Purpose

A significant amount of public trust has been placed on school A-F ratings without analyzing whether it is well-founded, independent of how school ratings are used. This study evaluates public school designations, A-F Letter Grades, and applies statistical analysis to better understand what the designations really tell us about schools. Specifically- what conclusions can we safely draw from a school's designation? It makes explicit unspoken assumptions underlying the use of school designations in *any* way— primarily that the designations do, indeed, reflect school quality and that the distinctions made by the designations (an *A* vs. a *B,* for example) are meaningful. However, given the consequential nature of these labels, this assumption must be tested and validated to ensure that schools are accurately evaluated. This study addresses the following research question:

> *Are the A-F Letter Grades measuring a school's contribution to learning or simply repackaging other indicators, such as poverty, and presenting them as measures of school quality?*

Perspective

With at least a dozen states utilizing accountability models that assign A-F Letter Grades to schools little research has been conducted about the validity of these designations and their methods for evaluating schools. A southwestern state, used in this study, adopted the A-F accountability system in 2011, modeled after Florida, as the latest iteration of a school accountability law initially passed in 2000 which provided additional funding for education in exchange for increased accountability for schools and teachers; teachers became subject to performance pay plans and schools became subject to school accountability designations (X.R.S. § 15-241). Regardless of the state, it is clear that even though school accountability ratings may evolve over time their purpose remains the same; to convey a judgment as to the school's quality or effectiveness at educating students. The message is quite simple: *A* schools are doing a great job at educating students and *F* schools are failing their students altogether.

While the general message may be simple, the subsequent uses of these school labels are not. In this southwestern state, school accountability designations are more than simply informational; they are an explicit tool for school reform with the school designation used to trigger the school improvement process, despite little research evidence that this improves student outcomes (Hanushek, E.A. & Raymond, M.E, 2005). School improvement policies have been developed based on the belief that the school rating captures meaningful information about school quality and the publication of the rating will elicit the desired behavior from parents, students, teachers, and/or administrators. Given this, it was necessary and appropriate to determine whether the state's A-F rating system truly provided valuable information to stakeholders about school quality.

The relationship between poverty and measures of achievement (e.g., percent of students passing the state's assessment) has long been a limitation of measurement of student

achievement in education (Kane, T.J. & Staiger, D.O., 2002). Unless otherwise mitigated, poverty can have a significantly deleterious effect on learning. Students growing up in poverty often arrive at school with challenges that adversely affect their ability to learn or to learn at the same rate as their wealthier counterparts. This relationship between poverty and learning has been widely known (and acknowledged) since the 1960's and has resulted in the federal Title I program providing additional resources to schools that serve a significant percentage of students in poverty.

The contribution that a school itself makes to student learning is difficult to determine, when we know from the start that the wealthier the students in a school are, the higher that school's scores will be, simply as a function of the student population. In other words, we expect wealthier students to score higher on standards-based proficiency measures than low-income students. This is not to say that schools do not make a difference; they do, but it is often difficult to measure their effect.

This is, in part, the reason why so much attention has been given to measures of student "growth," such as value-added models (VAM) and student growth percentiles (SGP)—a relatively new advancement in educational measurement made possible by data systems that allow student-level data to be linked across multiple years (Betebenner, 2008). VAM and SGP analyses allow the researcher to "control for," or set aside statistically, the relationship between poverty (and potentially other variables) and measures of achievement, in order to measure the effect that a school or teacher has had on an individual student's learning.

The state's A-F Letter Grade models utilize a combination of student achievement (percent passing) and student growth (most often using SGP) data for schools, though the exact calculations vary by model (XDOE, 2012). The 50/50 distribution of points between achievement and growth communicates to educators that both of these factors are equally important. Students must show growth, but they must also demonstrate proficiency on academic standards. The inclusion of growth in the accountability formula is meant to give credit to schools for their effect on student learning, and to address the limitations that we know exist when using only proficiency results to measure school quality. In other words, when combined with proficiency rates, growth is meant to erase, or mitigate, the negative relationship that exists between poverty and measures of student achievement.

Methods

In order to determine whether the A-F accountability model was accurately reflecting school proficiency and growth, as intended in the model specification, correlation statistics were used to determine the relationship between poverty, using the percent of students eligible for the National School Lunch Program (FRL), and the different types of points (Composite, Growth, and Total) resulting from the accountability model. Analyzes were conducted for the traditional,

small, and alternative models for the elementary, high school and combination (spanning multiple grade configurations) school configurations.

Researchers utilized the publicly released 2012 state accountability results for all public schools receiving classifications (N=1,968) obtained from the Department of Education's web page. These excel data files contained school level information regarding the type of school/accountability model, charter/district designation, number of total, composite and growth points, and the overall letter grade. These files were merged with publicly available National School Lunch Program (FRL) data for each school.

Lastly, researchers used masked student level SGP data files obtained from the XDOE to model an alternative growth point calculation. Rather than using medians and averages, researchers created alternative growth points utilizing the percentage of students' SGPs categorized as *low*, *typical*, or *high growth*. For this analysis, an individual SGP of 1-33 is considered *low*, 34-66 is *typical*, and 66-99 is *high*. In this analysis, Growth Points were earned based on the percentage of students who achieved high growth (or high and typical growth), much in the same way that schools earn points based on the percentage of students who pass the state assessment. These alternative points were calculated at the school level for all schools with existing SGP data.

Results

Figures 1a, 1b and 1c show the result of the analysis for Traditional elementary schools. The scatter plots graphically show the relationship between poverty and Composite Points, Growth Points, and Total Points. In each case, points are graphed along the Y (vertical) axis and Percent FRL is graphed along the X (horizontal) axis. In this discussion we use "Composite Points" and "percent passing" interchangeably because virtually all (100 out of 103) Composite Points available to elementary schools are based on the percent of students passing the state's standards based assessment.

Figure 1a, confirms the negative relationship we know exists—as poverty increases, Composite Points (or percent passing) decreases. The red line is the regression line, indicating the direction and strength of the relationship. This is a very strong statistical relationship at -.761.

Figure 1b, however, shows a much different relationship. While there is still a negative relationship between poverty and Growth, this relationship is statistically weak (-.176), indicating that there is no meaningful difference in the distribution of Growth Points among schools with varying degrees of poverty. A poor school is as likely to demonstrate the same amount of growth as a more wealthy school.

Figure 1c, shows the relationship between poverty and Total Points—the end result of combining Composite and Growth Points. While the negative relationship between poverty and Total Points (-.631) is a bit less dramatic than the -.761 of poverty to Composite Points, it is still strong and statistically meaningful. The bottom line: despite the attempt to "level the playing

field" through the use of growth points, schools with greater degrees of poverty are less likely to do well using the current school accountability system.

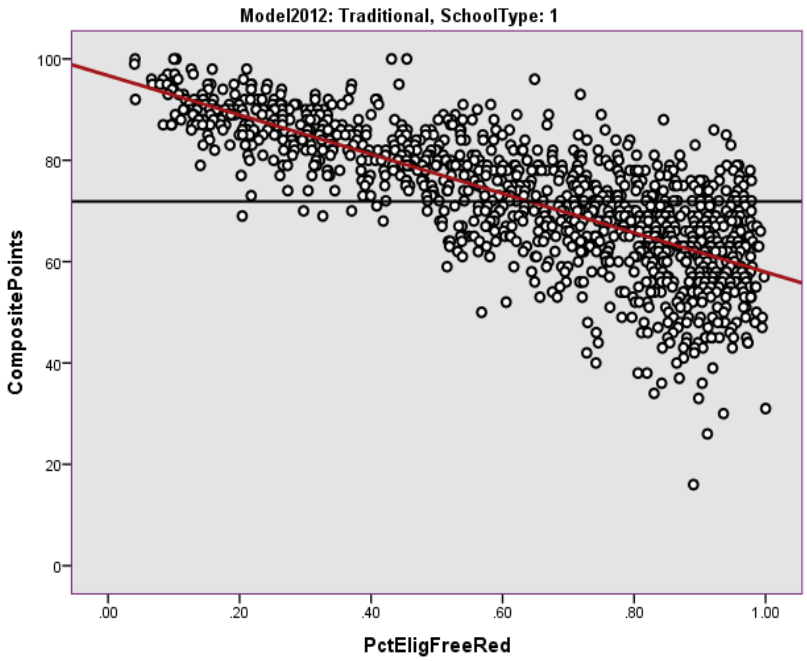

Figure 1a. Correlations between Composite Points and Percent Free and Reduced Lunch

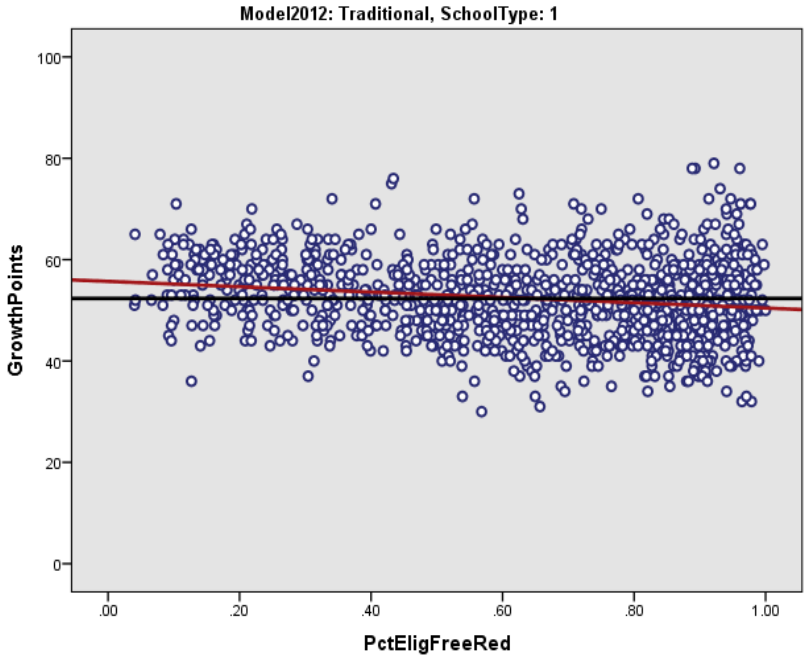

Figure 1b. Correlation between Growth Points and Percent Free and Reduced Lunch

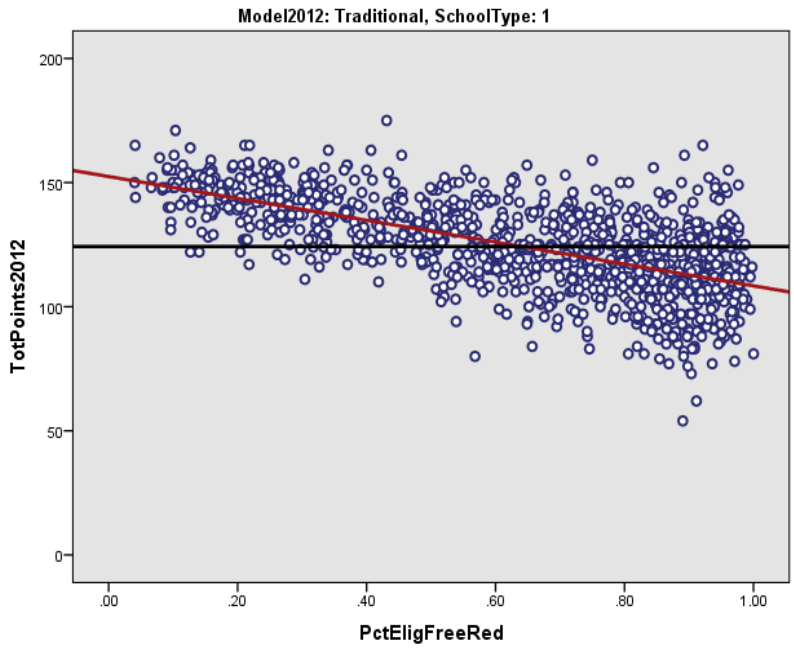

Figure 1c. Correlation between Total Points and Percentage Free and Reduced Lunch

When using the actual distribution of points, we see that what began as 50/50 weight to percent passing and growth turns into something approaching a 60/40 weight toward percent passing—clearly not the intention of the accountability model. Furthermore, the point differences resulting from the accountability formula as implemented are large enough to make a difference in the final rating of the school. Despite the inclusion of growth, the practical result of the accountability formula is yet another measure of performance that is closely related to poverty.

The current model fails to adequately control for the effect of poverty on indicators of achievement in order to measure the school's contribution to learning. This occurs not because of a limitation of the SGP model, but rather because of the methods utilized by the XDOE, specifically the calculation of a schools' median SGP. According to the XDOE's technical manual, a seven step calculation is used to compute the total growth points which are based on the calculation of four medians which are then averaged three times (XDOE, 2012, pg. 23-27).

Results of the alternative methodology for calculating growth points will be presented in the full paper.

Significance

Given the public nature of school accountability, it is clear that the A-F Letter Grade data is intended for stakeholders to use in order to make their decisions- sanctions, rewards or selection. Therefore, this finding is significant given the implications, positive and negative, that these letter grades have on schools. As the results indicate, these letter grades do not accurately

represent the performance of schools, in particularly those schools that serve a high population of students in poverty. However, no school- rich or poor- is receiving the appropriate allocation of points associated with student growth. This clearly violates the intent of the legislature and the State Board of Education's desires for a fair and balanced accountability system.

Additionally, given that over a dozen states utilize similar A-F letter grading systems that use some measure of student growth to "accurately and fairly" evaluate schools this study serves as a critical lens into potential limitations in their own models. For those states considering A-F systems, this study serves as a warning during the development phase to ensure that better methodological decisions are made to evaluate schools more effectively. Lastly, these results evidence that school accountability results are tied directly to methodological decisions made by agencies; therefore, it is critical that agencies model multiple methods to determine the most appropriate models.

References

Betebenner, D. W. (2008). *Norm- and criterion-referenced student growth.* Report of the National Center for the Improvement of Educational Assessment. Retrieved from Colorado Department of Education website: http://www.cde.state.co.us/cdedocs/Research/PDF/betebenner_norm_crit_measuresofgrowth.pdf

Hanushek, E.A. & Raymond, M.E. (2005). *Does school accountability lead to improved student performance?* Journal of Policy Analysis and Management, 24:2, pp. 297-327. dio: 10.1002/pam.20091

Kane, T. J. & Staiger, D.O. (2002). *The promises and pit-falls of using imprecise school accountability measures*. Journal of Economic Perspectives, 16:4, pp. 91-114. dio: 10.1257/089533002320950993

X Department of Education (XDOE). (2012). *A-F Letter Grade Accountability System Technical Manual* [Everything else pertaining to this citation has been removed to ensure blind peer review].