TAC Project

These topics and questions were taken from public feedback on the A-F plan and business rules.

Major Overarching Question: Why are high FRL schools mostly located within the lower grade rankings? Why are lower FRL schools located within the higher grade rankings? Is there a bell curve, and if there is, what is causing it? (Some have pointed to the growth weights)
R. Guyer

N-Count: T. Haladyna; A. Schlessman
What we know: N-size is 20, each subgroup needs to be constant, ELL can have a different n-count number
Questions to answer through data and bring on Monday:

1. How would lowering the N-count impact outcomes, say lowering the n-count to 10 or 15?
2. If the n-count was lowered, how many NR schools would now be included in the grading system?
3. What would be the potential outcome in terms of grade label of schools if more students were captured?
4. Could there be student privacy concerns with the N-count being lowered?

Growth: A. Beardsley; C. Bochna
What we know: More weight is given to partially proficient (PP) and minimally proficient (MP) students who grow, while less weight is given to proficient (P) and high proficient (HP) students who grow
Questions to answer through data and bring on Monday:

1. Does this have a negative impact on a highly proficient school?
2. Should all growth be weighted the same? Should P and HP students' growth be weighted differently? If so, how does this impact HP and P schools?
3. What happens to a student who is HP in the first year and P in the second year? The student is still technically still proficient. Should a threshold be put in place?
4. Statistically- how would different weights in the growth measure affect outcomes?
5. How does the SGP/SGT model affect the correlation to FRL? If it does not, how do you know? In order to close the gap, how would the weights/measurement need to be altered?
6. One concern that has been raised is that the current system is too complex, would there be valid way to simplify the growth part of the letter grade system? (For example, using only SGP or SGT)
7. Because the growth system is a relative norm comparison of growth, will the growth system have stability issues over time? If so is there a suggestion to how to address this type of issue.

Acceleration (K-8) **R. Guyer, D. Jordan**

What we know: n-count is too large for some schools to receive points in this section

Questions to answer through data and bring on Monday:

1. How many schools received the total amount of points that were available to them, even if the total points that they were eligible for was under ten?
2. How many schools were eligible for full points (10) but did not receive full points because they did not meet the criteria? Is there a way to assess the validity of the criteria?
3. How many schools exceeded 10 total points but only received 10 points due to capping?
4. How would a different calculation or policy effect outcomes?

Proficiency **C. Hovanetz, D. Jordan**

What we know: Proficiency is calculated by having greater weight given to those students who are with the school for longer periods of time (FAY- Full Academic Year)

Questions to answer through data and bring on Monday:

1. How does the current weighting on proficiency impact schools, both within the stability model and the weighting at different levels of proficiency?
2. Does this weighting negatively impact the correlation to FRL? Meaning does the impact grow.

ELL: **C. Bochna**

Questions to answer through data and bring on Monday:

1. Setting aside the n-count, does this measure discriminate enough throughout the plan? Meaning, how many schools received full points?
2. Because some schools do not have access to ELL or Graduation points, what would be the best way to calculate cut scores?

Non-typical Schools **ALL MEMBERS**

The SBE has given the TAC six options: the 4 from the slides from ADE, to take the higher of the two grades, and lastly, another option that the TAC comes up with. A popular idea for these schools is to average the two grades while placing more weight on the grade levels that have more students. For example: The school is a 7-12. The 7th and 8th grades would be calculated using the K-8 model. The 9-12 grades would be calculated using the 9-12 model. At this school however, more students are located at the 9-12, therefore, the 9-12 grade would have more weight when the two scores are averaged together. What does this look like when modeled?

If you can model data, please do so and bring to the meeting to show other members of the TAC.

Acceleration Readiness

How many schools received the total amount of points that were available to them, even if the total points that they were for was under ten?

Table 1: Number of schools per FRL and Accelerate readiness points earned.

Acceleration Readiness Points	Free/Reduced Lunch Distribution Number of schools										
	0	1	2	3	4	5	6	7	8	9	blank	Total
2		3		1	3	3	1	4	10	8	18	51
4		2		1	1	2	4	7	7	19	16	59
5					1	1					2	4
6		9	6	5	17	8	9	20	22	22	18	136
7				1	1	1	1	4	2	2	1	13
8	6	9	20	15	14	16	23	28	46	45	22	244
9			1	1	2	2	2	4	5	5	13	35
10	24	67	60	58	57	83	76	96	142	127	82	872
blank	2	1	3		3	1	5	3	5	5	25	53
Total	32	91	90	82	99	117	121	166	239	233	197	1467

Table 2: Percent of schools per FRL and Acceleration Readiness Points Earned

Acceleration Readiness Points	Free/Reduced Lunch Percent of Schools										All Schools Total
	0	1	2	3	4	5	6	7	8	9	blank	
2	0%	3%	0%	1%	3%	3%	1%	2%	4%	3%	9%	3%
4	0%	2%	0%	1%	1%	2%	3%	4%	3%	8%	8%	4%
5	0%	0%	0%	0%	1%	1%	0%	0%	0%	0%	1%	0%
6	0%	10%	7%	6%	17%	7%	7%	12%	9%	9%	9%	9%
7	0%	0%	0%	1%	1%	1%	1%	2%	1%	1%	1%	1%
8	19%	10%	22%	18%	14%	14%	19%	17%	19%	19%	11%	17%
9	0%	0%	1%	1%	2%	2%	2%	2%	2%	2%	7%	2%
10	75%	74%	67%	71%	58%	71%	63%	58%	59%	55%	42%	59%
blank	6%	1%	3%	0%	3%	1%	4%	2%	2%	2%	13%	4%

How many schools were eligible for full 10 points but did not receive full points because they did not meet the criteria? Is there a way to assess the validity of the criteria?

- There are limitations to the data file that was given to the TAC.
    - The data file does not include the actual number of students who tested on state assessments in grades 3 – 8. What is not included are any indicators as whether or not a school is eligible for each of the categories for the Acceleration Readiness calculations. With 0% and 0 points, is it zero because no students were at proficiency in that category or the n-count is less than 20?

    - For 5/8 EOC and Grade 3 MP reduction, where there are blank cells, schools are not eligible for the calculation probably because the school does not have a grade 3 or have any or enough students enrolled in EOC classes at grades 5 - 8.
    - The A to F data file does not have a field showing the grade configuration for each school in the state. The TAC was given a table listing schools with non-traditional grade configurations limited to schools that are K-12, 5-12, 6-12, 7-12 and 8-12. If the grade configurations for all schools are included then the determination of other Acceleration Readiness point may be determined.
        - For example, elementary and middle schools that do not have third grade classes cannot get points for decreasing the percent of student at minimally proficient on the grade 3 ELA AzMERIT.
        - Many schools that are K-5, K-6, or K-7 may not have any students taking advanced math classes. Cannot determine which school are eligible for the 5/8-EOC points.

- Using the AzMERIT MSAA 2017 data file, eligibility are estimated. The file includes total number of students tested reporting n-count equal to or greater than 10.
        - About 400 of 1300 schools do not get the Grade 3 MP Reduction.
        - About 1122 of 1300 do not get the EOC calculation
        - About 35 of 1300 schools get both EOC and Grade 3 MP Reduction.
    - Subgroups (Ethnicity, Special Education, English Language Learners, and Economically Disadvantage)
        - There are 10 total possible subgroups. Subgroups score points for both ELA and Math proficiency on the state assessments for ELA and Math
        - Using the AZMERIT/MPSS for All Schools released by the ADE per students tested. The following table shows that 120 schools have n-counts less than 10 and have 0 subgroups and corresponding points possible. 24 schools have 9 subgroups in which to score points. Points can be earned for ELA and Math.

# of Subgroups	# of Schools
0	120
1	103
2	202
3	257
4	449
5	299
6	253
7	181
8	49
9	24

- **58 EOC points** – 1074 schools were not eligible to get the points. 12 of 393 schools received 0 points of those who were eligible. 381 schools received the 5 points.

5/8 EOC Points	# of Schools
0	12
5	381
(blank)	1074
Grand Total	**1467**

58EOC Points	Free Reduced Lunch											
	0	1	2	3	4	5	6	7	8	9	FRL NR	All Schools
0 points	6%	0%	0%	0%	1%	0%	0%	0%	1%	2%	1%	1%
5 points	41%	38%	29%	33%	25%	36%	31%	21%	18%	15%	32%	26%
Not eligible	53%	62%	71%	67%	74%	64%	69%	79%	81%	83%	67%	73%

- **Grade 3 Decrease MP on ELA.** 467 schools were not eligible most likely due to the grade configuration of the schools. 482 schools earned the 5 points and 518 did not.

GR 3 ELA MP	# of Schools
0	518
5	482
(blank)	467
Total	**1467**

	Free Reduced Lunch											
Gr 3 ELA MP	0	1	2	3	4	5	6	7	8	9	FRL NR	All Schools
0 points	28%	42%	39%	38%	34%	40%	35%	37%	35%	39%	23%	35%
5 points	44%	42%	36%	40%	30%	32%	31%	28%	37%	35%	23%	33%
Not eligible	28%	16%	26%	22%	35%	27%	35%	35%	28%	27%	54%	32%

- **Subgroup Improvement.** The table is limited to the Subgroup Improvement of total actual points earned before the 10 point cap. 670 schools earned 12 points or more. 345 schools scored 4 points or less.

Total Points earned w/o capping at 10	# of schools
0	130
2	76
4	139
6	143
8	190
10	119
12	190
14	133
16	145

18	59
20	66
22	33
24	29
26	9
28	4
30	2

	Free Reduced Lunch										
Subgroup Point Totals	0	1	2	3	4	5	6	7	8	9	FRL NR	All Students
0	6%	3%	3%	4%	8%	4%	6%	8%	7%	7%	27%	9%
2	0%	4%	1%	1%	5%	4%	8%	3%	5%	8%	8%	5%
4	3%	8%	7%	5%	14%	9%	3%	10%	9%	11%	15%	9%
6	9%	7%	6%	7%	8%	11%	7%	13%	12%	10%	10%	10%
8	13%	10%	9%	13%	8%	13%	18%	10%	14%	17%	12%	13%
10	6%	8%	10%	10%	14%	11%	7%	6%	6%	8%	7%	8%
12	9%	23%	16%	12%	13%	15%	12%	14%	12%	12%	9%	13%
14	16%	12%	16%	16%	9%	12%	12%	7%	8%	6%	4%	9%
16	16%	10%	17%	15%	7%	11%	12%	10%	10%	10%	3%	10%
18	13%	5%	0%	6%	5%	3%	4%	6%	5%	3%	1%	4%
20	6%	3%	13%	7%	4%	4%	3%	4%	6%	3%	1%	4%
22	3%	4%	0%	2%	2%	2%	2%	3%	3%	3%	1%	2%
24	0%	2%	1%	1%	2%	0%	2%	4%	2%	2%	3%	2%
26	0%	0%	2%	0%	0%	2%	2%	0%	0%	0%	1%	1%
28	0%	0%	0%	0%	0%	0%	1%	1%	0%	0%	1%	0%
30	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%

Proficiency K – 8 (See PDF)

Concerns and Question About the Arizona A-F Accountability System

(Working Draft)

Dr. Thomas M. Haladyna
Professor Emeritus
Arizona State University

As a member of the Technical Advisory Committee for the Arizona Board of Education, this document is submitted as a partial fulfillment of my duties on this committee. My experience in test development and validation is extensive. Among the many clients I have served includes the Arizona Department of Education, the Arizona Attorney General, Equal Employment Opportunity Commission, and many high-stakes testing programs throughout the U.S.

My concern is that the Arizona A-F Accountability System is subject to criticism and even legal challenges unless and concerns are considered and questions are answered. Recently, the Office of Civil Rights challenged the Arizona Department of Education over the validity of how the AZELLA test classified English language learner students. I was involved in that conflict representing Arizona. The accountability system needs critical analysis and validation. The balance of this document asks questions and explains why good answers are important.

The document is organized by major topics: (1) causality, (2) curriculum, (3) instruction, (4) testing, and (5) validation. This document is a first draft. As I work with our committee, I plan to expand the scope and refine some of these questions.

Causality

1. Is the causal model plausible? The accountability system holds schools responsible for student learning. Does the causal model consider a student's readiness and preparedness to learn? Nearly half Arizona's student population includes students at risk. Many students come to school not ready to learn. Does the accountability system recognize this fact and consider it when determining how well the school is performing? Does the accountability system consider other factors influencing student achievement? For example, we have known that parents, home, family, and other social factors have a large, primary influence over student learning.

2. Do A schools get evaluated to learn exactly what they did to earn an A? In the past, one project in which I participated was funded by the legislature to evaluate innovative schools. It seems important for Arizona to study A schools to learn what curriculum and instructional strategies were used to achieve good results? This validates the wisdom of identifying a school as high-performing. Such information would also be crucial for improving education in Arizona.

Curriculum

1. Is the curriculum fully represented in the A-F Accountability System? Are all subject-matters covered in day-to-day instruction? For the pre high school grades, reading, writing, and mathematics is only a part of the school curriculum. In validity language, this is an issue of representation. Do the accountability results represent the entire curriculum or just part of the curriculum?

2. What curriculum do schools adopt and follow in guiding instruction? How is it used to drive instruction?

Instruction

1. Are teachers adequately prepared to teach? Many factors affect the providing of instruction that relate to the teacher. In the accountability system, are teacher absenteeism, mobility, qualifications and certification, and considered in the accountability system?

2. Do students have sufficient opportunity to learn and relearn? From learning theory and research, we know that if students are given adequate time to learn, they learn more. Many students, especially those considered at-risk, need extra time and remedial instruction that may include after school programs, summer school, or tutorial assistance. If test scores are truly a measure of student learning, the opportunity to learn and relearn should be provided and included in the accountability model.

3. Do teachers have sufficient resources to teach? As a former teacher, I remember how important it is to have resources to provide excellent instruction. Are Arizona teachers provided with these resources?

4. How much time is spent at each school on reading, writing, and mathematics? One elementary level teacher told me that if reading, writing, and mathematics is all that counts in accountability, that is all I will teach. A testing expert colleague with extensive experience in other states tells me that overloading instruction with what is tested is done and it works. Of course, this is blatantly dishonest. Does this happen in Arizona schools?

Testing

The heart of the accountability system is the state achievement testing program, AzMERIT. As a longtime advisor and consultant to the Arizona Department of Education, I can attest that this testing program is of very high quality and can be used dependably as a single indicator of student achievement. However, national standards for testing and evaluating student learning strongly recommend using two or more indicators for each subject matter. If validity is valued, other indicators should be included. For instance, student grades, school district tests, prior achievement, teacher assessment, parent assessment are all possible sources of student achievement. Also, there is a trajectory of prior achievement that is a useful predictor of achievement. It is a simple task to validate these sources.

Here are some specific questions of a technical nature:

Growth scores. What is the unit of analysis? Students, classrooms, schools? How reliable are growth scores? Theoretically, individual student growth scores are very unreliable. So knowing how reliable growth scores are is very important. What is the effect size of growth scores? Descriptive statistics of growth scores would also be very informative. Is there systematic error in growth scores? I have personally seen negative growth scores in AIMS data. Negative growth is hard to explain. Also, there is a long-term trajectory on student growth that is well predicted from non-school variables, known as risk. How much variability is there in these growth measures? How is individual growth analyzed with class growth and school growth? Is this method of analysis clearly documented and explained?

Proficiency scores. We know from the annual technical report of the AzMERIT that test scores are sufficiently reliable. How reliable are proficiency scores at the school level? What is the conditional standard error of measurement for the indexes used to decide a school grade? What is the distribution of scores leading to a school grade? The same questions asked about growth scores apply mostly to proficiency scores.

Adjustment. One goal is to eliminate the correlation between poverty and achievement so that schools can be fairly graded. Adjusting on poverty is insufficient. There are many risk factors that make up risk including poverty, learning English, disability, cultural isolation, homelessness, and behavioral problems to name a few. If you look at individual student profiles, you will see that some students have multiple risk factors. There is a strong correlation between degree of risk and performance on these achievement tests. If the accountability system only includes a poverty variable, much is missing that may bias results. Also, there are many adjustment methods. Which one is used? Has it had peer review? Different models for adjustment give different results based on different assumptions? Does this variability influence which schools get what grades?

Cut scores and random error. Around every cut score, there is a band of uncertainty that exposes the degree to which a score might be in random error. Random error can be positive or negative, high or low. The conditional standard error of measurement informs us as to the risk of misclassification. This is very important to schools, as random error might determine whether a school gets a higher or lower grade. If you add in the chance of systematic error corrupting the result, all schools are at risk of being misclassified. Has this risk been considered and properly explained?

Validity

Validity refers to the accuracy of test scores for any interpretation or use. Validation is a process for establishing evidence supporting validity. Every test use needs to be validated. Accountability is so important that validation should be taken very seriously. A third-party evaluation of an accountability system is desirable. More important, an annual technical report provides necessary explanation and evidence to support the valid use of the accountability system. Is there a technical report published annually? Along side a technical report is full documentation. Every document that provides validity evidence should be collected and cited. Validity studies are a special breed of validity evidence that answers an important question, solves a problem, or provides important validity evidence.

Unique School Configurations and A-F Letter Grades

Definitions

- Unique School Configurations are those that serve grades that span across the K-8 and 9-12 models
  - Enrollment data is pulled to determine the grades that a school serves
- SBE voted in June was to give these uniquely configured schools 2 letter grades (e.g., if it’s a 6-12 school grade 6-8 students were evaluated on the K-8 model and the 9-12 students on the 9-12 model)

Impact

- 108 traditional schools received 2 letter grades
- 71% were charter (*n* = 77)
- Grade configurations consisted of:
    - K-10 (*n* = 1)
    - K-12 (*n* = 40)
    - 1-12 (*n* = 1)
    - 2-12 (*n* = 1)
    - 3-12 (*n* = 2)
    - 4-11 (*n* = 1)
    - 4-12 (*n* = 2)
    - 5-12 (*n* = 7)
    - 6-10 (*n* = 1)
    - 6-11 (*n* = 2)
    - 6-12 (*n* = 20)
    - 7-11 (*n* = 2)
    - 7-12 (*n* = 28)

Impact

	A	AB/	AC/	A NR/	B	BC/	BD/	BF/	B NR/	C	CD /	CF/	C NR/	D	DF/	D NR/	F	F	NR
	A	BA	CA	NR A	B	CB	DB	FB	NR B	C	DC	FC	NR C	D	FD	NR D	F	NR	NR
K-10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
K-12	2	2	2	1	4	2	1	-	3	2	2	1	4	2	1	1	1	-	9
1 to 12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
2 to 12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
3 to 12	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4 to 11	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-
4 to12	1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5 to 12	3	3	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6 to 10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-
6 to 11	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-
6 to 12	1	4	2	-	4	3	-	1	-	-	-	1	1	-	-	-	1	-	3
7 to 11	-	-	-	-	-	2	-	-	-	-	5	-	-	-	-	-	-	-	-
7 to 12	2	-	-	-	2	7	-	-	2	2	-	-	2	2	2	-	-	-	1

Options to create one letter grade for these schools

1. Unique models for each school configuration
2. Merge the “outlier” grades into one model
3. Average the two letter grades
4. Prorate the two letter grades

5. A combination of the above
6. Additional alternatives

Option 1: Unique models for each school configuration

Definition: every school configuration has it’s own “model”

- Create a 6-12 model, 7-12 model, 1-10 model, etc.

Pros:

- Respects and values the different school configurations
- Is a fair accountability system for the grades they serve

Cons:

- Hard to sustain annually with new/different configurations
- Very challenging and time consuming to code and create a new model for every school configuration
- Would require a new ADEConnect platform (this is where schools are shown their letter grade and subsequent data) be built to accommodate these schools
- Would require a cut score set for every new model

Timeline:

- If TAC completes modeling by February
- ADE to release letter grades May / June to these schools assuming the modeling is approved at the March board meeting

Option 2: Merge the “outlier” grades into one model

Definition: use the existing models and place the “outlier” grades into one of the two models

- K-10, 6-10, 6-11, 7-11, and 4-11 schools could use the K-8 model because they don’t have CCRI or graduation rate data
- 4-12, 5-12, 6-12, and 7-12 could use the 9-12 model
- K-12, 1-12, 2-12, and 3-12 could use the 9-12 or the K-8?

Pros:

- Benefits certain configurations, for example 6-10, 6-12, 7-11, 7-12, who don’t have access to most of the acceleration readiness points due to minimum n size
- Easier to calculate and release in ADEConnect
- Current cut scores can be applied

Cons:

- For some of the configurations it forces the schools into one model type neglecting either acceleration/readiness or graduation rate/CCRI points
- Could be hard to sustain annually with new/different configurations

Timeline:

- If TAC completes modeling by December
- ADE to release letter grades February to these schools assuming the modeling is approved at the January board meeting

Option 3: Average the two letter grades

Definition: use the existing data as is and take an average

- Could look at high level letter grades – A and C = B
- Could calculate based on points – add total points earned for both models and divide by total points eligible for both models

Pros:

- Easy to calculate
- Sustainable with new configurations in future years

Cons:

- Less fair, doesn't take into account student enrollment numbers (e.g., what if the 9-12 grades serves 80% of the students compared to the 6-8 grades)
- How do you average an NR?
- Use of points to calculate the average could require a new cut score
- Would ideally want to build additional info into the ADEConnect platform

Timeline:

- If TAC completes modeling by November
- ADE to release letter grades January to these schools assuming the modeling is approved at the December board meeting

Option 4: Prorate the two letter grades

Definition: use the existing data as is and prorate the letter grades based on FAY enrollment numbers in each model

- The 6-12 school has a K-8 letter grade and a 9-12 letter grade. Determine how many FAY students were enrolled in grades 6-8 and how many FAY students were enrolled in grades 9-12. If 20% of the school's population is in grades 6-8 then the K-8 grade is only worth 20% while the 9-12 grade would be worth 80%. If the K-8 earned a percentage of 40% and the 9-12 earned a 90% the prorated grade would be: 80% (40% * 20% + 90% * 80%)

Pros:

- Relatively easy to calculate
- Sustainable with new configurations in future years

Cons:

- Schools without access to particular points (i.e., acceleration readiness, grad rate, CCRI points) on the current models still suffer
- How do you prorate an NR?
- Use of points to calculate the average could require a new cut score – what does the prorated percentage mean?
- Would ideally want to build additional info into the ADEConnect platform

Timeline:

- If TAC completes modeling by November
- ADE to release letter grades January to these schools assuming the modeling is approved at the December board meeting

The **95 percent tested requirement** is not applied to the 2017 school grades. What is the number and percent of schools that did not meet the 95% testing requirement. Understanding the impact of the requirement is important to gauge the validity of results from 2016-17 and know the potential impact for when it is applied in 2017-18.

Calculating academic achievement for elementary and middle schools based on one, two, and three years of full academic year data to determine proficiency for schools. In K-5 schools is three years determined based starting in K, 1, or 2, so by grade 3 the student was enrolled for three consecutive years? In grades 6-8 schools is three years starting with grade 6 as year 1?

- If the three-year requirement means grade 5 and 8 are more weighted, then scores are depressed as these are generally lower performing based on assessment results and would include fewer students.
- How many schools use a three-year FAY calculation compared to the two- and one-year calculations?
- How many students are included in each of the three-, two- and one-year calculations?

ELA 2017 (increase over 2016)

- Grade 3: 43 percent (+3 percentage-point gain since 2015)
- Grade 4: 48 percent (+6 percent)
- Grade 5: 44 percent (+12 percent)
- Grade 6: 41 percent (+5 percent)
- Grade 7: 44 percent (+11 percent)
- Grade 8: 34 percent (-1 percent)

Math 2017 (increase over 2016)

- Grade 3: 47 percent (+5 percent)
- Grade 4: 47 percent (+5 percent)
- Grade 5: 47 percent (+7 percent)
- Grade 6: 41 percent (+8 percent)
- Grade 7: 34 percent (+3 percent)
- Grade 8: 28 percent (-6 percent)

High school proficiency calculation does not require all student take a high school assessment in ELA and math because participation is based on the grade 12 FAY students. Schools could not test students in and not ever be accountable for the student performance if the student moves during the senior year or exits high school early. How many students from the grade 9 cohort do not have a test score four years later – this may be okay now, but is setting up a perverse incentive for schools to not test lower performing kids they think may drop out or move before their senior year. Delaying the testing could also delay graduation.

High school participation calculation using only full academic year 12th graders challenges the validity of the indicator because students may not be full academic year the year they tested (entered grade 10 in

January and took EOC in the spring so they are not accountable in proficiency but are accountable in participation two years later), the denominator includes only students who made it to grade 12 and were full academic year students while in grade 12 (excludes dropout in previous grades and early exiters), students may have tested in a different school than the school where they are enrolled in grade 12 (wrong school accountable for the student). How many students are excluded from the calculation compared to the grade 9 cohort?

Provide the **SGP/SGT scale impact** of all achievement level index points were the same using the Highly Proficient achievement level points of 0 for Below Target, 0.5 for At or Near Target and 1.0 for Exceeds Target.

Awarding different points based on achievement level is not statistically sound because these students are compared to their peers the 'difficulty' of achieving Exceeds Expectations is the same regardless of achievement level because it compares only to peers and guarantees 33 percent will Exceed. Awarding 2.0 points for Minimally Proficient for Exceeds Expectations and 1.0 points for Highly Proficient for Exceeds Expectations provides a positive systemic bias to schools with larger populations of Minimally Proficient students and a negative systemic bias to schools with Highly Proficient students when the accomplishment is equal.

On the SGT, how many kids who scores the -10 percentage points for At or Near Target get to proficiency in 3 years or by graduation?

How is the secondary target or Highly Proficient used for the current Highly Proficient and Proficient students in the calculation? Highly Proficient students are awarded 1.0 points for three consecutive years of decline, to maintain Proficient, this is not a valid growth indicator.

How many schools are eligible for **SPED bonus points?** How many schools earn SPED bonus points? How many schools changed a letter grade because of the SPED bonus points? Must determine if this is categorially fair to ensure letter grades are valid.

How many schools, analyzed by the number of subgroups, **earned bonus points for subgroup improvement**. More subgroups will set forth more opportunities to earn bonus points. Impact data on the number of points earned by school based on the number of subgroup at the school will help determine if this is categorially fair to ensure letter grades are valid.

Validity of the **CCR indicator** cannot be determined without the numerator and denominator.

Face validity checks on school grades. Provide the most disparate schools (three highest and three lowest ranked schools) on the Achievement Index, Percent Proficient, SGP, and SGT earning an A, B, C, D, and F. For example, provide the report for six A schools, three with the highest SGP and the three with the lowest. Compare to the six A schools, three with the highest SGT and the three with the lowest to determine if schools are being reasonably graded. These analysis can be used for further questions and suggested revisions to calculations and weighting to support a more valid and reliable system.

Correlation between Free/Reduced Lunches and A-F Accountability Scores:

R. Guyer

Data and Definitions

The data for free and reduced (FRL) lunches and the A-F accountability scores were provided by the Arizona State Board of Education (AZSBE).

The free and reduced lunch variable ranges from 0 to 9 with the following interpretations:

0 = 0 to 9.9% of students receive FRL

…

9 = 90 to 100% of students receive FRL

The accountability scores are expressed as the ratio of “Total Points Earned” out of “Total Points Eligible.” For the correlational analyses we will express these scores as proportions.

Correlations

The correlation between FRL and K-8 Total Accountability equals **-.560**

The correlation between FRL and 9-12 Total Accountability equals **-.503**

Free and Reduced Lunch Sampling

118 Grade K-8 schools did not provide FRL data. Of these 106 or 89.8% are charter schools. (246 of 1,262 schools in total were charter)

15 Grade 9-12 schools did not provide FRL data. Most (11 out of 15) are charter schools. (30 of 201 schools in total were charter)

Most schools that dropped out due to missing FRL data were charter. A cause of this is that some charter schools do not provide meals during school. A different economic indicator that does not require the school to serve lunch on campus would benefit the study.

Figure 1. Scatterplot of FRL by Accountability Score for K-8 Schools with best fit line

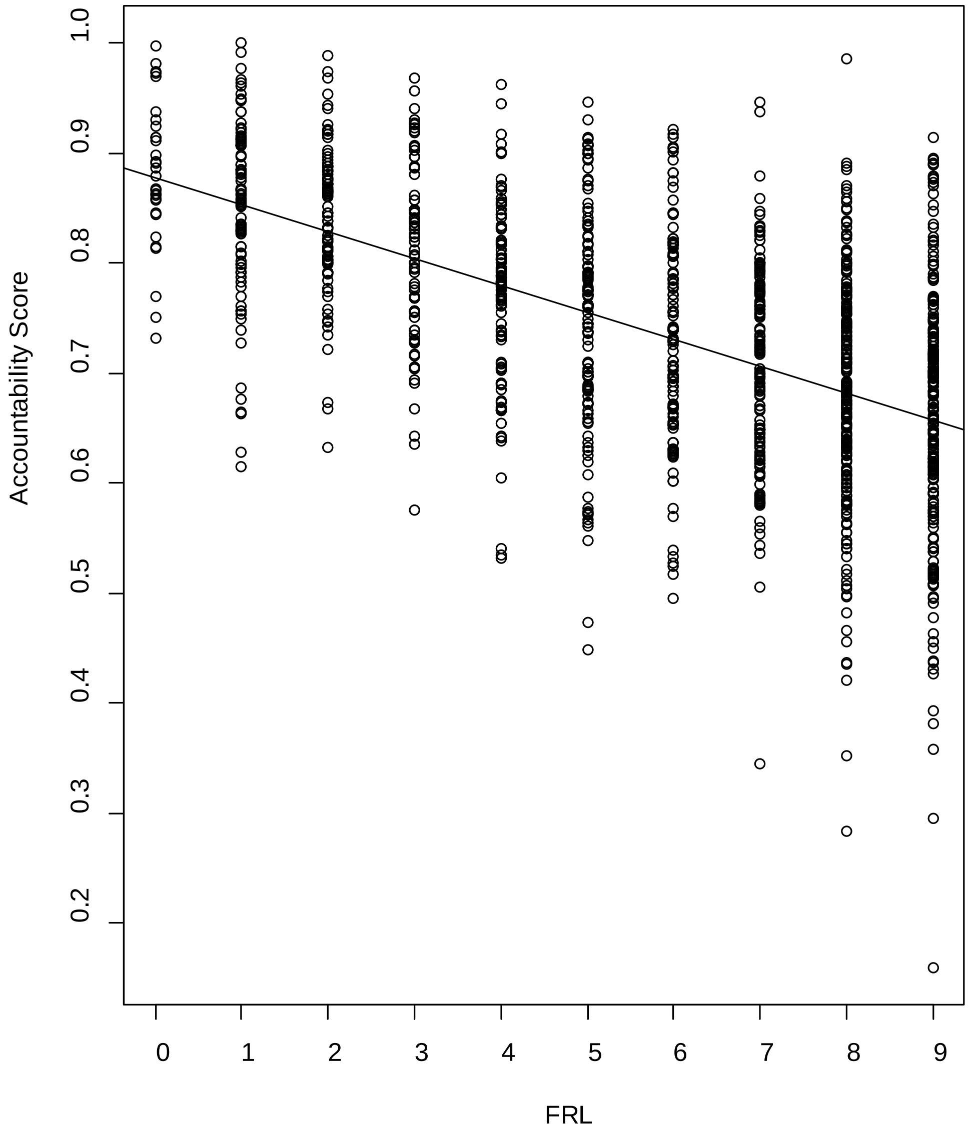
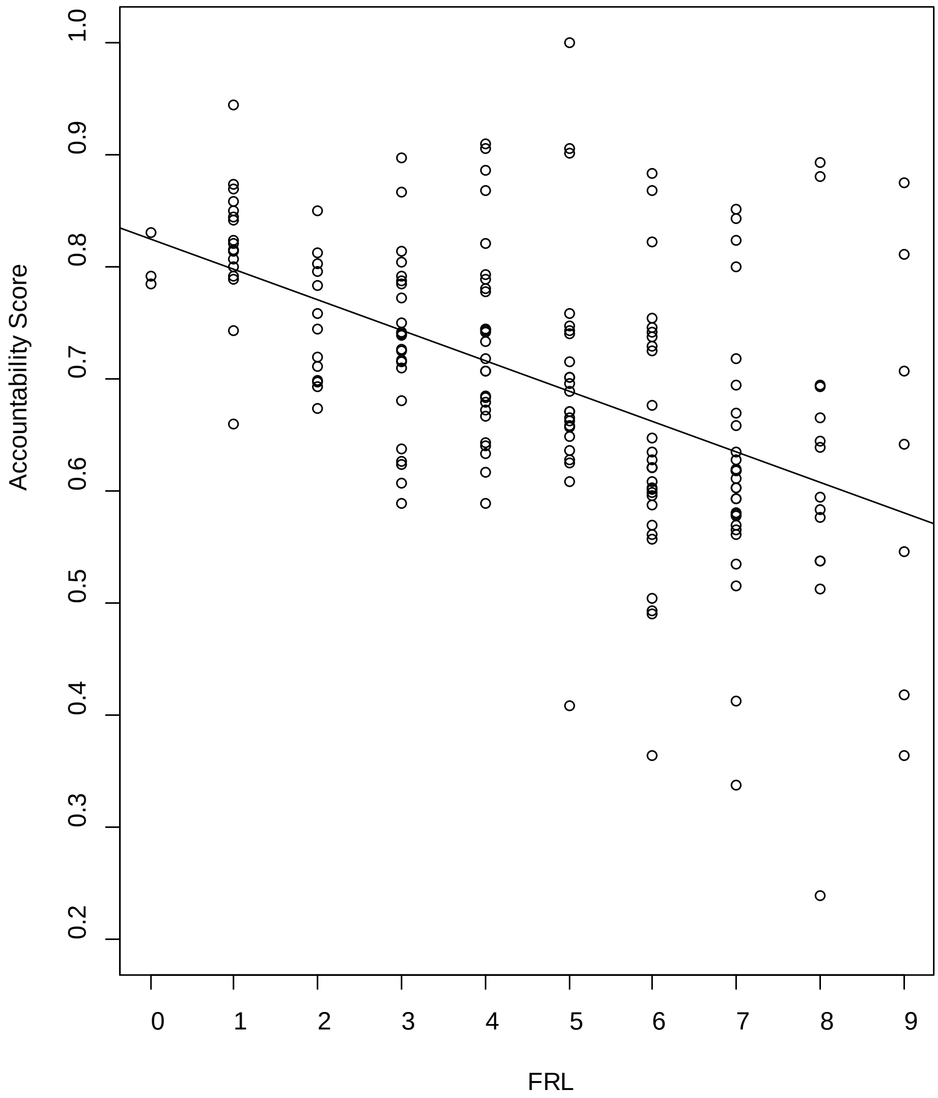


Figure 2. Scatterplot of FRL by Accountability Score for 9-12 Schools with best fit line



Growth: A. Beardsley; C. Bochna

What we know: More weight is given to partially proficient (PP) and minimally proficient (MP) students who grow, while less weight is given to proficient (P) and high proficient (HP) students who grow

1. Does this have a negative impact on a highly proficient school?
   What we know from the research, in general but not per our state's use of the SGP/SGT (yet) is that regression to the mean typically impacts all such models, whereas students like those above (i.e., the PPs and MPs AND students who are the highest achieving regress to the mean). So, the answer to this might be that students in highly proficient schools are being penalized already because of the models' ceiling effects. The only way to investigate this and also to make a more informed decision, however, would be to examine the extent to which schools with populations of students are able to demonstrate growth over time that is comparable to other schools, AND THEN discuss possible weights to help offset model weaknesses.

   In looking at the data, schools that exhibited both high growth and high proficiency received As. Schools that demonstrated high proficiency but not as much growth received Bs. It was possible for schools achieving the highest proficiency levels to receive an A, however, proficiency alone would not guarantee an A. High proficiency schools need to also earn points on growth and acceleration/CCRI.

2. Should all growth be weighted the same?
   This, of course, assumes that all schools have equal opportunities to demonstrate growth, and that their growth is not distorted by model limitations likely to cause bias. For example, if schools' growth performance is in fact correlated with the demographics of the students in the schools, which I believe most of us on the committee suspect (and hopefully we will have some answers re: this tomorrow), this would suggest that differential weights would likely be necessary to offset the statistically uncontrollable bias present, likely, in both models. If bias is present and uncontrollable in our most sophisticated value-added models, given the SGP and likely the SGT models are less sophisticated, this is something with which we must contend even more. Should P and HP students' growth be weighted differently? If so, how does this impact HP and P schools? See my responses above, also as per Q1 above.

   The AAG conducted models varying the growth weights throughout this year. They found that if growth is weighted the same for all performance levels the correlation to student poverty increased and higher growth weights for proficient students decreased the potential for every school in the state to earn an A. Perhaps the conversation shouldn't be about growth weights but instead around allowing the 50/30 percentage to use the higher of the two for a school (growth or proficiency). Or taking the cap off proficiency points.

3. What happens to a student who is HP in the first year and P in the second year? The student is still technically still proficient. Should a threshold be put in place?

   Rule of thumb on this is you MUST have three years of (more or less) RELIABLE or consistent data before anyone makes any decisions about labels, and especially any consequences attached to such labels. So, if you have a student who is HP, then P, then HP you could likely go with the mode, BUT given the issues with reliability or lack thereof across such models (including the SGP), things will likely not be so easy. Hence, and in order to make an informed decision, we would need research on the three-year consistency rates at the student proficiency levels, in that we would likely have many students presenting as "all over the place" which is also quite common. This is critically important, as well, in that without adequate reliability, which is statistically calculable, we can never get to validity, or rather the valid inferences "we" would like to draw from these data. Do also note that these issues occur at the teacher and school levels as well, so we'd want to be sure to define our unit of analysis when analyzing our own data to make such informed decisions.

The weights of SGP/SGT need to be reviewed going forward when more data is available. What define's "one year's growth?" Where is the greatest need to differentiate growth? Currently **all** SGTs are to Proficient. Should they be? The business rule for SGT needs to be evaluated for students at the highest end of SGT. Currently students who earn an SGT of 89 or higher only earn an "at/near" score point because they can't be considered for "above" score points.

4. Statistically- how would different weights in the growth measure affect outcomes?
This, too, is something we would need a consultant, should we be able to successfully hire/retain one, to analyze as per different simulations. The key here, I think, for all of us is to make research-informed decisions, and without somebody helping us with answering such questions, we cannot make solid decisions with the greatest intended and fewest unintended consequences.
I concur with Audrey's point and add that the modeling work done by the AAG and ADE already around this issue should stand until a data analyst is hired to assist the TAC.

5. How does the SGP/SGT model affect the correlation to FRL? If it does not, how do you know? In order to close the gap, how would the weights/measurement need to be altered?
See response to Q2 above, also with hopes that somebody from ADE is coming with at least some preliminary answers on the first question tomorrow. Note, also, Dr. Haladyna's note last week re: "risk factors." We might expand our correlational analyses to capture "risk factors," v. just equating this with one variable (i.e., FRL).
The K-8 model 30/50 proficiency/growth correlation to FRL is .42 as presented in April. This was the lowest correlation to poverty of those modeled. There will never be a weighting/measurement that will reduce this correlation to zero but it could be improved through the analysis of additional risk factors as proposed by Dr. Haladyna.

6. One concern that has been raised is that the current system is too complex, would there be valid way to simplify the growth part of the letter grade system? (For example, using only SGP or SGT)
Yes. Get rid of the letter grade system entirely. Keep the statistics in their simplest of forms, compare school-level statistics to state averages, note limitations as possible/needed, and let others (including consumers of public websites) be the judge. Take, for example, the National Assessment of Educational Progress (NAEP) – the Nation's Report Card and what is widely considered the nation's best test. This is what they do, to which they add other important indicators (left in descriptive form) so that others can be the judges of what might be good or bad. Here, though, what we would need to define is what counts as a "good" or "effective" school, and simply put the comparative statistics up for public consumption. It's when we get ourselves into oft-arbitrary discussions about weights, cut scores, and the like, just so that we can make summative/summary judgments about (As-Fs) that get us into trouble. This is also when, as the arbitrary decision increase, that validity of inference decreases.
The model has become complex to mitigate the impact of a single approximately 40 item test score on a school's letter rating. If one reduces the number of ways that schools can earn points, that test takes on even more meaning. As long as letter grades remain a part of legislative policy the model should remain complex with even more opportunities for schools to earn acceleration/CCRI points and more weight should be placed on those factors.

7. Because the growth system is a relative norm comparison of growth, will the growth system have stability issues over time? If so is there a suggestion to how to address this type of issue.
Nope. I know of nobody who has satisfactorily addressed any value-added or growth models' "stability issues" over time. This is important to note, also, in that all models suffer from what I would term are pretty extreme inconsistencies over time. However, we still do not know what our state's reliability (i.e., stability) coefficients are, at the state-level in particular as this is with what I believe we are most concerned. See also my response to Q3 above.
My answer on this question was, "Not sure" so Audrey's is far more eloquent.

ELL: C. Bochna Questions to answer through data and bring on Monday:

1. Setting aside the n-count, does this measure discriminate enough throughout the plan? Meaning, how many schools received full points?
   246 K-8 schools received the full points for this measure.

2. Because some schools do not have access to ELL or Graduation points, what would be the best way to calculate cut scores?

To address this issue, the number of points total/possible was decreased for those schools to compensate for the lack of data. A lack of CCRI and/or Acceleration points puts more emphasis on AzMERIT.

A. Schlessman TAC meeting, 11/20/2017

Accountability Information Regarding Student Number Count (n-count) in State Plans

California: https://www.cde.ca.gov/re/es/documents/caessastateplansubmitted.doc.

i. Minimum N-Size *(ESEA section 1111(c)(3)(A))*:

a. Provide the minimum number of students that the State determines are necessary to be included to carry out the requirements of any provisions under Title I, Part A of the ESEA that require disaggregation of information by each subgroup of students for accountability purposes.

California's accountability system will be applied to all schools, including charter schools, and all student groups with 30 or more students. The same minimum n-size of 30 will be applied to alternative schools when the alternative indicators are produced for the fall 2018 California School Dashboard release.

a. Describe how the minimum number of students is statistically sound.

Given the confidence level and margin of error, a sample size of 30 is needed to appropriately estimate the population. A sample size of 30 produces a standardized normal distribution, where the distance between the variance is normal/standard, resulting in statistically significant results (based on the central limit theorem), which is well documented in many statistics textbooks (Cohen, 2001; Cohen and Lea, 2004; Mendenhall and Ott, 1980; Urdan, 2001; Vogt, 2005).

a. Describe how the minimum number of students was determined by the State, including how the State collaborated with teachers, principals, other school leaders, parents, and other stakeholders when determining such minimum number.

Statistical research overwhelmingly supports a minimum n-size of 30 to produce a mean, range, standard deviation, and even distribution (Mendenhall and Ott, 1980; confirmed in later years by Cohen, 2001; Cohen and Lea, 2004; Urdan, 2001; Vogt, 2005). Based on this research, the California Legislature established the n-size for accountability purposes in California *Education Code* (*EC*) Section 52052. There was support from educational stakeholders and a general consensus regarding the established n-size of 30 when the legislation was introduced. In preparation for submission of the State Plan, over 400 comments were received on the accountability section through the 30-day public comment period through 13 stakeholder meetings, a public survey, and submitted written comments via letters and e-mails. These comments represent feedback from education administrators, teachers, parents, advocacy groups, and members of the public. The CDE's Technical Design Group also concurred that the n-size required under *EC* Section 52052 was statistically valid and reliable.

a. Describe how the State ensures that the minimum number is sufficient to not reveal any personally identifiable information.

To preserve student anonymity, the CDE has a long-established practice to not report data if a student group has less than 11 students. For reporting purposes only, California provides Status/Change data for student groups with 11 to 29 students in the group.

a. If the State's minimum number of students for purposes of reporting is lower than the minimum number of students for accountability purposes, provide the State's minimum number of students for purposes of reporting.

The minimum size for reporting is 11.

Colorado:

We use an n-count of 16 for achievement and a n-size of 20 for growth.

The currently adopted n-size for achievement is related to concerns regarding Personally Identifiable Information. For growth, an n-size of twenty addresses the technical reliability of the derived median growth percentiles. For n-sizes below this threshold the estimated values are less stable. Also, an n-size of twenty serves to adequately address any issues surrounding PII.

Jessica Watson
Accountability and Policy Principal Consultant

Dan Jorgensen, Ph.D.
Accountability Support Manager

Accountability and Data Analysis

COLORADO
Department of Education