education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 24 Number 4 January 11th, 2016 ISSN 1068-2341

An Empirical Test of Oklahoma's A-F School Grades

Curt M. Adams
Patrick B. Forsyth
Jordan Ware
University of Oklahoma
&
Mwarumba Mwavita
Laura L. Barnes
Jam Khojasteh
Oklahoma State University
United States

Citation: Adams, C. M., Forsyth, P. B., Ware, J. K., Mwavita, M., Barnes, L., & Khojasteh, J. An empirical test of Oklahoma's A-F grades. *Education Policy Analysis Archives, 24*(4). http://dx.doi.org/10.14507/epaa.v24.2127

Abstract: Oklahoma is one of 16 states electing to use an A-F letter grade as an indicator of school quality. On the surface, letter grades are an attractive policy instrument for school improvement; they are seemingly clear, simple, and easy to interpret. Evidence, however, on the use of letter grades as an instrument to rank and improve schools is scant at best. We address the gap in the literature by using student test scores to evaluate the use of Oklahoma's A-F grades as a school quality indicator. Achievement differences between letter grades were small and in most cases not statistically significant when student and school characteristics were held constant. School grades did not reveal large achievement gaps in the highest ranked schools. Additionally, free/reduced lunch and minority students in D and F schools outperformed peers in A and B schools.

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 18/5/2015
Revisions received: 28/12/2015
Accepted: 29/12/2015

Keywords: school accountability; A-F accountability grades; achievement equity; next-generation accountability

Resumen: Oklahoma es uno de los 16 estados que eligen utilizar calificaciones con las letras A-F como indicadores de calidad de una escuela. Supuestamente, las calificaciones con letras son un instrumento político atractivo para el mejoramiento escolar; son aparentemente claras, sencillas y de fácil interpretación. Sin embargo, la evidencia es escasa acerca de el uso de calificaciones con letras como instrumento para clasificar y mejorar las escuelas. Atendemos la brecha en la literatura usando resultados de los exámenes para evaluar el uso de las calificaciones A-F de Oklahoma como un indicador de la calidad escolar. Las diferencias de rendimiento entre las calificaciones con letras fueron pequeñas y en la mayoría de los casos no fueron estadísticamente significativos cuando las características escolares y de los estudiantes se mantuvieron constantes. Las calificaciones de las escuelas no revelaron grandes brechas de académicas en las escuelas de mejor desempeño. Además, alumnos que reciben asistencia en forma de almuerzos gratuitos y de minorías en escuelas calificadas como D y F superaron a sus pares en las escuelas con calificaciones A y B.
Palabras clave: responsabilidad escolares; calificaciones A-F de rendición de cuentas; equidad de logros; nuevos modelos de rendición de cuentas

Um teste empírico das notas A-F em Oklahoma
Resumo: Oklahoma é um dos 16 estados que optam por utilizar notas com as letras A-F como indicadores de qualidade de uma escola. Supostamente, notas são uma ferramenta política atraente para a melhoria da escola; eles são aparentemente claras, simples e fácil de interpretar. No entanto, há pouca evidência sobre o uso de notas com letras como uma ferramenta para classificar e melhorar as escolas. Nós atendemos a brecha na literatura utilizando os resultados dos testes para avaliar o uso das notas A-F de Oklahoma como um indicador da qualidade da escola. As diferenças de desempenho entre as escolas eram pequenas e na maioria dos casos não foram estatisticamente significativas quando foram mantidas constantes características de escolas e alunos. As pontuações para as escolas não revelaram grandes diferencias nas escolas com melhor desempenho acadêmicas. Além disso, os alunos que recebem assistência sob a forma de almoços gratuitos e de minorias em escolas classificadas como D e F superaram seus pares nas escolas com notas A e B.
Palavras-chave: responsabilidade da escola; notas A-F de responsabilização; equidade nos logros acadêmicos; novos modelos de prestação de contas

An Empirical Test of Oklahoma's A-F School Grades

Under No Child Left Behind (NCLB), test-based accountability became the primary policy instrument to fix what many perceived as a failing public education system. Most states now operate under NCLB waivers, granting them some flexibility in the design and use of test-based accountability. Oklahoma is one of 16 states electing to use a system that assigns an A-F letter grade as an indicator of school quality (Howe & Murray, 2015; Polikoff, McEachin, Wrabel, & Duque, 2014). Letter grades identify schools to be rewarded for their effectiveness and schools in need of interventions when performance standards are not achieved. High grades, A's and B's, designate schools as making excellent academic progress as measured from student test scores in reading, math, science, and social studies; low grades, D's and F's, designate low performing schools that are in need of mandated interventions, such as school choice options, reconstitution, or closure (Oklahoma State Department of Education, 2012).

On the surface, letter grades are an attractive policy instrument for school improvement; they are seemingly clear, simple, and easy to interpret. People generally share a consistent idea about the meaning of an A or an F, making it easier, proponents argue, for citizens to know how schools are doing. Evidence, however, on the use of letter grades as an instrument to rank and improve schools is scant. Much of what we know is limited to the consequences of grades on the housing market and teacher mobility in Florida. Figlio and Lucas (2004) found that housing prices and home sales were greater in communities with schools that received A's and B's. Feng, Figlio, and Sass (2010) found greater mobility in D and F schools compared to higher ranked schools. Although informative, these findings do not address the utility of using grades to improve schools.

Limited evidence makes it difficult to mount a case for or against the use of A-F letter grades in achieving policy objectives set forth by the US Department of Education. We address the lack of empirical evidence by using student test scores from the 2011-2012 and 2012-2013 school years to evaluate the validity of using Oklahoma's A-F grades as a policy instrument. Appropriately placed recognition, reform pressure, and interventions depend on having a trustworthy indicator of school quality. Oklahoma's A-F grades, at a minimum, need to accurately distinguish between schools that are raising learning outcomes for all students and those failing to move the performance needle. We proceed with a conceptual framing of test-based accountability before turning to validity criteria.

The Function of Test-Based Accountability

Test-based accountability policy falls within Schneider and Ingram's (1991) category of authority and incentive tools. These types of tools use external pressure in the form of hierarchical controls and/or inducements to provoke behavioral change within a social system. Authority and incentive tools derive from the rational person argument in which individuals are viewed as utility maximizers who pursue actions that are likely to maximize gain and/or avoid punishment (Schneider & Ingram, 1991). Consistent with this reasoning, it is claimed that greater incentives and heightened external pressure are needed to induce school agents to raise educational quality.

Authority and incentive tools contrast with capacity tools. Capacity tools assume school members have the motivation and desire to improve, but require proper supports to generate knowledge on how their collective actions are or are not meeting the learning needs of all students (Adams, 2013; Darling-Hammond, 2005). The design and use of performance information would have to change for test-based accountability to be used as a capacity tool (Adams, Forsyth, Ware, & Mwavita, in press; Schneider & Ingram, 1991). As it stands, Oklahoma's A-F system, like accountability under NCLB, uses performance information in a targeted way: to hold schools accountable for past achievement outcomes (Adams, Forsyth, Ware, & Mwavita, in press; Linn, 2005, 2008; Rothstein, Jacobson, & Wilder, 2008; Oklahoma State Department of Education, 2012; Smith & Fey, 2000). Performance information used as an authority and incentive tool relies on political and institutional pressure to raise and equalize learning opportunities and outcomes.

Support for using accountability indicators as an external tool comes primarily from behavioral frameworks like agency theory and expectancy theory (Polikoff, McEachin, Wrabel, & Duque, 2014; Ryan & Weinstein, 2009). In these theories, clear achievement standards and accurate performance measures function as external motivators for goal attainment. In agency theory accountability information is a mechanism used by principals (i.e. school administrators, community members, legislators, tax payers, etc.) to ensure school agents (i.e. teachers) deliver high student achievement (Polikoff, McEachin, Wrabel, & Duque, 2014). Expectancy theory assumes that rewards and threats will motivate teachers to improve achievement as long as standards are clear and performance indicators are accurate (Finnigan & Gross, 2007). Through an agency and expectancy

lens, perceived legitimacy and valence (i.e. attractiveness) of accountability systems affect the behavioral response of school members.

Legitimacy and valence are two hurdles NCLB struggled to clear. Unreasonable yearly progress expectations, high stakes attached to proficiency scores, and classification error affected the perceived legitimacy of accountability indicators (Forte, 2010; Ho, 2008; Linn, 2008; Raudenbush, 2004). Legitimacy was also affected by the limited instructional relevance of aggregated achievement scores. Valance could not overcome the punitive nature of state sanctions and the dearth of evidence that prescribed improvement plans, turnaround models, and school reconstitution could actually make schools better (Forte, 2010; Mintrop & Sunderman, 2009).

For Oklahoma's A-F accountability system to work as an authority/incentive tool, letter grades must overcome the legitimacy problems that plagued performance indices under NCLB. Incentives, threats, and sanctions lose their appeal and potency if accountability indicators cannot be trusted (Finnigan & Gross, 2007). As an external control, it is reasonable to expect that A-F grades should identify schools that are effective for all students, not just schools that benefit from a socially advantaged student population. Validity for using A-F grades to improve schools erodes if the indicator distorts school quality.

Validity Criteria for Use of A-F Grades

The standards for educational and psychological testing call for accountability systems based on student test scores to be examined for their valid use (AERA, APA, & NCME, 2014). Validity refers to the ability of a measure or indicator to yield truthful judgments about the object it purports to measure (Messick, 1995; Miller, 2008). While there are different types of validity (content, discriminate, convergent, consequential, etc.), they share the same fundamental logic – that validity exists to the degree that the indicator represents the underlining theoretical construct and informs credible judgments about the phenomenon of interest (Cronbach, 1971; Messick, 1995). Validity is not a property of the measure itself; but rather, a property of how tests and measures are used (AERA, APA, & NCME, 2014; Baker & Linn, 2002; Kane, 2006). For instance, state curricular assessments, the primary source of Oklahoma school grades, may have adequate validity for measuring student proficiency against content standards, but limited evidence to support their use as an indicator of school effectiveness.

From a policy perspective, our interest is in the validity of using A-F grades as a tool to achieve policy objectives set forth by the US Department of Education. In reference to accountability, the US Department of Education (2012) argues, "Fair, flexible, and focused accountability and support systems are critical to continuously improving the academic achievement of all students, closing persistent achievement gaps, and improving equity" (p.1). Oklahoma's accountability policy has more features than just A-F grades, but the grade is the engine behind the entire system. Letter grades place a quality label on schools, determine State recognition, and identify schools for State intervention. Interventions include prescriptive planning frameworks, school choice options, replacing the school principal and staff, closing the school and reopening it as a charter school, turning operations over to the State Board of education, or closing the school (Oklahoma State Department of Education, 2012). Appropriately placed recognition and intervention depend on the validity of using A-F grades as a quality standard.

Validation, as a process, is argument-based and requires that the assumptions and inferences be clearly stated. A first step is an examination of the inferences drawn from the indicator (Messick, 1995). Letter grades are intended as a quality indicator defined by specific policy objectives: improved achievement, closed achievement gaps, and improved equity. These objectives come from

the US Department of Education's call for new systems of school recognition and accountability. In theory, A-F grades should distinguish schools that have achieved the policy aims from those that have not. School quality rankings should not be confused with achievement status; these are distinct constructs that respond to different questions and yield different interpretations of school performance.

Achievement status primarily describes the percent of students meeting or exceeding the proficiency standard (Carlson, 2006). For instance, achievement status can describe the percentage of 3rd grade students in "X" school that scored proficient on the state math exam. Or, the percent of students in "J" school that scored in the advanced category on the 5th grade reading test (Carlson, 2006; Forte, 2010). Percent proficiency may be used as a descriptive indicator of achievement levels and patterns in schools, but it is not useful for making evaluative inferences about school quality (AERA, APA, & NCME, 2014; Carlson, 2006; Forte, 2010). Letter grades rank schools: A schools are assumed to be higher quality than B schools, B schools higher quality than C schools and so forth. Judgments of school quality are held to a higher validity standard than measures of achievement status (Forte, 2010; Harris, 2011).

Quality indicators should distinguish between achievement variation attributed to school context (e.g., social and economic characteristics of the community and composition of the student body) and variation attributed to school effects (Linn, 2008; Raudenbush, 2004; Raudenbush & Willms, 1995). Raudenbush and Willms (1995) refer to estimates of school attribution as a Type A effect. They note, "Type A effect is the difference between a child's actual performance and the performance that would have been expected if that child had attended a typical school" (p. 309). Raudenbush and Willms use statistical theory about causal inferences to define "typical school" as one where differences in student and school compositions have been controlled through random assignment. Because random assignment to schools is not possible, valid judgments of schools must be based on evidence that controls for factors affecting the propensity of students to attend a certain school, such as differences in student ability levels, minority composition, and poverty (Raudenbush, 2004; Raudenbush & Willms, 1995).

Applying the standard of Type A effects to our test of school grades requires that letter grades provide accurate information about the contribution schools make to student achievement and achievement equity. Moreover, school grades need to be an indicator of what schools and teachers control, not simply reward schools that serve socially advantaged students while penalizing schools serving children and families with limited learning opportunities. Harris (2011) refers to this as the cardinal rule of accountability: schools should be held accountable for what they do. For a single, summative letter grade to be 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 the highest rated schools (A and B" schools) should be considerably higher than lower rated schools (C, D, and F schools). Differences in student test performance should be substantial and persist after controlling for factors unrelated to teaching effectiveness or school practices. Finally, letter grades should reflect achievement gaps within and between schools.

Calculation of Oklahoma's A-F Grades

Oklahoma's formula converts test scores into categorical data, categorical data back into a continuous index, and a continuous index into a summative letter grade. For the 2011-2012 school year the final composite grade was derived from three components: (1) student achievement, (2) student growth, and (3) whole school performance (Ayers, 2011; OCEP & CERE, 2012; Oklahoma

State Department of Education, 2012). Measurement components changed in 2012-2013 to include only the student achievement and student growth components with some whole school performance indicators included only as bonus points (Table 1).

Table 1.
Changes in A-F Calculation Formula

A-F Components	2011-2012	2012-2013
Student Achievement	33% of final grade	50% of final grade
Student Growth	33% of final grade	50% of final grade
Whole School Performance	33% of final grade	Counted as bonus points

The student achievement component counted for 33% of the school grade for the 2011-2012 school year. Student achievement scores from state math, reading, science, and social studies exams were used to calculate a school's Performance Index (PI). The PI in 2011-2012 was calculated by assigning a score to each of the four proficiency levels. Unsatisfactory was assigned a score of 0, limited knowledge 0.2, proficiency 1.0, and advanced 1.2. Each proficiency score was weighted by the number of students with test scores in the proficiency band. Weighted scores for proficiency levels were aggregated across all four content tests and all grades in the school (e.g. 3rd, 4th, and 5th). The aggregated value was divided by the number of examinees across all content tests and multiplied by 100 to achieve the final PI, which had a range of 0 to 120. PI ranges where then categorized into letter grades and reassigned a numeric value ranging from 0 to 4, which when multiplied by .33 formed the overall school GP for the student achievement component (OCEP & CERE, 2012; Oklahoma State Department of Education, 2012).

In 2012-2013 the student achievement component increased from accounting for 33% of the overall school grade to 50%. A change was made to the weighting scheme for the PI calculation. The weights assigned to each of the four proficiency categories were collapsed into a simple binary scale. Students scoring below proficiency were assigned a score of 0 and students proficient or above a score of 1. PI scores continued to include test scores aggregated across reading, math, science, and social studies exams.

The student growth component accounted for 33% of the composite school grade in 2011-2012 and was calculated from math and reading tests. Schools with fewer than 30 students tested did not receive a student growth score. Instead, student achievement and whole school performance were each worth 50 percent of the school grade. Growth was divided into two parts: (1) overall student growth (17%) and (2) growth of the bottom 25 percent of students in a school (17%). For overall growth, students received points if they advance a proficiency level from one year to the next, or if they were proficient in the prior school year and remained proficient or above in the academic year used to calculate school grades. Two points were awarded to students who advance two proficiency levels and three points for a three level improvement. Students who decreased in proficiency, or students remaining at unsatisfactory or limited knowledge, earned zero points for growth (OCEP & CERE, 2012; Oklahoma State Department of Education, 2012).

Growth of the bottom 25% of students is how Oklahoma satisfies the achievement gap reporting requirement of the NCLB waiver. In 2011-2012 schools with 30 or more students in the bottom quartile were included in this growth calculation. Schools with fewer than 30 students in the bottom quartile had their individual growth score counted twice. Scores for growth of the bottom 25% were calculated the same way as overall student growth. Students advancing a proficiency level earned one point. Students advancing two levels earned two points, and a three level advance earned

three points. As with the student achievement component, growth was not calculated or reported by student subgroups.

There were three primary changes to the growth component in the 2012-2013 school year. First, student growth increased from 33% of the total school grade to 50%, with the overall student growth index and growth of the bottom 25 % contributing equally. Second, any increase in proficiency level was awarded one point regardless of how many levels the student improved. Finally, the minimum sample size required for computation of growth of the bottom 25% of students decreased from 30 to 10 students.

The whole school performance component counted for 33% of the composite school grade in 2011-2012, and the following year it was used as bonus points for schools. While the indicators within whole school performance were essentially the same over the two years, their absolute weights changed. Attendance rates determined the entire whole school performance score for elementary schools both years. For middle schools, attendance was the primary indicator, accounting for 90% of the score the first year and 60% the second. Advanced course participation and dropout rate made up the remainder of the score. For high schools, the dropout rate accounted for 79% of the score the first year and 50 % the second. Participation in advanced courses, college entrance exams, and graduation rates of the lowest achieving 8th grade students accounted for the remainder (OCEP & CERE, 2012; Oklahoma State Department of Education, 2012).

Methods

Messick (1995) argues, "validity is nothing less than an evaluative summary of both the evidence for and the actual as well as potential consequences of score interpretation and use" (p.5). Therefore, our empirical test examines evidence by which we can evaluate the usefulness of A-F grades for achieving the policy objectives of improved achievement, closed achievement gaps, and improved equity. For grades to be a useful authority and incentive tool, they need to differentiate schools that are raising learning outcomes for all students from those failing to do so. We tested the use of grades for this purpose by (1) estimating achievement effects between students in A/B schools and students in either C, D, or F schools, and 2) estimating school-level differences in achievement gaps.

Sample

We analyzed student test scores and school compositional data from a convenient sample of 61 elementary and middle schools in an urban district. Sampling urban schools allowed us to test the accuracy of school grades in a context where it is especially critical for performance information to differentiate effective schools from ineffective ones. Relative to the student population in Oklahoma, our sample had a higher percentage of FRL students (62 % FRL in the state), a larger representation of minority students (47 % in the state), average reading and math test scores near the state average of 700, and a larger percentage of D and F schools (15 percent of schools received a D and percent an F in 2013; 8 percent of schools received a D in and 1 percent received an F in 2012).

We obtained valid reading scores from 14,978 3rd through 8th grade students in the 2011-2012 school year and 15,116 3rd through 8th grade students in the 2012-2013 school year (Table 2). In 2011-12, approximately 64 % of the students identified as minority and 36 % as non-minority, Caucasian. Nearly 77 % qualified for Free or Reduced Lunch (FRL). In 2012-13, approximately 71 % of students identified as minority and 29 % non-minority, Caucasian. Nearly 74 % of the students qualified for FRL.

Table 2.
Descriptive Student and School Data

Student Sample	Mean	SD	Min	Max
Reading Sample 2011-12				
Minority	.64	.50	0	1.0
Free/Reduced Lunch	.77	.42	0	1.0
Reading Scale Score	701.32	90.44	400	990
Reading Sample 2012-13				
Minority	.71	.45	0	1.0
Free/Reduced Lunch	.74	.44	0	1.0
Reading Scale Score	702.53	91.81	400	990
Math Sample 2011-2012				
Minority	.64	.45	0	1.0
Free/Reduced Lunch	.77	.42	0	1.0
Math Scale Score	698.87	96.84	400	990
Math Sample 2012-2013				
Minority	.71	.45	0	1
Free/Reduced Lunch	.74	.44	0	1.0
Math Scale Score	698.61	94.59	400	990
School Sample				
2011-12				
Free/Reduced Lunch Rate	88%	21.5%	17%	100%
Minority Rate	70%	18%	24%	98%
Prior School Average Reading Achievement	701.32	45.19	636	844
Prior School Average Math Achievement	702.12	49.79	635	823
A Schools	5%	--	--	--
B Schools	13%	--	--	--
C Schools	21%	--	--	--
D Schools	54%	--	--	--
F Schools	8%	--	--	--
2012-2013				
Free/Reduced Lunch Rate	88%	21.5%	17%	100%
Minority Rate	70%	18%	24%	98%
Prior School Average Reading Achievement	696.18	42.51	625	817
Prior School Average Math Achievement	697.89	49.47	607	839
A Schools	10%	--	--	--
B Schools	10%	--	--	--
C Schools	5%	--	--	--
D Schools	23%	--	--	--
F Schools	53%	--	--	--

Note. N= 14,978 reading scores in 2011-12 and 15,116 in 2012-13; N = 15,105 math scores in 2011-12 and 15,151 in 2012-13. Scale scores range from 440 – 990.

Student scale scores from the Oklahoma Core Curriculum Tests (OCCT) in reading and math were used to operationalize achievement. OCCT tests are criterion referenced exams designed to measure student knowledge of content standards. Reading standards for students in 3rd-8th grade

covered vocabulary, comprehension and critical literacy, literature, and research and information. Math standards included number patterns and relationships, number sense and operation, geometry, measurement, and data analysis. Exams were administered in April of 2012 and April of 2013. Tests consisted of 50 multiple choice items that were scaled to a minimum score of 400 and a maximum score of 990 with 700 set as the proficiency cut score (CTB/McGraw Hill, 2013).

For the sample of students, the average reading scale score in 2011-12 was 701 with a standard deviation of 90. In 2012-2013 the average reading scale score was 702 with a standard deviation of 91. The average math scale score in 2011-12 was 701 with a standard deviation of 97. In 2012-13 the average math scale score was 698 with a standard deviation of 94.

The school level average FRL rate was 88% and the minority rate was 70%. Prior school average reading achievement was 701 in 2011-12 and 696 in 2012-13. Prior school average math achievement was 702 in 2011-12 and 697 in 2012-13. In 2011-12, 5% of the schools earned school grades of A; 11% earned grades of B; 21% earned grades of C; 55% earned grades of D; and 8 % grades of F. For 2012-13, 10% of schools earned grades of A, 10% earned grades of B, 5% earned grades of C, 23% earned grades of D, and 53% earned grades of F.

Analysis

We used a conventional multilevel model building process in HLM 7.0 (Hox, 2010). The purpose was twofold: First, to compare achievement differences between letter grades when controlling for differences in school composition. School-level controls were needed to rule out the possibility that A-F grades resulted from different beginning achievement levels and different student compositions (Linn, 2008). School controls included variance attributed to prior average school achievement in reading and math and student composition. Prior achievement was calculated as the average reading and math score for students from the previous testing year. Second, we estimated differences in FRL and minority achievement gaps across letter grades with a Random Coefficient Intercept and Slopes as Outcomes model. Our interest was in evaluating the degree to which school grades expose inequitable achievement distributions.

Random effects ANOVA. We first decomposed achievement variance 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: $\text{Ach}ij = \beta 0j + rij$

Level 2: $\beta 0j = \gamma 00 + \text{u}oj$

$$\rho = \frac{\sigma^2_{uo}}{\sigma^2_{uo} + \sigma^2_{ri}}$$

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 (Enders & Tofighi, 2007; Raudenbush & Bryk, 2002). Dummy coding was used for minority status and FRL status. Minority and FRL status were allowed to vary randomly across schools.

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

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

$\beta_{1j} = \gamma_{01} + u_{0j}$

$\beta_{2j} = \gamma_{02} + u_{0j}$

Random coefficient intercept and slopes as outcomes. The final step was to test a random coefficient slopes and intercepts as outcomes model with all significant student and school variables. We again used grand-mean centering for the student covariates. Our interest in estimating achievement effects attributed to letter grades after holding compositional school differences constant led to our choice of grand-mean centering (Ender & Tofighi, 2007). To ensure slope estimates were not biased by grand-mean centering, we also tested the model using group-mean centering. We found no substantive differences in slope estimates between grand-mean and group-mean centering. A and B schools were set as the referent, allowing us to compare achievement differences between students in the highest rated schools against students in C, D, and F schools[1]. The slopes and intercepts as outcomes models had less error and best fit with the data. In other words, they provided an unbiased assessment of the main effects and cross-level interactions. Estimates represent the actual difference in scale scores after controlling for factors not related to teaching practices and school performance.

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

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(C_j) + \gamma_{02}(D_j) + \gamma_{03}(F_j) + \gamma_{04}(\%\ \text{Minority}_j) + \gamma_{05}(\text{Prior Achievement}_j) + \gamma_{06}(\text{FRL Rate}_j) + u_{0j}$

$\beta_{1j} = \gamma_{10} + \gamma_{11}(C_j) + \gamma_{12}(D_j) + \gamma_{13}(F_j) + u_{0j}$

$\beta_{2j} = \gamma_{20} + \gamma_{21}(C_j) + \gamma_{22}(Dj) + \gamma_{23}(F_j) + u_{0j}$

β_{0j} = is the school achievement mean for math achievement
β_{1j} = Minority achievement
β_{2j} = FRL achievement
γ_{00} = grand mean for achievement
γ_{01} = difference in average achievement between A/B schools and C schools
γ_{02} = difference in average achievement between A/B schools and D schools
γ_{03} = difference in average achievement between A/B schools and F schools
γ_{04} = effect of school % Minority on achievement
γ_{05} = effect of prior average achievement on student achievement
γ_{06} = effect of % FRL rate on student achievement
γ_{11} = Difference in the minority gap between A/B and C schools
γ_{12} = Difference in the minority gap between A/B and D schools
γ_{13} = Difference in the minority gap between A/B and F schools

[1] We combined A and B because of the relatively small numbers of A schools in 2011-12. Though there was an equally small number of F schools in 2011-12, we did not combine D and F because of the large percentage of D schools.

γ_{21} = Difference in the FRL gap between A/B and C schools
$\gamma 22$ = Difference in the FRL gap between A/B and D schools
$\gamma 23$ = Difference in the FRL gap between A/B and F schools

Results

We start by comparing the distribution of school grades for 2011-12 and 2012-13 and reporting results of the unconditional random effects ANOVA. As seen in figure one, changes to the grading formula appear to have influenced grade distributions in our sample of schools. For both years, D and F grades predominate (62 % and 76 % respectively). Striking is the increase in F schools and the concomitant decrease in D and C schools. The numbers of A and B schools were similar in the two years (18 % and 20%). Had the achievement test scores of students declined, the increase in the percentage of F schools, and decrease in D and C schools, would be expected. However, average reading achievement increased from 701 to 702 over the two year period while average math achievement remained stable at 698 (see Table 1).

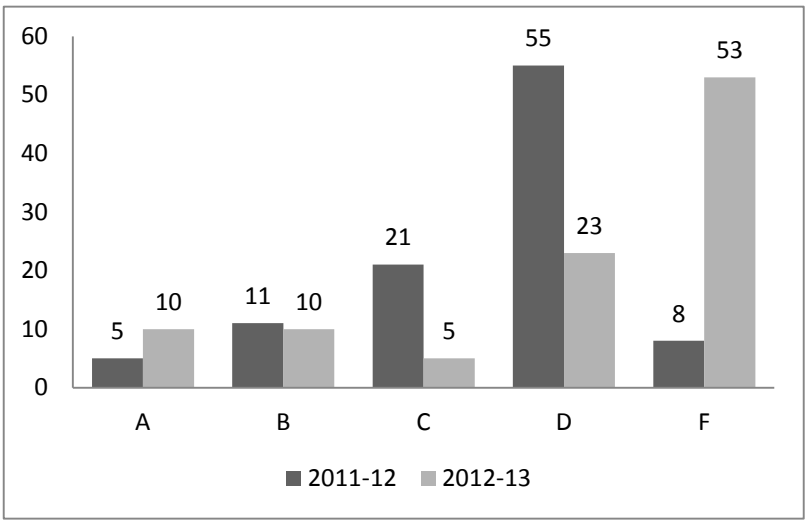

Figure 1. Distribution of school grades for 2011-12 and 2012-13

Table 3 reports variance components at the student (σ_i^2) and school levels (τ), as well as ICC's for math and reading. In 2011-12 achievement variation due to school differences was statistically significant with 21% of variance in reading achievement and 24% of variance in math achievement existing between schools. In 2012-13, 21% of the variance in reading also was attributed to school differences and 27 % in math.

Table 3.
Variance Components and IntraClass Correlation Coefficients

Variable	σ_2	τ	ICC(1)	Chi Square
2011-12 Reading Achievement	6411.00	1657.59	.21	4168.14**
2011-12 Math Achievement	7249.40	2244.45	.24	4432.06**
2011-12 Reading Achievement	6698.26	1783.53	.21	3972.53**
2012-13 Math Achievement	6683.29	2594	.27	69.27**

Note. ** p < .01.

Achievement Differences

Our first validity concern is with achievement differences between students in A/B schools and C, D, and F schools. Results are reported in table 4. For reading, results in 2011-12 indicate that reading averages in C schools were approximately 15 scale points lower than averages in A/B schools. This difference was not statistically significant and fell well within the measurement error of the test (CTB/McGraw-Hill, 2013). Average reading differences between D schools and A/B (γ_{j02} = -40.09, p <.01), and F schools and A/B (γ_{j03} = -41.03, p<.01) were statistically significant with small effect sizes (Cohen's d = .44). Approximately 25 scale points separated students in C schools from students in D and F schools. There was virtually no difference between students in D schools and F schools.

Reading results in 2012-13 report smaller achievement differences and effect sizes compared to 2011-12 estimates. Although not statistically significant, reading averages in C schools were 2 scale points higher than reading averages in A/B schools. Reading averages in D schools were 5 scale points lower than A/B schools. The only statistically significant difference was between F schools and A/B schools (γ_{j03} = -16.92, p<.05). This difference falls within the measurement error of the test (SEM = 33) with a small effect size (Cohen's d = .18).

Table 4.
Differences in Reading Achievement Between Students in A/B Schools and C, D, and F Schools for the 2011-2012 and 2012-2013 School Years

Fixed Effect	Reading Achievement 2011-12	Reading Achievement 2012-13
Intercept	696.00 (1.6)**	699.60 (1.4)**
C	-15.00 (9.3)	2.21 (8.7)
D	-40.09 (7.8)**	-5.34 (6.6)
F	-41.03 (7.9)**	-16.92 (7.04)*
%Minority	-0.27 (0.2)	-0.02 (0.1)
Prior Reading	0.15 (0.05)*	0.57 (0.1)**
% FRL	-0.39 (0.15)*	-0.12 (0.1)
Minority Slope		
Intercept	-15.37 (1.9)**	-22.22 (2.07)**
C	6.25 (5.0)	2.98 (11.1)
D	13.40 (3.6)**	4.92 (5.9)
F	7.47 (9.0)	0.43 (5.1)
FRL Slope		
Intercept	-12.01 (2.2)**	-18.59 (2.2)**
C	6.93 (6.5)	17.96 (12.4)
D	26.24 (5.8)**	9.58 (6.2)
F	48.69 (7.1)**	18.03 (5.3) **
Deviance	173516	175600
Δ Deviance	581	757
Explained Between School Variance	94%	95%

*Note. * p<.05, **p<.01. We had valid reading data for 14,978 students in the2011- 2012 school year and 15,116 in the 2012-2013 school year. Unstandardized regression coefficients come from random intercept and slopes as outcomes models. Standard errors are reported in parentheses. Student controls include FRL status and minority status. Contextual controls include prior reading achievement, percent minority, and FRL rate. All variables were grand-mean centered and restricted maximum likelihood estimation was used. Dummy coding was used for school letter grade. A and B schools were set as the referent group; thus, regression coefficients represent the mean difference between students in C schools and A/B school; students in D schools and A/B schools; and students in F schools and A/B schools. ⊿ Deviance represents the difference between the null model and random intercepts and slopes as outcomes model. All changes were statistically significant from the null model.*

For math, results for 2011-12 report a mean difference between C and A/B of 23 points (γ_{j01} = -23.69, p<.01) with a small effect size (Cohen's d = .23). A mean difference of 41 points between D and A/B (γ_{j02} = -41.64, p<.01) with a small effect size (Cohen's d = .42). And a mean difference of 60 points between F and A/B (γ_{j03} = -60.43, p<.01) with a medium effect size (Cohen's d = .62).

Achievement differences were smaller in 2012-13 than 2011-12. Average math scores in C schools were 3 scale points lower than averages in A/B schools. Students in D schools averaged 24 points lower than students in A/B schools (γ_{j02} = -24.61, p<.01) with a small effect size (Cohen's d = .25). Students in F schools averaged 30 points lower than students in A/B schools with a small effect size (Cohen's d = .31). The average measurement error of the test was around 22 scale points (CTB/McGraw-Hill, 2013).

Table 5.
Differences in Math Achievement Between Students in A/B Schools and C, D, and F Schools for the 2011-2012 and 2012-2013 School Years

Fixed Effect	Math Achievement 2011-12	Math Achievement 2012-13
Intercept	697.21 (1.8)**	696.57 (1.4)**
C	-23.69 (9.3)**	-3.15 (8.3)
D	-41.64 (10.3)**	-24.61 (6.3)**
F	-60.43 (12.9)**	-30.33 (6.7)**
%Minority	-0.63 (0.2)**	-0.31 (0.1)*
Prior Math	0.27 (0.06)**	0.67 (0.1)**
% FRL	-0.04 (0.18)	-0.13 (0.1)
Minority Slope		
Intercept	-23.04 (1.8)**	-24.13 (1.9)**
C	10.72 (6.0)	3.19 (10.5)
D	14.03 (5.3)*	2.73 (5.4)
F	9.48 (8.0)	3.22 (4.6)
FRL Slope		
Intercept	-12.75 (2.4)**	-17.91 (2.0)**
C	10.16 (7.0)	12.42 (11.5)
D	29.01 (6.3)**	14.10 (5.5)*
F	34.82 (10.2)**	19.57 (4.6) **
Deviance	176850	176687
Δ Deviance	632	722
Explained Between School Variance	91%	96%

*Note. * p<.05, **p<.01. We had valid math data for 15,105 students in the2011- 2012 school year and 15,151 in the 2012-2013 school year. Unstandardized regression coefficients come from random intercept and slopes as outcomes models. Standard errors are reported in parentheses. Student controls include FRL status and minority status. Contextual controls include prior reading achievement, percent minority, and FRL rate. All variables were grand-mean centered and restricted maximum likelihood estimation was used. Dummy coding was used for school letter grade. A and B schools were set as the referent group; thus, regression coefficients represent the mean difference between students in C schools and A/B school; students in D schools and A/B schools; and students in F schools and A/B schools. ⊿ Deviance represents the difference between the null model and random intercepts and slopes as outcomes model. All changes were statistically significant from the null model.*

FRL and Minority Achievement Gaps

Our second validity concern was related to achievement gaps within and between schools. We estimated differences in FRL and minority achievement gaps between A/B and C, D, and F schools (Table 4 and Table 5). Negative intercepts indicate that on average minority and FRL students scored lower than non-minority, non-FRL peers. Positive parameter estimates for the grade comparisons of slopes indicate a smaller achievement gap for schools receiving lower letter grades.

Achievement gaps in reading. Parameter estimates for reading achievement in 2011-12 show that the minority gap was smaller in C, D, and F schools than in A/B schools. The only statistically significant difference was between D and A/B (γ_{j12} = 13.40, p<.01) with the gap being

nearly 13 points smaller in D schools. For 2012-13, the differences in minority achievement gaps were not statistically significant, but the gaps were still smaller in C, D, and F schools compared to A/B.

The FRL reading gap moved in a similar direction as the minority gap. The FRL gap was lower in C, D and F schools compared to A/B schools. In 2011-2012, statistically significant differences were found between D and A/B ($\gamma_{j22} = 26.24$, $p<.01$) and between F and A/B ($\gamma_{j23} = 48.69$, $p<.01$). In 2012-2013, the only statistically significant differences were found with F schools ($\gamma_{j23} = 18.03$, $p<.01$).

Not only were reading gaps lower in C, D, and F schools, but minority and FRL students in the lower ranking schools outperformed their minority and FRL peers in higher ranked schools. Figures 2 and 3 illustrate this result. The figures report the relationship between reading performance and school GPA in 2011-2012 and the index score in 2012-13 for both FRL and non FRL students. Both GPA and the index scores are continuous variables used to derive the school grade. As illustrated, the FRL gap increases considerably as GPA and index scores increase. Further, FRL students in the lowest performing schools actually had higher average achievement than their FRL peers in the highest ranked schools.

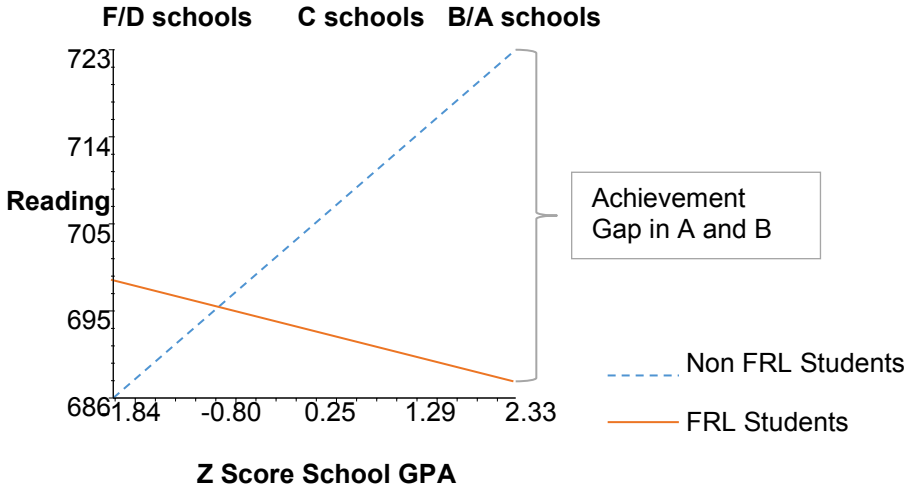

Figure 2. Cross level interaction for FRL and reading achievement, 2011-12

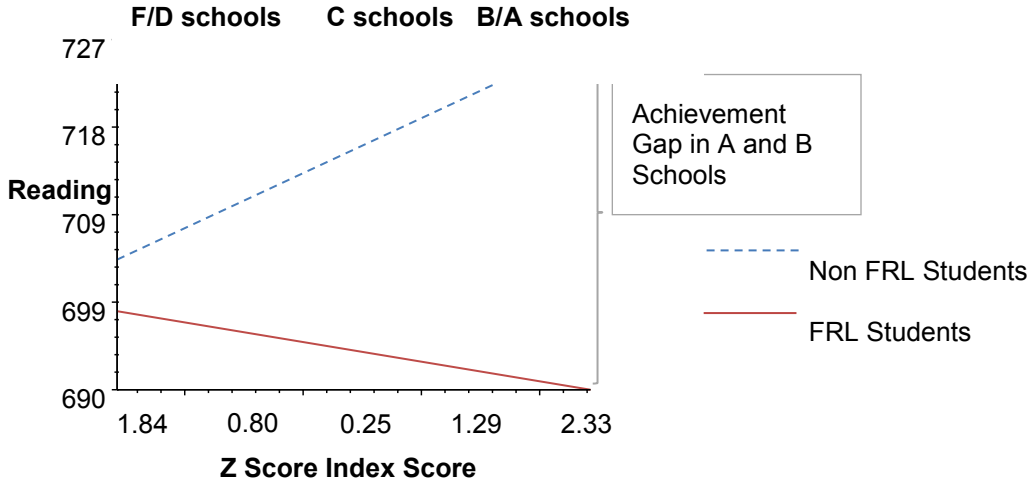

Figure 3. Cross level interaction for FRL and reading achievement, 2012-13

Achievement gaps in math. Parameter estimates for math achievement reveal a similar pattern as with reading. Negative intercepts show generally lower math achievement for minority and FRL students. As with reading achievement, positive parameter estimates show that the minority gap was lower in C, D, and F schools than in A/B. In 2011-2012, the only statistically significant relationship was between D and A/B ($\gamma_{j12} = 14.03$, $p<.01$). Similar to reading, lower minority gaps in math achievement for 2012-13 were not statistically significant. Math gaps for FRL students were also lower in C, D, and F schools compared to A/B schools. In 2011-2012, statistically significant differences were found between D and A/B ($\gamma_{j22} = 20.10$, $p<.01$) and F and A/B ($\gamma_{j23} = 34.82$, $p<.01$). For 2012-2013, math gaps in D schools ($\gamma_{j22} = 14.10$, $p<.05$) and F schools ($\gamma_{j23} = 19.57$, $p<.01$) remained statistically significant but the difference was less than the differences in 2011-2012.

Figures 4 and 5 illustrate the relationship between math achievement and GPA and Index score for FRL and non FRL students. As with reading, achievement gaps increased as GPA and index scores increased. The largest achievement gaps existed in the highest rated schools. Additionally, FRL students in the lowest rated schools outperformed their FRL peers in the highest rated schools.

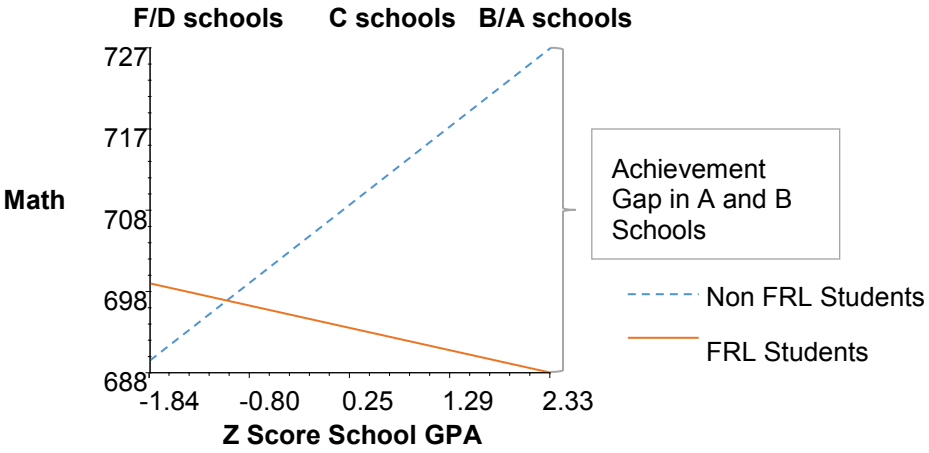

Figure 4. Cross level interaction for FRL and math achievement, 2011-12

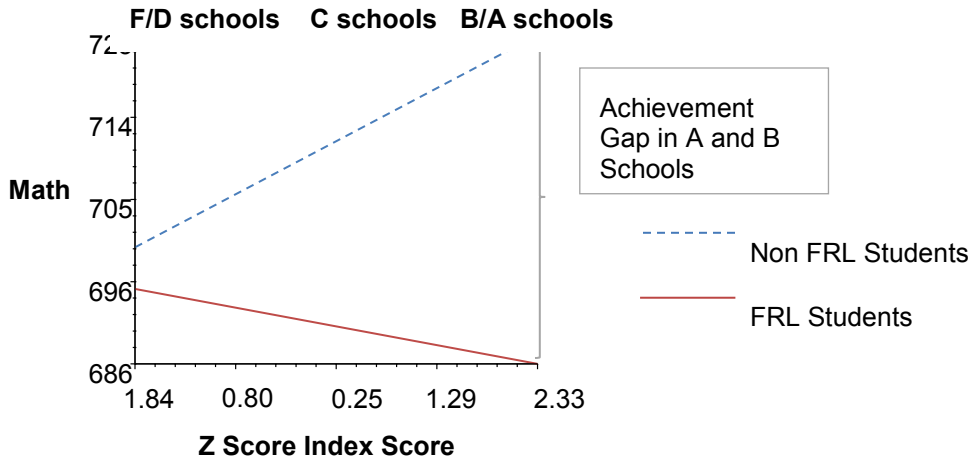

Figure 5. Cross level interaction for FRL and math achievement, 2012-13

In summary, actual achievement differences between letter grades were small and in most cases not statistically significant when student and school characteristics were held constant. School grades did moderate achievement gaps, but gaps moved in a direction opposite from what would be desired of an accountability system that measured achievement equity. To account for achievement equity, the FRL and minority gaps need to depreciate as letter grades improve. This, however, was not the effect of letter grades. Achievement gaps were larger in schools with higher grades.

Discussion

A-F letter grades are the engine of Oklahoma's accountability system. Schools receiving an A or B receive recognition and status as high quality places to teach and learn. In contrast, schools earning lower grades are judged to be ineffective and face increased reform pressure and State intervention. We agree that states need information by which they can identify and support lower

performing schools. At issue is the degree to which assigning schools a single A-F grade can result in decisions and action that improve learning opportunities and outcomes. If consequential decisions are predicated on a single letter grade, it is necessary for school grades to meet a standard of validity that supports its use (Latham & Wexley; Messick, 1995). We address the validity question first at a conceptual level and in light of our findings.

Conceptual Problems with Use of A-F Grades

Conceptually, it is hard to envision how a single letter grade, or any single indicator for that matter, could support the policy objectives of the US Department of Education. A single, composite letter grade elicits more questions than it provides clear answers about student performance. For example, does an A mean that students demonstrated proficiency in all tested subjects, grades, and classes? Or, did students perform better in certain subjects, classes, and grades but not others? Can we interpret a letter grade as a percentage of students scoring proficient? Can we conclude that an A means 90% or more of the students met or exceeded proficiency? Does an F mean only 50% of the students exceeded proficiency? These ambiguities effectively increase uncertainty about student achievement, leaving decisions and actions on what to improve or how to improve to conjecture or chance.

Assumptions behind letter grades simply do not correspond with the dynamic nature of schools and student learning. For instance, we know from large-scale school and teacher effects studies that the majority of achievement variance is within, not between schools (Heck, 2009; Nye, Konstantopoulos, & Hedges, 2004). Educators, policymakers, and the public need knowledge of within-school variance in teaching and learning, but instead of measuring and reporting this variability, grades collapse achievement differences into a single indicator. As a result, lower achieving students receive the same performance status as higher achieving students, essentially ignoring variance that can help schools recognize and respond to unmet student needs.

The inability of a single grade to reflect achievement variance has a profound effect on studying and improving achievement gaps. Not accounting for subgroup performance in the calculation of grades effectively hides the achievement of poor and minority students. By hiding achievement differences, grades make it difficult to sort schools by those with high average achievement and narrow achievement gaps from those with high achievement but large achievement gaps. A-F letter grades are not designed to be diagnostic, but at the very least they need to reflect the performance of all student groups. Rewarding schools for large FRL and minority achievement gaps and penalizing schools whose poor, minority students outperform peers in more affluent schools, is neither fair nor helpful in reducing achievement gaps and improving equity.

In regards to continuous improvement, A-F grades deliver very little, if any, instructional value to teachers and administrators. They hide achievement differences, they cannot be disaggregated by content standards, and they do not measure student growth toward college, citizenship, and career ready expectations. Furthermore, school grades cannot be used to measure the effectiveness of improvement strategies or interventions; any change from one year to the next is just as likely attributable to factors outside school control than to what happens within schools and classrooms. Conceptually, grades have little information value for improvement. Our empirical results raise additional questions about the use of A-F grades as a policy instrument.

Empirical Results

Evidence presented in this paper makes a weak case for using a grade as a tool to meet the policy objectives of improved learning, closed achievement gaps, and improved equity. To be useful as an external control, grades need to tell us something meaningful about school quality. We cannot conclude from our evidence, however, that the average student in a C, D or F school would have performed any better in math and reading than in an A or B school. Rather than providing a simple and clear indication of school performance, grades increase uncertainty about student achievement and achievement gaps.

For the schools and students we studied, letter grades were not capable of sorting schools by achievement or achievement gaps. For 2011-2012, reading differences between students in C schools and A/B schools were not statistically significant and well within the measurement error of the test. Differences with D and F schools were statistically significant but the effect sizes were small by Cohen's standards. These effects sizes translate to roughly a 73% overlap in the distributions of scale scores between the highest ranked and lowest ranked schools (Cohen, 1988), meaning the distributions are more similar than discrete. Raw score differences amounted to about four questions on a 50 item test. It is important to also note that there was no difference in reading achievement between D and F schools.

Results for 2012-13 demonstrate that school grades are incapable of differentiating achievement differences in reading. Although not statistically significant, students in C schools had higher average reading achievement than students in A and B schools. The difference between A/B and D was not statistically significant and was only five scale points, fewer than one question on a 50 item test. Additionally, students in F schools only scored 17 scale points lower than students in A and B schools. This difference is well within the average standard error of the reading test (CTB-McGraw-Hill, 2013), and translates to about a 2 ½ question difference between schools that qualify for state rewards and those that face state sanctions, a small margin for such high stakes.

As for math, in 2011-12 the mean difference between students in A/B schools and C schools fell within the measurement error of the test (Pearson, 2012). Additionally, effect sizes for C and D schools were small. There was about an 86% overlap in the distribution of math scores for C schools and A/B schools, and about a 73% overlap with D schools. The effect size between A/B and F was medium with about a 62% overlap. Similar to reading, achievement differences in 2012-13 were considerably smaller than 2011-12. There was no statistically significant difference between students in C schools and A/B. The difference with D was small and resulted in an 86% overlap in achievement scores. Likewise, the difference with F was also small with about a 79% overlap.

Letter grades also failed to reflect achievement gaps in the 2011-2012 and 2012-2013 school years. Grades did not capture achievement gaps when holding differences in student composition constant. Achievement gaps in reading and math were larger in A and B schools than C, D, and F schools. Also, reading and math achievement of minority and FRL students remained higher in the lowest ranked schools (D and F) compared to the highest ranked schools (A and B). D and F schools were on average more effective with FRL and minority students.

In sum, the evidence raises serious concerns with the use of Oklahoma's letter grades as a quality indicator. School grades do not accurately represent achievement patterns within schools, nor are they suitable for distinguishing between higher performing and lower performing schools. Some A and B schools earned high grades at the expense of their low income, minority students. This is not the intent of NCLB waivers. Accountability measures need to support high achievement and achievement equity, but grades effectively conceal within-school achievement differences between student subgroups and across subjects and grades. FRL and minority students in D and F schools

had higher average achievement than comparable peers in A and B schools. The imprecision of a letter grade weakens the viability of the accountability system.

Comparison of Grading Formulas

The empirical test also provided some comparative evidence related to the original and revised formulas used to calculate school grades. Letter grades in 2011-12 were based on three components (achievement status, achievement growth, and whole school performance) and used a weighting scheme of 0, 0.2, 1.0, and 1.2. The whole school component was eliminated in 2012-13 and the weighting scheme was reduced to a binary 0 or 1. Changes to the formula appear to have affected the distribution of school grades. In 2012-13 the number of C schools dropped from 21% to 5 % and the number of F schools increased from 8% to 53%. Changes to grade distributions occurred even though school demographics remained similar and average math and reading achievement were stable. Thus, it is unlikely that the considerable increase in the number of F schools was the result of lower math and reading achievement in the 2012-13 school year.

Changes to the grading formula in 2012-13 did not improve accuracy; in fact, letter grades in 2012-13 performed worse as an indicator of school effectiveness. Grades did not correspond meaningfully with the achievement differences across schools. Claims that A and B schools perform better, or at least are more effective, than C schools, or even D schools, are indefensible given the absence of achievement differences between these groups. Additionally, achievement distributions of students in F schools were more similar to students in A/B schools than different. Small effect sizes between these groups do not correspond to conventional understanding of an A or an F.

We conjecture that the decision to move to a binary scoring scheme for the achievement and growth index scores exacerbated validity problems found in 2012-13 grades. Dichotomizing achievement compounds grouping error and leaves achievement variance unexplained. We illustrate the measurement problem with two common scenarios. First, two students, one who scores just above the proficient threshold and one who scores advanced, receive the same score for their school even though the difference in their average achievement is substantial. Second, having many students score around the proficiency threshold biases estimates as well. Average achievement among two schools may be similar in this case, but the school with more students above the threshold will score better than the school with more students slightly below. Simply, letter grades do not portray meaningfully the variance profile in student achievement when scale scores are collapsed into a dichotomous category. This problem is compounded when test scores from multiple subjects are combined to form an overall composite grade.

Letter grades in both years did not expose existing achievement gaps. The problem appears to be more severe in 2011-12 when minority and FRL achievement gaps were considerably larger in A/B schools than C, D, and F schools. Even though achievement gaps were smaller in 2012-13, letter grades remained largely unaffected by the lower achievement of minority and FRL students in A and B schools. Evidence that FRL and minority students had higher achievement in D and F schools than A and B impugns the formula used to calculate school grades. The distribution of letter grades would change quite drastically if the state assigned achievement gaps the same weight it assigns to achievement status. In our sample, several D and F schools would become C or B schools, and many A and B schools would become C or D schools. Poor, minority students end up being left behind when grades obscure achievement differences within schools.

Conclusion

With new accountability systems taking effect across the country, researchers and policy analysts have raised concerns about measurement flaws that jeopardize the ability of accountability indicators to advance high and equitable achievement (Hall, 2013; Polikoff, McEachin, Wrabel, & Duque, 2014). Without empirical evidence from student test scores, it is hard to know if concerns based on reviews of design features signal legitimate threats to the use of indicators for improvement purposes. We leveraged a unique two year data set to test the valid use Oklahoma's A-F school grades to achieve the broad policy objectives set by the federal government. It turns out that for our sample of schools, grades did a poor job of sorting schools by different effectiveness levels. The data we analyzed demonstrate quite dramatically that letter grades have very little meaning and certainly cannot be used legitimately to inform high-stakes decisions about schools.

Given limitations of accountability indicators used as authority and incentive tools, we argue that next-generation accountability systems need to be designed with capacity building and continuous improvement in mind. Accountability used to induce change by pressuring schools to improve does not align with new, more challenging expectations that schools deliver deeper learning experiences that result in college, career, and citizenship ready students (Darling-Hammond, Wilhoit, & Pittenger, 2014). The goal of readiness for meaningful participation in a post-industrial economy lays out an educational vision that goes far beyond test score proficiency as the standard to judge schools. Higher learning standards set the right expectations, but in setting this vision states have a responsibility to design accountability policies so that schools receive the support needed to enact reforms that make teaching learning measurably better.

As demonstrated with our findings, a single accountability indicator used to determine rewards and interventions fails to provide information that is suitable for generating useful knowledge about teaching, learning, and learning opportunities afforded students. A new approach to accountability is needed. In moving toward a framework supportive of capacity building, we advance three principles as foundational for knowledge creation and improvement: shared accountability, adaptive improvement, and informational significance.

Shared accountability recognizes that school effectiveness is a collective responsibility of all actors in the educational system, and as such, performance information needs to cover processes and resources that lead to school outcomes. Adaptive improvement reflects the fact that schools differ by their capacity to achieve desired outcomes, necessitating a system that provides information useful no matter where schools are at in their improvement journeys. Informational significance aims to place relevant and useful information in front of responsible actors who are expected to make intelligent decisions about the developmental needs of students and the future direction of teaching and learning. Accountability policy designed with these three principles in mind would replace a single, composite indicator with multiple indicators that combine to present a holistic profile of school resources, teaching and learning processes, and student outcomes.

A-F grades do not function as a capacity tool, and the measurement limitations we identified in our sample shows that they are even problematic as an authority/incentive tool. Even though our sample is limited, we know Oklahoma's accountability system shares similar components with other states. Other states use proficiency bands without reporting results by subgroups, use achievement of the bottom 25% to fulfill the achievement gap requirement, and use a composite indicator to judge school effectiveness. These three components are likely to behave the same way in a different context with different schools. It is not variability in schools that presents a problem, but rather weaknesses of the components to measure achievement variance within schools, as well as the failure to account for resources and processes that make schools engaging places to teach and learn.

References

AERA, APA, NCME (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Adams, C. M. (2013). Collective trust: A social indicator of instructional capacity. *Journal of Educational Administration*. 51(3), 1-36. http://dx.doi.org/10.1108/09578231311311519

Adams, C. M., Forsyth, P. B., Ware, J., & Mwavita, M. (in press). The informational significance of A-F school accountability grades. *Teachers College Record*.

Ayers, J. (2011). *No child left behind waiver applications: Are they ambitious and achievable? Center for American Progress*, Washington, DC. Retrieved from http://files.eric.ed.gov/fulltext/ED535638.pdf

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 from 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. http://dx.doi.org/10.3102/00028312042002231

Carlson, D. (2006). *Focusing state educational accountability systems: Four methods of judging school quality and progress*. Dover, NH: The Center for Assessment. Retrieved from http://www.nciea.org/publications/Dale020402.pdf.

CEP (2012). *Accountability issues to watch under NCLB waivers*. The George Washington University, Center on Education Policy. Retrieved from http://files.eric.ed.gov/fulltext/ED535955.pdf.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 443-507). Washington, DC: American Council on Education.

CTP McGraw Hill (2013). *Oklahoma school testing program, Oklahoma core curriculum tests: Grades 3 to 8 assessments, 2012-2013 technical report*. Retrieved from http://ok.gov/sde/accountability-state-testing-results

Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). Accountability for college and career readiness: Developing a new paradigm. *Educational Policy Analysis Archives, 22*(86), 1-25.

Domaleski, C., & Perie, M. (2013). *Promoting equity in state education accountability systems.* National Center for the Improvement of Educational Assessment, Center for Educational Testing and Evaluation, University of Kansas.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121-138. http://dx.doi.org/10.1037/1082-989X.12.2.121

Figlio, D. N., & Getzler, L. S. (2002). *Accountability, ability and disability: gaming the system*. Working paper 9307. Cambridge, MA: National Bureau of Economic Research. http://dx.doi.org/10.3386/w9307

Finnigan, K. S., & Gross, B. (2007). Do accountability policy sanctions influence teacher motivation? Lessons from Chicago's low-performing schools. *American Educational Research Journal, 41*(3), 594-630. http://dx.doi.org/10.3102/0002831207306767

Forte, E. (2010). Examining the assumptions underlying the NCLB federal accountability policy

on school improvement. *Educational Psychologist, 45*(2), 76-88. http://dx.doi.org/10.1080/00461521003704738

Haladyna, T. M., Nolen, S. R., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 42*(17), 2-7. http://dx.doi.org/10.3102/0013189X020005002

Hall, D. (2013). *A step forward or a step back? State accountability in the waiver era.* The Education Trust. Retrieved from http://files.eric.ed.gov/fulltext/ED543222.pdf

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. http://dx.doi.org/10.1108/09578231311325659

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know.* Cambridge, MA: Harvard Press.

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. http://dx.doi.org/10.3102/0162373708317689

Ho, A. D. (2008). The problem with proficiency: Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*(6), 351-360. http://dx.doi.org/10.3102/0013189X08323842

Howe, K. R., & Murry, K. (2015). Why school report cards merit a failing grade. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/publication/why-school-report-cards-fail

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed. pp 17-64). Westport, CT: American Council on Education/Praeger.

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. http://dx.doi.org/10.1257/089533002320950993

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

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. http://dx.doi.org/10.14507/epaa.v13n33.2005

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. http://dx.doi.org/10.3102/01623737024001029

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into school meaning. *American Psychologist, 50*(9), 741-749. http://dx.doi.org/10.1037/0003-066X.50.9.741

McNeil, M. (2012). States punch reset button with NCLB waivers. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2012/10/17/08waiver_ep.h32.html?tkn=NSLFJ%2BWQnkqPlIMGUAUBakJda6JiHNTaJZDt&intc=es.

Miller, D. M. (2008). Data for school improvement and educational accountability: reliability

and validity in practice. In K. Ryan & L. Shepard (Eds.), *The future of test-based educational accountability* (pp. 249-262). New York: Routledge.

Mintrop, H., & Sunderman, G. L. (2009). Predictable failure of federal sanctions-driven accountability for school improvement and why we may retain it anyway. *Educational Researcher, 38*(5), 353-364. http://dx.doi.org/10.3102/0013189X09339055

Neal, D. & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics, 92*(2), 263-283. http://dx.doi.org/10.1162/rest.2010.12318

Nye, B., Konstantopoulos, S. & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237-257. http://dx.doi.org/10.3102/01623737026003237

OCEP & CERE. (2012). *An examination of the Oklahoma State Department of Education's A-F report card.* The Oklahoma Center for Education Policy, University of Oklahoma, and The Center for Educational Research and Evaluation, Oklahoma State University.

Oklahoma State Department of Education. (2012, April). *A-F Report Card Guide*. Retrieved from http://ok.gov/sde/f-grading-system

Pearson, Inc. (2012) *Technical Report of the Oklahoma School Testing Program, Oklahoma Core Curriculum Tests, Grades 3 to 8 Assessments*, Pearson, inc.

Polikoff, M., McEachin, A., Wrabel, S. Duque, M. (2014). The waive of the future? School accountability in the waiver era. *Educational Researcher*. Retrieved from http://www-bcf.usc.edu/~polikoff/Waivers.pdf

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

Raudenbush, S. W., & Willms, D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*(4), 307-335. http://dx.doi.org/10.2307/1165304

Rothstein, R. (2009). Getting accountability right. *Education Week*. Retrieved from 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.

Ryan, K. E. (2008). Fairness issues and educational accountability. In K. Ryan and L. Shepard (Eds.), *The future of test-based educational accountability* (pp. 191-208). New York, NY: Routledge.

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

Schneider, A., & Ingram, H. (1990). Behavioral assumptions of policy tools. *Journal of Politics, 52*(2), 510-529. http://dx.doi.org/10.2307/2131904

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

Sirotnik, K. A. (2002). Promoting responsible accountability in schools and education. *The Phi Delta Kappan, 83*(9), 662-673. http://dx.doi.org/10.1177/003172170208300908

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement Issues and Practice, 16*(2), 5-24. http://dx.doi.org/10.1111/j.1745-3992.1997.tb00585.x

US Department of Education. (2012). *EASA Flexibility – Request*. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

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

Willingham, W. W., & Cole, N., S. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum.

About the Authors

Curt M. Adams
University of Oklahoma
Curt.Adams-1@ou.edu
Curt Adams is the Linda Clarke Anderson Presidential Professor at the University of Oklahoma and co-director of the Oklahoma Center for Education Policy.

Patrick B. Forysth
University of Oklahoma
Patrick.forsyth@ou.edu
Patrick Forsyth is a professor of education at the University of Oklahoma and co-director of the Oklahoma Center for Education Policy.

Jordan K. Ware
University of Oklahoma
jordware@gmail.com
Jordan K. Ware is a post-doctoral fellow at the University of Oklahoma with the Oklahoma Center for Education Policy.

Mwarumba Mwavita
Oklahoma State University
mwavita@okstate.edu
Mwarumba Mwavita is an assistant professor of Research, Evaluation, Measurement, and Statistics at Oklahoma State University and director of The Center for Educational Research and Evaluation.

Laura L. Barnes
Oklahoma State University
lbarnes@okstate.edu
Laura L. Barnes is an associate professor of Research, Evaluation, Measurement, and Statistics at Oklahoma State University and a research scientist with The Center for Educational Research and Evaluation.

Jam Khojasteh
Oklahoma State University
jam.khojasteh@okstate.edu
Jam Khojasteh is an assistant professor of Research, Evaluation, Measurement, and Statistics at Oklahoma State University and a research scientist with The Center for Educational Research and Evaluation.

education policy analysis archives

Volume 24 Number 4 January 11th, 2016 ISSN 1068-2341

archivos analíticos de políticas educativas
consejo editorial