Kristin Blagg

School Accountability System Design Competition

Design Objectives

Measurement and Assessment Clarity

As a former high school and middle school math teacher, I know first-hand that school accountability systems can often appear complex and intimidating. To outsiders, these systems seem to function like a black box, taking various student- and school-level inputs and generating (sometimes inscrutable) score or performance level outputs. My priority is to create a system that parents, teachers, and administrators can easily understand. Schools and teachers should be able to clearly see and navigate a path towards improvement on metrics like test scores, student growth, and other measures.

Focus on Quantitative and Qualitative School Quality

When monitoring and assessing charter schools, many state authorizers use in-depth school site visits, in addition to state test scores and other data, to assess school quality. I argue that similar site visits should be employed, at minimum, as part of a state assessment of schools that serve students attending kindergarten through third grade. Academic achievement is typically first assessed by states in the third grade, at which point a student is expected to read and answer questions based on a given text, conduct short research projects on a given topic, multiply and divide numbers up to 100, and show a basic understanding of fractions (based on Common Core Standards). The early elementary grades are foundational for success, and it is crucial that we use additional qualitative measures to ensure that all schools provide students with a high-quality start.

Design Categories

Academic Achievement (40 Points)

In line with the requirements of the Every Student Succeeds Act (ESSA), my system would require an annual summative math and reading comprehension test for elementary schools in grades 3-5. ESSA also mandates that a science test be given at least once during the elementary school period. Evidence suggests that scientific thinking and inquiry are developmental skills that grow substantially in the elementary school years, and can be considered independent of the skills of text comprehension and general reasoning (Koerber et al 2015, Zimmerman 2007). To satisfy the ESSA requirement, and promote the development of scientific reasoning throughout elementary school, I would include an annual science test for grades 3-5, based on state science standards and age-appropriate use of the scientific method.

Not much is known about the effect of state test performance level structure and cut-off points on the behavior of schools and districts. However, there is evidence that accountability pressure, particularly when schools or districts are on the margin of adequate yearly progress (AYP), is associated with neutral to positive achievement gains (Chiang 2009, Dee and Jacob 2011, Reback et al 2014). I would use a five-tier system to assess the levels of student achievement, banking on the idea that four rather than three cut-offs would spur more schools

to coach "bubble" students into the next achievement level. Accordingly, I would increase the "value" of each achievement level, with a premium placed on the movement of students into the Basic (Tier 4) and Proficiency (Tier 3) levels . In keeping with my goal of a simple metric, the maximum score for this portion would be forty points (forty percent of the overall score), calculated as a weighted average of student scores (Table I).

Table I

	Points Awarded	Example Percentage of Students	Example Points Awarded
Tier 1 (Advanced)	40	10%	4
Tier 2 (Mastery)	35	25%	8.75
Tier 3 (Proficient)	30	35%	10.5
Tier 4 (Basic)	15	20%	3
Tier 5 (Below Basic)	0	10%	0
Total	-		26.25 points of 40 possible

This calculation would be done for all three tests, and the result would be averaged together for a cumulative score (for example, a school that earns 30 points from math, 20 points for reading, and 25 points for science would have an overall score of 25).

Overall Student Growth and Subgroup Growth (30 Points)

There are numerous value-added and growth models, and the jury is still out on which models might best capture the impact of the teacher and school on student growth. A value-added and growth metric is often used as an adjustment to "even the playing field" - a way of accounting for the socio-economic sorting that occurs between schools and districts.

My priority of developing simple, easy-to-understand metrics pushes me to opt for using a student growth percentile (SGP) approach in my accountability system. The SGP allows for teachers and parents to articulate how well an individual student is performing conditional on her prior state test performance, and allows for easy and consistent aggregation up to a class- or school-level measure, typically a median growth percentile (Betebenner 2011). Compared to value-added measures, growth percentiles are reportable at the student level and understandable: a student with a SGP of 75 performed in the 75th percentile on the 4th grade math test, compared to students who scored similarly to her on the 3rd grade math test. At both the student and school level, a score higher than 50 means faster-than-average growth, while a score below 50 means relatively slower growth.

In addition to its relative simplicity, the SGP approach allows the state to provide projections of whether a student is on-track to attain or maintain a given achievement level. These growth projections are useful for assessing the narrowing of proficiency gaps for student subgroups. The accountability requirement for subgroups in ESSA states only that a measure "take into account the improvement necessary on such measures to make significant progress" for subgroups who are behind. In particular, English learners are required to show progress towards English proficiency (I use the state reading comprehension test as this measure).

The state would develop subgroup growth targets for individual schools (dependent on the schools' unique mix of prior achievement in those groups) that would put the median subgroup student on track towards proficiency (Tier 3) within four years. For example, based on prior scores, the state might give an school a subgroup median growth percentile target of 52 for free and reduced-price lunch students, and a subgroup target of 55 for English language learners, allowing schools to develop and prioritize goals and strategies for these groups in the coming year. One drawback to this approach is that the SGP metric becomes unstable with subgroups of less than 30 individuals (Goldschmidt et al 2012). Nonetheless, a minimum subgroup size of 30 students is in line with current minimum n-sizes of other states, such as Oklahoma (Le Floch 2013).

The scoring of student growth would grant up to 20 points for the overall median growth percentile of the school, and up to 10 points for movement towards subgroup median growth percentile goals (in both cases, points are allocated based on the overall distribution of school growth scores, see an illustrative example in Table 2). If no subgroups can be identified at a given school due to subgroup size, then all 30 points are allocated using the school's overall median growth percentile.

Table 2

Overall Median Growth Percentile Value	Overall Growth Points Awarded	Average Difference from SGP Subgroup Goal	Subgroup Growth Points Awarded
40 or less	0	-5 or below	0
41	1	-4	0
42	2	-3	1
43-44	4	-2	3
45-46	6	-1 to +1	5
47-48	8	+2	7
49-51	10	+3	8
52-53	12	+4	9
54-55	14	+5 or above	10
56-57	16		
58	18		
59	19		
60 or more	20		

School Observation (30 Points)

The final, more radical leg of my state accountability plan is the development of an annual school observation system, conducted by qualified outside observers. At the elementary school level, this introduces a critical aspect of accountability for quality in the non-tested grades. The development of a system like this is not without some precedent. Many states now use a Quality Rating and Improvement System (QRIS) to differentiate among early childcare and education providers. Some states require the completion of an on-site teacher and school observation, such as the Classroom Assessment Scoring System (CLASS), in order to be eligible for the highest QRIS rating levels. In addition, state charter boards frequently use school site observations as a means of monitoring charter schools.

My system would be functionally similar the Washington DC IMPACT teacher observation system, where roughly forty successful master teachers are employed in the District full-time as classroom observers. In my proposed state accountability system, regional master teachers and administrators would provide observations and feedback at the school level. The site visit score would be an aggregate of classroom observations, focus group conversations with administrators and teachers, and general school climate.

It is important that these observations are as transparent and collaborative as possible. The site visits will be broken down into three equal sections, adding up to 30 points: 1) quality of student-teacher interactions (measured primarily through classroom observation), 2) teacher and administrator responses to issues of school culture, discipline, and professional development, and 3) performance on a school-selected measure chosen from a list of options (such as implementation of a given curriculum, or the development of supplementary services for given subgroups). Schools will be notified in advance of the day of their observation, and should expect a written feedback from their observers as well as a rubric-based overall score.

While qualitative feedback is more likely to introduce a level of bias into school ratings, even with the use of outside observers, I believe that annual school-level observations are a critical next step for accountability. If necessary, that some bias could be eliminated using an adjustment of scores based on student demographics (Whitehurst et al 2014). Even with some remaining bias in the measurement, school site visits, similar to teacher observations, are an important, and largely unused, measure of student instruction and school climate, especially in non-tested grades and subjects.

School Assessment and Comparison

Summative Assessment (100 Point Scale)

The goal of a state accountability system is to accurately identify school quality, with a focus on rewarding strong performance and remedying poor performance. Under No Child Left Behind (NCLB), many states ended up over-identifying schools for improvement, creating confusion as how to best assist a large swath of "low performing" schools (Hyslop 2013). NCLB waivers have allowed states to move to a framework based on relative levels of school performance, where at least five percent of Title I schools would be identified as priority schools, while ten percent are identified as focus schools. I would retain this fifteen percent framework, as it appears to provide states with a manageable portfolio of schools to assist.

Aside from identification of priority and focus categories, schools would be assigned overall scores on a 0 to 100 scale. Because of the difficulty of achieving a perfect score in any

one category, I anticipate that most schools would fall into a middle tier (I estimate this tier as anywhere between 45 and 65 points), with relatively thin tails of low- and high-achieving schools. My system does not use categories to classify schools by performance tier or a given cut score. Instead, in keeping with]the goal of a simple system, schools would assess their performance relative to their previous or historical scores; an increase or positive trend in a school's overall score would indicate improvement, while a decrease would generate concern.

Absolute levels of academic achievement have the most weight in my accountability system, ensuring that schools which produce strong results are recognized, and providing a high water mark of school quality for the state. However, as demonstrated in Table 3, it is possible for a school with high achievement to be ranked on par with a low-achieving school, based on differences in student growth and school observation measures.

School	Description	Academic Achievement	Overall Median Growth Percentile	Subgroup Growth	School Observation	Overall Score
	Total Possible Points	*40*	*20*	*10*	*30*	*100*
Washington	*High-achieving school with average growth and strong school climate*	31	12	5	24	**72**
Lafayette	*Mid-achieving school, with strong growth for subgroups (especially ELL), moderate school climate*	28	7	10	20	**65**
Hamilton	*Low-achieving school, with strong growth, moderate school climate*	20	16	8	18	**62**
Adams	*High-achieving school with below-average growth (especially for subgroups), moderate school climate*	35	6	1	20	**62**
Burr	*Low-achieving school, with low growth and weak school climate*	20	5	3	14	**42**

Table 3

Conclusion

ESSA and the Department of Education have set forth a challenge to the states to design their own accountability systems. Many states have already begun to pursue this challenge under NCLB waivers. I believe that the ideal system should be easy for teachers and administrators to comprehend and manage towards, and should incorporate a qualitative component that allows for individual school feedback and goal-setting.

Citations

Betebenner, Damian W. "A Technical Overview of the Student Growth Percentile Methodology: Student Growth Percentiles and Percentile Growth Projections/Trajectories." August 12, 2011. The National Center for the Improvement of Educational Assessment. Dover, New Hampshire. Accessed at http://www.nj.gov/education/njsmart/performance/SGP_Technical_Overview.pdf.

Chiang, Hanley. 2009. "How accountability pressure on failing schools affects student achievement." *Journal of Public Economics.* 90 (9-10); 1045 -1057.

Dee, Thomas, Brian Jacob. 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management.* 30 (3): 418-446.

Hyslop, Anne. "It's All Relative: How NCLB Waivers Did — And Did Not — Transform School Accountability." December 2013. New America Education Policy Program. Accessed at https://www.newamerica.org/downloads/ItsAllRelative-12-17-2013.pdf.

Koerber, Susanne, Daniela Mayer, Christopher Osterhaus, Knut Schwippert, Beate Sodian. 2015. "The Development of Scientific Thinking in Elementary School: A Comprehensive Inventory." Child Development. 86 (1): 327–336.

Le Floch, Keristin Carlson. "Adequate Yearly Progress and Highly Qualified Teachers." Published in *No Child Left Behind and the Reduction of the Achievement Gap: Sociological Perspectives on Federal Educational Policy*. 2013. Sadovnik, Alan R. (ed). Routledge.

Reback, Randall, Jonah Rockoff, Heather Schwartz. 2014. "Under Pressure: Job Security, Resource Allocation, and Productivity in Schools Under NCLB." *American Economic Journal: Economic Policy.* 6(3): 207–241.

Whitehurst, Grover J. "Russ", Matthew Chingos, Katharine M. Lindquist. "Evaluating Teachers with Classroom Observations: Lessons Learned in Four Districts." May 2014. Brookings. Accessed at http://www.brookings.edu/research/reports/2014/05/13-teacher-evaluation-whitehurst-chingos.

Zimmerman, Corinne. 2007. "The development of scientific thinking skills in elementary and middle school." *Developmental Review.* 27 (2): 172-223.