poverty divided by the total number of school years the students had attended state schools. I used student records from 2007 through 2015. Poverty was defined as having a value of one for each year a student qualified for free lunch, and 0.5 for each year the student qualified for reduced lunch. Years of family-paid lunch were valued at zero.

Percentage of Students with Disabilities (SwDs). Because some schools specialize in services to SwDs, they may have higher proportions of SwDs. They also will have a concomitant higher risk of lower academic performance. Using state assessments, which are based on college-ready standards, I separately examined subgroup performance on state assessments. Among NCLB subgroups, SwDs had the highest proportion of very low-scoring students. The contradiction of identifying students as having a disability that impairs their academic performance, and then

evaluating the performance of schools based on the academic performance of SwDs, has no current remedy in the ESSA. Nor does the SEA have measures that might alleviate the contradiction. The SEA can calculate the length of time SwDs continue as SwDs, but without beginning measures that control for severity of impairment at the time of disability identification, nor long-term, individual goals, it is not possible to measure school or district influences on SwDs outcomes at the state level.

Percentage of English Learners (ELs). ESSA emphasizes EL performance over that of other subgroups. Schools are required to include an EL measure in the academic measures used to identify CSI schools, and in whatever design the SEA chooses to differentiate between schools. ELs are also a subgroup, and as such, if their performance is persistently low, they can trigger the identification of the subgroup for targeted assistance. In state-level simulations, subgroups of ELs had the second-highest risk of lower academic performance among traditional subgroups. Like the SwDs, there is a Catch-22 in current use of EL academic performance being used as school and district performance measures. EL students are identified as such because they don't know enough English to function normally in an English-speaking classroom, yet school and district performance is evaluated based on the academic performance of ELs. ESSA does open the door to a new remedy to this contradiction—a measure of progress-to-English proficiency.

The workgroup decided to include the percentage of EL students as a risk factor.

Rate of Suspensions and Expulsions. I aggregated individual student in-school suspensions, out-of-school suspensions, and expulsions. This total, the number of discipline events, was divided by the total number of students then converted to a percentage. There are other, non-suspension and non-expulsion disciplining events, and truancies, but the work group chose not to include them.

Percentage of Migrant Students. The count of migrant students over the count of total students was also converted to a percentage and included as a risk factor.

Demographic Distance between Teachers and Students. The workgroup identified segregation, by ethnicity and poverty, as a risk factor that predicted poor school performance. One problem posed by segregation is the proportionate difference in ethnicity and class between teachers and students. Teachers are typically White, middle-class females. Students in large urban districts, or districts with meat-processing industries, can have large proportions of Hispanic or African-American students from low-income families. A representative from a large urban district said that middle-class White female teachers are often afraid of their male minority students, which causes problems her district had recognized and was working to alleviate.

According to the Kansas' Licensed Personnel Summary Report and its unaudited enrollment reports in the 2013-2014 school year, 97.1 percent of the state's teachers are White and 75.1 percent are female. But Hispanic and African-American students are 26 percent of the State's student population, with high percentages in some districts. If students learn more through models with whom they share gender and ethnicity, then minority males are especially lacking access to models with whom they can identify in schools.

In light of this discussion, the workgroup asked for a measure of the demographic distance between students and teachers. I developed the following experimental measure:

In each school, I grouped and counted the teachers and students into ten gender-ethnic groups: African-American females, African-American males, American Indian females, American Indian males, Asian or Pacific Islander females, Asian or Pacific Islander males, Hispanic females, Hispanic males, White females, and White males. If a student claimed more than one group—for example, that she is female, Hispanic, and White, she was counted in both the Hispanic female group and the White female group. To be included in the calculation, there had to be at least 30 students in a gender-ethnic subgroup within the school.

For all student gender-ethnic groups with at least 30 students in a school, I calculated the following percentage: ((count of teacher in gender-ethnic group A / total number of teachers in the school) / (count of students in gender-ethnic group A / total number of students in the school)) * 100. A perfectly congruent teacher-student match would produce a score of 100 for that group. So an absolute difference of each gender-ethnic percentage from 100 would be a measure of demographic distance between teachers and students for that group.

I will spare the reader the details of the calculation because the measure proved problematic in several ways. In my formula, gender-ethnic subgroups that are theoretically advantaged by having a similarly composed teaching staff were treated the same as those who were theoretically disadvantaged by being under-matched. My measure of demographic distance is also very superficial. It captures no actual classroom interactions between students and teachers (Dee, 2004). Teachers of whatever gender or race who have achieved impartiality are assumed to be biased. Small numbers of minority students—those with less than thirty members—are also ignored.

Percentage of New Teachers. Comparative studies of international education systems have sometimes pointed to high teacher turnover in the United States as an important factor in comparatively lower student performance (Auguste, et al., 2010). Domestic studies also note the high costs of teacher turnover and use teacher mover and leaver rates as measures (Kukla-Acevedo, 2009; Ingersoll, 2001). I prepared the following building-level counts for the workgroup: new teachers, teachers who had been in the same school for two or more years, teachers who rotated between schools or districts, teachers who left teaching after a year or less, teachers who had left teaching or left the state, teachers who had left a school or district for a year or more and then returned, teachers who had moved from one school to another within the same district, teachers who changed schools within a district after taking a year or more off, teachers who changed districts, and teachers who changed districts after taking a year or more off. The group chose to measure a

five-year average of new teachers. They thought the five-year average would better identify systemic teacher turnover rather than identifying schools with a single year of high teacher turnover. In smaller schools, a single-year with a high number of retirements and new teachers could miss-identify a school as high-risk.

Below are the statistics and bivariate correlations between the nine variables (Tables 1 and 2).

Table 1

Means, Standard Errors, and Standard Deviations of the 9 Variables

	N	Median	Mean	SE	SD
Academic Performance Index	1,263	300.5	303.0	1.83	64.9
% absent 10 or more days	1,283	25.7	26.3	0.38	13.6
% changing schools within the year (mobile)	1,283	3.6	4.5	0.11	3.8
Cumulative poverty rate	1,283	43.6	44.8	0.61	21.8
% Students with Disabilities	1,295	13.8	14.5	0.19	6.8
% English Language Learners	1,283	2.2	9.6	0.44	15.9
% Suspensions and Expulsions	1,295	0	1.7	0.11	3.8
% Migrant Students	1,295	0.17	2.1	0.12	4.8
Demographic Distance Index	1,256	73.4	71.5	1.16	41.2
% New Teachers (5-year mean)	1,246	13.2	15.3	0.27	9.5

Table 2

Bivariate Correlations Between 9 School Risk Factors

		1	2	3	4	5	6	7	8	9
1	% absent 10 or more days	1								
2	% changing schools within the year	.220**	1							
3	Cumulative poverty rate	.250**	.493**	1						
4	% Students with Disabilities	.102**	.281**	.234**	1					
5	% English Language Learners	.064*	.138**	.568**	-.198**	1				
6	% Suspensions and Expulsions	.191**	.342**	.334**	.195**	.152**	1			
7	% Migrant Students	.070*	.000	.288**	-.066*	.598**	-.009	1		
8	Demographic Distance Index	-.083**	.231**	.504**	-.084**	.607**	.162**	.258**	1	
9	% New Teachers (5-year mean)	.009	.080*	.139**	.032	.097**	.122**	-.005	.153**	1

* $p < .05$ (2-tailed), ** $p < .01$ (2-tailed)

Statistical Model

We want to direct attention to causes effecting long-term student outcomes, not to correlates of causes, which, when they are the targets of intervention, may or may not improve student outcomes. Testing and comparing the relative influence that school staff identify as predictors of academic achievement is a good first step toward identifying causes. In SPSS, I used a simple linear regression with all nine predictors entered into the model at the same time. The objective was to discover the power of the model to explain school performance, and to quantify the relative strength of the independent variables to predict school performance. The factors that actually depress school academic performance could then be used to identify the highest-risk CSI schools.

To verify the model, I also tested the data with a different statistical method. The method I used has various names—relative importance analysis, conjoint analysis, Shapley value regression or dominance analysis. Shapley used it to identify optimum outcomes in game theory. Advertisers use it to identify the most important characteristics of a product by repeatedly tinkering with the features of a product and then measuring consumer reactions. I used the relative importance technique within SPSS to confirm the variables identified by regression and their relative influence over school performance.

Finally, I also used the original regression model to generate a second set of school performance measures. In examining the individual school risk factors and the schools each identified, it was clear that the model would identify mostly high-poverty urban schools as CSI schools. But SEAs have an obligation to cultivate the best possible student outcomes for students in all schools. In a continuous improvement system, an SEA will want to provide constructive feedback to all schools so that even schools with comparatively advantaged students can identify their weaknesses and make improvements. Thus, controlling for all the risk factors identified by the workgroup and Title I status, the first statistical model predicted API scores. I then compared the

predicted API scores to the actual API scores. This method identified a diverse set of schools—not just high-poverty urban schools—as performing lower than expected or better than expected, given their risks and populations.

Results

The adjusted r-squared was 0.580 and the standard error was 41.2, so the school risk factors picked by the workgroup explained more than half of the variance in school API scores.

Did the data need corrections? In the correlations table above, we see some relatively high correlations. Collinearity between predicting variables can distort measures of their relative influence. Converting the predictors to z-scores removes this threat of distortion so all of the predictors were standardized to z-scales.

One more correction was made. Sometimes outliers—schools with extreme values—can distort coefficients. I examined scatterplots of the leverage individual cases had over the coefficients. I removed three extreme outliers. The model's accuracy improved a tiny amount to an adjusted r-squared of 0.592 and a standard error of 40.5. The standardized coefficients did not change significantly. ESSA is asking us to identify the schools with the extreme needs, so we do not want to unnecessarily remove any schools from the analysis. Since removing the extreme outliers only made tiny improvements in the model, no more outliers were removed.

Some variables acted as expected (see Table 3 below). Cumulative poverty is the second strongest predictor. For every percentage increase in students' cumulative poverty at the school level, the model predicts a school's API will drop about a half-point. For every percentage increase in the EL population, a school's API score is predicted to drop a little more than a quarter point. Each

percentage increase in suspension and expulsion, and in the percentage absent ten or more days, are also associated with nearly a quarter-point drop in a school's API score.

Table 3

Summary of the Linear Regression After 3 Outliers were Removed

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	302.917	1.190		254.534	0.000
% absent 10 or more days	-14.470	1.287	-0.228	-11.245	0.000
% mobile (moved in school year)	-1.672	1.483	-0.026	-1.127	0.260
students cumulative years in poverty	-29.663	1.968	-0.471	-15.070	0.000
% who are currently SwDs	1.734	1.839	0.021	0.943	0.346
% who are currently ELs	-17.467	2.159	-0.281	-8.091	0.000
% suspended and expelled	-18.229	1.598	-0.240	-11.411	0.000
% migrant students	-1.676	1.584	-0.026	-1.058	0.290
demographic distance between teachers and students	35.792	1.612	0.573	22.200	0.000
% new teachers, 5 year-average	-0.891	1.270	-0.014	-0.702	0.483

Predictors have been standardized to the z-scale.

Four of the variables aren't significant predictors—the percentage of new teachers, the percentage of migrant students, the percentage who are Students with Disabilities, and the percentage who are mobile. The workgroup expected all four of these to predict significant declines in a school's API. They didn't.

That the SwDs did not show up as a predictor is especially telling. As noted above, in a separate analysis, using 2015 data and the API in an examination of which traditional subgroups would perform poorly as measured by a college-ready assessment, the SwD subgroup had the highest proportion of low-scoring subgroups. Why didn't the percentage of Students with Disabilities show up as a school-level risk factor? In large part, it was because cumulative poverty has absorbed the variance that, in the absence of cumulative poverty, would have been associated with the SwD subgroup. When I removed cumulative poverty from the regression above, the percentage of SwDs became a significant predictor of school academic performance. So did student mobility. Chronic absence and the percentage of ELs became stronger predictors. This suggests that chronic poverty is a developmental driver of disability status and a preceding driver of student mobility. This result fits with the emerging developmental knowledge of poverty's influence on the developing brain and its association with a powerful set of parenting risks (Brooks-Gunn, & Markman, 2005; Luby, et al., 2013; Noble, Houston, Kan, & Sowell, 2012).

This result may have been slightly influenced by the way the Dynamic Learning Maps (DLM) assessment was scored. Designed for the one-percent of SwDs who are most impaired, the DLM is a different assessment than the general college-ready assessment. It was generously scored to 4 performance levels. Collapsing the more precise scale of individual scores more grossly to four levels with an upward bias may have blunted a signal from the Student with Disabilities.

The biggest surprise here is the predictive strength of the experimental variable measuring the demographic distance, in gender and ethnicity, between teachers and students. Many studies have

shown mixed results on similar measures, in part because very general models like the one we're using can't separate out the biases created by the class and ethnic sorting and segregation that takes place between neighborhoods and then again, within schools and classrooms. We have no measures of specific interaction effects between teachers and students (Dee, 2004).

But demographic distance is the *strongest* predictor in the model and it is *positive*. Why? One clue is that the student subgroups theoretically advantaged by being over-represented, usually White females, were treated in the same way as the demographically disadvantaged groups. The over-representation of the White female teachers contributed more to the demographic distance measure than the under-representation of the White males. We may be measuring the advantage of the majority group rather than the disadvantage of minority groups.

Does the application of the relative importance technique give us the same results?

Applying the relative importance technique to the same data, I confirmed the same factors as predictors, but the relative importance of the factors shifted slightly. Cumulative poverty is affirmed as the most significant predictor while the demographic distance between teachers and students becomes the second most important predictor. Chronic absence becomes the third most influential variable and the percentage of students who are ELs the fifth most influential.

Table 4

Comparison of the Influence of Predictors under 2 Statistical Techniques

	relative importance partitioning:		linear regression:	
	portion of variance explained	order of importance	order of importance	standardized coefficients
% absent 10 or more days	0.117	3	5	-0.228
% mobile (moved in school year)	0.028			
students cumulative years in poverty	0.158	1	2	-0.471
% who are currently SwDs	0.005			
% who are currently ELs	0.056	5	3	-0.281
% suspended and expelled	0.087	4	4	-0.240
% migrant students	0.018			
demographic distance between teachers and students	0.122	2	1	0.573
% new teachers, 5 year-average	0.001			

Predictors are standardized to z-scale. Relative importance measures are not scaled to 100%.

How can these two slightly different results be used to identify the highest-risk CSI schools? The rankings largely agree and neither is known to be superior to the other. The workgroup asked that the risk factors that were identified as depressing school performance be used in proportion to their influence. Thus, I averaged the proportionate influences from both models. Chronic absence had a weight of 0.2442, the percentage of ELs, a weight of 0.1615, cumulative poverty a weight of 0.4509, and the percentage suspended or expelled a weight of 0.1362. ESSA requires that the identification process give a greater weight to academic performance than to qualitative measures, so an overall risk index was calculated by applying a 60-percent weight to each school's API and a 40 percent weight to the proportioned risk factors. The five-percent of Title I schools with the highest overall risk index were identified as the CSI schools.

Which of these new measures can be applied to subgroups?

ESSA requires that the measures used to identify CSI schools eventually be applied to measure subgroups' performance. Some school risk factors, like high percentages of ELs, SwDs, and migrant students, are school-level descriptors of specific subgroups, not performance measures appropriate for all subgroups, so they will not meet this ESSA requirement.

But some measures could be applied to all subgroups: the cumulative poverty rate, the percentage suspended and expelled, the percentage chronically absent, and the percentage mobile. Would reporting these measures by subgroup be a constructive addition to traditional outcome measures like assessment means and growth, graduation, attendance, and participation? Publically reported, they might help the public acquire a more realistic understanding of how chronic poverty is an important driver of academic gaps and disability status. But their greater value may be as independent variables. The influence of cumulative poverty, the percentage of ELs, and other

factors over which schools have little control, can be measured. Could the variance between schools that remains after controlling for these factors be a useful measure of school performance?

Two for One: Identifying Lower-Than-Expected Performance

The predicted API of each school was a product of linear regression when the nine risk factors, and Title I status, were used as independent variables. How can we know if the results are due to random variation around the mean, or to differences in unmeasured influences excluded from the model, or the results of factors under the control of staff, the local school board, or policymakers controlling funding? And how does this second model inform the SEA and the field?

We can't yet attribute the differences to causes. We can only ask more questions. If the difference between the actual API and the predicted API is strongly positive, and sustained over time, this might suggest that the school is systematically cultivating better student outcomes than we would expect given the school's population and risk factors. The other extreme, schools with an actual API that is consistently far below the predicted API given a school's risk factors, would suggest that something associated with the school, students, or neighborhood might be systematically suppressing school performance. The shape of the distribution may be informative—long tails on the high or low performance side of the distribution more strongly insistent that something is unusual about the schools at the extremes. Do observations with consistent trajectories—sustained improvement, decline, or maintenance at high or low levels across three or more years—suggest that the pattern is not due to natural variation around the mean? Our current model is only explaining about 41 percent of schools variation in academic performance. The discovery of additional predictors could change the interpretation of schools' relative contribution to students' academic performance.

Nevertheless, adding predicted APIs and residuals to the model efficiently adds a way of identifying high and low performing schools among schools with diverse populations. The nine-

factor highest-need CSI model strongly tends to identify urban schools with very high levels of cumulative poverty. This second model, comparing actual school academic performance to predicted school performance, has the potential to identify high and low performing schools across diverse communities—from low poverty, suburban, to high poverty and urban or rural, or any combination in between.

Discussion

The method described here for identifying the five percent of Title I CSI schools has some strengths other states may find useful. It was developed in consultation with district staff, who identified factors suspected to be most predictive of school risk. It validated some factors and quantified their relative contribution to suppressed school academic performance. It used the validated results to identify the highest-risk schools as CSI schools. It also efficiently produced a secondary identification of schools that were not among the highest-risk schools, but were performing below expectations when risk factors were used as control variables.

But the method developed here also has a number of weaknesses that perhaps other states can address. Academic growth was not incorporated as a dependent variable despite being identified as a required performance measure by ESSA. For example, the dependent variable could have incorporated both a performance index and a growth index based on the Student Growth Percentiles developed by Damian Betebenner.

The model may extend ESSA's overemphasis of the academic performance of ELs. ESSA identifies EL proficiency as one of the five required performance measures. ELs are also a subgroup. When subgroups perform poorly, ESSA requires that they be identified for targeted assistance. The model here also identified the percentage of ELs as a predictor of depressed school performance. Compared to other subgroups, ELs have been given greater importance by both ESSA

and the model described here. Depending on a state's accountability model, this could lead schools and districts to put a greater emphasis on EL performance than on the performance of other subgroups.

Implications for Policy

While this experiment is only a case study from one SEA, it does illustrate the lack of infrastructure to support a continuous improvement model and the lack of measures that are causally linked to depressed school performance. Under NCLB the federal and state governments began building the infrastructure to support a continuous improvement model in education, but it was miss-directed toward testing a single hypothesis, that poor teaching was the cause of poor academic performance. ESSA also does not cultivate a science of school improvement, or experiments in the practices suggested by new developmental knowledge, or by international comparisons of teacher selection and training. It implicitly expects that states, districts, and schools will successfully conduct experiments that accurately identify causes and best practices.

As this paper is being written, a new administration, Congress, and the U.S. Department of Education should consider how to help states build the infrastructure to support true continuous improvement, preferably with designs which will prevent partisan political goals from subsuming a broad set of goals well-founded in research. NCLB, under both Republican and Democratic administrations, demonstrated the futility of ideologically driven reform. Federal support for the regional education labs and for research in schools of education could be reconfigured to support continuous improvement models. The brief literature review above suggests experiments in better teacher selection, training, and retention, and in improving the quality of child-rearing and the social environments shaping development. If the day-to-day administrative data within schools could be re-designed to support strong evaluations of programs, curricula, teacher training, social inclusion,

and more, it would greatly facilitate large-scale evaluation and continuous improvement. If early childhood health providers and early childhood records were included in these experiments, some specific developmental vulnerabilities of children could be identified and the dangers they pose avoided or reduced. In the current environment where the internet and social media have eroded privacy, this suggestion may seem dangerous. But many parents, maybe most, will want to know the specific vulnerabilities of their children, and how their developmental trajectories can be optimized. There appears to be a strong public interest in realizing the same.

A final policy implication comes out of the finding that cumulative poverty supersedes disability as a school risk factor. Disproportionality in disability rates by ethnic groups is sometimes taken as evidence of ethnic-based discrimination. But if poverty and stress are more pervasive among ethnic minorities, especially during early childhood, we should expect disproportionality in disability rates by ethnicity. This finding suggests that interventions that reduce the developmental risks associated with poverty may be more effective in reducing disproportionality in disability rates than punitive actions against schools and districts.

Conclusions

Under the emerging federal requirements of ESSA, one SEA, and representatives from districts, identified and tested nine school-level risk factors. Four were predictive of school-level academic performance: cumulative poverty, the percentage of students who were English Learners, the rate of suspension and expulsion, and the percentage of students chronically absent. The percentage of SwDs was *not* a predictor when student cumulative poverty was included in the model but was when cumulative poverty was removed from the model. Contrary to expectations, an experimental measure of the demographic distance between teachers and students strongly

predicted positive school performance but for reasons that were not clear. While the SEA included district representatives in the identification of risk factors and then tested their validity, this experiment also illustrated states' limitations in identifying and testing causal models. Although ESSA is less punitive than NCLB, it continues to emphasize assessment scores over causal models. ESSA missed an opportunity to cultivate the supporting science, technical expertise, and longitudinal data to develop a science of school improvement, and to institute continuous improvement models in education.

References

Auguste, B., Kihn, P. and Miller, M. (2010). Closing the talent gap: Attracting and retaining top-third graduates to careers in teaching. Washington, DC: McKinsey. http://www.mckinsey.com/App_Media/Reports/SSO/closing_the_talent_gap_september_2010.pdf

Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., …Shepard, L.A. (2010). Problems with the use of student test scores to evaluate teachers. Retrieved from http://www.epi.org/publication/bp278/ .

Brooks-Gunn, J. & Markman, L. (2005). The contribution of parenting to ethnic and racial gaps in school readiness. *The Future of Children*, 15, 1, Spring, 139-168.

Caspi, A., Sugden, K., Moffitt, T.E., Taylor, A., Craig, I.W., Harrington, H., McClay, J., Mill, J., Martin, J., Braithwaite, A., & Poulton, R. (2003). Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science,* 301, 386-389. Retrieved from http://usdbiology.com/cliff/Courses/Advanced%20Seminars%20in%20Neuroendocrinology/Susceptibility%20and%20Resilience/Caspi%2003%20Sci%20Stress%20Depression%205-HTTLPR%20s.pdf .

Dee, T.S. (2004). A teacher like me: Does race, ethnicity or gender matter? *American Economic Review*, 95(2), 158-165.

Duckworth, A.L., Yeager, D.S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher,* 44, 4, 237-251.

Heckman, J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science, 312* (5782), 1900-1902.

Hochbein, C. (2012). Relegation and reversion: Longitudinal analysis of school turnaround and decline. *Journal of Education for Students Placed at Risk,* 17: 92-107.

Ingersoll, R.M. (2001). Teacher turnover, teacher shortages, and the organization of schools. University of Washington: Center for the Study of Teaching and Policy. Retrieved from http://depts.washington.edu/ctpmail/PDFs/Turnover-Ing-01-2001.pdf.

Johnson, R.C. & Schoeni, R.F. (2007). Early-life origins of adult disease: The significance of poor infant health and childhood poverty. Retrieved from http://www.researchgate.net/publication/228640445 .

Kautz, T., Heckman, J.L., Diris, R., ter Weel, B., & Borghans, L. (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. OECD Education Working Papers. Retrieved from http://www.oecd-ilibrary.org/education/fostering-and-measuring-skills_5jxsr7vr78f7-en .

Kim-Cohen, J., Caspi, A., Taylor, A., Williams, B., Newcombe, R., Craig, I.W., & Moffitt, T.E. (2006). MAOA, maltreatment, and gene-environment interaction predicting children's mental health: New evidence and a meta-analysis. *Molecular Psychiatry,* 11, 903-913. Retrieved from http://www.nature.com/mp/journal/v11/n10/pdf/4001851a.pdf .

Klute, M., Cherasaro, T., & Apthorp, H. (2016). Summary of research on the association between state interventions in chronically low-performing schools and student achievement (REL 2016-138). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Eduation Evaluation and Regional Assistance, Regional Education Laboratory Central. Retrieved from http://ies.ed.gov/ncee/edlabs.

Kukla-Acevedo, S. (2009). Leavers, movers, and stayers: The role of workplace conditions in teacher mobility decisions. *The Journal of Educational Research,* 102, 6.

Kutash, J., Nico, E., Gorin, E., Rahmatullah, S., & Tallant, K. (2010). The school turnaround field guide. FSG Social Impact Advisors. Retrieved from

http://www.wallacefoundation.org/knowledge-center/Documents/The-School-Turnaround-Field-Guide.pdf.

Lerner, R.M., (2006). Developmental science, developmental systems, and contemporary theories of human development. In W. Damon & R.M. Learner, (Eds.), *Handbook of Child Psychology*, (pp. 1-17), Hoboken, New Jersey: John Wile & Sons.

Luby, J., Belden, A., Botteron, K., Marrus, N., Harms, M.P., Babb, C., Nishino, T., & Barch, D., (2013). The effects of poverty on childhood brain development. *JAMA Pediatrics*, 169 (10), 938-946.

Noble, K.G., Houston, S.M., Kan, E., & Sowell, E.R., (2012). Neural correlates of socioeconomic status in the developing human brain. *Developmental Science*, 15, 4, 516-527.

Office for Civil Rights (2016). 2013-2014 civil rights data collection: A first look. Retrieved from http://www2.ed.gov/about/offices/list/ocr/docs/2013-14-first-look.pdf.

Rothstein, J. (2008). Teacher quality in educational production: Tracking, decay, and student achievement. National Bureau of Economic Research Working Paper No. 14442.

Sawchuk, S. (2011). What studies say about teacher effectiveness. Education Writers Association. Retrieved from http://www.ewa.org/topic-essa .

Sahlberg, P. (2011). Finnish lessons: What can the world learn from educational change in Finland? New York: Teachers College Press.

Trujillo, T., & Rivera, M. (2016). Review of measuring school turnaround success. Retrieved from http://nepc.colorado.edu/files/reviews/TTR%20Trujillo%20Turnaround_0.pdf

Tucker, M. (Ed.) (2011). *Surpassing Shanghai: An agenda for American education built on the world's leading systems*. Cambridge, MA: Harvard Education Press.

West, M.R. (2016). Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts. *Economic Studies at Brookings, Evidence Speaks Reports, 1, 13.*