GREAT LAKES CENTER
FOR EDUCATION RESEARCH & PRACTICE

DOCUMENT REVIEWED:	"Stuck Schools"
AUTHOR:	Natasha Ushomirsky and Daria Hall
PUBLISHER/THINK TANK:	Education Trust
DOCUMENT RELEASE DATE:	February 2010
REVIEW DATE:	April 7, 2010
REVIEWER:	Jaekyung Lee, University of Buffalo, SUNY
E-MAIL ADDRESS:	jl224@buffalo.edu
PHONE NUMBER:	(716) 645-1132

Summary of Review

The Education Trust research report *Stuck Schools* suggests a framework for identifying chronically low-performing schools in need of turnaround. The study uses Maryland and Indiana to show that some low-performing schools make progress while others remain stagnant. The report has four serious problems of reliability and validity, however. First, the norm-referenced methodology guarantees "failed" schools independent of any true performance or improvement level by the school. Second, the report's reliance on state assessment data is misleading, and some schools' reported growth may be an artifact of regression to the mean and ceiling effects as well as instructional and testing practices. Third, the use of a linear growth model is questionable, since schools may not follow a strictly linear pattern of improvement. Fourth, the label of "stuck" becomes problematic given that there is no research-based guidance on how to improve schools other than vague prescriptions. In conclusion, the report's methods are so simplistic, arbitrary and ill-fitting with its own assumptions that it is more harmful to sound policymaking than helpful. There remains an outstanding question of how to help struggling schools after identification, but we need to know first whether the identification is based on reliable and valid measures, and if so, what school factors account for these differences.

Review

I. INTRODUCTION

Stuck Schools, a research report from the Education Trust, written by Natasha Ushomirsky and Daria Hall, is aimed at providing a framework for identifying chronically low-performing schools that are arguably in need of school-turnaround interventions.[1] The study selects two states, Maryland and Indiana, as showcase examples to demonstrate how to use the currently available state assessment database to identify such schools.

The authors' chosen classification is based on two variables, performance (status) and improvement (growth). First, they sort schools into three categories—high-performing, average-performing and low-performing—based on "status": the baseline status of achievement, or how well students perform, on average, over the first three years of the five-year period under study. (Slightly different study periods are used for the two states.) The report then classifies the same schools into another three categories based on "growth": how much the schools improve their proficiency rates over the five-year study period.

Next, the report cross-classifies schools according to these two dimensions and examines how many schools are simultaneously low-performing and low-improving. Schools that fall within cell D in Figure 1 are designated "stuck" schools. The study also defines and identifies "chronically low-performing schools" as those where performance in the final three years of the study period fell consistently below the bottom 5% bar; several "stuck" schools are classified as chronically low-performing as well.

This study's focus on the schools that are both low-performing and low-improving is best understood in the context of a recent policy paradigm shift in the American school-accountability system, from a narrow focus school performance to a dual focus on both performance and improvement. The Education Trust study is designed to inform the Obama Administration's new school accountability policies, which concentrate on state interventions in chronically low-performing schools that do not show signs of improvement.

As the report acknowledges, its analysis does not fully align with the current accountability and school identification policies in place through NCLB. While NCLB relies pri-

		Performance (Status)	
		High	Low
Improvement	High	High on both Performance and Improvement (A)	High on Improvement and Low on Performance (C)
(Growth)	Low	Low on Improvement and High on Performance (B)	Low on Both Performance and Improvement (D)

Figure 1. Classification of schools by the level of performance (status) and improvement (growth)

marily on the status model of school performance evaluation, this report builds upon the model of combined status and growth. It is similar in intent to the turnaround provisions of the administration's Race to the Top program.

Moreover, while NCLB takes a criterion-referenced (standards-based) approach, with predetermined performance targets and timelines for all schools, the study takes a norm-referenced approach (i.e., a comparison of relative performance rankings) to identify schools needing improvement.

II. FINDINGS AND CONCLUSIONS OF THE REPORT

The study examines trends in overall performance and improvement over time. Specifically, it raises the following four related research questions: (1) What did performance look like several years ago? (2) How big were the annual gains at high-improving schools? (3) How about at low-improving ones? (4) Among the lowest-performing schools, how many remained stuck, how many made extraordinary gains, and how many fell somewhere in between?

For the first question, the study finds enormous variations among schools in baseline performance. The baseline (2005-07) reading proficiency rate in Maryland (using the state's own accountability ratings) is 79% on average, but it ranges from 27% to 99% among Maryland's 1,066 schools serving any combination of grades 3-8. The baseline (2004-06) reading proficiency rate in Indiana is 73% on average. But it ranges from 26% to 96% among Indiana's 1,477 schools serving any combination of grades 3-8. The study then identifies the bottom 25% of schools in both states. In those lowest-performing schools, about 58% of students meet the proficiency standard in Maryland and 57% in Indiana.

For the second and third questions, the study reports different patterns of school academic improvement in the two states. In Maryland, the study shows that schools at all different levels of performance generally made progress, with low-performing schools making the biggest gains. In Indiana, the study shows that the average rate of reading proficiency remained stagnant from 2004 to 2008, while there are few variations in gains between high-performing and low-performing schools.

For the fourth question, the study reports fewer stuck or chronically low-performing schools in Maryland (4% of the state's elementary and middle schools) than in Indiana (15% of the state's elementary and middle schools). For Maryland, there were 44 stuck or chronically low-performing schools in total, with 22 of those identified for reading and 31 for math (9 schools fell into both categories). For Indiana, there were 228 stuck or chronically low-performing schools, 155 identified for reading and 147 for math (74 schools fell into both categories). The key finding of this report is that among initially low-performing schools, "some schools are improving; others are stuck" (p. 1). Further, the authors emphasize that some schools persistently produced worse results than 95% of schools in their states, even as they managed to make some gains. In conclusion, they recommend differentiated approaches: benchmarking of practices from low-performing schools that made significant progress, and targeted support and interventions for low-performing schools that have made little, no, or negative improvement.

III. RATIONALES SUPPORTING THE FINDINGS AND CONCLUSIONS

The study grows out of the new federal policy movement and the urgent need for empirical research for policy guidance. The authors note:

> In recent months, the federal government has put billions of dollars on the table with a demand for real action in turning around our country's lowest performing schools. At the same time, federal and state leaders are considering future directions for education policy. In this context, understanding recent patterns of school improvement is particularly important (p. 1).

At the bottom line, the new policy supports the rationale of the study to separate two different kinds of low-performing schools, (1) schools that are chronically low-performing without any indication of major improvement (cell D in Figure 1), and (2) schools that are low-performing initially at the baseline but show great improvement over the course of five years (cell C in Fig 1).

The study is based on the premise that there are good schools and bad schools in terms of academic performance and improvement, and that we can and must identify the bad schools for the sake of children and the society. In lieu of "bad," the report uses terms such as "stuck."

There is, of course, an underlying logic to such categorizations, but in order to identify those stuck schools and turn them around, the measures and methods used for identification must be valid, reliable, and fair. Unfortunately, the study does not address any of the key psychometric and statistical issues that may threaten the validity of its findings and conclusions. This decision may be understood in light of the report's target audience of the policy community rather than the research community. However, this serious omission of important scientific and technical issues can undermine the very rationale and purpose of the study. This report offers policy guidance without engagement in important methodological issues. The most important problems concerning validity and reliability are discussed below, in the review of research methods and findings.

IV. THE REPORT'S USE OF RESEARCH LITERATURE

This report is identified as the first of a four-part "Stuck Schools Series" intended to "provide educators, policymakers, and the public with a framework for using data to identify schools and districts that are making academic progress or that desperately need help" (p. 1). Notwithstanding this goal, the report's conceptual and analytical framework does not build on any established theory or prior research, and it fails to capitalize on recent advances in value-added growth model experiments in several states.[2] The study's idea of differentiating schools by two separate dimensions (performance and improvement) is not new, but it doesn't learn from earlier efforts. The article does not provide any references to the extensive prior research on this topic. The current (NCLB-linked) school accountability systems in most states rely heavily on the indicators of schools' academic status rather than their progress, although they may combine the two pieces of information for a final decision. Previous studies have found that the relationship between the status and progress of school achievement is generally tenuous.[3] The Education Trust report shows this same pattern; among low-performing schools, the authors observe both fast and slow rates of improvement. By using lessons from earlier research, the report's authors could have examined whether such differentiation of school improvement levels is reliable and valid.

The scope and depth of data analysis in the report is highly limited. This study is selective in the sense that it focuses attention on

that one particular category of schools that are low-performing and low-improving at the same time. The study further attempts to refine that category by identifying higher-risk schools that are not only stuck but also chronically low-performing. However, as discussed below this further classification is highly arbitrary, and the report's operational definitions are not based on research literature. What, for instance, is the rationale for targeting the bottom 25% or the bottom 5%? Finally, although the report attempts to discriminate between "stuck" schools and "chronically low-performing" schools, the underlying basis for finding schools in both low performance and low improvement is similar.

V. REVIEW OF THE REPORT'S METHODS

The report's designation of some schools as "stuck" suggests that the schools themselves—rather than the structural and resource issues within which those schools carry on—should be blamed and "turned around." That is the authors' clear position, which would explain why they eschewed a more neutral term such as "struggling" schools. No matter the term, however, the key issue is how the study identifies such schools. Although the report never makes its methods explicit, the description that is provided in the appendix shows that it uses a linear regression method to estimate the slope of regression (i.e., annual growth rate) and used that estimate of a regression coefficient to classify schools into three levels of improvement.

This approach immediately raises questions. What if schools had showed curvilinear pattern of growth rather than linear pattern of growth? The assumption of linear growth means that schools make an equal increment of proficiency gains every year (e.g., 3 point gain per year over the five-year period = 3 times 5 = 15-point gain total). The assumption may be reasonable given the limited number of years available for tracking school performance trends, and it actually may fit most cases. However, the imposition of this particular growth model across all schools has the risk of misestimating growth rates and dismissing other possible patterns of growth.

The report also does not consider (or simply does not report) the reliability of estimating growth rate through a time-series regression method. To explore this issue, I conducted a re-analysis of the same Maryland data and generated regression coefficients with standard errors and indicators of statistical significance. The exercise reveals that many of the schools classified as high-improving or low-improving by this study are not really showing a consistent "linear" pattern of improvement. This calls into question the reliability of the report's measures of school aggregate performance trend reliable.

In Figure 2, I illustrate this issue with the same data from schools in Maryland 2005-09.[4] The Figure focuses on just one school in Maryland, showing tremendous instability. The school has a generally upward performance trend until 2008, followed by an unexpected large decline in 2009. The linear regression method (the approach that the Education Trust study used) would identify the slope of regression, giving an annual growth rate of -1.19. As shown by the direction of fitted regression line in Figure 2, the school's performance trend looks negative despite earlier positive gains. However, the standard error of the regression coefficient (information that the Education Trust study did not consider or report) is 4.45, and the growth rate is not statistically significant: the 95% confidence interval of this slope ranges from a low of -10 points to a high of

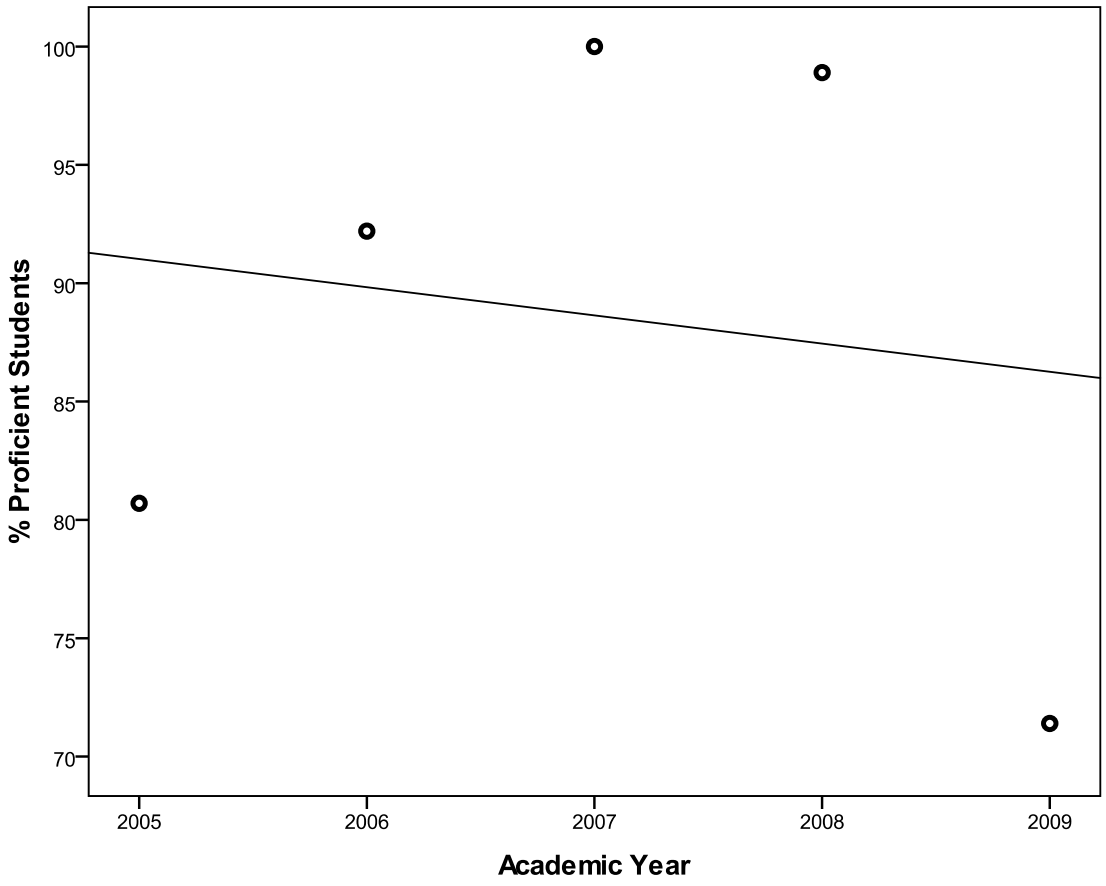

Figure 2. One Maryland school's reading proficiency rate trend during 2005-2009 (the line is estimated through a simple linear regression of the percent-proficient variable on the academic-year variable)

+8 points. In other words, the linear regression model does not fit this particular school's data, and there is no systematic "linear" pattern of growth that we can draw from these five years of data due to the outlier (i.e., idiosyncratic test result in 2009). Despite the uncertainty of the growth pattern, this school would be classified as a low-improving school by the method used in the study. As one would expect, part of the inconsistency seems to be related to school size; the smaller a school, the more inconsistent or unstable its average proficiency over time. There were only about 95 students in the Figure 2 school who took the test, and this uncertainty is likely to worsen if we break down the school by subgroups.

In the report, after a growth rate has been identified, the authors classified schools into quartiles, with a focus on the bottom quartile. Their use of quartiles as a reference point of classification is justifiable by statistical analysis convention, but the rationale for their choice is not explained to their policy audience. In order to decide how much academic growth is good enough, the study chose to use a norm-referenced classification scheme as opposed to a criterion-referenced classification. A major problem

with this approach is that it does not recognize broad improvement. Schools or students are pitted against one another, and no matter how much positive growth they make, some of them will, by definition, still be below the norm. Accordingly, their progress will not be recognized. In contrast, a criterion-referenced approach involves setting desired standards for growth based on externally determined criteria such as curricular-based or age- or grade-based expectations for student performance (e.g., value-added growth models adopted by states like North Carolina and Tennessee). This requires setting performance standard for each grade and connecting them across grades. The approach used in the report also raises an unanswered question as to how much growth is sufficient to warrant proficiency in the future, and how soon.

VI. Review of the Validity of the Findings and Conclusions

The authors argue that

> progress differs drastically among the states. For example, on average, reading and math proficiency rates in Maryland have improved substantially in recent years, yet in other states—Indiana, for example—average performance has remained flat. (p. 2)

Although this statement can be true in general, simple interstate comparisons of achievement trends based on their own state assessment results can be misleading. As noted in the report, the two states' performances on NAEP as well as their own state tests are very similar; this indicates that the rigor of their proficiency standard is comparable. However, the state-assessment trends in Indiana and Maryland diverge, despite common flat trends on NAEP (see Figure 3). Given the flat NAEP trends in both states, it appears that Maryland's state test score gains might be an artifact of something not related to authentic educational progress. From a psychometric perspective, these are extraneous factors not transferrable to an

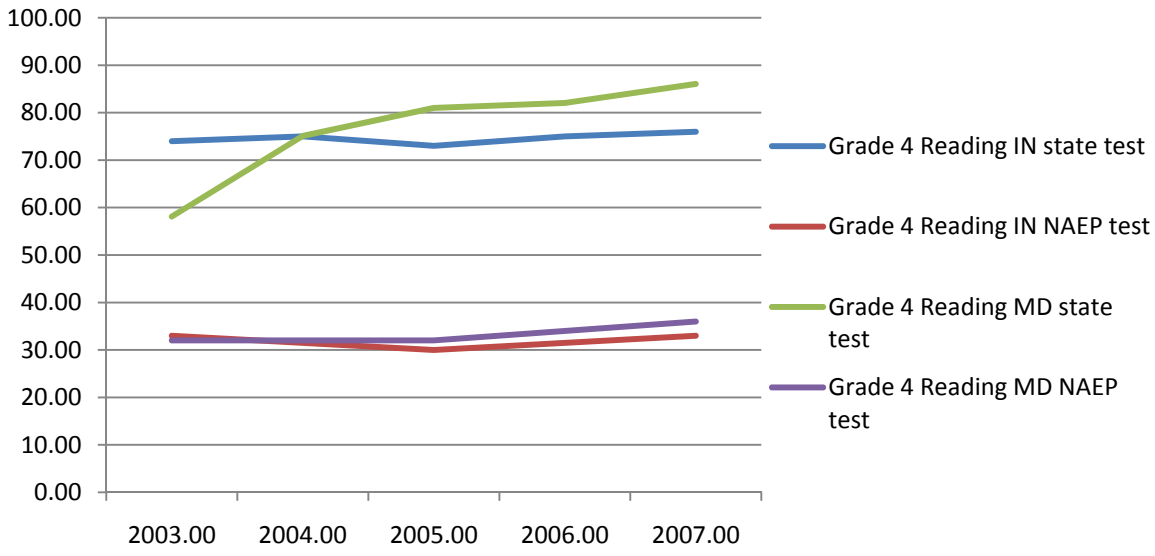

Figure 3. Grade 4 Reading Proficiency Rate Trends on NAEP versus State Assessments in Indiana and Maryland

independent, low-stakes NAEP test (e.g., narrowing the curriculum and teaching to the test). Although the report does not give any clues about the specific practices in these particular states, prior research suggests that this kind of evidence raises a question about the validity of using state test score results as a sole measure of academic progress under NCLB.[5]

The report shows an improvement gap between high-improving and low-improving schools of about 4-6 percentage points. For Indiana, the gap is from 2.2 to -1.9 percentage points; for Maryland, it's from 0.5 to 5.6 percentage points. But the report did not explore possible causes for the variability of improvement levels. It does show that high-improving schools have significantly more minority and low-income students than low-improving schools in Maryland, whereas the student demographics are very similar between the two groups of schools in Indiana. But the reader is left to ponder why this might be.

In fact, the reason why Maryland's high-minority and low-income schools showed a greater degree of improvement is most likely a statistical artifact known as regression to the mean.[6] My re-analysis of the same data from Maryland confirms that pattern; the correlation between initial status and growth rate is -0.72 in reading and in math.

Beyond this likely regression phenomenon, there is no information in the report that can help differentiate high-improving and low-improving schools. Unless we understand the school mechanism (such as school input, context, or process variables) that facilitates or constrains the differential pattern of growth, a simple presentation of demographic differences can be misleading.[7] While school-related effects may vary from state to state, it is worth investigating those factors that contribute to value-added academic growth beyond the effects of student and family background characteristics.

VII. USEFULNESS OF THE REPORT FOR GUIDANCE OF POLICY AND PRACTICE

The overall conceptual framework of the report helps bring more attention to the issues of validating and using school-level performance trend data for accountability. The report's idea of measuring and recognizing growth in addition to status is part of a general improvement to the current exclusive and narrow focus on a year-by-year snapshot model of school evaluation under NCLB. However, the report's methods are so simplistic, arbitrary and poorly fitting to the report's own assumptions that it is more harmful to sound policymaking than helpful.

The report's norm-referenced model guarantees failed schools independent of their true performance and improvement levels. There will always be winners and losers when the calculation is based on comparisons of schools' relative performance or improvement against percentile ranks rather than absolute benchmarks. In fact, this purely norm-referenced approach may pose potential conflicts in the real policy world, since it goes against the spirit of setting common standards for all. We need further research and policy discussion with regard to setting desirable and feasible goals of school performance and improvement targets.

There remain outstanding questions about the validity and reliability of the measures and methods used by the study. The difference shown in Figure 3 above, between the improvement patterns based on national versus state assessment results, suggests that the report's sole reliance on state assessment data can be misleading. Further, many

schools do not follow a strictly linear pattern of improvement (i.e., same incremental gains each year), and thus the report's imposition of a linear growth model on all schools is questionable. This is more problematic in small schools where the school-level aggregate performance patterns are not highly consistent and stable over time. Since the Education Trust plans to report the analysis of school subgroup performance as part of this series of publications, it should seriously consider the reality that it becomes more challenging to reliably measure growth for student subgroups in small schools.

The utility of the current report is also limited since it did not examine the school characteristics associated with differences between low-improving versus high-improving schools that had low initial performance status. Consequently, the authors are vague about what specific strategies—such as benchmarking, funding, reconstitution, and capacity-building—are more viable and effective options for identified schools. This question should, in fact, remain unanswered until we know whether differences in growth rates are based on reliable and valid measures and if so, what school factors caused these differences. Using the framework shown in Figure 1, how can we help struggling schools move from cell D to cell C, and then ultimately to cell A?

Notes and References

[1] Ushomirsky, N. & Hall, D. (2010). *Stuck Schools*. Washington, DC: The Education Trust. The authors declare: "Identifying these schools and either helping them improve or forcing them to close, with students removed to higher performing schools, is an economic and moral imperative" (p.2).

[2] Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review, 16*(3), 283-301.

Bryk, A., Thum, Y., Easton, J., & Luppescu, S. (1998). Assessing school academic productivity: The case of Chicago school reform. *Social Psychology of Education, 2*, 103-142.

Raudenbush, S. W. (2004). *Schooling, Statistics, and Poverty: Can We Measure School Improvement?* Angoff Lecture No. 9. Princeton, NJ: ETS.

Lee, J. (2007). *The Testing Gap: Scientific Trials of Test-Driven School Accountability Systems for Excellence and Equity.* Charlotte, NC: Information Age Publishing.

Commission on No Child Left Behind (2006). *Growth models: An examination within the context of NCLB.* Aspen Institute Commission Staff Research Report.

[3] School improvement levels are not strongly related to school performance levels. See:

Lee, J. (1998). *Assessing the Performance of Public Education in Maine: Factors Influencing School Differences*. Occasional Paper No. 29. Orono, ME: University of Maine Center for Research and Evaluation.

Raudenbush, S. W. (2004). *Schooling, Statistics, and Poverty: Can We Measure School Improvement?* Angoff Lecture No. 9. Princeton, NJ: ETS.

[4] The correlation between adjacent year scores for reading and math proficiency are in the range of .9, and they become a little weaker for remote years. These results indicate that school-level aggregate proficiency rates as a proxy measure of school performance are reliable. However, the problem lies with the reliability of the school improvement measure (i.e., growth rate).

[5] See Lee, J. (2010). Trick or Treat? New Ecology of Education Accountability System in the United States. *Journal of Education Policy, 25*(1), 73-93.

[6] See Lubienski, C. (2008). *Review of "Promising start: An empirical analysis of how EdChoice vouchers affect Ohio public schools" from the Milton & Rose D. Friedman Foundation.* Boulder and Tempe: Education and the Public Interest Center & Education Policy Research Unit. Retrieved March 26, 2010, from http://epicpolicy.org/thinktank/review-promising-start.

[7] See Lee, J. (1998). *Assessing the Performance of Public Education in Maine: Factors Influencing School Differences*. Occasional Paper No. 29. Orono, ME: University of Maine Center for Research and Evaluation.

Raudenbush, S. W. (2004). *Schooling, Statistics, and Poverty: Can We Measure School Improvement?* Angoff Lecture No. 9. Princeton, NJ: ETS.

Lee' study (1998) of Maine school achievement trends identified factors that differentiate between the most- and least-improving schools and the highest- and lowest-performing schools; it found that school poverty is much more powerful in explaining the gap in status than the gap in growth. Raudenbush's study (2004) of a national sample also shows that school accountability systems using mean proficiency would disparately and unjustifiably identify high-poverty schools to be failing. Switching from mean proficiency to the value-added approach would produce better results for high-poverty schools. Lee's study also showed that the schools with the most improvement also have better per-pupil expenditures, teacher education and experience, and instructional resources than the least improving schools. This finding may not hold in high-stakes testing situations, which that often lead to test score inflation without authentic gains.

The Think Tank Review Project is made possible by funding
from the Great Lakes Center for Education Research and Practice.