Arizona State Board of Education
Technical Advisory Committee

NOTICE OF PUBLIC MEETING

Pursuant to Arizona Revised Statutes (A.R.S.) § 38-431.02, notice is hereby given to members of the State Board of Education Technical Advisory Committee (the “Committee”), and to the general public, that the Committee will hold a meeting open to the public on **Tuesday, November 28, 2017, at 4:00 PM, at the Arizona Department of Education, Room 122, 1535 W. Jefferson, Phoenix, Arizona 85007.** A copy of the agenda is attached. The Committee reserves the right to change the order of items on the agenda, with the exception of public hearings. One or more Committee members may participate telephonically.

Pursuant to A.R.S. § 38-431.02(H), the Committee may discuss and take action concerning any matter listed on the agenda.

Pursuant to A.R.S. § 38-431.03(A)(3), the Committee may vote to convene in executive session, which will not be open to the public, for legal advice concerning any item on the agenda.

Persons with a disability may request a reasonable accommodation such as a sign language interpreter, by contacting the State Board Office at (602) 542-5057. Requests should be made as early as possible to allow time to arrange the accommodation.

DATED AND POSTED this 27th day of November, 2017.

By: ______________________

Alicia Williams
Executive Director
(602) 542-5057

AGENDA

TECHNICAL ADVISORY COMMITTEE
Tuesday, November 28, 2017
4:00 PM
Arizona Department of Education, Room 122
1535 W. Jefferson
Phoenix, AZ 85007

4:00 PM CALL TO ORDER AND ROLL CALL

1. Presentation, discussion, and possible recommendation to the State Board of Education on the methodology of a single letter grade for schools with non-typical school configurations

2. Presentation and discussion regarding student number count (n-count) issues within the A-F accountability plan and business rules

3. Presentation, discussion and possible action on the final report to go to the State Board of Education on the issues within the A-F accountability plan and business rules including:

   a. N-Count Issues
   b. Growth Issues
   c. Proficiency Issues
   d. English Language Learners (ELL) Issues
   e. Acceleration Issues within the K-8 model

4. CALL TO THE PUBLIC: This is the time for the public to comment. Members of the Committee may not discuss items that are not specifically identified on the agenda. Therefore, pursuant to A.R.S. 38-431.01(H), action taken as a result of public comment will be limited to directing staff to study the matter, responding to any criticism or scheduling the matter for further consideration and decision at a later date.

5. FUTURE MEETING DATES AND ITEMS FOR FUTURE AGENDAS. The Executive Director or a member of the Committee may discuss future meeting dates and direct staff to place matters on a future agenda.

Adjourn

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

For the TAC meeting, 11/28/2027

Student Number Count (N-count) Issues
within the A-F Accountability Plan and Business Rules

Methodological, Statistical, and Technical Concerns/Questions:

Several TAC members have expressed concern that:

1. There are methodological and statistical concerns about the stability of using N-counts less than 20.
2. That focusing on N-count detracts from bigger methodological concerns regarding calculation of Growth, conditional standard error of measurement, cut scores, and random error.

A lower N-count results in a wider margin of error. Even descriptive statistics such as mean and median become more uncertain.

Questions to answer through data:

1. How would lowering the N-count impact outcomes, say lowering the n-count to 10 or 15?

By lowering the N-count to 15, these were the results for schools that had student eligibility in various categories that were reported in the file with student counts.

This researcher could not in good conscience run data for an N-count lower than 15.

K-8 Total NR schools = 78

ProficiencyFAYsum	7
TotalNumberELFayStudents	0
FAY_SGP	10
FAY_SGT	10

9-12 Total NR schools = 69

ProficiencyFAYsum	10
TotalNumberELFayStudents	0
FAY_SGP	8
FAY_SGT	8
numcohort4	2

It was not possible to calculate if lowering the N-count to 15 would capture more students in the College and Career Readiness Category. ADE was advised that College and Career Readiness is all self-report and therefore did no validation of submissions. If the school met the N-count of 20, they should have submitted. If the school did not, then they should have selected NA.

2. If the n-count was lowered, how many NR schools would now be included in the grading system?

K-8

With a threshold of 80 points to earn a letter grade for 2016-17 (p. 25 of 2017 *A-F Letter Grade Accountability System: Traditional Schools Business Rule*s), it is most likely that seven (7) K-8 schools currently receiving a NR rating would receive a letter grade.

9-12

The threshold for 9-12 schools is 50 points (p. 32 of 2017 *A-F Letter Grade Accountability System: Traditional Schools Business Rule*s). It is possible that the ten (10) schools with students meeting the adjusted N-count of 15 for Proficiency would receive the additional 20% required to receive a letter grade. If so, 14% (10/69) would receive a letter grade. The State Board of Education would probably want to consider the policy question of whether they would be comfortable assigning a letter grade to a traditional school that did not have enough students with Proficiency scores to meet an N-count of 15.

3. What would be the potential outcome in terms of grade label of schools if more students were captured?

It might seem desirable to lower the student count threshold and capture more students. The seven (7) K-8 schools and possibly ten (10), 9-12 schools that would most likely receive a letter grade, rather than an NR, for 2016-17 might be pleased this year.

According to these simple calculations, seven (7) K-8 schools is only 9% (7/78). As mentioned, 9-12 predictions are not precise. We do not know how many more schools would report College and Career Readiness if N-count were lowered to 15 graduates. Using current data - If the ten (10) schools with enough FAY Proficiency students, also earn another 20% via SGP, SGT, or Graduation, 14% (10/69) might receive letter grades.

TAC members have expressed that lowering the N-count results in achievement profile framework calculations that are subject to instability of the system. It will be uncertain if year-to-year fluctuations in a school's letter grade are due to the work of the school or instability of the framework calculation(s).

Lowering the student number count might be a Band-Aid fix that increases the number of schools with letter grade labels for FY 2017 yet becomes a tourniquet in subsequent years.

4. Could there be student privacy concerns with the N-count being lowered?

Absolutely. If someone knows how to mine information from public records, privacy could be compromised.

Privacy might be more protected if the N-count were lowered to 15. Anything less than 15 exacerbates the issue.

AZSDE Questions and Guyer Responses

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?

Based on my review all schools were eligible to receive maximum points due to the Subgroup Improvement variable which accounted for 6 potential points. Of the 1374 schools that had valid grades and acceleration calculated, 872 (63.4%) received the maximum of 10 points. More information is required to determine whether a school qualified for the maximum points.

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?

502 based on the data information

3. How many schools exceeded 10 total points but only received 10 points due to capping?

736 of the 872 schools (84.4%) exceeded the 10 point cap.

4. How would a different calculation or policy effect outcomes?

Small cell size exclusion errors

I identified a large number of cells with the value **0** in the current year (CY) to prior year (PY) comparisons. As 20 students are required for inclusion in the comparison, values of 0 must represent missing data. This makes it possible to automatically receive full points if the “zero” cell lands in the correct location of the comparison. (And alternately receive no points if it falls in the wrong side.) Providing valid percentages regardless of group size would remediate this problem – provided one of the two cells meets the minimum N of 20 rule.

Subgroup Improvement and cell sizes

The Subgroup Improvement variable divides students into seven groups based on race/ethnicity and three based on other conditions. The number of schools that had sufficient N to participate (out of 1372) are listed below:

Table 1. School Participation by Subgroup Improvement Category

Category	**ELA**	**Math**
White	1103	1103
African American	419	422
Hispanic	1252	1253
Asian	201	204
Native American / Alaskan Indian	187	188
Pacific Islander	2	2
Two or more races	213	215
English Learner	468	484
Special Education	1038	1043
Economically Disadvantaged	1188	1188

Schools have between 0 and 20 opportunities to earn the maximum allotment of six points. The more cells with N of at least 20, the more likely a school is to achieve maximum credit. For this reason, the Subgroup Improvement calculation favors schools with large enrollment (i.e., urban schools).

General methodological concerns:

All comparisons are all-or-nothing with no standard for acceptable improvement. Achieving an improvement of 0.01% is equivalent to 10% improvement. Likewise falling 0.01% short earns the school zero points. Point allocation based on “below”, “at-or-near”, and “above” target would reduce the arbitrariness of the measures.

Measuring Academic Growth as Part of an Accountability System

Dr. Thomas M. Haladyna
Professor Emeritus
Arizona State University
November 26, 2017

This is a working paper written as part of my work on the
Technical Advisory Committee to the Arizona Board of Education.

A major part of the accountability system in Arizona is student growth in reading, writing, and mathematics. As my contribution to this committee's work, I continue to ask questions and offer advice on issues affecting validity. The focus of this report is the validity of using growth measures in this high-stakes environment.

Technical Report and Documentation

As stated in an earlier paper (Haladyna, November 26, 2017), technical documentation is highly recommended for any high stakes testing program (AERA, 1999; AERA, APA, & NCME, 2014; Ferrera & Lai, 2016). AzMerit is used as a major part of the Arizona A-F Accountability System. Is there a technical report? Does it address questions and concerns stated in an earlier paper and this one? Have validity studies been done (Haladyna, 2006)? Documentation supplies validity evidence and an argument that using AzMerit test scores to measure growth in student achievement is fair and accurate in an accountability system. Without documentation, the state is subject to criticism and, even, legal challenges should some party feel that the accountability system has injured parents, students, and teachers.

What Do We Know About Academic Growth?

Without exception, all standardized achievement tests will show steady growth of students from grade three to grade eight. However, as we approach these higher grades there is an asymptotic tendency—that is, growth reaches a ceiling. Growth in grades 6, 7, and 8 may not be very much. How are teaching and school effects measured when variability is so greatly reduced in these grade levels? How can we make refined, accurate judgments when growth is so limited. A standard measure of differences is effect size (Cohen, 1988). What is the effect size of school effects at all grades? A box plot of growth by school for grades 3 to 8 would answer one question.

Growth in What Subject Matters

As state previously, Haladyna (November, 26, 2017), only reading, writing, and mathematics are used in a growth model. The most important question is why isn't the complete curriculum measured? If the accountability system only addresses growth in three subject matters, doesn't this invite school leaders and teachers to game the system and emphasize all three. We also know that the measurement of writing has validity issues and that teachers can game writing tests. Thus, schools and classes within schools can game the system by emphasizing these three subject matters and forgo instruction on those subject matters not measured in the accountability system.

Unit of Analysis

The unit of analysis is an important concept in an accountability system. One unit of analysis is the student. We measure academic progress from one year to the next. Another unit of analysis is the classroom. A third unit of analysis is the school. Sans a technical report or some other documentation, what is the unit of analysis?

If the student is the unit of analysis, growth scores may be very unstable. If the class is the unit of analysis, class composition varies from year to year. A cohort (class) seldom remains intact. Thus, bias is introduced by having new students come in and some students leaving. The class mean may be meaningless. A unit of analysis based on the entire school may be reliable if mobility is not a major factor. Random error (reliability) and systematic error (construct-irrelevant variance) are discussed in another section but it relevant to the issue of the unit of analysis.

Outliers

When looking at the indexes that comprise a total score for the accountability grade, outliers are very important concerns. These outliers may be a form of contamination. Each outliers should be investigated to see if there are tendencies associated with other factors in a school or teachers profile.

Personally, I have seen AIMS test growth data with negative gain scores in one published study (Mahoney, McSwan, Haladyna, & Garcia, 2010). These students were English language learners. This is very troubling because it suggests that either the data is contaminated or that these ELL students did not participate with the same motivation that is expected. I have also seen item responses by students with omits and not-reached blanks in AIMS test data at grades 3 to 8, These blanks are scored incorrect. Thus, their scores are lower. Test wise students know to fill in the blanks even if they have to guess. Some school leaders emphasize this and others don't. This is called test preparation. In other for this not to threaten validity, all students should have the same test preparation otherwise those receiving it will have artificially higher scores.

A good example of outliers in data is shown in a paper by Betebrenner and Linn (2000, p. 6). Growth is plotted against prior achievement with many, unexplained outliers. Before any action is taken where grades are assigned, outliers should be investigated to determine what might have caused such a serious disturbance. It is difficult to explain an outlier.

Stability of Growth Measures

For quite some time, we have known that student gain scores can be unreliable (Cronbach and Furby, 1970). A recent paper address the fact that growth scores in accountability system are often unstable (Griffith & Petrilli, November 10, 2017) When combined to form class means, gain scores can have increased reliability. When combined by school, we should know that reliability estimate. Also, we should have descriptive statistics, reliability estimates for various units of analysis. Amrein-Beardsley (2014) presents arguments and observations about residual and interaction effects that may distort an interpretation of teacher effects. One of these is that teachers may swap teaching assignments. I have done this in my elementary school teaching. Thus, a teacher effect is distorted. We know that some schools and teachers offer incentives whereas other do not. This too is a type of distortion. As we can see, growth measures do not seem to be too stable yet they play an important role in the accountability system.

Reliability and the Use of Cut Scores to Produce Achievement Classifications

Reliability is a useful concept. However, it is only a gateway to a more important statistic: Conditional standard error of measurement (CSEM). We use adjusted test scores and three cut points on the test score scale to classify students as Minimal Proficiency (MP), Partial Proficiency (PP), Proficient (P), and High Proficiency (HP). These are graded categories. Underlying these categories is a continuous scale representing student achievement. Two highly relevant and related questions/concerns arise:

1. The CSEM is an estimate of how much random error exists around each cut score that separates MP and PP, PP and P, and P and HP. Graphically it looks like this:

```
             xxxxx                   xxxxx                xxxxx                 xxxxx
---------------MP-------------------PP---------------P-----------------HP--------------------------
```

Each x represents a school near the cut point.

   Schools will be distributed close to cut points. The margin of random error is important because schools might be misclassified simply due to random error. If the margin of random error is large, than many schools are at risk of misclassification. It is important to know this and to have a strategy for minimizing the risk of misclassification.

2. The four categories for classifying students is artificial. It is based on human judgment, which we know is subjective and is possibly biased. Moreover, the categories are crude approximations of an underlying score scale that is more informative than these categories. Various panels of persons setting these cut scores produce different recommendations. When we measure growth, a school may have a sizable improvement yet still remain in the same category. Another school may have very small growth, but due to the fact that they reside at the cut point will be reclassified as improving. This fact exposes a danger in using categories instead of actual test scores. Moreover, states vary considerably in the stringency of their cut scores (Betebrenner & Linn, 2000).

Construct-irrelevant Variance (Bias)

The idea that test scores can be corrupted has been around for a long time. In 1988, Haladyna, Nolen, and Haas (1989) were invited by the Arizona legislature to conduct a study of how Arizona teachers regard the administration and use of the state's standardized achievement test. This was in the day of low-stakes testing. The results showed considerable cheating on the test. These results were confirmed by another, independent study (Smith, 1989) where interview data supplanted the survey data. With high-stakes accountability, cheating is an industry and scandal that has touched many cities, including New Orleans, Dallas, Houston, Atlanta, Chicago, Washington, DC to name a few. We have many strategies for improving test scores without relying on real student learning. It is very important to study threats to validity, which technically are called construct-irrelevant variance (Haladyna & Downing, 2004). Bias is a term that more popularly describes contamination in test scores.

A study by Briggs and Weeks (2009) revealed that different growth models produced different results. Thus the model itself comes into question in producing bias in test scores. Which model is most accurate, truthful? A related issue is that any growth measure is subject to its size versus how it compares with other similar students. This issue collides with the use of poverty or risk as a control. As stated previously (Haladyna, November 26, 2017), poverty is a poor control variable because risk is much more complicated and representative of the hardships students at-risk face in schools.

Discontinuity From Grade to Grade

Examination of content standards for each grade level (grade 3 to 8) will show considerable variation. The test used for each grade level represents what is supposed to be taught. However, most teachers will adapt teaching to the status of the class, what they can actually do as opposed to what they should do at that grade level. Low achieving fourth graders will likely be reviewing third grade work, and high-achieving sixth graders may enjoy enrichment activities. Both groups may suffer when growth is considered on a single test that is not well aligned with instruction. One way to discover these inequities is a forensic item analysis where each item is subjected to independent analysis for bumps and falls. Teachers can actually predict which items their students will perform well and which items they probably will not perform well.

Another point, made by Betebrenner and Linn (2000) is that the subject matter tested may change from grade to grade, thus a smooth continuum in a vertical scale may not represent actual student growth but jumps due to the structure of the subject matter. For instance, mathematics is cited where algebra may be emphasized in one grade, and geometry emphasized in an adjacent grade.

Conclusion

We have many challenges in measuring growth and then using it appropriately to improve student learning. Betebrenner and Linn (2010, p. 4) stated: “Growth analyses based upon impoverished measures of student achievement are themselves necessarily impoverished.” Linn (2008) further states that growth results in an accountability system should not be used for causal inference but should be use to flag schools for investigation.

Given the high-stakes use of growth scores, validity studies, technical documentation, and peer review are important activities that help validate the use of growth scores in this accountability system.

Before embracing growth measures and embedding them in complex statistical models, we need some assurance that the data provided is validly interpreted.

References

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Betebrenner, D. & Linn, R. L. (2010). *Growth in Student Achievement: Issues of Measurement, Longitudinal Data Analysis, and Accountability.* Princeton, NJ: Educational Testing Service.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74*(1), 68-80.

Ferrera, S., and Lai, E. (2016). Documentation to support test score interpretations and use. In S. Lane, M. Raymond, and T. M. Haladyna (Eds). *Handbook of Test Development*. New York: Routledge.

Griffith, D., & Petrilli, M. J. (November 10, 2017). *How states should redesign their accountability system.* Washington, DC: The Thomas B. Fordham Institute.

Haladyna, T. M. (November 26, 2017). *Questions and Concerns About the Arizona A-F Accountability System.* Phoenix: Author.

Haladyna, T.M. (2006). Roles and importance of validity studies in test development. In S.M. Downing and T.M. Haladyna (Eds) (pp. 739-755). *Handbook of Test Development*, Mahwah, N.J.: Lawrence Erlbaum Associates.

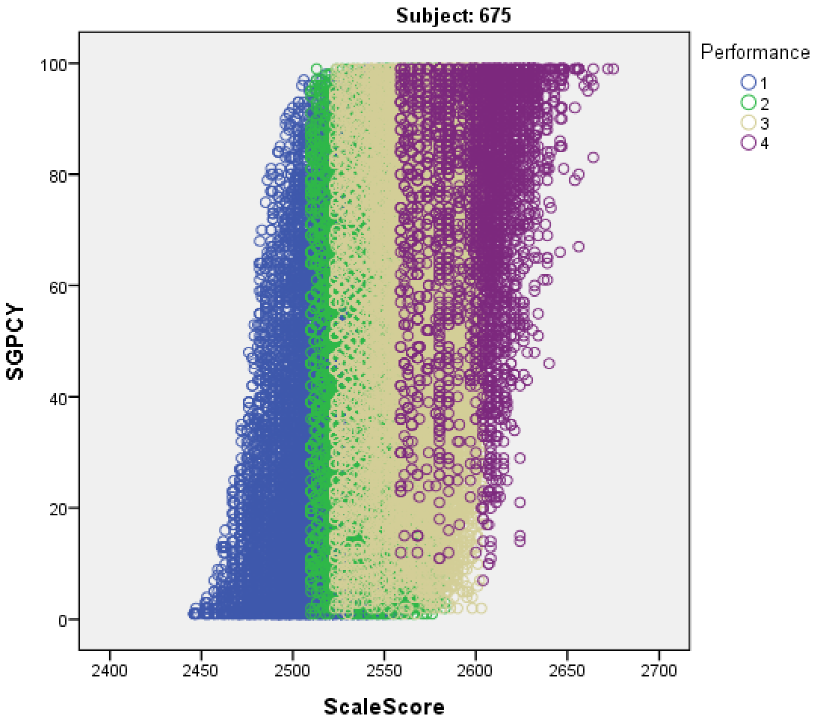
Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.

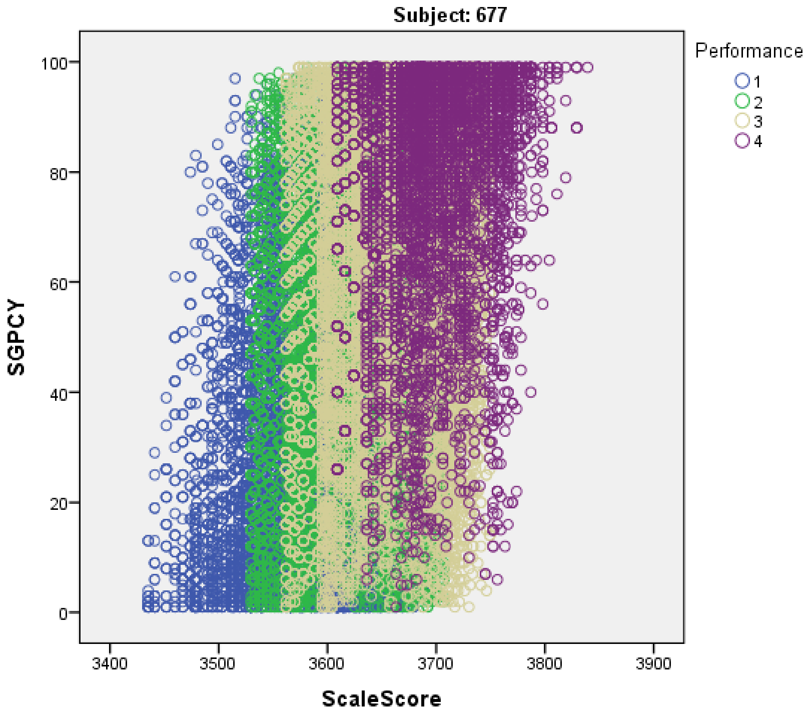
Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20,* 2-7.

Linn, R. L. (2008). Educational accountability systems. In K. E. Ryan & L.A. Shepard (Eds.), The future of test - based educational accountability (pp. 3–24). New York, NY: Routledge.

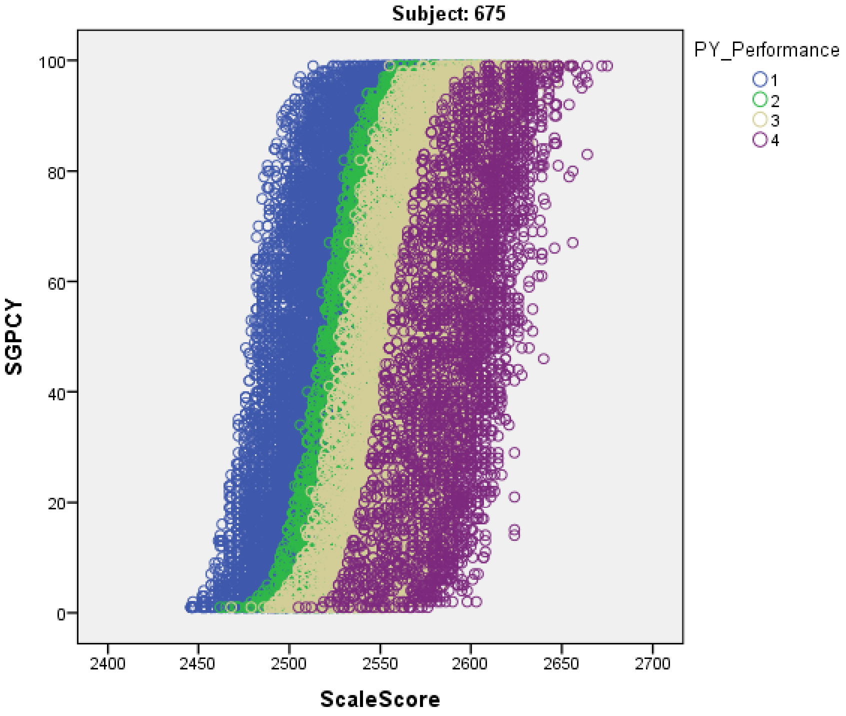
Mahoney, K., MacSwan, J., Haladyna, T., & García, D. (2010). Castañeda's third prong: Evaluating the achievement of Arizona's English learners under restrictive language policy. In P. Gandara & M. Hopkins (Eds.), Forbidden languages: English learners and restrictive language policies (pp. 50-64). New York: Teachers College,

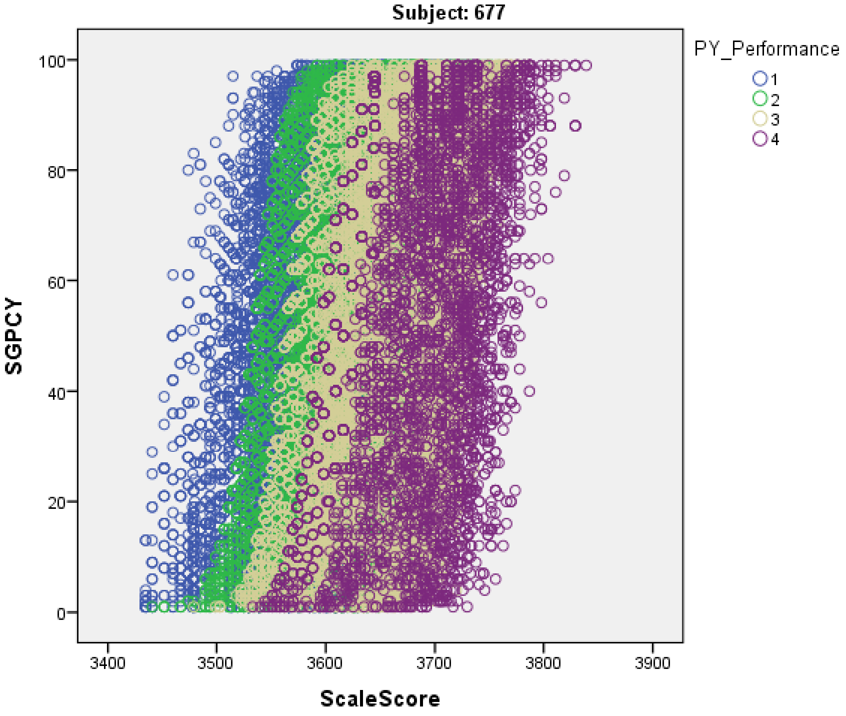
Smith, M. L. (1991). Put to the test. The effects of external testing on teachers. Educational Research, 20(5) 8-11.



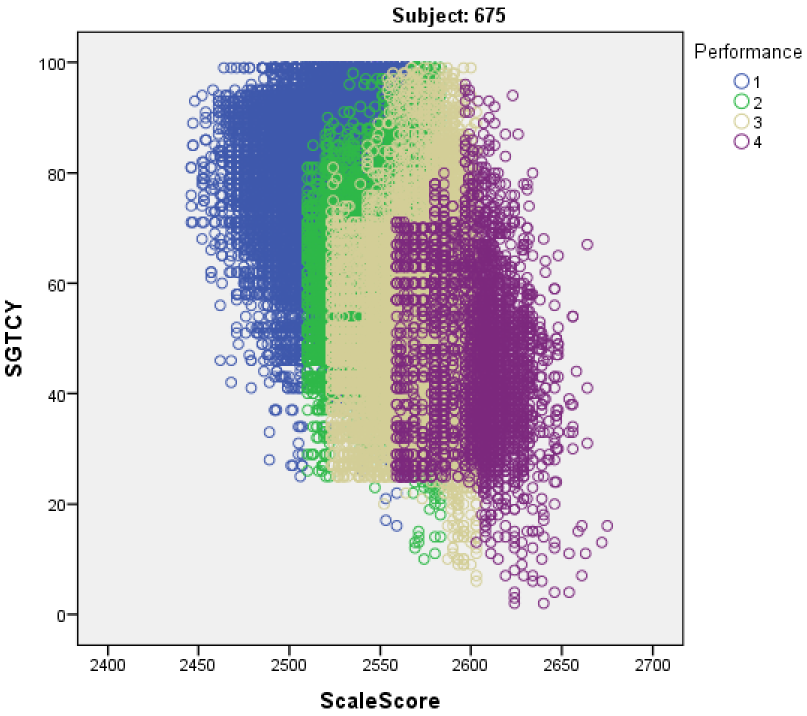


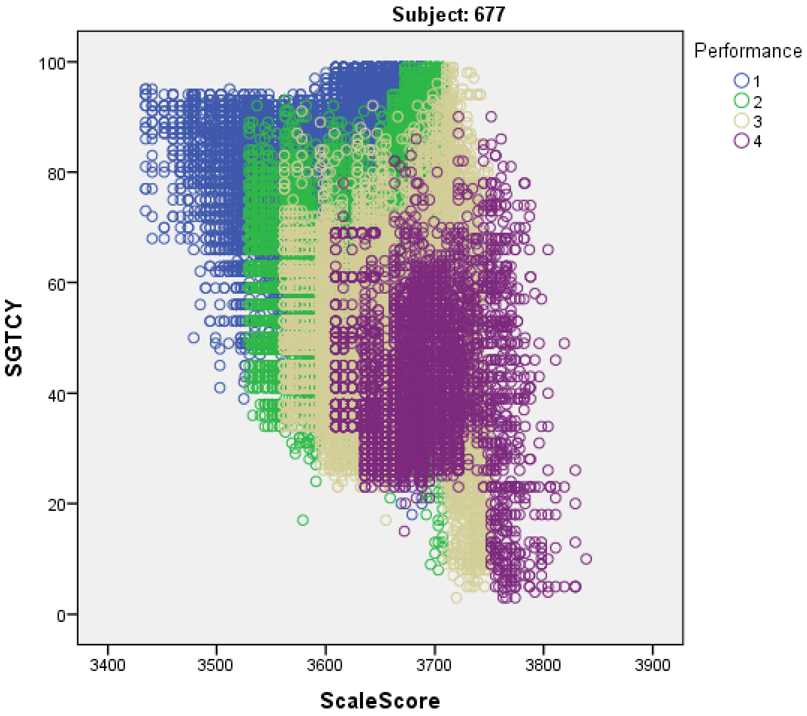
Here is an analysis done of the MPS student level A-F static file posted in October 2017. It is broken out by subject. 675 is ELA, 677 is Math. SGPCY represents the Current Year Student SGP, Scale Score is also current year Scale Score. Performance levels are 1 for Minimally Proficient, 2 for Partially Proficient, 3 for Proficient, and 4 for Highly Proficient. MPS students across all proficiency levels are demonstrating growth on AzMERIT.





Here is an analysis done of the MPS student level A-F static file posted in October 2017. It is broken out by subject. 675 is ELA, 677 is Math. SGPCY represents the Current Year Student SGP, Scale Score is also current year Scale Score. PY_Performance levels are for Prior Year Performance and are 1 for Minimally Proficient, 2 for Partially Proficient, 3 for Proficient, and 4 for Highly Proficient. MPS students across all prior year proficiency levels are demonstrating growth in the current year on AzMERIT.





Here is an analysis done of the MPS student level A-F static file posted in October 2017. It is broken out by subject. 675 is ELA, 677 is Math. SGTCY represents the Current Year Student SGT, Scale Score is also current year Scale Score. Performance levels are for Current Year Performance and are 1 for Minimally Proficient, 2 for Partially Proficient, 3 for Proficient, and 4 for Highly Proficient. MPS Minimally Proficient students have a much higher SGT than their Highly Proficient counterparts. The cause of outliers would need to be investigated in the future.

ELPoints – K-8 Dataset

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	16	1.1	2.5	2.5
	4	8	.5	1.2	3.7
	5	40	2.7	6.2	9.9
	6	46	3.1	7.1	16.9
	7	78	5.3	12.0	29.0
	8	89	6.1	13.7	42.7
	9	125	8.5	19.3	61.9
	10	247	16.8	38.1	100.0
	Total	649	44.2	100.0	
Missing	System	818	55.8		
Total		1467	100.0		

In the K-8 Dataset (TAC_K8_schooldata2.0.xls), 38% of schools eligible for the EL points earned the full points available.

ELPoints – 9-12 Dataset

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	3	.8	4.5	4.5
	5	1	.3	1.5	6.0
	6	1	.3	1.5	7.5
	7	6	1.6	9.0	16.4
	8	22	6.0	32.8	49.3
	9	21	5.7	31.3	80.6
	10	13	3.5	19.4	100.0
	Total	67	18.2	100.0	
Missing	System	301	81.8		
Total		368	100.0		

In the 9-12 Dataset (TAC_912_schooldata2.0.xls), 19% of schools eligible for the EL points earned the full points available.

Options for Schools with Unique Grade Configurations

R. Guyer

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.

Proposed Approach:

The outlier merge option is best suited to situations when there *are* outlier grades. Schools that cover nearly the full range of grades (e.g., K-12) should use one of the other options as the K-8 or 9-12 models alone cannot provide a fair account of performance.

- Schools that do not include Grade 12 adopt the K-8 model
- Schools that include Grade 12 but begin with the 5th grade or higher adopt the 9-12 model
- Schools with Grades K/1/2/3/4 to 12 use both K-8 and 9-12 models

Option 3: Average the two letter grades

Definition: Use the existing data as-is and take an average.

Considerations:

As discussed in the slides provided by AZSBE, this approach does not take into account student enrollment numbers.

There were 75 schools with valid grades for both the K-8 and 9-12 models. Let’s consider the ratio of Full Academic Year (FAY) student enrollment between the K-8 and 9-12 models.

Table 1. Ratio of K-8 to 9-12 FAY Enrollment for Common School Configurations

Configuration	**Schools**	**Avg. K-8 Enroll**	**Avg. 9-12 Enroll**	**Ratio**
7-12	24	141.9	271.0	0.524
6-12	16	196.4	364.3	0.539
5-12	7	371.4	267.6	1.388
K-12	22	345.3	134.4	2.569

These results show that the 6/7-12 configurations were majority 9-12 students based on the enrollment data. Likewise the K-12 configurations were majority K-8 students.

The disparity in the enrollment numbers for the common school grade configurations provides evidence against Option 3: Simple Average as the solution.

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.

Considerations:

This model updates Option 3 by weighting the two letter grades by FAY enrollment. As shown by Table 1, the ratio of enrollment (K-8 to 9-12) differs depending on the school configuration. These results provide support for Option 4 over Option 3 as it is fairer to all school configurations.

Option 4 still cannot account for schools that:

1) Exclude Grade 12 and miss graduation rate and/or CCRI points
2) Include too few K-8 grades to obtain acceleration / readiness points

Option 4 alone provides a better solution than Option 3. It still suffers from the above cited flaws (and more) that limit its applicability to all configurations.

Option 5: Hybrid Approach

This model adopts components of Options 2 and 4 with a goal of improving the fairness (and reducing the cons) of the approach.

See the proposed business rules defined below:

A. Schools without Grade 12 adopt the K-8 model **(Merge to K-8)**
B. Schools that include Grade 12 but begin with the 5th grade or higher adopt the 9-12 model **(Merge to 9-12)**
C. Schools with Grades K/1/2/3/4 to 12 use both K-8 and 9-12 models **(Prorate K-8 and 9-12 grades using FAY enrollment)**

Schools that receive a prorated grade (Option 5.C) would receive a grade based on **all** K-8 and 9-12 schools. This is the most appropriate population to compare such schools to since the letter grade is a weighted average of both models.

Schools given an A-F grade under Option 5 would be included in the standard deviation model once for their Option 5 A-F grade. If a school receives an A-F letter grade for Grades 6-8 *and* 9-12 then the grade would be merged and included in the 9-12 model only (excluded from the K-8 model).

Missing data (NR ratings)

There were 33 schools with unique grade configurations that did not receive a grade for the K-8 and/or 9-12 models.

Missing High School Rating:

There were **nine** K-12 schools missing the high school rating. These schools had small FAY enrollments in high school (average 33) compared to grades K-8 (average 173).

Four schools with different configurations (6-10, 4-11, 6-11, and 6-12) were missing just the high school rating. They generally had low FAY student counts (K-8 ranged from 37-76) and high school included only between 20 to 86 students.

Missing K-8 Rating:

Four schools were missing the K-8 rating and were all Grades 7-12. These schools featured large high school enrollments (552, 1542, 2087, and 2106) and only one included more than zero FAY K-8 students (just 13).

Missing Both Ratings:

There were **16** schools missing both K-8 and 9-12 ratings. These schools included few FAY students in K-8 (average 13) or 9-12 (average 22).

The solution to the missing data problem depends on the type of school configuration. Some missing data could be resolved through use of a hybrid model (see Option 5) to merge the less represented grades together. The schools that were missing both grades fell into a different category due to low FAY enrollment, and likely require alternative modeling to evaluate their performance.

Unique School Configurations and A-F Letter Grades

AZ State Board of Education – A-F Technical Advisory Committee

November 26, 2017

D. Jordan

Breakdown of schools that serve elementary grades and high school grade levels

- Note: Of the 5 schools that do not have grade levels 11 and 12 for SY2016-17, it is reasonable to expect that schools that were limited to grade 11 last year may have their first graduation classes for 2018.

Table 1: Grades Served and Number of Schools

Distribution: Number of Schools			
Grades Served	**Public**	**Charter**	**Total**
K-10		1	1
1-12	1		1
K-12	10	30	40
2-12	1		1
3-12		2	2
4-11		1	1
4-12		2	2
5-12		7	7
6-10		1	1
6-11		2	2
6-12	5	15	20
7-11		2	2
7-12	14	14	28
Total	**31**	**77**	**108**

- More background of the 108 unique grade level configured schools by FRL.
- 46 of the 108 schools do not appear to report FRL.

Table 2: Breakdown of Unique Configured Schools by FRL

Number of Schools by Grades Served and Percent FRL												
Grades Served	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**	**70-80**	**80-90**	**90-100**	**(blank)**	**Total**
K-10											1	1
1-12											1	1
K-12		2		3	3	1	6	8	4	3	10	40
2-12								1				1
3-12											2	2
4-11						1						1
4-12											2	2
5-12			1								6	7
6-10								1				1
6-11											2	2
6-12		1	1	1			2		3		12	20
7-11				1					1			2
7-12	1	1	1	2	6	1	1	2		3	10	28
Total	**1**	**4**	**3**	**7**	**9**	**3**	**9**	**12**	**8**	**6**	**46**	**108**

- 42 of the 46 schools that do not report FRL are charter schools.

Table 3: Distribution of Unique Configured Grade Level Schools by FRL and School Type

Number of Schools with Different Grade Configurations			
Percent FRL	**Public**	**Charter**	**Total**
0-10	1		1
10-20	4		4
20-30		3	3
30-40	2	5	7
40-50	6	3	9
50-60	2	1	3
60-70	3	6	9
70-80	3	9	12
80-90	4	4	8
90-100	2	4	6
(blank)	4	42	46
Total	**31**	**77**	**108**

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

- Applying this option, 92 of the 108 schools are evaluated with a letter grade for either K-8 or 9-12 model.
- Some schools have so few students at 9-12 or k-8 that a letter grade may not be calculated for either one of the two models.
    - However, there are enough students within one of the two models in which a letter grade can be calculated.
    - For these schools, whichever model can be applied that is the letter grade for the school
- 33 schools are not rated with the k-8 model, the 9-12 model, or both models (16 schools).
    - Options 3 and 4 require calculated points for both models.
    - For consistency, apply the model most appropriate
        - Example, most of the 6-12, 7-12, and 8-12 configured schools have a large percentage of grade 8 students taking the Algebra 1 EOC test. For schools such as these, it is reasonable to apply the 912 model for the letter grade.
- Of the schools that are 3-12, 2-12, 1-12, and K-12 students the number of students in K-8 are more than the number of students in grades 9-12.
    - It is logical that with four high schools grade levels in comparison to the number of elementary grade levels that the high school level has far fewer students.
    - As students move from the elementary grade levels through the high school grade levels, are there indicators that a high percentage of the students are graduating with strong indicators of being college and career ready? If that is observed in the data, it is logical to apply the 912 model to the unique configured schools.

Table 4: Distribution of Letter Grades for Unique Grade Configured Schools (higher of the two models or only one model that can be calculated)

Option 2: Whichever is higher (K-8 or 9-12)							
Grades Served	**A**	**B**	**C**	**D**	**F**	**(blank)**	**Total**
K-10						1	1
1-12						1	1
K-12	7	12	8	4	1	8	40
2-12						1	1
3-12	2						2
4-11			1				1
4-12	2						2
5-12	6	1					7
6-10					1		1
6-11	1	1					2
6-12	7	7	2		1	3	20
7-11			1			1	2
7-12	2	11	9	4	1	1	28
Total	**27**	**32**	**21**	**8**	**4**	**16**	**108**

Table 5: Applying the 9-12 model for schools that have a calculated 9-12 letter grade regardless of the K-8 calculated grade.

Option 2: When appropriate apply the 9-12 letter grade.							
	A	**B**	**C**	**D**	**F**	**(blank)**	**Total**
-1-10						1	1
1-12						1	1
-1-12	5	11	8	6	2	8	40
2-12						1	1
3-12	2						2
4-11			1				1
4-12	2						2
5-12	6	1					7
6-10					1		1
6-11	1	1					2
6-12	7	7	2		1	3	20
7-11			1			1	2
7-12	2	9	11	4	1	1	28
Total	**25**	**29**	**23**	**10**	**5**	**16**	**108**

Option 3: Average K-8 points and 9-12 points

- 33 schools do not get letter grades because the numbers are too few for a letter grade calculation at either K8 or 9-12 or both.
- Which grade scale is applied: K-8 or 9-12?

Table 6: Straight up average of the K8and 912 models.

Number of Schools	**Averaging K8+912 Points (No Prorating or Weighting)**					
	9-12 Cut Point Scale					
K-8 Cut Point Scale	**A**	**B**	**C**	**D**	**(blank)**	**Total of K8 Cut**
A	19					19
B		22				22
C		5	11			16
D			6	6		12
F				6		6
(blank)					33	33
Total of 912 cut	**19**	**27**	**17**	**12**	**33**	**108**

Option 4: Weighted/Prorated per number of students K-8 and number of student 9-12.

- Calculated the percent of all FAY students K-8 and calculated percent of FAY 9-12 students. Basis to determine weighting.
- Steps calculating prorating/weighting for letter grade points.
    1. # of K8 FAY students/# of FAY students K-12 = % K8 FAY
    2. # of 912 FAY students/# of FAY students K-12= % 912 FAY
    3. Multiply K8 Earned % of total eligible by the % K8 FAY = weighted K8 points
    4. Multiply 912 Earned % of total eligible by the % 912 FAY = weighted 912 points
    5. Sum weighted K8 points and weighted 912 points = Total points for letter grade.
    6. Which scale to use for school letter grade: K-8 Scale or 9-12 scale?

Prorated/Weighted Per # of Students K-8 and 9-12						
	912 Cut Scale					
K8 Cut Scale	**A**	**B**	**C**	**D**	**(blank)**	**Total**
A	12	5				17
B		20				20
C		7	12			19
D			4	6		10
F				7		7
(blank)		2			33	35
Total	**12**	**34**	**16**	**13**	**33**	**108**

- Note: K8 student FAY number and 9-12 student FAY number are calculated for any or all grade levels in the two separate grade bands. Example: For a 6-12 school the prorated calculation for K-8 is limited to students enrolled in grades 6 – 8.
- Which grade scale is applied: K-8 or 9-12?
- 33 schools do not get grades.