An Examination of the Oklahoma State Department of Education's

A-F Report Card

Commissioned by OSSBA and CCOSA

A report produced by staff of

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

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

January 2013*

The analyses reported in this publication were supported by OSSBA and CCOSA.

Contributing authors:

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

- Curt M. Adams, Senior Research Scientist
- Ellen Dollarhide, Research Associate
- Patrick B. Forsyth, Senior Research Scientist
- Gaetane Jean-Marie, Senior Research Scientist
- Phillip Garland, Research Associate
- Ryan Miskell, Research Associate
- Jordan Ware, Research Associate

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

- Laura L.B. Barnes, Senior Research Scientist
- Mwarumba Mwavita, Senior Research Scientist

Two internationally known experts reviewed and critiqued the analyses and recommendations of the staff report:

Robert Lee 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.

Robert J. Sternberg

Robert Jeffrey Sternberg is an American psychologist and psychometrician; currently he serves as George Kaiser Family Foundation Professor of Ethical Leadership, Provost and Senior Vice President, and Regents Professor of Psychology and Education, Oklahoma State University. Previously he was Dean of Arts and Sciences at Tufts University, IBM Professor of Psychology and Education at Yale University, and the President of the American Psychological Association. He is a member of the editorial boards of numerous journals, including American Psychologist. Sternberg has a BA from Yale University and a PhD from Stanford University. He holds ten honorary doctorates from one North American, one South American, and eight European universities, and additionally holds an honorary professorate at the University of Heidelberg, in Germany. He is currently also a Distinguished Associate of The Psychometrics Centre at the University of Cambridge. Among his major contributions to psychology are the Triarchic Theory of Intelligence, several influential theories related to creativity, wisdom, thinking styles, love and hate, and is the author of over 1000 articles, book chapters, and books.

Table of Contents

An Examination of the Oklahoma State Department of Education's A-F Report Card

Executive Summary

There is an undisputed need to assess public schools to determine their effectiveness. Oklahoma responded to this need by legislating an assessment system intended to be comprehensive, while at the same time, understandable and transparent, using A-F grades as the reporting outcome. Strengths of the assessment system are its inclusion of student achievement in a breadth of content areas, a measure of growth for low achieving students, and consideration of multiple artifacts in whole school performance. However, the effort of the state to report school quality using the familiar letter grade, while laudable, falls short of providing a clear and credible picture of individual school performance for a variety of reasons outlined in this report.

It is our goal to support the good intentions of Oklahoma policymakers in their school improvement efforts by identifying methods of the grading system that may be potentially misleading (Baker & Linn, 2002). Some problems with the A-F Report Card are unique to methods used by the Oklahoma State Department of Education (OSDE) to calculate student achievement, student growth, and school performance indices. Other problems are longstanding conceptual and methodological constraints associated with aggregated test scores as measures of school performance. Although achievement data are obviously important for assessing schools, an accountability grade based almost exclusively on test scores does not account for numerous critical factors that contribute to school performance.

Performance measurement and accountability systems in many sectors take a more balanced approach to assessment. In other enterprises, executives and managers are not expected to make strategic decisions based on outcome data alone. Healthcare, for example, has made great strides adopting scientific process and outcome measures for evaluation. Manufacturing forecasts future profitability using measures of customer satisfaction, demand, internal processes, and innovation and growth, relying on much more than past financial reports. It is now standard accounting practice to evaluate companies as much by their intangible resources as by their physical capital. Sole reliance on outcome indicators produces biased assessment and does not depict fairly or accurately how school leaders and teachers respond to the dynamic needs of students.

Accountability systems are only useful if their measures are credible and clear. Despite good intentions, the features of the Oklahoma A-F grading system

produce school letter grades that are neither clear, nor comparable; their lack of clarity makes unjustified decisions about schools. Further, A-F grades are not productive for school improvement because they do not explain the how or why of low performance. Building on what has already been done, Oklahoma can and should move toward a more trustworthy and fair assessment system for holding schools accountable and embracing continuous, incremental improvement.

Statistical Trustworthiness

All three components of the A-F System have statistical limitations that jeopardize their validity, reliability, and usefulness. Moreover, combining flawed indicators creates a misleading measure in the form of a single grade for each school. The most troubling concerns found with the letter grade approach are summarized below and described in more detail in the full report.

Student Achievement Component

- The scores assigned to represent proficiency levels (0, .2, 1.0, 1.2) do not seem to correspond to any recognizable metric.
- The metric does not justify the mathematical manipulations performed in the A-F scaling.
- The use of proficiency levels rather than test scores in these computations introduces grouping error.
- Information is lacking regarding classification consistency (reliability).
- Basis for letter grade conversion is undocumented.

Student Growth Component

- Within proficiency level improvement is not recognized.
- Student mobility within and across districts affects interpretation of growth.
- The metric behaves unpredictably as a basis for assigning grades.
- Unclear mathematical properties of the index.
- Information is lacking regarding gain score reliability.
- Unclear conceptual meaning of the index.

Whole School Performance Component

- Overreliance on attendance and graduation rates.
- Attendance and graduation rates are known to be correlated with socioeconomic status.
- Graduation rate calculation is not in compliance with the Federal guidance formula and the State Law (SB 2), thus current rates are erroneous and misleading.

Practical Consequences of the Evaluation System:

The current A-F Report Card design produces several practical challenges for districts and schools. Criteria of effective assessment for accountability identified in the proposed Standards for Educational Accountability Systems (Baker, Linn, Herman, & Koretz, 2002) and the Standards for Educational and Psychological Testing (AERA et al., 1999) expose specific elements of the accountability system that prevent or inhibit its use in capacity building and school improvement.

- By not making explicit threats to the validity of report card grades, the OSDE misinforms the public about the credibility and utility of the A-F accountability system.
- Performance information from the current A-F Report Card has limited improvement value; particularly, it is not useful for diagnosing causes of performance variation.
- The summative aspects of the accountability system overshadow formative uses of assessment and performance.
- High stakes testing, as a cornerstone of school assessment and accountability, corrupts instructional delivery by focusing effort on learning that is easily measured.

Recommendations:

1) Report school performance like school report cards that provide indicators of performance periodically and in multiple areas over the course of a year. **Action: Develop a report card format that uses multiple school indicators that more adequately reflect a school performance profile.** Eliminate the single grade, which cannot be composed without adding together unlike elements and promoting confusion and misunderstanding.

2) Develop a balanced performance measurement plan that aligns with strategic goals of schools. Track school indicators (i.e., inputs, process, and performance outputs) longitudinally to understand growth or stasis. **Action: Rely on trend lines of both process and outcome indicators over the year and multiple years to determine growth in school performance.** Growth indicators for different subject content should not be co-mingled to create single growth estimates for a whole school.

3) Include valid and reliable measures of school climate, motivation, and the dispositions of school role groups longitudinally. **Action: Promote the use of valid and reliable measurement of process variables at the district and school level, to be used by schools in their improvement plans.**

4) Any accountability system should be upfront about its limitations and uses.
Action: Make explicit the limitations of the accountability system and warn of its inappropriate use for high-stakes decision-making.

5) Embed assessments in instruction throughout the year.
Action: Legitimize the process by embedding assessment throughout the school year. Embedded, assessment data can be used to change course and make adjustments when needed; such assessments will be viewed as having a formative (improvement) purpose instead of a summative judgment about the past year.

6) The accountability system is important to our State and it is essential that it be credible, accurate, and supported by policy-makers, school professionals, and the people of Oklahoma.
Action: Take the time to enlist the services of assessment and evaluation experts who can objectively build an exemplary Oklahoma accountability system directed at incremental and continuous school improvement.

An Examination of the Oklahoma State Department of Education's A-F Report Card

I. Introduction

Oklahoma, like other states, has introduced a system for reporting individual school performance as part of a general accountability plan outlined in its ESEA waiver application. In response to a request from the Oklahoma School Boards Association (OSBA) and the Cooperative Council for Oklahoma School Administration (CCOSA), research staff from the Oklahoma Center for Education Policy (OU) and the Center for Educational Research and Evaluation (OSU) have collaborated to examine the State's A-F Report Card assessments for purposes of seeking to improve them. Baker and Linn (2002) cogently outlined the rationale for this task:

> What is expected to focus the energy of the people in classrooms and schools to do now what they have been unwilling or unable to do before—that is, to systematically improve learning for students who have done poorly in the past? There is a belief that the power of incentives and sanctions will come into play and organize attention in the desired direction. Of concern to us as observers is that the rewards and sanctions may indeed focus attention on the bottom line, but not on needed steps or processes to get there. A lack of capacity (whether through selection, turnover, or inadequate professional development and resources) cannot be directly remedied by increased motivation to do well, especially over a short period. The central notion of the validity of accountability systems herein resides. **Accountability systems intending to promote real learning and improved effectiveness of educational services must themselves be analyzed to ensure that changes in performance (the proverbial bottom-line) are real, are due to quality instruction plus motivation, and sustainable, and can be attributed to the system itself. (pp. 3-4)** (bold added).

In Part II, we examine the evaluation system's indicators and formula for calculating a performance index and grade. The three primary components of the system (student achievement, individual student growth, and whole school performance) are scrutinized for their fidelity to sound principles of measurement, including reliability and validity. The methods used are described and concerns are outlined. Lastly, the formula for composing the school grade is examined with an eye to improving its trustworthiness and utility for school improvement.

In Part III, the practical implications and consequences of the enacted

accountability plan are explored. Shulman's criteria (Pillars) for effective assessment for accountability are applied to the Oklahoma A-F Report Card as an analytical template for discovering weaknesses and opportunities for improvement. The criteria call for clarity and transparency, use of multiple measures, embedded assessments, and a focus on formative rather than a high stakes use of accountability measures.

In Parts IV and V, implications, recommendations, and conclusions are presented.

II. Statistical and Measurement Analysis

The foundation of the A-F Grade is the Oklahoma State Testing Program. Sixty-seven percent of the grade is based on achievement testing. The remaining 33% is based primarily on attendance for elementary and middle school, and on graduation rate for high school. According to materials published by the State Department of Education, these tests include Oklahoma Core Curriculum Tests (OCCT), End-of-Instruction Exams (EOI), Oklahoma Modified Alternative Assessment Program (OMAAP), and Oklahoma Alternative Assessment Program (OAAP). The latter two programs are designed for special needs students—the OMAAP is a somewhat modified version of the OCCT and the OAAP is a portfolio assessment system for students with profound needs. Every tested content area is included in the assessments —Reading, Math, Science, Social Studies, History, Geography, Writing, Algebra 1, Geometry, Algebra 2, English 2, English 3, Biology, and US History Exams.

There are 3 components to the A-F Grade: Student Achievement, Student Growth, and Whole School Performance. Achievement and Growth are test-based. In the following sections we provide a description of the methodology for calculating each grade component and identify issues concerning the methodology.

Component 1: Student Achievement—33%

Methodology

Student achievement on all tests is reported by proficiency levels based on previously established cut scores. There are 4 levels: Unsatisfactory, Limited Knowledge, Proficient, and Advanced. The Achievement component begins with the Performance Index (PI), which is computed in the following steps:

Step 1. Assign a score to represent test performance within each proficiency level for each test. A proficiency level score of zero is assigned to represent the range of test performance within the Unsatisfactory category level, a score of 0.2 to all scores associated with Limited Knowledge, 1.0 to Proficient scores, and 1.2 to Advanced scores.

Step 2. Weight each proficiency level score by the number of students in that proficiency level aggregated across all content areas and across all grades (e.g., 6^{th}, 7^{th}, 8^{th}) in the school. For example, Table 10 reproduced from the A-F Report Card Guide shows that the numbers of students with Limited Knowledge in a hypothetical middle school were 15 on the math test, 45 on the reading test, 5 on the science test, 20 on US History, 15 on Geography, 5 on Writing, and 5 on Algebra 1. Thus, 110 students received a weight of 0.2 and the contribution to the Performance Index of the Limited Knowledge scores is .2*110=22.

Table reproduced from OSDE *A-F Report Card Guide*, April 2012.

TABLE 10: Example of Middle School Performance Index Calculation

Subject	Number Tested	Number Limited Knowledge	Number Proficient	Number Advanced	Index Calculation	Letter Grade
Mathematics	300	15	220	60	((15 * 0.2) + (220 * 1) + (60 * 1.2)) / 300	98 = A
Reading	300	45	195	40	((45 * 0.2) + (195 * 1) + (40 * 1.2)) / 300	84 = B
Science	90	5	75	10	((5 * 0.2) + (75 * 1) + (10 * 1.2)) / 90	98 = A
US History	90	20	60	3	((20 * 0.2) + (60 * 1) + (3 * 1.2)) / 90	75 = C
Geography	110	15	80	10	((15 * 0.2) + (80 * 1) + (10 * 1.2)) / 110	86 = B
Writing	90	5	80	5	((5 * 0.2) + (80 * 1) + (5 * 1.2)) / 90	97 = A
Algebra I	30	5	23	2	((5 * 0.2) + (23 * 1) + (2 * 1.2)) / 30	88 = B
Performance Index	**1010**	**110**	**733**	**130**	**((110 * 0.2) + (733 * 1) + (130 * 1.2)) / 1010**	**90 = A**

Step 3. Perform these computations for each proficiency level and sum the products across all proficiency levels. Numbers of Unsatisfactory students are not shown in the tables though they are included in the calculations with a score of zero.

Step 4. Divide this sum by the total number of examinees across all the content areas resulting in an index ranging from 0 to 1.2

Step 5. Multiply this value by 100 to yield the Performance Index (PI) with a range of 0 to 120.

The following 2 steps show the conversion of the Performance Index for use in computing the report card grade.

Step 6. Categorize the PI into a Letter Grade as follows: 90-120 = A; 80-89=B; 70-79=C; 60-69=D; below 60=F.

Step 7. Convert the letter grade to a point value A=4, B=3, C=2, D=1, F=0.

This point value is then multiplied by the achievement category weight of .33 in computing the overall school GPA.

Concerns

The scores assigned to represent proficiency levels (0, .2, 1.0, 1.2) do not seem to correspond to any recognizable metric. We could locate no written documentation for the origin of these score values. It seems likely that the original plan was to dichotomize the proficiency levels for the A-F system, collapsing unsatisfactory and

limited knowledge into one unsatisfactory category with a score of zero, and proficient and advanced into a proficient category with a score of 1. Perhaps a decision was made to award some amount of extra credit to the advanced category, thus the score of 1.2 and a similar .2 bonus to the limited knowledge category, though one wonders why .2 was selected rather than some other value.

The metric does not justify the mathematical manipulations performed in the A-F scaling. For data to be mathematically manipulated in the way described in the A-F Report Card Guide (OSDE, 2012) it needs to be on an interval or ratio scale. An equal interval scale is defined when "equal numerical differences in scores represent equal differences in the property being measured" (Thorndike & Thorndike-Christ, 2010, p. 27). However, the scale values are not equidistant and there is no evidence that the achievement difference between the categories matches the assigned point-value difference between the categories (Crocker & Algina, 1986). There is no clear justification for the achievement difference between the four proficiency levels. Why is the difference between limited knowledge and proficient (difference=.8) four times greater than the difference between limited knowledge and unsatisfactory (difference=.2) or between proficient and advanced (difference=.2)? Why is the one point difference between advanced and limited knowledge 1.25 times greater than the .2 point difference between proficient and limited knowledge? The implications for this differentially weighted scoring system are tremendous but supportive evidence is lacking.

The use of proficiency levels rather than test scores in these computations introduces grouping error. The decision to use proficiency levels apparently arose from the need to have a common metric for aggregating across content categories. The reliance on assigning score values to proficiency levels introduces grouping error (King & Minium, 2003). Grouping error occurs when data are aggregated into categories or levels and all score values within the category or level are treated the same. The use of proficiency levels when continuous scores are available amounts to throwing away information about examinee test performance. The utility of the disaggregated test scores for analysis depends to some extent on the test specifications and method of construction. If the test was designed specifically to permit the establishment of cut scores, there should be somewhat greater score variability between proficiency groups and somewhat less within the groups. Nevertheless, the score variability within proficiency groups ought to be taken into account by using the original test scores or some derived function of them that preserves the within group variability.

The use of proficiency levels rather than test scores in these computations ***also*** *introduces measurement error due to classification inconsistency.* According to the Standards for Educational and Psychological Testing, "When a test or composite is used to make categorical decisions, such as pass/fail, the standard error of measurement at or near the cut scores has important implications for the

trustworthiness of these decisions." (AERA et al., 1999 p. 35). The percentage of consistent classifications should be made through the use of a repeated-measurement approach. This information was not available in the A-F Guide nor, to our knowledge, in materials released to the public.

Basis for letter grade conversion is undocumented. The Performance Index ranges from 0 to 120. Letter grades are assigned such that 90-120 is "A", 80 is the cutoff for "B", and so forth. This may be grounded in a perception that the index scores should correspond to percentage grades which are commonly derived by assigning an "A" to a score that represents 90% of perfection (usually operationalized as total points possible); 80% of perfection is the cutoff for a B, etc. However, here the total points possible are not 100; they are 120. So, if this were the basis for the grade assignments, the letter grade cutoffs are set too low.

Component 2: Individual Student Growth—17%

Methodology

In Math and Reading, students' performance levels on this year's and the most recent prior year's tests are compared. Students are awarded points depending on whether their proficiency level increased, decreased, or stayed the same. Zero points are awarded to cases that decreased proficiency levels or to cases of Unsatisfactory or Limited Knowledge that stayed the same. One point is awarded for a one-level gain or to staying in the same level of Proficient or Advanced; 2 points for a 2-level gain; and 3 points for a 3 level gain. Also, students who score in the Unsatisfactory and Limited Knowledge levels whose OPI is above the state average are awarded a point.

TABLE 17: Calculation of Points for Mathematics

Calculation of Points for Mathematics	Number of Students	Point Value	Calculation	Points
Number Proficient or Advanced Remaining Proficient or Above	150	1	150 x 1	150
Number of Unsatisfactory Improving to Limited Knowledge	10	1	10 x 1	10
Number of Unsatisfactory Improving to Satisfactory or Proficient	6	2	6 x 2	12
Number of Unsatisfactory Improving to Advanced	0	3	0 x 3	0
Number of Limited Knowledge Improving to Satisfactory	20	1	20 x 1	20
Number of Limited Knowledge Improving to Advanced	4	2	4 x 2	8
Number with OPI Growth Greater than State Average	8	1	8 x 1	8
Total Points				**208**
Total Number of Students	**246**			

Table reproduced from OSDE A-F *Report Card Guide*, April, 2012

The point value is multiplied by the number of exam pairs having that point value. This product is summed across the different categories of improvement and aggregated across the Math and Reading contents. The result is divided by the total number of exam pairs to form the school's growth index. The index can range from 0-300 and is then categorized into letter grades as follows: 90-300=A, 80-89=B, 70-79=C, 60-69=D, below 60=F. The letter grades are then converted to a GPA scale for purposes of including it in the weighted calculation of the overall school GPA. This is illustrated in the OSDE Table 17 above.

Concerns

Within proficiency level improvement is not recognized. Students who improve their test scores enough to change performance levels are awarded points. Students whose improvement is just short of that required to change to a higher proficiency level are treated as if they had **not** improved under this system. Further, because no change and negative change are accorded the same zero point value, the rules would theoretically treat as equivalent, the performance of students who improved from the bottom score to the top score within a proficiency level and students who declined from the top score in Advanced to the bottom score in the Unsatisfactory level. Though an unlikely occurrence, this eventuality is not a desirable feature of a scoring system.

Student mobility within and across districts affects interpretation of growth. The OSDE states that, "The previous test scores can come from any school in the state. Students do not need to be in the same school two consecutive years to be included in the growth calculations" (p. 14). Schools with high mobility often find their students less prepared to benefit from instruction due to large gaps in their prior learning. For these schools, the individual student growth measure reflects the school's ability to overcome in less than one school year the cumulative effects of these learning gaps.

The metric behaves unpredictably as a basis for assigning grades. This is an index that can range from 0-300 but it is not clear why 90 is the cutoff for an A. Possibly the OSDE realized that this metric could penalize high achieving schools. For example, a school tests 300 students all of whom score advanced in year 1 and advanced in year 2. Their score is: (300*1)/300=1.00 *100=100. A 90% cutoff for an "A" would seem reasonable given this scenario. However, there are multiple ways a school can earn 100 on this metric. An alternative scenario is a school of 300 that has all students in the Unsatisfactory category. Half of them become Proficient (+2) and half don't change at all (0). The score is [(150*2)+(150*0)]/300=300/300=1.00*100=100. This school receives an A even though half of the students remained unsatisfactory. The lack of movement for half of the low performers might be picked up in the "growth of bottom quartile" unless

the bottom 25% was in the group that improved by 2 categories! Then this school would receive an A for both growth measures. Even worse would be a situation where a school has half the students in Limited Knowledge and half in Unsatisfactory. If all of the Limited Knowledge students move down to Unsatisfactory and the Unsatisfactory students gain 2, this school also receives an A because a loss and no gain are treated identically.

What are the mathematical properties of this index? There is a nonlinear relationship between proficiency level and growth since growth is restricted at the top. Assigning step numbers to these categories assumes equal intervals (i.e., that it takes the same degree of improvement to go from U to LK as from LK to P and so on). This may be a tenable assumption, but is unlikely. Further, how would this index account for increased standards reflected in the cut score? An index, like any scoring system, needs to behave in predictable ways to be reliable, valid, and useful. This index is entirely unpredictable. It operates differently in different contexts, depending on the spread and pattern of scores in a school. This growth index is very problematic.

The measure of student growth based on simply increasing a proficiency category from year to year assumes the category change score is a reliable and valid indicator of achievement growth. This would require a demonstration that a category change on two different tests at two different points in time is replicable and largely not random (i.e., is reliable), and is strongly reflective of actual differences in achievement (i.e., is valid). The variety of measurement conditions employed in this endeavor (multiple pre-tests and post-tests, varying growth periods) presents a major challenge to the ability of the OSDE to construct a chain of evidence supporting the use of this index. But, really this direct comparison method is fraught with conceptual and statistical problems and holds little promise of being a useful measure of growth.

What does this index mean conceptually? One of the primary conceptual flaws in the growth component of the A-F grading system is that it is based on measuring proficiency-level change for individual students and *attributing that change to the school/classroom most recently attended*. There are a number of threats to the validity of the argument that student change from a pretest to a posttest score should be attributed to the school associated with the post-test scores. Gall, Borg, & Gall (2003) require for this type of causal attribution that "All extraneous factors can be estimated with a high degree of certainty or can be safely assumed to be minimal or nonexistent" (p. 391). This is clearly not the case. There are many alternative hypotheses not involving school-level attributions. Some have been alluded to above. One hypothesis is that the student did increase in proficiency but that the measurement was not sensitive enough to detect it. The student may have increased his/her score on the proficiency test but not enough to increase proficiency level. Or, the tests aren't designed to measure change that occurs from year to year

because they measure different content at different years. So a student may have increased proficiency in the previous year's content but that content wasn't covered in this year's test. Or, a student learned the new content just as well as he/she learned the old content—is that not growth even if it wasn't at a higher level than the previous year? Baker and Linn address this: "For example, students scoring at the 50th percentile in the fourth grade who, in the following year, score at the 50th percentile in the fifth grade did not stand still; they learned a considerable amount of new material. But often such results are used as evidence that the educational system is not making progress" (2002, p. 13).

More importantly, students grow or don't grow for reasons beyond the control of the school. Students' ability to benefit from educational opportunity depends to a large degree on how well their basic needs are met. Students who are hungry, tired, poorly clothed, have little family support or security, are arguably just as capable as their wealthier counterparts, but they have fewer "disposable cognitive resources" to spend on studying. To attribute the lower achievement of these students to the schools they attend is an interpretation that goes way beyond the data collected in the A-F study and is inconsistent with a large body of research.

Growth of bottom quartile—17%

Methodology

The growth of the bottom quartile is based on pretest scores in math and reading that are in the Unsatisfactory or Limited Knowledge level. The lowest scoring 25% of students in a school whose pretest scores are U or LK are included in the calculations if this group includes at least 30 students in a school. Schools with fewer than 30 qualifying students are exempt from this component and have their entire growth component based on the overall student growth. Selection of students for this computation when more than 25% are in the U and LK categories, is based on a state percentile conversion of the actual test scores. Once the student scores in the U and LK categories are identified, the growth index is computed in the same way as for overall growth. The index can range from 0 to 300 with grades assigned in the same manner as with the overall student growth.

Concerns

Some issues with this bottom 25% growth measure have already been addressed above. Further, schools with many low achieving students are penalized twice in the calculation of growth. Secondly, longitudinal studies of achievement gains have shown that associated error is highest for small schools because of their small sample size (Baker & Linn, 2002). The problem is even worse when considering the growth of the lower quartile, which can be based on as few as 30 students. Oklahoma policy makers are to be commended for recognizing this issue and requiring a minimum sample size though 30 may be inadequate in the face of the multiple sources of measurement error (measurement error associated with the

continuous test scores, grouping error, errors of classification inconsistency, and errors associated with gain scores). As noted by Baker and Linn (1992), "Whatever the level of precision of school-level results, the results for schools should be accompanied by information about the dependability of those results as required by the *Standards for Educational and Psychological Testing* (AERA et al., 1999). This might best be done where schools are placed into graded performance categories by reporting information about the accuracy of classifications" (p. 17). This information is currently lacking in the OSDE documentation.

Another issue is that the state's calculation of the state-wide average of the Oklahoma Performance Index (OPI) needs to be explained. Students in the bottom quartile who score above the average OPI are awarded a point. Therefore, the manner in which the state average is calculated is important; however, the basis for that computation is not evident in the available documentation.

Component 3: Whole School Performance—33%

Methodology

In the OSDE literature, this is variously referred to as Whole School Performance and Whole School Improvement. For elementary schools, attendance determines this component. Bonus points are added for various other things such as advanced course work, community engagement and turning in a school climate survey. For middle school, attendance is 90% of this "improvement" measure, dropout rate and advanced course work contribute the remaining 10%. Bonus points from completion of a school climate survey, and parent/community engagement contribute a maximum of 12 points. For high schools, graduation rate counts 79% with the remaining 21% coming from participation and performance in advanced coursework, exams associated with advanced coursework, college entrance exams, graduation rate of low achieving 8th graders, and the bonus points. "Every indicator receives a letter grade of A-F. The indicators are combined to create a weighted grade point average." (p.20). High schools are awarded credit for having students in AP or IB classes, but they are severely penalized for having students in these advanced classes who do not attempt the AP/IB exam. "Schools with students enrolled in AP or IB courses that do not attempt the exam will be given an "F" (SDE, p.25).

Concerns

Concerns associated with using school attendance or graduation rates as a measure of a school's effectiveness are relevant since these indicators carry the lion's share of the grade in this component. These indicators have been shown to be correlated with SES. Further, Oklahoma continues to be out of compliance in its methodology for calculating graduation rates.[1] The small weight assigned to all the other indicators of whole school effectiveness suggests they are not valued highly in this assessment system.

Determining Report Card Grade

Methodology

Each component GPA is multiplied by its weight and summed together to form a School GPA. The School GPA is assigned back to a letter grade as follows:

3.75 to 4.0=A
2.75t to 3.74=B
1.75 to 2.74=C
0.75 to 1.74=D
0-.74=F

Concerns

The purpose for developing the A-F School Grading System was to put in place a comprehensive accountability system that would be transparent. It would provide clear communication to "replace past systems that were too complicated for most parents to understand" (OSDE, 2012, p. 7). A great deal of thought and effort has gone into making this system comprehensive. It is based on more than just math and reading which has been a flaw in other state accountability systems. However, in its current form, the A-F Report Card Grading system is neither transparent nor uncomplicated. If it seems easy to understand, it is only because the use of a single indicator to represent something complex is familiar. We are used to letter grades. A truly comprehensive evaluation system is best not boiled down to a single value because it masks the very complexity it is trying to capture.

A number of concerns have been identified in the A-F Grading system. One is the reliance on test scores to assess school quality. Previous reports have identified the problems associated with the use of student test scores to evaluate teachers (Barton, Darling-Hammond et al., 2010). Some of those problems have been discussed here and apply equally to the problem of evaluating schools. A major concern throughout the system is the creation of metrics and derived indices that are psychometrically insupportable as the basis for grading schools. In general, the system in its current state is severely flawed and, in our opinion, should not be utilized for decision-making.

III. The Practical Implications and Consequences of A-F

There are two main purposes attached to assessment; one is "summative" and it refers to assessment that informs high stakes decision-making. In public education, that might mean closing a school, firing a teacher, or moving one's child to a different school. The other is "formative" and it refers to assessment that informs incremental and continuous change and improvement. The approaches and indicators selected to assess systems and performance often limit their use to specifically summative or formative applications.

Although Oklahoma educators hoped the accountability system would provide them with information to improve schools, the assessment choices enacted appear to favor summative decision-making over incremental school improvement (see Barresi in *Oklahoma State Department*, 2012). By definition, summative assessments try to simplify in order to enable confident and incisive decision-making. Incremental improvement, the goal of formative assessment, is fostered by detailed analysis about how things work in general and in particular. Formative assessments are difficult to communicate since they require an understanding of a multitude of conditions, data, and contingencies; they do not make good political fodder. Nonetheless, the State's primary motive is to improve schools through its assessments and we review the capacity of the A-F accountability system to achieve this objective.

Pillars of Effective Assessment for Accountability

Shulman (2007) argues "the great promise of assessment is its deployment in the service of instruction, its capacity to inform the judgment of faculty and students regarding how they can best advance the quality of learning." To guide our discussion of implications and consequences of the A-F system, we adapt Shulman's "pillars" of effective assessment for accountability. The "pillars" are based on Standards for Educational Accountability and Educational and Psychological Testing. Effective accountability systems share the following characteristics:

1. They make the performance story told by the accountability system explicit and **clear**.
2. They design and use **multiple measures** so as not to base consequential decisions on a single instrument.
3. They **embed** assessment in ongoing instruction by assessing early and often.
4. They use information from accountability systems for **improvement**, not for high stakes consequences.

We explore the practical implications of the A-F system using the four adapted pillars of effective assessment for accountability.

Pillar 1: Explicit Performance Story

An explicit performance narrative makes clear its limitations and how and what kinds of performance are being measured. At its simplest level, the clarity Shulman advocates refers to the ease with which the question "What does it mean?" can be answered. Contrary to the intentions of the A-F system's designers, that question cannot be answered with ease, or at all, when considering the single letter grade assigned as a summary of a school's annual performance. The complexity of the formula for calculating the grade, which reduces many kinds of non-comparable measurements through a set of arbitrary rules to a single letter, makes it impossible to understand a particular grade's meaning. Moreover, the single grade offers no guidance with respect to the "how" or "why" of a school's performance. However, it is the information about "how" and "why" that can inform a school's efforts to change and improve.

The State's claims that its grading system is simple and clear are not justified. Simplicity, or parsimony, refers to the ability to portray the truth with the fewest words. Two criteria are required for clarity: brevity and truth. The letter grade is not simple, but simplistic, that is, it meets the brevity criterion, but fails the truth criterion. It does not present adequate information and thus portrays a partial truth. Instead of informing and empowering the public about school performance, it provokes the public with grades whose meaning is unclear, moving it to conclusions that are unjustified.

To be a clear representation of a school's performance, the grade must capture performance that is the consequence of what the school does and not other things. It should not, for example, vary as a result of conditions the school does not and cannot control, for example concentrated neighborhood poverty. If it cannot be claimed that the grade is the result of the school's efforts only, then the measure contains error, contributing to its lack of clarity.

Not only do the arbitrary rules for calculating the formula confuse the meaning of the grade, but given the conceptual and statistical limitations of the composite indicator, aggregation of individual achievement scores to the school level are questionably valid measures of school performance. The A-F accountability system is susceptable to forms of "test score pollution" (Haladyna, Nolen, & Haas, 1991). Test preparation and instructional practices designed to increase scores on achievement tests and other types of assessments in response to high stakes testing conditions introduce contaminants that threaten the validity of interpretations drawn from achievement data. There are multiple pathways to achieving a high grade that may have nothing to do with enhanced teaching or school effectiveness. Similarly, there are many causes for low grades that may have nothing to do with instructional practices or teaching quality. To illustrate, changing enrollment

boundaries or grade configurations of schools can affect test scores. Mobility also affects achievement results.

In short, letter grades are familiar to most people, but when this familiarity produces false confidence in the meaning of the assessment, it is harmful to school improvement. The threats to the validity of A-F grades and the limitations of their meaning should be made explicit. Further, the public should be informed that inferences about school performance should not be based on an accountability grade alone. Providing a summative grade, which is not a clear and complete indicator of school performance misleads the public.

Pillar 2: Multiple Measures

Physicians do not determine the health of a patient based on a single diagnostic procedure such as a blood pressure reading. Economists do not forecast economic growth from a single GDP measure. Accounting practices are no longer based solely on financial indicators. Mysteriously however, we expect education policy makers and leaders to diagnose the health of schools with only outcome measurements such as achievement scores and attendance rates. Shulman argues "it is dangerous to permit highly consequential decisions of policy and practice to rest on the results of a single instrument" (p.4).

The A-F accountability system is based on a measurement model that does not align with the information needs of educators or the knowledge-driven work processes of schools (Mehta, Gomez & Bryk, 2011). Manufacturing, healthcare, and technology sectors, in contrast to public education, have anchored continuous improvements in balanced performance systems that account for structures, processes, and practices that drive outcomes. In fact, balanced measurement has become standard practice in most industries as managers have realized the limitations of making decisions on outcome data alone (Kaplan & Norton, 2005). The A-F accountability system stands in contrast with more comprehensive understandings of measuring organizational effectiveness.

The science of quality school performance is robust, yet accountability systems like A-F operate as if outcome data are the only drivers of improvement. We know, for example, that collective trust facilitates learning in schools (Forsyth, Adams, & Hoy, 2011); that autonomy-support and competence-support enhance student engagement and learning (Assor, Kaplan, & Roth, 2002; Jang, Reeve, & Deci, 2010); and that a school's instructional capacity is related to teaching effectiveness (Harris, 2011; King & Bouchard, 2011). If improved school performance is the goal, the accountability system should support the development of conditions that have the most potential to improve teaching and learning.

To support continuous improvement, measurement systems need to balance process and outcome indicators. The A-F system is not balanced. As a result,

Oklahoma educators are expected to make consequential decisions about school improvement without evidence linking school strategies to performance. Public education should emulate the manufacturing, healthcare, and technology sectors and develop measurement systems capable of quantifying effective and efficient processes. Diagnoses of performance gaps can improve with better process data, thereby reducing dependence on commercially produced "canned" education interventions that often profit their corporate distributers more than schools. Process information is more likely to restore control, design, and implementation of continuous school improvement to school practitioners at the school site.

The OSDE should become a statewide resource, helping schools to develop balanced measurement systems that provide useful input, process, and outcome information. The current approach to process information, that is, using the result of a climate survey as bonus points in the formula is both inadequate and methodologically questionable. Useful process data need to match improvement strategies constructed within districts and schools. As challenging as the task may seem, it is doable. High quality measurement systems in healthcare, for example, use scientifically developed process indicators (Nolan & Berwick, 2006). Accounting practices now assess companies' future profitability in part on intangible resources like intellectual capital (Stewart, 1999).

Pillar 3: Embeddedness

In assessing the work of schools and holding educators accountable for student learning, Shulman (2007) argues that assessment should be embedded within the ongoing work of a school. Embedding such practices results in low-stakes and high-yield forms of assessment (formative assessments), administered repeatedly throughout the year. The results of these assessments are transparent and immediately communicated to those who can use them for change. Embedded assessments allow educators to monitor student performance and afford schools the opportunity to create and adjust environments where students themselves take ownership of their education. Transparent assessments coupled with prompt reporting also allow stakeholders to be informed continuously about school performance. Embedded assessment facilitates incremental and continuous improvements.

The intention of Oklahoma's A-F School Grading System is to hold schools accountable for learning while providing stakeholders with school performance information so that families and communities can work with schools to improve learning. However, Oklahoma's accountability assessments are not embedded in ongoing instruction throughout the school year. Instead, the school performance grade is based primarily on a single set of tests administered at the end of the year. Past performance indicators are part of most school accountability systems, but to

be useful for improvement purposes, performance needs to be assessed and reported during the school year.

Reforms that do not embed assessment and accountability initiatives within the ongoing work of a school often use high-stakes and low-yield forms of assessment (summative assessments) to punish or reward schools for achievement or the failure to achieve at specific levels. Such summative assessments do compare performance to accountability standards, but they do little to improve and inform teaching and learning. There are exceptions to the predominant summative approach. Connecticut, for example, has recognized the incomplete picture provided by end-of-year tests and is exploring additional metrics for school accountability and assessment that can be embedded within instruction and throughout the year. This more nuanced understanding of how assessment data can inform teaching and learning will emphasize trend data to provide a more complete picture of student performance.

In short, finding ways to provide useful information about learning processes and performance throughout the school year is necessary for continuous improvement. Just as Connecticut is attempting to design an accountability system that can also generate formative performance information, so too should Oklahoma consider how assessment can be embedded within the work of schools. Oklahoma has the opportunity to reexamine its grading system and lead other states in demonstrating how assessment can be used to hold schools accountable and simultaneously inform school improvement.

Pillar 4: High Stakes Consequences

Using a single letter grade to summarize a school's performance certainly suggests the intent of "high-stakes" use. Shulman outlines some of the unanticipated consequences of making assessment "high-stakes" (Shulman, 2007). He cites the tendency of test-makers to emphasize only objective facts, excessive teaching to the test, and school dishonesty in test management. This corruption of assessment's purpose quickly reduces the goals of schooling to their most simplistic, easily measured outcomes, a far cry from outcomes we might hope for such as job readiness, intellectual creativity, and enlightened citizenship.

The assumption behind the use of high stakes accountability as a lever for change is that schools are unwilling or unmotivated to reform. Even if this were the case, external inducements to task performance reliably undermine motivation (Ryan & Weinstein, 2009). Districts and schools are not adverse to improvements that have real potential to enhance learning. Many simply lack the capacity to learn from their experiences and to adapt practices to unmet learning needs of students. The A-F Report Card does not provide a framework supportive of capacity building. In fact, grading schools based on outcomes seems to turn a blind eye to

the systemic problems contributing to achievement gaps and low educational attainment.

Capacity building, a fundamentally different approach from high stakes assessment, offers the best chance for accountability systems to support districts in moving from poor to fair, fair to good, or good to great (Hargreaves, 2011; Harris, 2011; King & Bouchard, 2011). The A-F Report Card does not support districts and schools in developing capacity. As described in the previous three pillars, the Report Card conceals threats to the validity of school grades, it ignores processes and conditions that promote learning, and it uses data for summative not formative purposes. These weaknesses actually contribute to performance problems by fostering practices that contaminate test scores and undermine rich learning opportunities for all students (Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010).

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.

Summary of Implications for Practice

Effective education policy facilitates capacity building in districts and schools. In its current design and use, the A-F Report Card presents several practical challenges for achieving the above objective. Shulman's pillars of effective assessment for accountability expose specific elements of the A-F accountability system that act as barriers to capacity building and school improvement. A few of the more salient issues are summarized here.

1) The familiarity of letter grades conceals the conceptual and analytical problems of using a single letter grade to measure school performance. By not making threats to the validity of report card grades explicit, the OSDE misinforms the public about the credibility and utility of the A-F accountability system.

2) Outcome data alone do not provide a comprehensive or adequate story of school performance. Effective measurement systems balance process and outcome data to describe the inner workings of schools; the Oklahoma system relies primarily on achievement data. Performance information from the

current A-F Report Card has very limited value for informing school improvement; it is incapable of being used to diagnose causes performance variation.

3) For assessment data to have some positive effect on learning they need to be available when they can be useful and embedded in instruction. The A-F Report Card is neither timely nor embedded in instruction. Students are assessed at the end of the school year and school grades are not reported until the following year. If improvement is the goal, the summative aspect of the accountability system should not overshadow the formative uses of assessment and performance information.

4) High stakes testing, as a cornerstone of assessment and accountability, corrupts instructional delivery by reducing the goals of schooling to their most simplistic, easily measured outcomes, a far cry from outcomes we might hope for such as job readiness, intellectual creativity, and enlightened citizenship.

IV. Recommendations

1) Report school performance like school report cards that provide indicators of performance periodically and in multiple areas over the course of a year. **Action: Develop a report card format that uses multiple school indicators that more adequately reflect a school performance profile.** Eliminate the single grade, which cannot be composed without adding together unlike elements and promoting confusion and misunderstanding.

2) Develop a balanced performance measurement plan that aligns with strategic goals of schools. Track school indicators (i.e., inputs, process, and performance outputs) longitudinally to understand growth or stasis. **Action: Rely on trend lines of both process and outcome indicators over the year and multiple years to determine growth in school performance.** Growth indicators for different subject content should not be co-mingled to create single growth estimates for a whole school.

3) Include valid and reliable measures of school climate, motivation, and the dispositions of school role groups longitudinally. **Action: Promote the use of valid and reliable measurement of process variables at the district and school level, to be used by schools in their improvement plans.**

4) Any accountability system should be upfront about its limitations and uses. **Action: Make explicit the limitations of the accountability system and warn of its inappropriate use for high-stakes decision-making.**

5) Embed assessments in instruction throughout the year. **Action: Legitimize the process by embedding assessment throughout the school year.** Embedded, assessment data can be used to change course and make adjustments when needed; such assessments will be viewed as having a formative (improvement) purpose instead of a summative judgment about the past year.

6) The accountability system is important to our State and it is essential that it be credible, accurate, and supported by policy-makers, school professionals, and the people of Oklahoma. **Action: Take the time to enlist the services of assessment and evaluation experts who can objectively build an exemplary Oklahoma accountability system directed at incremental and continuous school improvement**

V. Conclusion

The work of schools and school leaders might be compared to gardening, that is, tending to the growth of a great variety of life. Gardeners are not preoccupied only with the harvest alone. They bring to bear all kinds of knowledge, skill, and information, adjusting what they do constantly to enrich the environment of the garden, providing nurture and protection from everything that might harm it. Gardeners know that the harvest at hand is important, but that care for soil conditions, monitoring surrounding vegetation, and assuring availability of supplementary water and fertilizer are just as important; future harvests will benefit from the enhanced general conditions of the garden. The metaphor suggests that accountability in schools cannot be defined in the same way quality assurance is attained in manufacturing. Schooling more resembles what Thomson (1967) calls an intensive technology, in which the processing of nonstandard raw material relies on constant response to new information. The metaphor and the theory both point to accountability for process elements and capacity building as well as outcomes; a focus on outcomes alone would not adequately serve the complexities of schooling or the long-term goals of our society. The collaboration among Oklahoma's education stakeholders could benefit by a metaphor that reminds us of the importance a long-term perspective has for effective school improvement.

VI. References

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

Assor, A., Kaplan, H., & Roth, G. (2002). Choice is good, but relevance is excellent: Autonomy-enhancing and suppressing teacher behaviours predicting students' engagement in school work. *British Journal of Educational Psychology, 72,* 261-278.

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.* Washington: Economic Policy Institute.

Baker, E.L. & Linn, R.L. (2002). Validity Issues for Accountability Systems, CSE Technical Report 585. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (UCLA).

Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. (2002, Winter). Standards for Educational Accountability Systems. *Policy Brief 5*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

Crocker, L., & Algina, J. (1986). *Introduction to Classical and MOSDErn Test Theory*, Belmont: Wadsworth.

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.

Darling-Hammond, L. (2010). Performance Counts: Assessment Systems that Support High-Quality Learning. Washington, DC: Council of Chief State School Officers.

Forsyth, P.B., Adams, C.M., & Hoy, W.K. (2011). *Collective Trust: Why Schools Can't Survive Without It*. New York: Teachers College Press.

Gall, M. D., Borg, W. R., & Gall, J. P. (2003*). Educational Research: An Introduction* (Seventh ed.). Boston: Pearson.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2-7.

Hargreaves, A. (2011). System redesign for system capacity building. *Journal of Educational Administration*, 49(6), 685-700.

Harris, A. (2011). System improvement through collective capacity building. *Journal of Educational Administration*, 49(6), 624-636.

Jang, H., Reeve, J., & Deci, E.L. (2010). Engaging students in learning activities: It is not autonomy support or structure but autonomy support and structure. *Journal of Educational Psychology, 102,* 588-600.

Kaplan, R. S., & Norton, D. (2005). The balanced scorecard: Measures that drive performance. *Harvard Business Review, 83*(7/8), 172-180.

King, B. M., & Bouchard, K. (2011). The capacity to build organizational capacity in schools. *Journal of Educational Administration*, 49(6), 653-669.

King, B., & Minium, E. (2003). *Statistical Reasoning in Psychology and Education*, 4th ed., Hoboken: Wiley.

Mehta, J. D., Gomez, L. M., & Bryk, A. S. (2011). Schooling as a knowledge profession. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2011/03/30/26mehta_ep.h30.html?tkn=MWSFT6mRTR0xt7SA8DRVsYCyhFBZ%2BCMuLYaF&cmp=clp-edweek

Nolan, T., & Berwick, D. M. (2006). All-or-none measurement raises the bar on performance. JAMA, 295(10), 1168-1170.

Oklahoma State Department of Education. (2012, April). *A-F Report Card Guide.* Rest of citation.

Oklahoma State Department of Education.(2012, October). *REAC³H Network: State longitudinal data system meeting.* Tulsa, OK.

Reeve, J., Ryan, R., & Deci, E. L., & Jang, H. (2008). Understanding and promoting autonomous self-regulation: A self-determination theory perspective. In D. H. Schunk & B. J. Zimmerman (Eds.), Motivation and self-regulated

learning: Theory, research, and applications. New York, NY: Erlbaum. pp. 223-244.

Thomson, J.D. (1967). Organizations in Action: Social Science Bases of Administrative Theory. New York: McGraw-Hill.

Ryan, R.M. & Weinstein, N. (2009). Undermining quality teaching and learning: A self-determination theory perspective on high-stakes testing. *Theory and Research in Education*, 7(2), 224-233.

School Graduation Rates Act. S.B. 2 Ok. § 3-151.2 (2011).

State Department of Education. (2012, April). *A-F Report Card Guide.* Rest of citation.

Shulman, L. Counting and recounting: Assessment and the quest for accountability. Change, January/February 2007. Vol 39, No. 1. http://www.carnegiefoundation.org/change/sub.sap?key=98&subkey=2169

Stewart, T. A. (1999). Intellectual capital: The new wealth of organizations. New York: Doubleday.

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and Evaluation in Psychology and Education.* (Eighth ed.). Boston: Pearson.

U.S. Department of Education, National Center for Educational Statistics (2006). *User's guide to computing high school graduation rates* (NCES Publication No.2006-604).

U.S. Department of Education (2012). *ESEA Flexibility Request: Connecticut* (OMB Number: 1810-0708). Retrieved from http://www2.ed.gov/policy/eseaflex/approved-requests/ct.pdf

U.S. Department of Education, A uniform, comparable graduation rate 2008). Retrieved from http://www2.ed.gov/policy/elsec/guid/hsgrguidance.pdf

Notes

[1] The Oklahoma Department of Education (OSDE) was directed by the United States Department of Education (USDOE, 2008) and the Oklahoma Legislature (2011) to develop a cohort-based high school graduation rate by FY 2012. With such a high level (79%) of the Whole School Performance Index assigned to high school graduation rate determinations it is imperative that the OSDE develop the capacity to provide highly accurate graduation rates for Oklahoma high schools.

The OSDE is well behind other states in the provision of cohort-based/longitudinal graduation rates as mandated by USDOE guidance, and OSDE's failure to implement the cohort-based graduation rate leaves it out of compliance with existing Oklahoma law SB 2 in (2011). Instead of using a cohort graduation rate formula as currently mandated in federal guidance and state law, the OSDE continues to use the Graduation Leaver Indicator (GLI) formula to determine the graduation rate of Oklahoma schools. The USDOE has described the GLI as follows: "Most importantly this rate differs in that it is a leaver rate, rather than a graduation rate" (NCES, 2006, p. 16). OSDE's continued lack of progress in the development and promulgation of rules associated with the implementation of an accurate cohort-based graduation rate formula for Oklahoma schools brings into question the fidelity and credibility of the existing graduation data associated with each Oklahoma high school in the most recent A-F school calculation.

At the October Reach Conference in Tulsa OSDE staff released the following information, "Oklahoma is out of compliance with federal law. The OSDE cannot currently calculate a cohort graduation rate." The OSDE has posted in A-F grading scales graduation percentages in the mid to high 90's for districts/sites with known graduation rate problems. An effective OSDE cohort-based longitudinal graduation rate (it is after all a four year process) determination would have indicated these problems well ahead of the students' senior year of high school. OSDE should develop the capacity to create accurate longitudinal measures for critical state data including both graduation and academic growth data.

Appendix A: Comments of Robert L. Linn

Comments on "A Examination of the Oklahoma State Department of Education's A-F Report Card" commissioned by OSBA and CCOSA dated December 2012

Robert L. Linn

December 17, 2012

The evaluation of Oklahoma's A-F report card system by the staff of the Oklahoma Center for Education Policy and the Center for Educational Research and Evaluation provides a clear and accurate description of the components and calculations that lead to the A to F grades for schools. Overall, the criticisms of the A-F report card system are carefully reasoned and well justified. As is indicated in the evaluation report, the A to F grades are based on combining various sorts of information about student achievement in arbitrary and complicated ways that obscure the meaning of the grades. The grades do not provide schools with information that they can use to improve instruction and student learning.

The achievement component of the index starts by assigning values of 0, .2, 1.0, and 1.2 respectively to the four achievement levels (Unsatisfactory, Limited Knowledge, Proficient, and Advanced). These numerical values for the four achievement levels appear to be arbitrary and lack any stated rationale. They do not distinguish among students who are at the high end of an achievement level from students who are at the low end of that category. The 0 to 1.2 scores are averaged over subject areas and then multiplied by 100 to yield an index score with a possible range from 0 to 120. Index scores are then categorized into letter grades using arbitrary and unjustified cutoffs. The cutoff scores of 90, 80, 70 and 60 may have been chosen because teachers sometimes use those cuts in grading student work, but that is normally done on a scale that has a possible range of 0 to 100 not 0 to 120.

Importantly, the achievement component, which is intended to count for a third of a school's grade, is not a valid measure of school effectiveness because status scores are influenced by many factors that are not under the control of a school.

The "growth" component is based on the number of students that maintain or improve their achievement level from one time to another. It ignores what may in some cases be substantial growth within a category. The tests used at the two times may or may not be measures of the same construct and may vary in the time between the pre and post measures, making the measure of "growth" impossible to interpret. Furthermore, the measure may have nothing to do with the learning at a particular school since some students may have changed schools during the school year.

The whole school performance component of a school's grade is based on factors such as attendance or graduation that are poor indicators of school effectiveness. Combining the whole school performance component with the other components does nothing to enhance the interpretation of the A to F grades.

All in all, Oklahoma would be well advised to scrap the A to F report card and replace it with a system along the lines of the recommendations made in the evaluation report prepared by the Oklahoma Center for Education Policy and the Center for Educational Research and Evaluation.