**A Healthier Mess: State Accountability Options under ESSA**

**Sherman Dorn**

**Design objectives**.

**Citizen judgment as part of the process**. A grand-jury structure for identifying the worst and best schools in addressing inequality is a way to guarantee credibility for judgments in a way that algorithmic accountability has rarely had.

**A combination of measures to avoid large weights for any individual statistic**. Since ESSA requires the use of proficiency rates, one design objective is a combination of measures on academic achievement to reduce both the short-term gaming around “bubble kids” (both real and perceived) and also the long-term incentive to lowball cut-scores for various achievement bands on statewide tests. Where improvement is included, this proposed set of measures gives schools and districts incentives to pay attention to all vulnerable subgroups by including both the *most* and *least* improved vulnerable subgroup scores. In some areas, I propose using data from multiple years, and I also mix up the type of measure depending on what I thought was the worst side-effect to avoid. This is an explicit tradeoff: we lose quantitative simplicity to gain balance.

**An incentive for long**-**term ambitions**. Baking in multiple stretch goals for what happens to students after they leave for middle school is effectively a set of bonuses to keep long-term ambitions on the radar of elementary schools. This will require tracking students after they leave elementary school in relatively unique ways, and will stretch state data systems.

**Structure**.

**Indicator(s) of cross-sectional academic achievement**. Roughly equal weighting to the following (plus any way to work in easier units, such as multiplying everything that follows by 10):

* + 1. **Logit (natural log-odds) of the percentage proficient for all students, reading and math, across grades. (Two measures, one each for reading and math, combined across grades. Logit used to make extremes on proficiency rates more important than differences in the middle of the range. With a logit transformation, the gaming-the-system “value” of setting low cut scores for achievement bands is diminished.)**
		2. **Logit of the lowest percentage proficient for vulnerable sub-groups, reading and math, across grades (two measures).**
		3. **Logit of the highest percentage proficient for vulnerable sub-groups, reading and math, across grades (two measures).**
		4. **A measure of “distance from the middle” for the lowest-performing students: Scale-score differences between the 10th and 50th percentile students in reading and math, each grade third grade and up, in standard deviation units for the state for that year (for the grade and test), weighted so that together they are equivalent to two other measures (one each in reading and math), using a constant\* minus the 10-50 scale-score difference.
		\* I would try 1.25 or 1.3 as the starting constant. This may depend on the assessment used in a particular state.**
		5. **For any other subject the state chooses, mirror i-iv in that subject.**

**Indicator(s) of student growth or an alternative**. One of the following, or a mix:

* + 1. **Measures of changes in proficiency percentages as follows:**
			1. **For each tested subject that counts (math, reading, and any other subject the state chooses), some constant\* times the natural log of the following quantity for all students: one plus the absolute change in percentage proficient over the past three years (e.g., for a 7% positive change, the natural log of 1.07).**
			2. **For each tested subject that counts, some constant times the natural log of one plus the absolute change in percentage proficient over the past three years for the sub-group that has made the *least* improvement.**
			3. **For each tested subject that counts, some constant times the natural log of one plus the absolute change in percentage proficient over the past three years for the sub-group that has made the *most* improvement.
			*\* The constant should be chosen so that the growth components as a whole are weighted equally with the status components.***
		2. **If the state has a computer-adaptive testing system for one or more subjects and a vertically-scaled score for consecutive grades, a value-added measure for both the general student population and subgroups.**
		3. **It may be appropriate to have a mix of i. and ii. depending on the assessments in a state.**

**Indicator(s) of progress toward English language proficiency**: Trimmed mean of scale scores from WIDA ACCESS for ELLs, for fourth and fifth grade English language learners who have been in the United States for at least three years. The scale scores will need to be transformed based on a state-level goal for accomplishments. (Trimmed mean to avoid problems with ceiling and floor effects.)

**Other indicator(s) of student success or school quality.** Most of these will be lagging indicators or require sufficient investment that they will be available only every two or three years, or will require several years to develop the needed infrastructure.

* + 1. CLASS ratings in sampled K-3 classrooms (one of Bob Pianta’s classroom observation instruments), or another classroom observation system for primary grades with equal or greater evidence of reliability and validity to judge the quality of classroom interactions, with sampling and observation windows designed to draw school-level (not class-level) inferences—see for example the use of CLASS in the Head Start Designation Renewal System. This may need to be sampled across multiple years or only applicable every two or three years. See advice to Department of Education at the end.
		2. A reliable and valid uniform survey of parents and, for fourth- and fifth-graders, of students. This would require the development and validation of surveys in multiple languages and some guidelines for response rates.
		3. Alumni completion of challenging courses in middle school (by the end of eighth grade), as defined by the state – this is certain to include Algebra I but may include other courses, or even non-curricular achievements if sufficiently well-defined (such as the International Baccalaureate Middle Years Programme assessment, or proficiency in a foreign language).
		4. Alumni participation in challenging extracurricular activities in middle school, as defined by the state – this could include individual or team achievement in various competitions such as robotics competitions, math leagues, juried music festivals, art competitions, etc.

**Calculating summative school ratings**.

* + 1. **Weights**: At first, the messy first group of measures should be around 80% of the score. English language proficiency 10%, and the mix of “other” indicators at 10%. Those weights can change with greater experience for English language proficiency assessments and the gathering of primary-grade classroom-observation, survey, and alumni data. See iii below.
		2. **How many summative calculated ratings:** Global, reading, math.
		3. **How to handle subgroups that are small in each school**, both in terms of subgroup performance in general and the requirement to include English language proficiency assessment as a component. Recommendation (see advice to Department of Education below): Use a moving average over several years to accumulate enough student numbers, with the latest year having a greater weight than earlier years. In contrast to Mike Petrilli’s suggestion to have *lower* *weights of unstable estimates*, I recommend *maintaining the same weight and adding stability with weighted, moving averages*.
		4. **How to handle changes in state assessment systems**, such as this past year’s disruptions to state testing: ESSA does not provide an explicit option. See recommendation below.

**Schools with low-performing subgroups:** This system would use the civil investigative role of a grand jury (e.g., as used in California and Georgia county juries, among other states) to identify both the schools with especially low-performing subgroups and also schools deserving commendation for addressing equity issues. The grand jury system will identify such schools every two or three years. For an average state this may be accommodated by a number of regional grand juries spanning several counties or single metropolitan counties (e.g., Philadelphia would be a single-county grand jury “region” in Pennsylvania). The state Department of Education shall provide all of the data used above to the grand jury, which will have subpoena authority to gather additional evidence.

This use of an investigative grand jury is the greatest conceptual departure of this proposal from current rating systems used by states. The greatest weakness of an algorithmic rating of schools is the sense of educators and communities that such calculated ratings omit critical context, especially around the judgment of schools for having low-performing vulnerable demographic subgroups. One remedy is to insert citizen judgment precisely around issues of educational equity, where a civil grand jury report will be a clear, official judgment by citizens, and where a grand jury has independent subpoena authority if its members feel that additional evidence is needed.

The grand jury will need to be given statutory guidance on the minimum and maximum proportion of schools that can be identified as having under-served vulnerable sub-groups or that have served them extraordinarily well. Grand juries will be empowered to draw broad findings and make recommendations for the region or state in addition to identifying low-performing and high-performing schools for vulnerable subgroups.

**School grades or ratings:**

1. **Labels. Administrators want to brag about high labels and avoid low labels—and there is little persuasive evidence that the form of the labels has mattered much beyond that human motivation**. So there needs to be three or more global labels. Beyond that (well, okay, nineteen levels are too many), the labels are unlikely to matter.
2. **Determining low-performing schools:**
	1. The global, reading, and math ratings should identify approximately five percent of elementary schools in the state as provisionally low-performing.
	2. The state board of education should have the opportunity to review the provisional listing and make uniform, rule-based adjustments that affect all schools in an equitable manner, so long as the roster of low-performing schools is not expanded or contracted more than 10%.
	3. In a year with civil grand-jury reports, the listing of schools with troubling inequity equity will be added to the public roster of low-performing schools. The state board of education shall not have the ability to remove the names of schools identified by the grand-jury process.

**Recommendations for the Department of Education**.

**Data over multiple years, lagged data. Several of the recommendations here use either multiple years of data or lagged data. There are important policy questions in the use of multiple years of data – but on the legal front it** is not clear from ESSA’s language whether states have flexibility in using multiple years of data for either technical or policy reasons. Making that option explicitly and transparently available to states would give states additional flexibility, especially in setting challenging goals for how elementary schools set up students for later success.

**Judgment calls**: Similarly, the recommendations here “push the envelope” by allowing two places for lay panels to make judgment calls. I am recommending that state boards of education should be allowed to tinker with accountability calculations in a *post hoc* fashion to take reality of year-to-year contexts into account, so long as the total number of schools identified as low performing is close to the original recommended number and as long as changes are applied uniformly to schools in the state. My guess is that this fits under ESSA, or at least the U.S. Department of Education would not choose to penalize states for this type of practical judgment by a state board of education.

What is less clear is whether the type of judgment proposed for schools with low-performing subgroups is legal under ESSA: would a state’s use of a grand jury system to make judgments be acceptable?

**A limited number of pause buttons**: a state should have the ability to hit a pause button for some parts of its rating system on occasion where the changes to assessment systems makes it very difficult for a range of reasons to assign labels to schools based on algorithms such as the recommendations here. I believe that this is not allowed under ESSA except where the U.S. Department of Education just ignores state action. One year out of 5-7 would probably be reasonable, as long as transparent reporting and some parts (such as the grand jury idea) continue during a pause year.