Oklahoma School Grades: Hiding "Poor" Achievement

A-F Report Card

A report produced by the staff of

The Oklahoma Center for Education Policy (University of Oklahoma) and

The Center for Educational Research and Evaluation (Oklahoma State University)

October 2013

Contributing authors:

The Oklahoma Center for Education Policy (University of Oklahoma)
- Curt M. Adams, Senior Research Scientist
- Ellen A. Dollarhide, Research Associate
- Patrick B. Forsyth, Senior Research Scientist
- Ryan C. Miskell, Research Associate
- Jordan K. Ware, Research Associate

The Center for Educational Research and Evaluation (Oklahoma State University)
- Laura L.B. Barnes, Senior Research Scientist
- Jam Khojasteh, Senior Research Scientist
- Mwarumba Mwavita, Senior Research Scientist

Corresponding authors:

Curt M. Adams, Curt.Adams-1@ou.edu
Patrick B. Forsyth, patrickforsyth@ou.edu

This paper was reviewed by internationally known measurement and accountability expert Robert Linn.

Robert Linn is Distinguished Professor Emeritus of Education in the research and evaluation at The University of Colorado. Linn has published more than 250 articles and chapters on a wide range of theoretical and applied issues in educational measurement. His research explores the uses and interpretations of educational assessments, with an emphasis on educational accountability systems. His work has investigated a variety of technical and policy issues in the uses of test data, including alternative designs for accountability systems and the impact of high-stakes testing on teaching and learning. Dr. Linn is a member of the National Academy of Education (NAEd) and a Lifetime National Associate of the National Academies. He has been an active member of the American Educational Research Association (AERA) for more than 40 years and served as vice president of the AERA Division of Measurement and Research Methodology, vice chair of the joint committee that developed the 1985 Standards for Educational and Psychological Testing, and as president of AERA. He is a past president of the National Council on Measurement in Education (NCME), past editor of the Journal of Educational Measurement and editor of the third edition of Educational Measurement, a handbook sponsored by NCME and the American Council on Education. He was chair of the National Research Council's (NRC) Board on Testing and Assessment and served on the NRC's Board of the Center for Education, and of the Advisory Committee for the Division of Behavioral and Social Sciences. He served as chair of the NAEd Committee on Social Science Research Evidence on Racial Diversity in Schools, and as chair of Committee on Student Achievement and Student Learning for the National Board of Professional Teaching Standards.

Table of Contents

Executive Summary 4

I. Introduction 7

II. System Weaknesses 8

III. Empirical Test 9

Data Source 9

Results 11

Concern 1: Small Achievement Differences 12

Concern 2: High Classification Error 14

Concern 3: Within School Achievement Gap 18

IV. Discussion 22

V. Conclusion 24

VI. Recent Legislative Changes 28

References 32

Appendix A: Robert Linn Comments 35

Appendix B: Technical Description of Analysis 36

Oklahoma School Grades: Hiding "Poor" Achievement

Executive Summary

We are among those who favor examining schools to determine how effective they are in their mission to maximize learning for all children. We are passionate about making the evaluation of schools a truthful and credible process. Oklahoma is one of many states that has chosen to report school performance using a single letter grade generated primarily from standardized test results. In a white paper published earlier this year, we examined Oklahoma's school evaluation system and discovered fundamental flaws that make letter grades virtually meaningless and certainly ineffective for judging school performance. Our analysis and conclusions were reviewed by two nationally renowned testing and evaluation experts who concurred with our claims. Subsequently, the State made some changes to the system, but the changes do not address the flaws; in fact, the likelihood is that they made them worse.

The pursuit of a defensible school evaluation system requires that the following question be addressed: Should *school* performance be based solely on *student* standardized test results? To some, it sounds reasonable that it should. However, multiple examinations of the sources of variation in student test performance reveal that more than 70 percent is due to non-school causes. Of course, schools do affect test results, but the effect size is routinely found to be between 20 and 30 percent (Heck, 2009; Linn & Haug, 2002; Nye, Konstantopoulos, & Hedges, 2004). Thus, composing school letter grades from student test performance alone will frequently give false credit or blame to schools for effects that are mostly unrelated to what they do.

Since ratings produced through the use of letter grade systems are often attached to high-stakes decisions (e.g. school closure, school leadership, teacher employment, and funding), we were surprised to find very few systematic evaluations of their use (Schwartz, Hamilton, Stecher, & Steele, 2011). The surprise is enhanced because potential problems with the composition of letter grades should be quite apparent to the measurement and evaluation personnel implementing these state-level projects. This work, then, is intended to stimulate a national and state-wide debate about the legitimacy of the single letter grade approach to school accountability. The analyses are limited by the fact that they are based on one state's system and data, although the technical flaws present in Oklahoma's system appear to be universal.

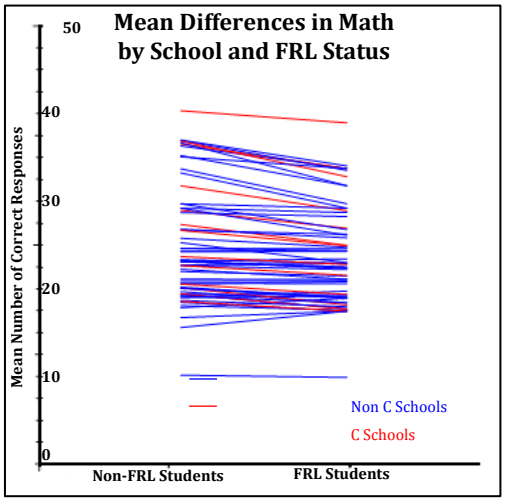
In this paper, we examine how the Oklahoma school grading system operates in practice. Our analyses and conclusions are based on actual state-assigned school letter grades and individual student test scores. The belief that standardized test results are the primary indicators of school performance is implicit in the letter grade evaluation approach. If this belief is true, state-assigned grades have meaning only if schools given high grades reflect a pattern of high academic achievement; conversely, schools given low grades should reflect a pattern of low academic achievement.

What we found:

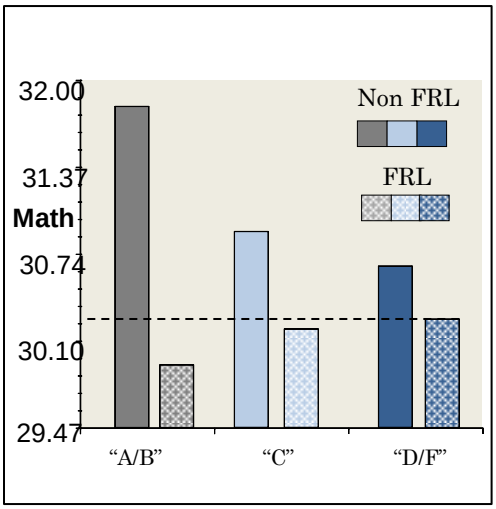
1) Very Small Differences Predicted by Letter Grades. When school raw scores for reading, math, and science were averaged, three to six correct responses separated "A" schools from "F" schools on 50 question tests. These are small effects on which to base significant decisions. Many of the achievement differences between letter grades were likely due to chance; even when they reached statistical significance they were of questionable practical utility, generating little confidence in grade distinctions.

Fixed Effects	Mean Differences in Reading by Number of Correct Responses
A-B	0.28
A-C	1.76
A-D	4.86
A-F	3.67

2) Classification Error. A classification error is the consequence of attempting to summarize relatively independent dimensions with a single indicator. Summarizing a school's test performance on math, reading, and science in a single letter grade is difficult because school test averages vary independently across subject areas. As an example, our analysis showed that math performance in some "D" and "F" schools was higher than that in some "B" and "C" schools; moreover, none of the seven schools with the highest math average were "A" schools. To be meaningful, the letter grade would have to represent a school's performance pattern, but it turns out that within-school variation across subject areas fluctuates a great deal. Thus, it is never clear what an A is or what an "F" is.



3) Achievement Gaps. Letter grades hide low test performance of poor and minority children. Consistently across the three subject areas (reading, math, and science), minority and poor children tested highest in "D" and "F" schools and lowest in "A" and "B" schools. Put differently, according to the State's own effectiveness grades, "A" and "B" schools are the least effective for poor and minority children; high scoring, affluent students in those schools produce averages that give the appearance of school effectiveness for all, essentially masking the especially low performance of poor and minority children.



In summary, the data we have analyzed demonstrate quite dramatically that the letter grade system for school evaluation has very little meaning and certainly cannot be used legitimately to inform high-stakes decisions. The letter grades hide important differences between schools rather than reveal them. This obscurity is the result of two basic flaws that we discussed at length in our earlier paper: 1) It attempts to summarize unlike dimensions with a single indicator, and 2) it utilizes proficiency bands in a complicated formula that transforms raw scale scores into categories and back again, losing precision at every turn; then bonus points are added. The resulting grade has practically no meaning or utility.

Based on our empirical testing, we urge policy makers to abandon the single letter grade approach. The fix is quite simple. A school's performance should be reported on multiple dimensions--a profile that includes scale scores for subject areas as well as other relevant school conditions (e.g. program coherence, social climate, and faculty and administrative stability). Scale scores are more easily understood and less susceptible to manipulation and distortion. A balance of process and contextual conditions helps portray a truer performance picture that provides clarity to parents and focuses the improvement efforts of school professionals. Decisions about intervention should take demographics such as poverty and neighborhood vitality into consideration. A bureaucratic evaluation system that produces nearly meaningless grades is no substitute for reasoned decision-making based on careful consideration of all credible evidence.

Oklahoma School Grades: Hiding "Poor" Achievement

I. INTRODUCTION

Like other citizens we are concerned with the quality of teaching and learning in Oklahoma's public schools. We also believe that accountability systems, if designed and used correctly, can support efforts to improve learning and close achievement gaps. What constitutes an effective accountability system, however, has become controversial. Different interest groups and individuals have come to define effectiveness by favored design features, not by objective evidence on the performance of accountability systems. We believe a litmus test of an effective system should be based on how accountability indicators operate in practice (Baker & Linn, 2002; Linn, 2005).

When letter grades were put to the test with actual student achievement data, it turns out that they do more to hide achievement differences than provide a clear understanding of school effectiveness. In our analysis of over 15,000 student test scores from 63 schools, results showed that school grades do not fulfill the intention of the State to provide parents and schools with a clear understanding of school performance. The following results raise serious concerns about the performance of a single letter grade:

- "A" and "B" schools were least effective for poor and minority students;
- A "C" school outperformed all "A" and "B" schools in math;
- Students in "F" schools had higher average reading and math achievement scores than students in "D" schools;
- There were virtually no achievement differences in reading, math, and science among "A", "B", and "C" schools.

We also examined recent legislative changes to the A-F reporting system and found that the fundamental problems contributing to inaccurate results and invalid interpretations have been compounded, not resolved. Despite initial missteps in the design of an effective accountability system, our objective, which we share with State leaders, is to get accountability right so all students have access to learning opportunities that prepare them for academic and personal success.

Where do we, as a State, go from here? Can we put partisan politics aside and design an accountability system that at the very least is capable of accurately reporting achievement differences within and between schools? Or, will we continue to use a framework that hides achievement gaps, reports inaccurate results, and fuels invalid interpretations of school effectiveness?

II. SYSTEM WEAKNESSES: CONCEPTUAL, MEASUREMENT, & USE

On the face of it, what is wrong with using letter grades to measure school performance? A persuasive case can be made, even before an empirical demonstration is mounted, that the accountability approach taken by the State of Oklahoma is not reasonable. Here's why. While basing the letter grade solely on student test performance and like indicators, the A-F policy ignores the fact that most achievement variation exists within schools not across schools. Standardized test scores used to measure outcomes of school and teacher performance are in fact mediated and moderated by individual student differences, family characteristics, and school contextual differences. The most strident research claims for school effects on student academic performance are based on findings that consistently fall below 30 percent; in other words, no more than 30 percent of the variation in student achievement is due to a school and its teachers (Linn & Haug, 2002). That means that over 70 percent of school variance is an indicator of non-school conditions.

The validity of a measure depends on the uses and interpretations of the measure rather than what it aspires to measure (Schwartz, Hamilton, Stecher, & Steele, 2011). In this instance that means isolating variation in student achievement that is the consequence of what the school does. School grades composed primarily of student achievement scores cannot do this because as measures of school performance they are unpredictably contaminated by variables not under school control (Rothstein, Jacobsen, & Wilder, 2008). Moreover, with status measures (categorical performance ranges or proficiency bands), schools can be identified as effective even if achievement is not improving or if achievement gaps exist. Similarly, schools whose students are making progress may be identified as needing improvement if achievement does not exceed specified thresholds.

A primary assumption of the A-F accountability system, that student test scores can be dissected and manipulated into valid indicators of school performance, is simply false (Linn, 2005; Rothstein, Jacobsen & Wilder, 2008). Student test scores are not a trustworthy measure of school performance. We are not advocating that accountability systems should ignore achievement; rather, we are simply highlighting problems associated with making inferences about school performance from tests designed to measure student achievement. The development of valid indicators of school performance is possible, but not through the use of measures that rely almost exclusively on standardized test data. A valid measure of school performance should be comprehensive, accounting for school processes, conditions, practices, and outcomes (Sirotnik, 2002).

A test of good policy is its utility and cost effectiveness. Unfortunately, test-based accountability systems "have mostly failed to translate to fundamental changes in teachers' pedagogy" (Hamilton et al, 2013, p. 457). This conclusion is not

particularly surprising; Oklahoma's Report Card is very costly, bureaucratically cumbersome, and seems not designed to improve learning in schools. We draw these conclusions because the State provides schools with one-shot test evidence delivered to them in the following school year. The reporting to schools is not useful because it isn't timely, nor does it provide information, direction, or resources that could make a difference in teaching practices.

In addition to resting on a false assumption and having low utility, the A-F accountability system is infested with measurement decay, a flaw in the compositional approach used by the State to produce letter grades for schools. The three components (now two with recent legislative changes) that constitute the letter grade go through multiple iterative cycles moving from interval data (raw scores on a continuous scale) to ordinal data (named or numbered clusters of scores), losing precision at every step. By the time the letter grade has been calculated for each school, further distorted by weightings and bonus points, it no longer carries with it the meaning contained in the original raw scores. This arbitrary and unorthodox manipulation of scores with little regard for accepted statistical and psychometric practice results in a letter grade that is virtually meaningless. To test our "meaningless" hypothesis, we subjected the 2011-12 Oklahoma letter grades to a test of their predictive validity, answering the simple question "Does a single letter grade accurately predict differences in student achievement?"

III. EMPIRICAL TEST: DO LETTER GRADES ACCURATELY PREDICT DIFFERENCES IN STUDENT ACHIEVEMENT?

If a single, summative letter grade is meaningful, each letter should designate some level of school performance that is distinct from other performance levels. Reasonably, the achievement distribution of student test scores in an "A" school should be higher than those in a "B" school. These differences in test performance by letter grade should be substantial and they should persist even after controlling for factors that are unrelated to teaching effectiveness or school practices. Finally, letter grades should not mask subgroup achievement gaps that may exist within schools. Effective schools, as designated by and "A" or "B", should be effective for all students, not just students with more learning opportunities or greater social resources.

DATA SOURCE

We examined the test scores of over 15,000 students in 63 urban schools by state-assigned letter grade. Our choice of schools was purposeful. Urban schools are more alike in the percent of students who qualify for the federal lunch program and minority/non-minority compositions than the general population of Oklahoma schools. And, the urban sample allowed us to test the validity and fairness of school

grades in a context where it is critical for performance information to differentiate effective schools from ineffective ones.

Table 1 contains descriptive data for the sample of students and schools. We had valid math scores from 15,315 students, valid reading scores from 15,380 students, and valid science scores from 4,935 students. The sample size difference between math/reading and science is explained by the fact that only 5th and 8th graders were tested in science while 3rd through 8th grade students were tested in math and reading. Approximately 54 % of the students identified as minority and 46 % as non-minority Caucasian. Nearly 77 % qualified for Free or Reduced Lunch (FRL). The average math scale score was 701, reading 699, and science 744. The school sample shows that the average FRL rate was 86 %; 5 % of the schools earned school grades of "A"; 13 % earned grades of "B"; 21 % earned grades of "C"; 54 % earned grades of "D"; 8 % grades of "F"; 49 schools were elementary; and 14 were middle schools.

Table 1.
Descriptive Student and School Data

Student Characteristics	
Percent Minority	54
Percent Non-Minority	46
Percent Free/Reduced Lunch (FRL) Status	77
Mean Math Scale Score	701.48
Mean Math Raw Score	32.74
Mean Reading Scale Score	699.30
Mean Reading Raw Score	30.73
Mean Science Scale Score	743.67
Mean Science Raw Score	27.47
School Sample	
Average School Free/Reduced Lunch Rate	86
Percent "A" Schools	5
Percent "B" Schools	13
Percent "C" Schools	21
Percent "D" Schools	54
Percent "F" Schools	8
Number of Elementary Schools	49
Number of Middle Schools	14

Note. N= 15,315 math; N = 15,380 reading; N= 4,935 science; N=63 schools; Raw scores ranged from 1-45 for science, 1-50 for reading, and 1-50 for math; Scale scores ranged from 400-990

ANALYTICAL PROCEDURE

We explored student differences in reading, math, and science achievement between schools graded "A", "B", "C", "D", and "F" by the State, after accounting for variance that cannot be attributed to teachers or school performance. A full technical description of the analytical approach is provided in Appendix B.

RESULTS

The combined results indicate that a single, summative letter grade cannot accurately identify school performance or the contribution schools make to student achievement. Letter grades obscure actual effectiveness or ineffectiveness of schools, identifying some schools as effective when they are not meeting the needs of their FRL and minority students and other schools as ineffective when they have higher achievement and smaller achievement gaps. One reason for this is the proportion of achievement variance attributed to student and school differences. Student differences accounted for 79 percent of variance in reading, 74 percent in math, and 73 percent in science. Schools, on the other hand, accounted for only 21, 26 and 27 percent of achievement variance in reading, math and science respectively (Table 2). The unaccounted for within-school variance is precisely the achievement difference that needs to be addressed in order to understand how student subgroups are faring in schools.

Table 2.
Decomposition of Achievement Variance

Variable	Achievement Differences due to Non School Factors	Achievement Differences due to Schools
Reading Achievement	79%	21%
Math Achievement	74%	26%
Science Achievement	73%	27%

Note. Variance decomposition was performed with an Unconditional Random Effects ANOVA

Failure to attend to within-school achievement variance yields a distorted picture of school performance. Results of the empirical test reveal three serious concerns arising from a letter grade that cannot measure sufficiently achievement differences within schools: (1) small achievement differences among school grades, (2) high classification error, and (3) the inability to detect achievement gaps within schools.

Concern One: Small Achievement Differences. Table 3 displays average differences in the number of test questions answered correctly by students of a school. For example, a one point difference means one group of students answered one more question correctly on the test than the comparison group. What stands out is that actual mean differences among students in "A", "B", and "C" schools were small and in most cases not statistically different than zero. This means differences were likely due to change and were unlikely to result from systematic performance differences across schools. Additionally, the margin separating "A" schools and "D"/"F" schools was much smaller than one would reasonably expect.

For reading achievement, there were virtually no significant differences among "A", "B", and "C" schools. The small differences we found were more likely the result of chance than systematic achievement differences across schools. On average, less than one correct question separated "A" school students from "B" school students. The average difference between "A" and "C" was fewer than 2 questions. Differences between "A" and "D" school students and "A" and "F" school students were statistically significant, but the margins were small. Students in "A" schools on average had around 4 more correct questions than students in "D" schools and fewer than 4 questions compared to students in "F" schools. Average achievement differences between "B" and "C" schools were not significantly different from zero. The most troubling finding was that students in "F" schools had higher average reading scores than students in "D" schools.

Letter grades performed only slightly better in predicting a school's average math score. There were virtually no achievement differences between students in "A" and "B" schools. Average math scores between "A" and "C" schools were significantly different from zero, but the difference was small. Nearly four questions separated the average student in an A school from the average student in a "C" school. The average math difference between students in "B" and "C" schools was approximately two points, and the difference between students in "C" and "D" was approximately 2.4 points. Students in "F" schools had slightly higher average math achievement scores than students in "D" schools.

School letter grades were least effective at predicting differences in science achievement. Achievement averages in "A", "B", "C", and "D" schools were not significantly different from zero. Additionally, students in "C" schools had higher average achievement scores than students in "B" schools. Students in "F" schools had lower achievement scores than other students. Generally, differences in average performance by letter grade in the three measured subject areas are very small, so small as to have little practical meaning.

These results demonstrate clearly that a single, summative letter grade does not discriminate effectively achievement differences across schools. There were virtually no differences in average reading and science achievement among "A", "B",

and "C" schools. In math the differences between "A" and "C" and "B" and "C" schools were statistically significant, but small. We also found higher achievement averages in schools with lower letter grades. Students in "F" schools, for example, had higher average reading and math scores than students in "D" schools. Students in "C" schools had higher science averages than students in "B" schools. If a letter grade, which is based primarily on standardized test scores, does not necessarily tell us anything about school differences in reading, math, and science outcomes, what does it tell us?

Table 3.
Raw score differences by assigned letter grade

Fixed Effects	Mean Reading Differences	Mean Math Differences	Mean Science Differences
Intercept	32.23 (.19)**	30.36 (0.22)**	27.69 (0.22)**
A-B	0.28 (0.95)	1.83 (1.11)	0.75 (1.02)
A-C	1.76 (1.11)	3.79 (0.93)**	0.62 (1.10)
A-D	4.86 (1.22)**	6.18 (1.01)**	1.67 (1.35)
A-F	3.67 (1.43)*	5.81 (1.31)**	3.87 (1.62)*
B-C	1.48 (1.02)	1.96 (0.75)*	-0.12 (0.64)
B-D	4.85 (0.83)**	4.34 (0.74)**	0.92 (0.82)
B-F	3.38 (0.96)**	3.97 (1.32)**	3.12 (1.27)*
C-D	3.32 (0.57)**	2.38 (0.52)**	1.05 (0.65)
C-F	1.36 (1.01)	2.01 (1.02)*	3.25 (1.02)**
D-F	- 1.19 (0.82)	- 0.37 (0.89)	2.20 (0.81)**
Deviance (-2 Log likelihood)	110468	110894	33792
Δ Deviance	-623 **	-1064*	361**
Explained Between School Variability	92%	92%	94 %

*Note. * p<.05, **p<.01. We had valid reading data for 15,380 students, valid math data for 15,315 students, and valid science data for 4,935 students. Estimates come from random intercept and slopes as outcomes models. Standard errors are reported in parentheses. Student controls include FRL status, minority status, grade, and gender. Contextual controls include prior achievement, percent minority, and percent FRL rate. Student and school variables were grand-mean centered, and full maximum likelihood estimation was used. Raw scores range from 1-50.*

Will recent legislative changes to A-F enable grades to predict differences in student achievement?

Legislative changes have not addressed the primary source of grade imprecision. Continued problems include:

- Reliance on proficiency scores makes grades sensitive to factors that are unrelated to school performance or teaching effectiveness.
- The use of dichotomized proficiency levels rather than test scores compounds grouping error.
- Student mobility within and across districts affects growth interpretations.
- Mathematical properties of the growth index are unknown.
- Conceptually the growth index is not meaningfully tied to growth.
- Arbitrary changes made to cut scores make the achievement measure meaningless

see pages 28-32

Concern Two: Classification Error. We used math scores to illustrate the classification error that results from using a single letter grade to summarize school performance. Math scores were selected because mean differences were slightly higher for math than reading and science, suggesting better measurement precision. Even with math scores, however, there were cases in which schools with lower letter grades had higher achievement than schools with higher letter grades. In addition, several schools with lower letter grades had smaller achievement gaps than schools with higher grades.

The extent of the classification error is observable in figures one through five. Figure 1 compares "A" schools (red) against other schools in the sample (blue). The highest math achievement did not belong to an "A" school. In fact, seven schools have higher math achievement than the three "A" schools. Figure 2 reports the variability in math achievement among "B" schools (red). Four "B" schools have high math achievement while three others have achievement lower than some "C" and "D" schools (see figures 3 and 4). Figure 3 shows that "C" schools (red) fall all over the distribution. One "C" school had the highest math achievement, three others had achievement comparable to "A" and "B" schools, and several scored lower than "D" and "F" schools.

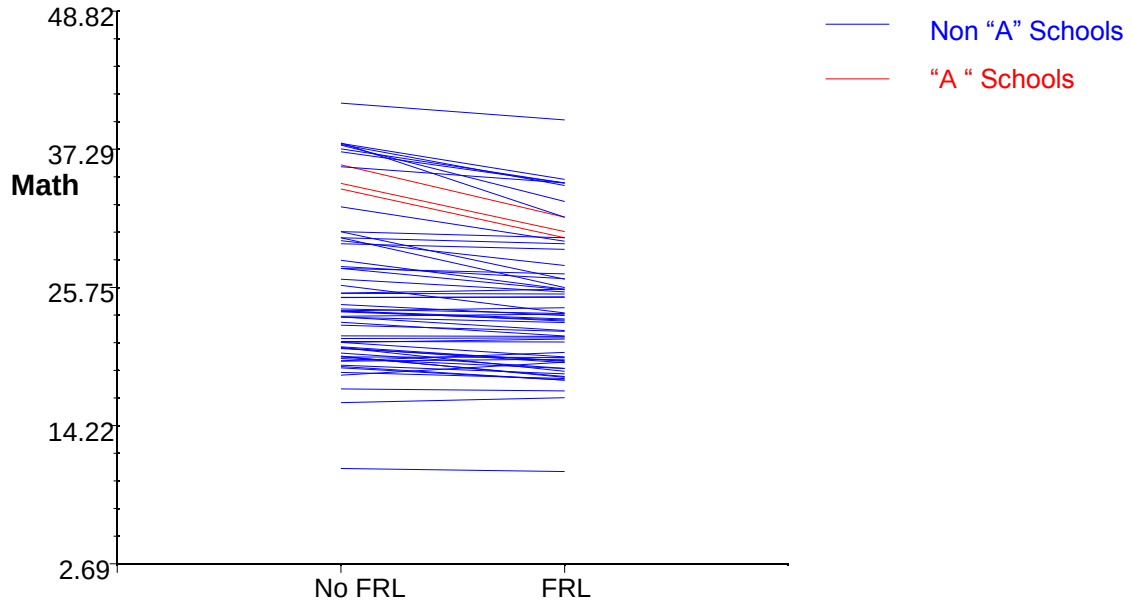


Figure 1. Math achievement of FRL and Non FRL by “A” schools (red) compared against all other schools in the sample (blue). The y axis reports raw math scores. The x axis identifies the FRL status of students. Results represent average math achievement by FRL status without holding constant differences in contextual school conditions and student characteristics.

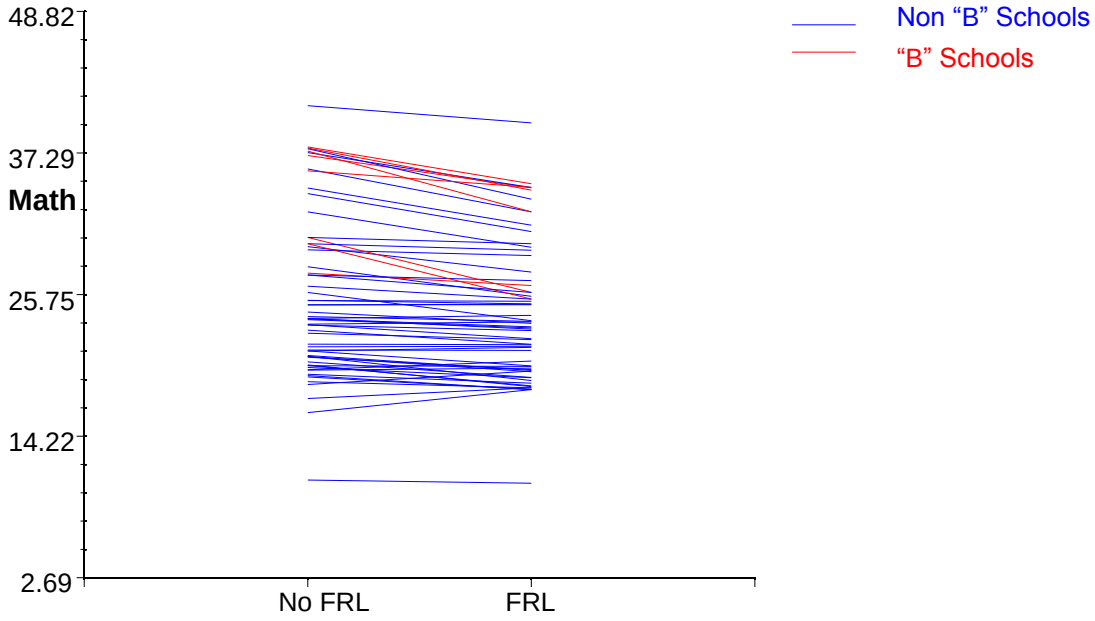


Figure 2. Math achievement levels of “B” schools (red) compared against all other schools in the sample (blue). The y axis reports raw math scores. The x axis identifies the FRL status of students. Results represent average math achievement by FRL status without holding constant differences in contextual school conditions and student characteristics.

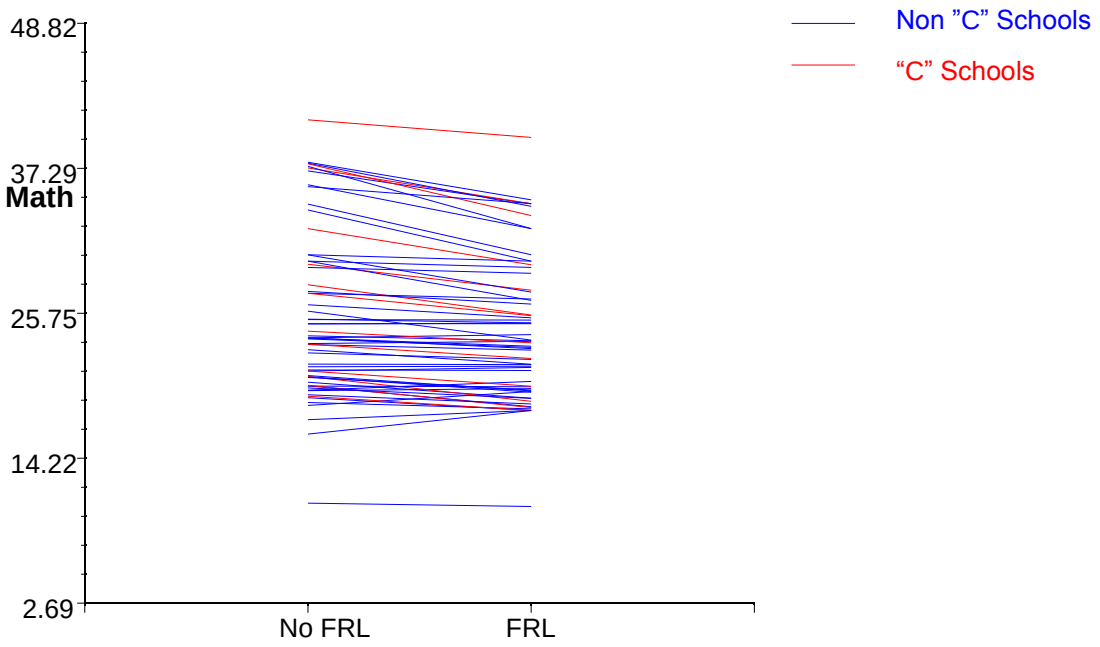


Figure 3. Math achievement levels of "C" schools (red) compared against all other schools in the sample (blue). The y axis reports raw math scores. The x axis identifies the FRL status of students. Results represent average math achievement by FRL status without holding constant differences in contextual school conditions and student characteristics.

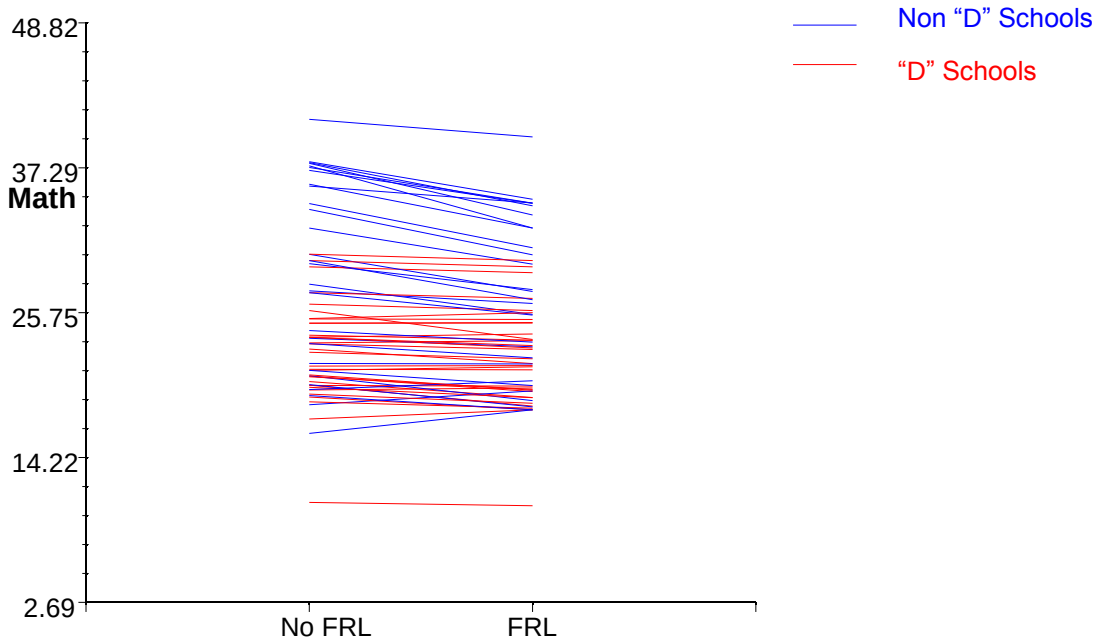


Figure 4. Math achievement levels of "D" schools (red) compared against all other schools in the sample (blue). The y axis reports raw math scores. The x axis identifies the FRL status of students. Results represent average math achievement by FRL status without holding constant differences in contextual school conditions and student characteristics.

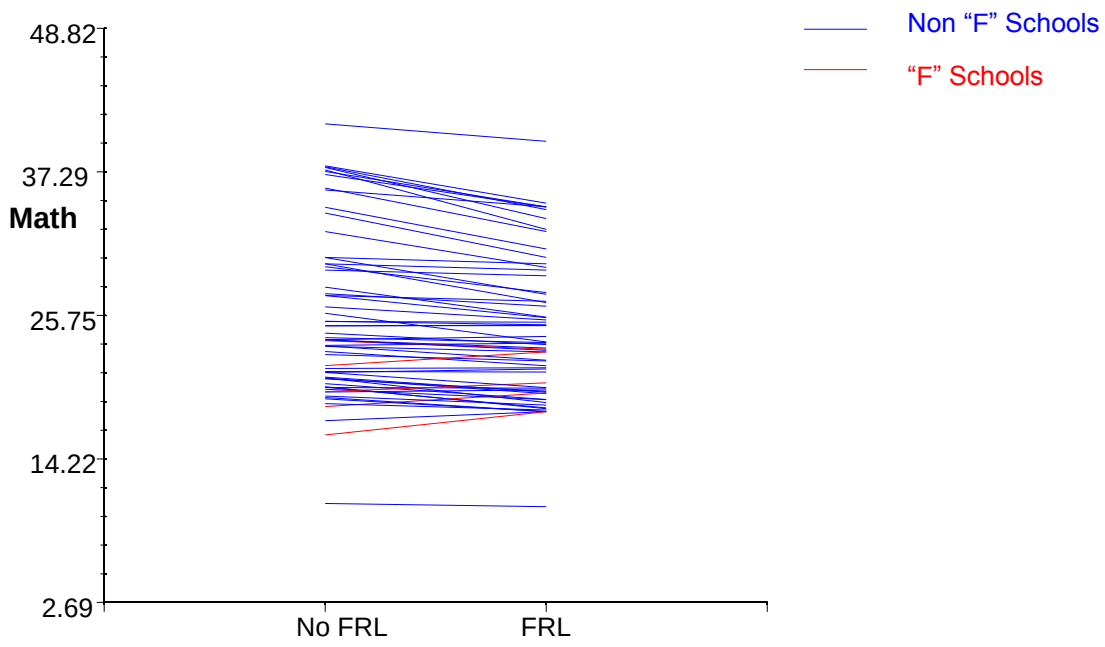


Figure 5. Math achievement levels of "F" schools (red) compared against all other schools in the sample (blue). The y axis reports raw math scores. The x axis identifies the FRL status of students. Results represent average math achievement by FRL status without holding constant differences in contextual school conditions and student characteristics.

As for "D" schools, several outperformed "B" and "C" schools, but there are several other "D" schools that had lower math scores than "F" schools (Figure 4). There are even a few "F" schools that outperformed "C" schools (Figure 5). In short, classification error is a consequence of using a single letter grade to summarize school performance. There are many cases where schools judged to be lower performing based on their letter grade scored higher in math than schools judged to be higher performing based on their grade.

When using a summary grade, it is inevitable that there will be instances where a school's achievement pattern is somewhat inconsistent. For example, a school with generally high achievement might have a slightly lower score in one subject area. We would not expect to find "C" schools outperforming "A" schools, or even cases were "D" and "F" schools had higher achievement averages than "B" and "C" schools. If public opinion and policy decisions are dictated by a school's summative grade, and yet that grade's meaning is distorted by classification error, how defensible or fair are such perceptions and decisions? For instance, "D" and "F" schools are perceived as ineffective, yet our analysis shows that math achievement was higher in a few "D" and "F" schools than some "B" and "C" schools. Conversely, "A" and "B" schools are perceived as effective even though for some schools their math achievement is lower than some "C" and "D" schools. This confuses us. Why is average math achievement higher in some "D" and "F" schools than some "B" and "C" schools? And, why is it lower in some "A" and "B" schools than some "C" schools?

Will recent legislative changes to A-F reduce high classification error?

Legislative changes do not address the fundamental problems contributing to high classification error. Continued problems include:

- The use of a single, summative grade does not account for the large within-school differences in achievement that occur within and between subject areas.
- Continued use of proficiency bands will hide achievement gaps within schools.
- Scoring of growth encourages gaming practices that target students near the "bubble" at the expense of lower or higher performing students.
- The selection of ten students as a minimum sample size is well below any accepted professional standard as a basis for group statistics, especially in high stakes-situations.
- Bonus points will be rewarded unequally depending on student composition and school configuration (e.g. elementary, middle, or high school)

see pages 28-32

Concern Three: Within School Achievement Gaps. Effective schools should promote excellence for all. That is, the criteria for designating an "A" or "B" school should include both high achievement and an absence of achievement gaps. On average, we did not find excellence and equity in the "A" and "B" schools in our sample. Within school achievement gaps among subgroups increased as school GPA's and grades increased, indicating that "A" and "B" schools were least effective for FRL and minority students.

On average, FRL students scored lower in reading, math, and science than their non-FRL peers (Figures 6, 7, and 8). The FRL/non-FRL achievement gap is largest in A and B schools. The most troubling finding is that FRL students in "D" and "F" schools had higher average reading, math, and science achievement than FRL students in "A" and "B" schools. This pattern across all three subject areas challenges the assumption that "A" and "B" schools are effective for all students. In fact, "D" and "F" schools were more effective for FRL students than "A" and "B" schools.

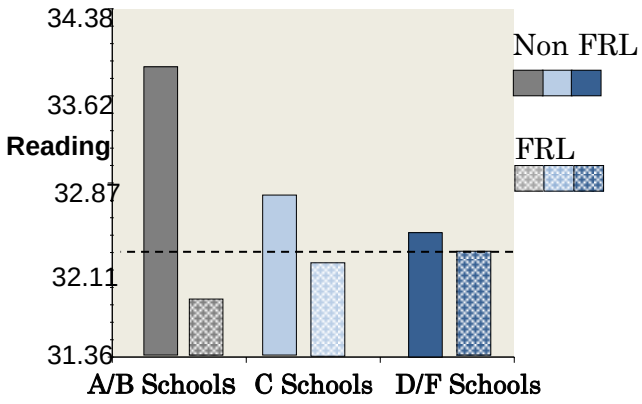


Figure 6. Graph from intercepts and slopes as outcomes model of Reading achievement. Results show lower average reading achievement for FRL students in A and B schools compared to C schools and D and F schools.

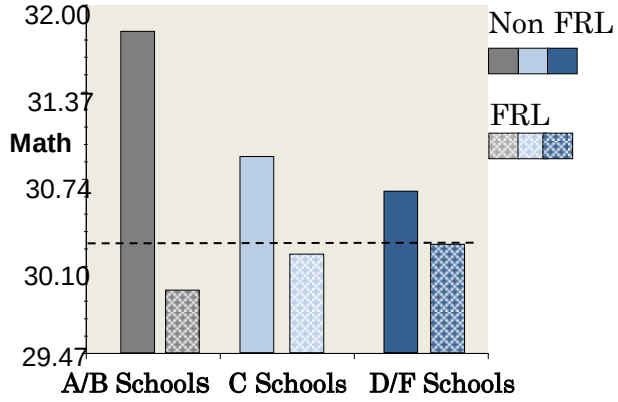


Figure 7. Graph from intercepts and slopes as outcomes model of math achievement. Results show lower average math achievement for FRL students in A and B schools compared to C schools and D and F schools.

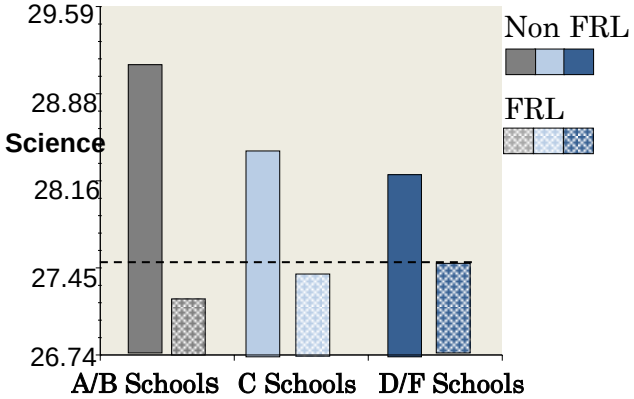


Figure 8. Graph from intercepts and slopes as outcomes model of Science achievement. Results show lower average science achievement for FRL students in A and B schools compared to C schools and D and F schools.

The achievement gap between minority and non-minority students was larger than the FRL/non-FRL gap. Minority students on average performed lower than non-minority students, but again we found that the size of the gap was largest in "A" and "B" schools. Similar to FRL, we found that average reading, math, and science achievement was higher for minority students in "D" and "F" schools than minority students in "A" and "B" schools. In short, the summative letter grade obscures the performance of subgroups within schools. Schools with minority/majority and FRL/non-FRL achievement gaps can earn "A's" and "B's" even though FRL and minority subgroups score below their peers in "D" and "F" schools. Thus, the letter grade exploits achievement levels that derive from wealth and social advantage, while obscuring a school's failure to serve all children.

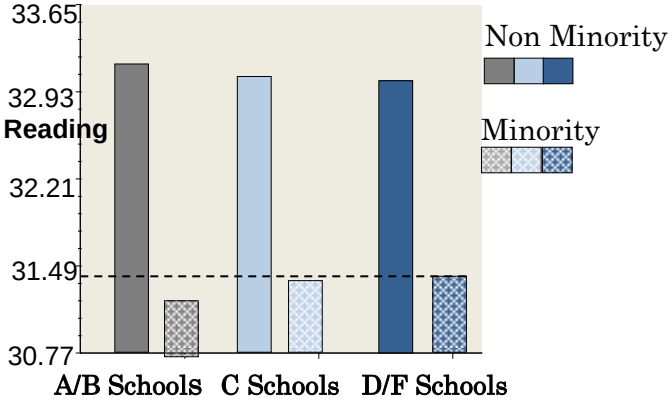


Figure 9. Graph of intercepts and slopes as outcomes model of reading achievement. Results show lower average reading achievement for minority students in A and B schools compared to C schools and D and F schools.

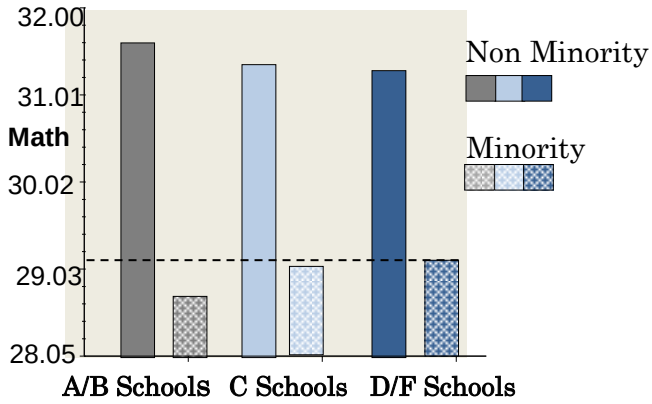


Figure 10. Graph of intercept and slopes as outcomes model of Math achievement. Results show lower average math achievement for minority students in A and B schools compared to C schools and D and F schools.

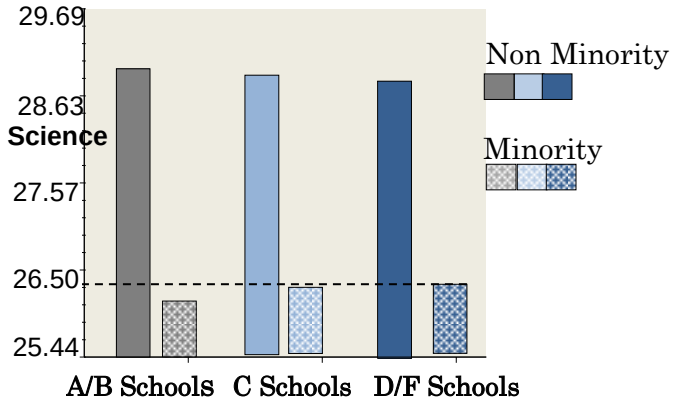


Figure 11. Graph of intercepts and slopes as outcomes model of Science achievement. Results show lower average science achievement for minority students in A and B schools compared to C schools and D and F schools.

Will the recent legislative changes to A-F continue to hide within school achievement gaps?

Legislative changes do not address the inability of grades to account for achievement gaps within schools. Continued problems include:

- Achievement scores are not reported by student subgroups.
- Percent proficiency measures used for student achievement and growth hide the performance of individual students and student subgroups.
- Dichotomizing achievement by proficient and non-proficient masks the distribution of individual student achievement within schools.

see pages 28-32

IV. DISCUSSION

The measure of a good school accountability system reflects its ability to help schools meet the learning needs of all students. At the most basic level, a fair and effective accountability approach must (1) produce accurate results, and (2) facilitate valid interpretations of results (Baker & Linn, 2002; Linn, 2005; Rothstein, Jacobsen & Wilder, 2008). Oklahoma's use of a single letter grade to measure school performance fails to achieve these minimum standards for effective use.

Accurate Results. We would expect to find achievement differences between "A", "B", "C", "D", and "F" rated schools analogous in meaning to the grades found on children's report cards for those same grades. We did not find this to be the case. What makes an "A" school better than a "B" or "C" school if the average achievement difference among them is virtually zero? Why are "F" schools lower performing than "D" schools when average reading and math achievement are higher in "F" schools? Why did a "C" school have higher math achievement than "A" and "B" schools, but other "C" schools had math achievement lower than some "D" and "F" schools? Letter grades are essentially value statements about school performance that are meaningless if grades are not indicative of school effectiveness.

Proposed action or intervention to improve student learning taken by the State, districts, or schools is unjustified when the accountability indicator cannot be trusted. Schools with poor achievement and large achievement gaps need to improve, but inaccurate and untrustworthy grades empower school leaders to resist meaningful efforts to restructure and redesign learning environments. Likewise, schools with high grades may use them as justification for preserving the *status quo* when the *status quo* may not be sufficient, or the *status quo* means FRL and minority students are not being served by the school. A meaningless or ambiguous measure of school performance even relieves policy-makers from responsibility for ineffective policies because of the difficulty in measuring the consequences of their decisions.

We have seen harmful unintended consequences result from decisions based on a faulty accountability system. State accountability systems prescribed by No Child Left Behind have not transformed schools, or even altered achievement trends (Baker, et., al. 2010; Forte, 2010; Rothstein, Jacobsen & Wilder, 2008; Schlechty, 2010; Witford & Jones, 2000). In fact, overwhelming evidence points to an educational system that has constricted, not advanced. Standardization, narrow performance expectations, gaming practices, and cheating are symptoms of schools in peril, not indicators of a healthy system capable of adapting, innovating, and improving (Rothstein, 2008; Schlechty, 2010). These realities are not likely to

change as long as a single letter grade is the standard by which school performance is judged and high stake consequences are applied.

Valid Interpretations of Results. The Standards for Educational and Psychological Testing state that validity is the most fundamental property for assessing the quality of assessments and measurements (AERA, APA, & NCME, 1999). It may appear from this statement that validity is a property of the test or measure itself, but this is not the case. Validity is a property of test and measurement use (Baker & Linn, 2002). This is an important distinction. Results of state achievement tests may yield valid interpretations of student competencies and skills, but validity erodes when aggregated achievement results become the indicator to judge school effectiveness (Linn, 2005).

Several validity concerns surfaced from our empirical results. First, a single letter grade makes student achievement less transparent and harder to interpret by hiding achievement patterns within schools. A school with a cluster of students scoring in the top quartile and a cluster of students scoring in the bottom quartile looks very different than a school where the majority of students score in the middle of the distribution. It is possible for both schools to earn an A, but their effectiveness is clearly not equivalent. The first school may be effective for a certain student population but ineffective for subgroups; whereas, the second school has a more equitable achievement distribution across students. With over 70 percent of achievement variance attributed to student differences, it is just as important to know about achievement differences within schools as it is between schools.

Second, claims that "A" schools are better, or at least more effective, than "B" and "C" schools seem indefensible given the small to virtually non-existent achievement differences among "A", "B", and "C" schools. Actual achievement differences may have more to do with chance, measurement error, or test score pollutants (e.g. gaming practices, difference in student and community characteristics, etc.) than how schools engage students in learning. With high classification error, it is even untenable to claim that all "D" and "F" schools are ineffective compared to "A", "B", or "C" schools. Letter grades used to classify school effectiveness assume achievement of schools earning the same grade is equivalent; this is clearly not the case.

Finally, letter grades that hide achievement gaps bias interpretations of school effectiveness. Effective schools promote high achievement and an equitable achievement distribution. Evidence that FRL and minority students had higher achievement in "D" and "F" schools than "A" and "B" schools calls into question the formulas used to calculate school grades. Letter grades would change if the state considered achievement gaps and achievement status equally. Several "D" and "F" schools would probably become "C" or "B" schools, and many "A" and "B" schools would probably become "C" or "D" schools. This does not mean that we excuse "D"

and "F" schools for lower average achievement, but perceptions of these schools change with evidence that "A" and "B" schools were the worst places for FRL and minority students.

Summary. One of the biggest problems affecting accurate results and valid interpretations is the use of proficiency bands to calculate student achievement and student growth. As simple as proficiency bands are to interpret, they are not a credible indicator of school performance (Forte, 2010; Ho, 2008). Changes in proficiency scores are contaminated by many variables and conditions beyond the control of schools, making it difficult to sort out contributions of schools from other factors affecting learning and achievement (Forte, 2010; Linn, 2005). Additionally, these measures have fueled unethical gaming practices like blatant cheating, targeting bubble students at the expense of lower and higher performing students, and manipulating testing samples by increasing suspensions of low achieving students, retaining lower performing students, or increasing special education placement (Figlio & Getzler, 2002; Kane & Staiger, 2002).

If letter grades are used to judge school performance, the State has a legal and ethical responsibility to ensure that grades accurately distinguish among different levels of school effectiveness. It seems unreasonable and imprudent to subject children, families, and educators to an accountability system that yields inaccurate results and fuels invalid interpretations of school performance. We would never see a new medical treatment approved if clinical trials produced mixed and inconsistent effects. Manufacturing companies could not compete locally and globally if they had high defect rates along with high customer complaints. Airplanes would be grounded if a gauge did not function properly. Policies that have consequences for the quality of learning experiences in schools should be held to higher standards than what is currently in practice, not lesser ones.

V. CONCLUSION

Letter grades promised to provide a clear and easy way to understand and measure school performance. However, when tested with individual student achievement data, we found that school letter grades do no such thing; instead they obscure achievement differences between schools and hide achievement gaps within schools. Three salient design flaws should to be addressed to develop an accountability system capable of fostering academic excellence and equity.

PROBLEM: USE OF PROFICIENCY BANDS

Proficiency bands are the consequence of cut scores that are invariably arbitrary. Across the Nation, the use of proficiency bands has been a source of extreme abuses and gaming practices. School districts have been known to assign staff members who identify students whose scores are near the cut score (known as bubble kids) so

they can coach those students to exceed the cut score (Booher-Jennings, 2005). Schools have suspended students unlikely to exceed proficiency, or placed lower performing students in special education to avert having their test scores count against the school (Heilig & Darling-Hammond, 2008). Schools have erased answers of marginal students and entered correct responses to exceed the cut score (Baker et al, 2010).

Ironically, there is no persuasive reason to use proficiency bands (Ho, 2008; Rothstein, Jacobsen & Wilder, 2008). They obscure information that is more precise in the raw test scores from which they are derived. They create arbitrary categories and promote "gaming the system" rather than deep learning. They encourage schools to focus attention on students in the middle of the achievement distribution and not the growth of all students. They are biased by school size, student differences, and other factors beyond school control (Ho, 2008, Kane & Staiger, 2002; Linn, 2005; Rothstein, 2009). Changes in percent proficiency are unstable and imprecise as a result of measurement and sampling error (Ho, 2008; Linn & Haug, 2002). In short, measures of school performance based on percent proficient are "scientifically indefensible for high stakes decisions" (Raudenbush, p. 1, 2004).

Alternative to Proficiency Bands. One alternative would be to report achievement averages and standard deviations for all students and student subgroups. Basic descriptive statistics tell us precisely how our children are performing on tests within schools and between schools (Ho, 2008). Two, the State should examine trends in achievement averages and standard deviations across multiple years to enhance the reliability of estimates and to improve interpretations of school performance (Kane & Staiger, 2002). Finally, the State should track achievement changes of individual students over time (e.g., from the beginning of the year to the end, and from year to year) (Linn, 2008).

Finland

Finland has not always had exceptional success on international assessments like the Program for International Student Assessment (PISA). Low achievement and achievement gaps were common 1970's and 1980's. Reform policies started in the 1980's have contributed to both higher achievement and a more equitable achievement distribution among students. Notably absent in Finland's reform policy is an accountability system than ranks students and schools by the percent of students who score in proficiency bands. Testing and accountability play a prominent role in the Finnish educational system, but use of assessment data differ dramatically from the United States. In Finland, students are evaluated on open-ended assessments that inform instructional and curricular decisions at the local level. Assessments emphasize problem solving and test higher order cognitive abilities. Students receive in-depth verbal and narrative feedback on their performance that is used to improve the delivery of learning for all students.

PROBLEM: USE OF A SINGLE LETTER GRADE

The notion of using the letter grade, iconic in American education, to grade schools is attractive because it appears to be a simple way to convey summary school performance. Unfortunately, unlike dimensions cannot be summarized meaningfully. Just as human height and weight cannot be summarized in a single indicator without grave error, the distinct dimensions of school performance cannot be combined. Simple indicators are truthful and accurate; simplistic indicators are just the opposite. In our analyses we found that the single letter grades reliably told us nothing about schools. However, they did classify students arbitrarily and obfuscate very important subgroup performance.

School performance is multidimensional; therefore, examining and reporting its performance must reflect this fact. Moreover, basing a school's evaluation almost entirely on test performance is deceitful, since schools are only partly responsible for the high scores of suburban schools and the low scores of urban schools (Rothstein et al, 2008). A responsible and effective accountability system uses multiple indicators, both quantitative and qualitative, to inform professional judgments of educators who are proximate to the performance needs of the school or district (Sirotnik, 2002).

Alternative to a Single Letter Grade. The solution to both of these issues is devising a school performance profile that includes indicators of standardized test performance, other outcome indicators, school process indicators, and school inputs. There can be no substitute for reporting indicators for these discrete dimensions of school effectiveness. We agree with Sirotnik (2002), " no modern organization would ever use a lone indicator to judge the worth of its operation" (p.665).

Singapore

Singapore is frequently recognized as one of the world's leaders in education. The Global Competitiveness Report (2011-12) ranked Singapore second among the world's leading education systems. According to the 2007 Trend in International Mathematics and Science Study (TIMSS) report, Singapore students were among the top in math and science scores. Singapore was also rated as one of the world's best-performing school systems by a 2007 McKinsey report.

The reformed accountability system in Singapore is based on a school self-assessment model. This model allows school leaders to facilitate the school improvement process. Schools are assessed on nine measures of quality which include five process criteria that enable school improvement: leadership, strategic planning, staff management, resource management, and student-focused processes; and four outcome criteria: administrative results, staff results, partnership and society results, and key student performance results (Ng, 2008).

Schools are required to submit detailed evidence to justify their self-assessment scores, and every five years the Ministry of Education conducts a review

to confirm the scores. Under the self-assessment system, the public ranking of schools was replaced with a banding scale that allows schools to compare their scores with similarly situated schools (Ng. 2008). The public's attention is also directed toward the school awards system. The awards are based on the various performance criteria, with some recognizing successes in the process criteria and others recognizing outcome criteria. This form of public reporting recognizes schools for their achievements, rather than their failings (Ng, 2008).

PROBLEM: HIGH STAKES USE OF LETTER GRADES

It is a myth to think that using student test scores to punish or reward schools is a driver of improvement (Baker et al., 2010). Countries scoring at the top of international assessments like PISA and TIMSS measure student and school outcomes, but results are not used to punish or reward teachers and schools (Fullan, 2011). Low achievement or a lack of improvement does not stem from a lack of will or motivation; instead, it results from limited capacity. State letter grades and interventions triggered by low grades fail to address the lack of capacity in many Oklahoma schools and communities. Even if grades were an accurate and valid indicator of school effectiveness, it is hard to understand how letter grades, with their narrow focus on outcomes, could reveal much valuable information about reasons for low achievement and/or existing achievement gaps.

Alternative to High Stakes Use. How can Oklahoma shift the focus of the A-F Report Card from high stakes to capacity building? Darling-Hammond (2005) advances three functions of effective reform policies that serve as a useful guide. First, the accountability system needs to facilitate extensive learning opportunities for school professionals, parents, and community members. Second, policies should allow for widespread engagement in the process of developing and enacting theories of change. Third, policies need to structure an effective balance between external pressure and local autonomy. It is hard to envision the current A-F Report Card being capable of carrying out the above functions without significant changes to its methods of calculating grades, the type of performance information gathered, and the use of data.

Ontario, Canada
The Education Quality and Accountability Office for the Ministry of Education in Ontario tracks cohorts of matched students from grades 3, 6, and 9 in reading, writing, and math. Achievement data are used to describe achievement status, achievement changes, and achievement trends. Moreover, longitudinal tracking of matched students enables analysis to determine if students maintain, advance or drop in achievement status between 3rd through 9th grades. Ontario's accountability system relies on less testing; they use testing as an improvement

strategy, not for punishment or reward; and they wrap results around professional capacity building initiatives.

VI. Recent Legislative Changes to A-F: What do they do?

A critique of the initial implementation of the A-F accountability calculations raised concerns regarding their statistical trustworthiness. While some concerns have been addressed in the new method, there has been little substantive, overall improvement. In our earlier critique, we pointed to problems with over-reliance on student standardized test scores; instead of addressing this difficulty, the new approach actually increases reliance on student achievement data from 67% to 100% (or 91% of you consider that bonus points make the total possible points 110, rather than 100).

Student Achievement Component. The Student Achievement component of last year's original A-F was based on a conversion of proficiency levels to scores of 0, .2, 1, and 1.2 for unsatisfactory, limited knowledge, proficient, and advanced, respectively. An index based on the weighted average was formed, which ranged from 0 to 120 points; the indices were then categorized into letter grades, which were then reassigned a point value. It was this final point value ranging from 0-4 that was ultimately used to represent the student achievement component in calculating the report card grade. Our criticism was leveled at the arbitrariness of the 0 to 1.2 scoring system and the over manipulation of the data. We also faulted the approach for its use of proficiency levels rather than continuous scores, ignoring the variability of student performance within the proficiency levels.

In contrast, the new system dichotomizes student achievement proficiency levels. The achievement index is based simply on the percentage of test scores that are proficient or advanced across all tests within a school. The packaging and repackaging of scores is eliminated and the calculation of the school grade is simplified. However, in this simplification lurks an old problem—the compounded grouping error resulting from treating test data in two proficiency categories as equivalent. Ignored within group variability is actually much greater than in the previous system. All students scoring as limited or unsatisfactory receive a score of zero; all students scoring as proficient or advanced receive a score of 1. So all achievement variability in the Performance Index is reduced to a simple percentage of proficient test scores across all tests.

Not only is grouping error compounded, but legislative changes continue to hide within-school achievement gaps and do not address classification error. The problems can be seen in the within-school achievement distribution of students in the example school depicted in Figure 12.

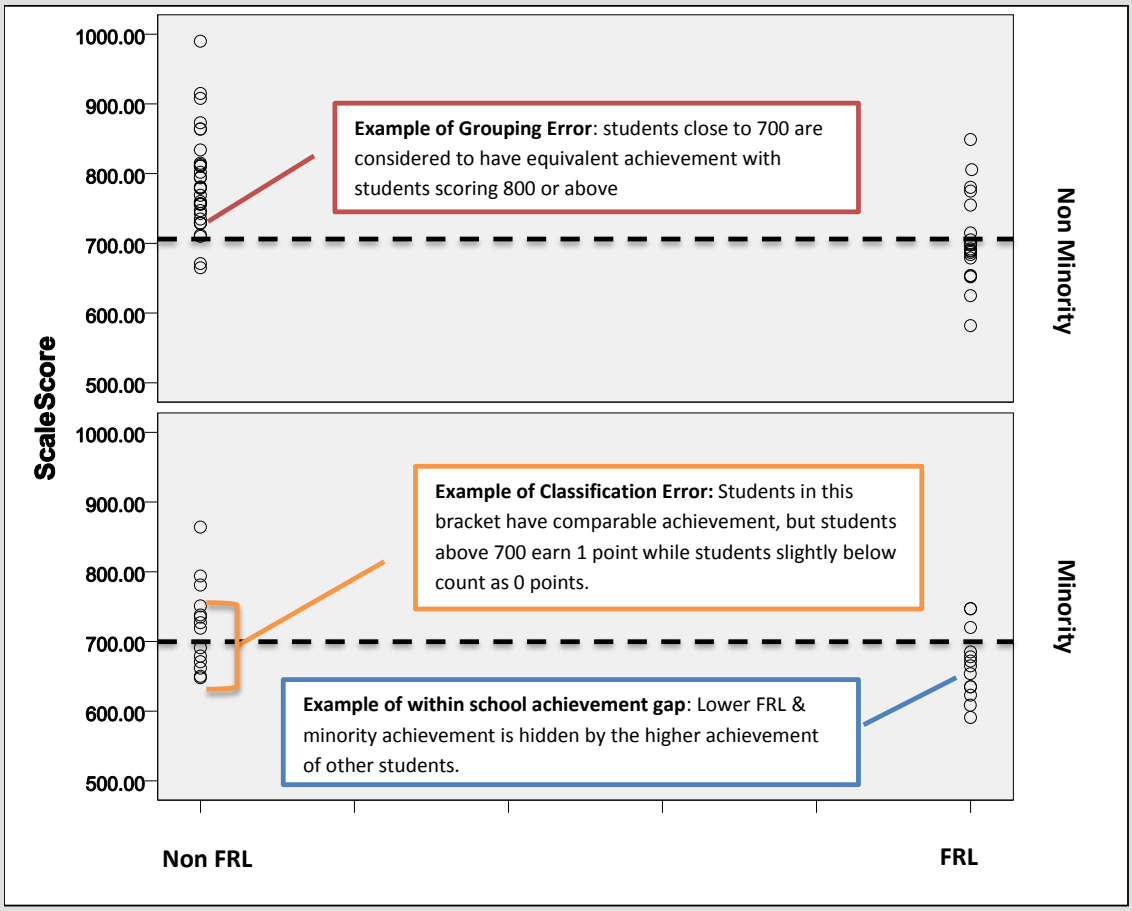


Figure12. Problems associated with using proficiency bands as the basis for letter grades. Circles represent individual student scores. The dashed line is the proficiency threshold

The following concerns in the Student Achievement component remain:

- Continued reliance on percent proficiency scores makes grades sensitive to factors that are unrelated to school performance or teaching effectiveness.
- The use of dichotomized proficiency levels rather than test scores compounds grouping error.
- Within proficiency level improvement is not recognized in the growth index.
- Student mobility within and across districts affects growth interpretations.
- Mathematical properties of the growth index are unknown.
- Conceptually the growth index is not meaningfully tied to growth.

Student Growth Component. Under the new system, the Growth Index, unlike the Student Achievement Index, distinguishes among the proficiency levels to some degree. Any increase in proficiency level is awarded 1 point even if the student increased two or three proficiency levels. So, a distinction is made between unsatisfactory and limited knowledge, as this increase would be worth 1 point.

However, students who increase from unsatisfactory to proficient are also awarded just 1 point, so in this case no distinction is made between limited knowledge and proficient. This inconsistency in the handling of proficiency level differences in the different parts of the grading system affects the interpretation of the final grade. The growth index is the percentage of test scores that show an increase of at least 1 proficiency level from the previous year in math and reading. Its calculation is simplified from last year. However the following, serious concerns remain:

- Within proficiency level improvement is not recognized.
- The vertical equating of the tests being compared has not been established.
- Student mobility within and across districts affects interpretation of growth.
- Mathematical properties of the index are unknown.
- Conceptually this index is not meaningfully tied to growth.

Handling of growth in the bottom 25% resembles the method used in the overall growth measure except for two things. If fewer than 10 matched pairs of scores can be found, the Overall growth measure will account for 50% of the entire report card grade. We pointed out last year that n=30 was probably an inadequate sample size given the multiple sources of measurement error that affect these growth indicators. Secondly, if students who scored unsatisfactory or limited proficiency on the previous test do not increase a level but have improved within a level more than the statewide average of positive growth on the OPI they earn one point. The calculation of the statewide OPI is still an issue. Nothing has changed in this regard.

In other matters, the legislation leaves implementation rule development up to the state department. As an example, the minimum number of tested students required for a school to receive a grade has been changed from 30 to 10. The legislation originally a required sample size to be "based on accepted professional practice for statistical reliability and prevention of unlawful release of personally identifiable student data." That requirement was stricken from the new legislation. The SDE was instead directed to "establish the lowest minimum sample size necessary to meet the requirements of this paragraph." The selection of N=10 as a minimum sample size is well below any accepted professional standard as a basis for group statistics, especially in high-stakes situations. The influence of each student's idiosyncratic test performance has an undue effect on the calculations, especially percentage-based calculations such as used in the A-F system. Further, any form of disaggregation may jeopardize confidentiality of individual student data.

Bonus Points. The use of bonus points mandated by Oklahoma's new A-F legislation regarding the A-F school rating system raises some concerns, especially with respect to equity. For example, schools are rewarded bonus points for attendance, graduation rates, and student performance (i.e., number of students in AP courses,

etc.). However, attendance and aforementioned variables are all correlates with SES (Hogrebe & Tate, 2012). Therefore, schools that are already graded higher (e.g., A or B) will likely also be rewarded with bonus points, whereas, schools graded lower (e.g., C, D, or F) will likely not receive points. Thus an inherent confounding variable (i.e., SES) exists and schools will be rewarded un-equally, due to factors beyond a school's control.

In addition, the attendance rate is not weighted equally across types of schools (i.e., elementary, middle, and high school configurations). That is, the only bonus points elementary schools can receive are from attendance, whereas middle and high schools receive six and five bonus points, respectively and other criteria govern the awarding of bonus points. Consequently, elementary schools are inequitably awarded bonus points for a single criterion, whereas middle and high schools can receive bonus points based on multiple criteria. A more equitable approach would allow attendance to be worth five bonus points in all schools; the other five points can be rewarded for meeting other criteria. We understand elementary and middle schools do not have AP courses and participation in college entrance exams, but other "school indicators" could be used to reward schools equitably in all configurations.

With the issue of attendance, we believe a gradient scale would be more practical than an "all or nothing" reward, allowing attendance rate classification to be equal across all types of schools. Elementary and middle schools have to have a 94% attendance rate to be rewarded 5 points whereas a high school only needs 90%. This should be equalized, or at least a strong rationale provided as to why the inconsistent rates exists. An example clarifies. If an elementary school has a 93% attendance rate it is NOT awarded any additional points, however school with a 94% attendance rate receives all of the points allowed. In some small schools this could be the result of a few children who do not attend, whereas in large schools it could be several more students. In this way, the threshold for attendance rate can negatively affect smaller schools more than larger ones. A possible approach would be that schools who have 95% or higher would receive all bonus points (5 as suggested earlier for all types of schools), schools 90%-94.9% receive 4 bonus points, schools 85%-89.9% receive 3 bonus points, etc. In summary, the bonus points might help schools achieve a higher grade, but further discussions are needed to determine if the technical and practical complexities surrounding the use of bonus points outweigh the potential inequities.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC.

Baker, E. L., & Linn, R., L. (2002). *Validity issues for accountability systems. CSE Technical Report 585*. National Center for Research on Evaluation, Standards, and Student Testing. University of California, Los Angles.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F. Linn, R. L., Ravitch, D. Rothstein, R. Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers.* The Economic Policy Institute (Retrieved: http://epi.3cdn.net/724cd9a1eb91c40ff0_hwm6iij90.pdf).

Booher-Jennings, J. (2005). Educational triage and the Texas accountability system. *American Educational Research Journal, 42*(2), 231-268.

Darling-Hammond, L. (2005). Policy and change: Getting beyond bureaucracy. In Hargreaves, A. (Ed.), *Extending Educational Change: International Handbook of Educational Change* (pp. 362-387). Dordrecht, Netherlands: Springer.

Figlio, D. N., & Getzler, L. S. (2002). *Accountability, ability and disability: gaming the system*. Working paper 9307. Cambridge, MA: National Bureau of Economic Research.

Forte, E. (2010). Examining the assumptions underlying the NCLB federal accountability policy on school improvement. *Educational Psychologist*, 45(2), 76-88.

Hamilton, L.S., Schwartz, H.L., Stecher, B.M., & Steele, J.L. (2013). Improving accountability through expanded measures of performance. *Journal of Educational Administration, 51*(4), 453-475.

Heck, R. H. (2009). Teacher effectiveness and student achievement: Investigating a multilevel cross-classified model. *Journal of Educational Administration, 47*, 227-249.

Heilig, V. J., & Darling-Hammond, L. (2008). Accountability Texas-style: The progress and learning of urban students in a high-stakes testing context. *Educational Evaluation and Policy Analysis, 30*(2), 75-110.

Ho, A D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*(6), 351-360.

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives, 16*(1), 91-114.

Linn, R. L. (2008). Educational accountability systems. In K. Ryan and L. Shepard (Eds.), *The future of test-based educational accountability* (pp. 3-24). New York, NY: Routledge.

Linn, R. L. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives, 13*(33), 1-20.

Linn, R. L., & Haug, C. (2002). *Stability of school building accountability scores and gains. CSE Technical Report 561*. National Center for Research on Evaluation, Standards, and Student Testing. University of California, Los Angeles.

Ng, P. T. (2008). Educational reform in Singapore: From quantity to quality. Educational Research for Policy and Practice, 7, 5-15.

Nye, B., Konstantopoulos, S. & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237-257.

Raudenbush, S. W. (2004). *Schooling, statistics, and poverty: Can we measure school improvement*? Princeton, NJ: Educational Testing Service.

Rothstein, R. (2009). Getting accountability right. Education Week. Available on-line at: http://www.csun.edu/~krowlands/Content/SED610/reform/Getting%20Accountability%20Right.pdf

Rothstein, R., Jacobson, R., & Wilder, T. (2008). *Grading education: Getting accountability right.* New York, NY: Teachers College.

Schlechty, P. C. (2010). *Leading for Learning: How to transform schools into learning organiztions*. San Francisco, CA: Wiley.

Schwartz, H. L., Hamilton, L. S., Stecher, B. M., & Steele, J. L. (2011). *Expanded Measures of School Performance.* Technical Report: Rand Corporation.

Sirotnik, Kenneth A. (2002). Promoting Responsible Accountability in Schools and

Education. *The Phi Delta Kappan*, 83(9), 662-673.

Tate, W. F & Hogrebe, M. (2012). Place, poverty, and Algebra: A statewide comparative spatial analysis of variable relationship. *Journal of Mathematics Education at Teachers College, 3, 12-24*.

Whitford, B. L., & Jones, J. (2000). *Accountability, Assessment, and Teacher Commitment: Lessons from Kentucky's Reform Efforts*. New York: State University of New York.

Appendix A: Robert Linn Comments

Overall Evaluation

The report clearly demonstrates several major limitations of the A-F report card system used by Oklahoma for school accountability. These limitations are serious and result in an accountability system with inadequate validity for the intended uses and interpretations. Although the inadequacies should be apparent to anyone who has studied testing and accountability systems, the report provides empirical evidence illustrating several of the major flaws in the system. The report's conclusions are well justified. The report provides a strong rationale for scraping the current A-F system and developing a more valid accountability system for Oklahoma. As the report suggests, some of the features of an improved accountability system would include:

1) a profile of achievement in the tested subject areas for a school rather than a single letter grade,
2) a profile of achievement gains
3) the use of test scale scores rather than dichotomized proficiency levels,
4) an appraisal of achievement gaps within a school, and
5) the inclusion of indicators of school characteristics and performance other than those based on student test scores.

Appendix B: Technical Description of Analytical Process

We used a conventional multilevel model-building process in HLM 7.0 to explore differences in student achievement between A, B, C, D, and F schools. The purpose was to evaluate achievement differences between students after accounting for variance that cannot be attributed to teachers or school performance.

Random Effects ANOVA. We first decomposed achievement differences to within school and between school components with an unconditional random effects ANOVA. Results were used to calculate the IntraClass Correlation Coefficient (ICC), the percent of achievement variance attributed to school and non-school factors.

Level 1: $Ach_{ij} = \beta_{0j} + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + u_{oj}$

$P = \sigma^2_{uo} / \sigma^2_{uo} + \sigma^2_{eo}$

Random Coefficient Regression. We tested the effects of student characteristics on achievement with a Random Coefficients regression. Student variables were grand-mean centered. Grand-mean centering has a computational advantage over group-mean centered or un-centered models in that it controls for any shared variance between individual and group level predictors. Dummy coding was used for minority status, gender, and FRL status. Significant student variables were retained and set to vary randomly across schools. Non-random student effects were fixed to the school level.

Level 1: $Ach_{ij} = \beta_{0j} + \beta_{1j}(\text{Minority Status}_{ij}) + \beta_{2j}(\text{Grade}_{ij}) + \beta_{3j}(\text{FRL Status}_{ij}) + \beta_{4j}(\text{Gender}_{ij}) + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + u_{oj}$

$\beta_{1j} = \gamma_{01} + u_{oj}$

$\beta_{2j} = \gamma_{02} + u_{oj}$

$\beta_{3j} = \gamma_{03} + u_{oj}$

$\beta_{4j} = \gamma_{04} + u_{oj}$

Random Coefficient Slopes and Intercepts as Outcomes. The final step was to test a random coefficient slopes and intercepts as outcomes model with all significant student and school variables. All student and school level predictors were grand-mean centered. School GPA was used to explain variation in the minority and FRL

slopes. These models had less error and best fit with the data. In other words, they provided an unbiased assessment of mean differences between A, B, C, D, and F schools. Estimates represent the actual difference in raw scores after controlling for factors not related to teaching practices and school performance.

Level 1: $Ach_{ij} = \beta_{0j} + \beta_{1j}(\text{Minority Status}_{ij}) + \beta_{2j}(\text{Grade}_{ij}) + \beta_{3j}(\text{FRL Status}_{ij}) + \beta_{4j}(\text{Gender}_{ij}) + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(B) + \gamma_{02}(C) + \gamma_{03}(D) + \gamma_{04}(F) + \gamma_{05}(\%\ \text{Minority}) + \gamma_{06}(\text{Prior Achievement}) + \gamma_{07}(\text{FRL rate}) + u_{0j}$

$\beta_{1j} = \gamma_{00} + \gamma_{11}(\text{GPA}) + u_{0j}$

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

$\beta_{3j} = \gamma_{00} + \gamma_{31}(\text{GPA}) + u_{0j}$

$\beta 4_{j} = \gamma_{00} + u_{0j}$

β_{0j} = is the school achievement mean for math achievement
β_{1j} = Minority achievement gap
β_{2j} = distributive effects of grade
β_{3j} = FRL achievement gap
β_{4j} = Gender achievement gap
γ_{00} = grand mean for achievement
γ_{01} = is the difference in average achievement between A schools and B schools
γ_{02} = is the difference in average achievement between A schools and C schools
γ_{03} = is the difference in average achievement between A schools and D schools
γ_{04} = is the difference in average achievement between A schools and F schools
γ_{05} = is the effect of school % Minority on achievement
γ_{06} = is the effect of prior school achievement on student achievement
γ_{07} = is the effect of FRL rate on student achievement
γ_{11} = cross-level interaction of minority achievement and school GPA
γ_{31} = cross-level interaction of FRL achievement and school GPA