TECHNICAL ADVISORY COMMITTEE

December 4, 2017 Report

A-F

Report by:

Amy Schlessman - Chair
Rick Guyer - Vice Chair
Audrey Amrein-Beardsley
Cindy Bochna
Thomas Haladyna
Christy Hovanetz
David Jordan

(Disclosure: TAC Report was finalized on November 30, 2017. Member Amrein-Beardsley and Member Hovanetz were not present at that meeting).

Background

On October 23, 2017, the State Board of Education (Board) directed the Technical Advisory Group (TAC) to review the A-F Accountability Plan, business rules and impact data for problematic issues.

To date, the TAC has met five times to discuss issues relating to the impact data, business rules and the A-F Accountability Plan. The pace of these meetings has been accelerated to meet deadlines. The TAC thinks there has not been enough time to consider, study and evaluate all issues thoroughly.

From reviewing the data, the TAC has identified some problematic issues:

N-Count:

The full academic year (FAY) n-counts for proficiency are aggregated across subject areas (ELA, Math, and Science). This makes proficiency points more accessible. In contrast, the FAY n-counts for Student Growth to Target (SGT) and Student Growth Percentile (SGP) are broken out by subject area (ELA and Math). By disaggregating the FAY n-counts for SGT and SGP, fewer schools have access to these points. One solution to this may be to adjust the SGT and SGP calculation to include both ELA and Math. This would have the added benefit of reducing model complexity by providing consistent treatment for proficiency and growth. This potential solution would require additional time and analysis.

The same problem outlined above applies to subgroup improvement as well because the n-counts are separated by subject area. Combining them would give more schools access to these points. However, if one combines across subjects for subgroup improvement, there are half as many categories in which to earn two points (20 reduced to 10). This may not be beneficial to all schools.

For smaller schools, one option would be to aggregate n-counts across school years within the school so that more schools have the potential to reach the n of 20. If this is done, the school's current year averages can be compared to their prior year averages for the measures to determine if they earn credit for improvement or not.

The Every Student Succeeds Act (ESSA) requires the same n-count for each category (e.g. proficiency, growth, CCRI) except for English Language Learners (ELL) and for both frameworks (i.e. K-8 and 9-12). The data files provided by Arizona Department of Education (ADE) Accountability did not include verification for College and Career Ready Indicator (CCRI) n-count because that data was self-reported. If the school met the n-count, 20, the school should have submitted. If the school did not, the school should have selected N/A. If the n-count were changed, to be fair to all schools, the window for 9-12 schools would need to be re-opened. Schools meeting a revised n-count could then self-report their data.

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).

Another issue with n-counts is the standard error of the mean. It increases, thus categorical consistency is lower. Also, there is a bias issue. By using a different n-count it is not known if the new n-count includes higher or lower achieving students.

Growth:

It is important to understand the differences between SGPs and SGTs. SGPs are normative. All students can benefit or not from the SGP calculation regardless of proficiency level. To obtain an SGP, every student is compared by scale score to their peers around the state. In other words, the highly proficient student (as determined by Scale Score) is compared to other highly proficient students and then ordered from 1 to 99 to determine their SGP. Every student has the potential to earn a 1 SGP up to a 99 SGP within that peer group.

In contrast, SGTs are not normative - every student has the potential to be on target regardless of their peer group. SGT is reported as the growth percentile a student needs to earn to be at the proficient scale score in three-years time or by high school graduation. This SGT target was set by Board policy and can be altered as the Board sees fit. The confusion comes because SGT is presented in the same format as SGP, on a 1 to 99 scale. It is the SGP needed to be on track to proficiency. A minimally proficient student has a much harder time of obtaining their target than a proficient student whose target will be lower, but all student SGTs are independent of all other students. This is contrasted with SGPs which are inter-dependent – within the peer group one student will get the 1 SGP and another student will earn the 99 SGP. For the SGP, approximately 33% of students will fall into the low growth category (1-33), 33% into the average growth category (34-66), and 33% into the high growth category (67-99).

Here is a hypothetical situation: A student who is proficient in year one AND proficient in year two may have an SGT of 40 for year two. If the student's SGP is 20 in year two, they will be in the low growth category for SGP (p. 14 of business rules) and below target for SGT (p. 16 business rules) because their trajectory indicates they are no longer on track to be proficient in three years. If the student's SGP is 40 in year two, then they will be in the average growth category for SGP and at/near target for SGT. They are not penalized for maintaining, points are awarded for maintenance per the business rules. If the student's SGP is 67 in year two, they will be in the high growth category for SGP and exceeds target for SGT. Proficient and highly proficient students can and do demonstrate growth – it is a misperception that they do not. However, the opposite is also true: students in all four proficiency categories may NOT demonstrate growth.

A proficient student's SGT should typically fall into the at/near target category as outlined on page 16 of the business rules**.** One identified issue is that there is, currently, an SGT ceiling effect related to students with an SGT of 89 or higher as presently the business rule indicates that the student can only receive credit for being “At/Near Target” as opposed to “Exceeds Target”. A solution for this may be to change the business rule to give all students who meet the 89 or higher SGT credit for “Exceeds Target”. This is similar to giving full points for having a 90% or higher graduation rate. This would need to be investigated through future analysis that is beyond the scope of the time constraints present in developing this report.

The weights assigned to SGT and SGP for proficient and highly proficient students have been identified as a point of contention. The argument for adjusting the SGT and SGP weights for proficient and highly proficient students to be higher is that this will allow proficient students access to all the points in the model. From a validity standpoint, it appears that current weights are unfair to high proficiency schools.

Weighting the SGPs and SGTs higher for the proficient and highly proficient students will lend validity to the model but increase the number of points earned by only proficient/highly proficient students in the model. These proficient and highly proficient populations already earn 30% of overall model points for proficiency that the minimally and partially proficient students do not. If the SGP and SGT weights for proficient students are increased, the majority of the points in the letter grade model will go to schools with high levels of proficiency. This will skew the letter grades in a way that will be more correlated to poverty. If the cut scores for letter grades are not then adjusted along with these weights, higher poverty schools will have less access to the higher letter grades. To balance these additional points, the weights for minimally and partially proficient SGT and SGP would also need to be adjusted up but this would ultimately result in higher point totals overall. Another consideration would be to have the model assign more points to the non-normative SGT; or to assign greater weight for either SGP or SGT, depending on the school's higher score. Which would add an additional layer of complexity.

Another analysis argues that the growth indicator negatively impacts schools with high proficiency, due to the weighting within the SGP/SGT model.

This can best be demonstrated using a simple example. Suppose School A was composed of nothing but highly proficient students that met the target on ELA and Math. That school would earn 30 + 25 = 55/80 points (equivalent to a C) for the K-8 model. School B was composed of only proficient students that met the target for ELA and Math. School B was composed of only proficient students that met the target for ELA and Math. School B earns 30 + 35 = 65/80 points which is the equivalent of a B. Note that in this example the schools are not eligible for acceleration/readiness points or CCRI points.

School A earned a lower A-F grade than School B despite having students at a higher proficiency. Evaluation and modeling to correct for the “ceiling effect” of SGT and/or concerns regarding the application of bonus points to cut score determinations will reduce these concerns.

In the current model using data provided by ADE, of the K-8 schools with high growth (40-50 points earned for growth): 222 received an A, 184 received a B, and 18 received a C. In the 9-12 file for high growth schools (15-20 points earned for growth): 38 received an A, 32 a B, and 10 a C. These demonstrate that high growth alone will not lead to a school receiving an A, that schools must also demonstrate proficiency and earn points for acceleration/CCRI to get an A. During the past year the Board appeared to make the policy decision that to be an A school one would have to be excelling across the measures. There is evidence the current approach does that. Adjusting the weights is more a policy decision than a psychometric decision.

Proficiency

In the current model using data provided by ADE, of the 305 K-8 schools demonstrating high proficiency (25-30 points earned for proficiency): 183 received an A, 104 received a B, 9 a C and 9 an NR. Of the 9 C schools, all were evaluated out of 90 points due to not having FAY EL students. 2 of 9 are part of the non-typical configuration schools and all earned less acceleration readiness points than the high proficiency A and B schools. In the 9-12 file for the 34 high proficiency schools (25-30 points earned for proficiency): 31 received an A, 1 a B, and 2 an NR. High proficiency alone does not lead to an A; a school must also show improvement in student growth and earn points for acceleration/CCRI to earn an A. This appears to be consistent with the Board's direction that an A school be truly excelling.

Acceleration Measures (K-8)/CCRI (9-12):

As mentioned under the n-count heading, n-count concerns in this area could be addressed by aggregating student counts across school years and then comparing the current year averages to the prior year averages to see if improvement was achieved. One problem is that new schools were not eligible for the majority of these points because most of the indicators are evaluated based upon improvement over the prior year. To give new schools access to these points they could be evaluated against the state average until they have two years' worth of data. Some schools are only eligible for a limited number of acceleration points due to the homogeneous natures of their populations. The Board may want to review the business rules to determine if schools should be graded upon the number of acceleration measures for which a school qualifies and meets instead of the total number of acceleration points in the entire model. For example, if a school only had 4 possible comparisons worth 2 points each and the school achieved two of those four, then their total would be four out of eight points.

There are 20 possible points in the Acceleration/Readiness Indicator, though schools are capped at 10 points. Due to the n-count or other eligibility deficiencies, some schools are not eligible to earn points in each metric (Grades 5, 6, 7, 8 HS EOC Math; Grade 3 ELA Minimally Proficient; Chronic Absenteeism; Subgroup Improvement; and Special Education Inclusion). The denominator in the calculation remains 10 regardless of how many points the school is eligible to earn. The Board may want to consider using the total points possible a school is eligible to earn, capped at 10, in the denominator of the calculation.

ELL

In analyzing the ELL Points, 38% of schools eligible for ELL points in the K-8 dataset earned the full 10 points while 19% of schools eligible for ELL points in the 9-12 dataset earned the full 10 points on this measure. Schools that do not meet the n-count in the current year could have their n-count combined with that from the prior year in order to reach the minimum n of 20 and have access to the points.

Free and Reduced Lunch (FRL)

TAC computed with a file provided by ADE its own correlation coefficient for the relationship between number of FRL and A-F Accountability scores. The accountability scores are expressed as the ratio of “Total Points Earned” out of “Total Points Eligible.” For the correlational analyses these scores are expressed as proportions. A negative correlation indicates that as the percentage of FRL students increases, the total points earned tends to decrease. The computations show a moderate correlation between FRL and accountability scores for both K-8, -.56, and 9-12, -.50. The correlation between FRL and Proficiency equaled -.805 for K-8 schools and -.620 for 9-12 schools. The correlation between FRL and Growth was -.277 for K-8 schools and -.218 for 9-12 schools. Most TAC members agree that some correlation is inevitable, and a moderate correlation is more desirable than a strong one.

One TAC member with agreement from several others observed that FRL alone is not the best measure of at-risk and a more sophisticated risk index could be used. At least one TAC member disagreed. The Board might want to task the TAC to review this further in the future.