**Accountability under ESSA: A Model Design**

**By: Chad Aldeman**

**Design Objectives**

In designing new accountability systems under the Every Student Succeeds Act (ESSA), states should strive for three over-arching goals:

* Simplicity: Accountability is a tool to show parents how their child’s school is performing. As such, the average parent should be able to read and understand how the system works.
* Clarity: Accountability systems should provide clear signals about which schools need to improve and in what ways. As such, the information must be clearly linked to desired student outcomes.
* Fairness: Schools should be held accountable for what they control, and not just the type of students it enrolls.

The system outlined below is not the best accountability system money could buy. Instead, it’s designed as the simplest, clearest, and fairest system that any state could quickly and easily adopt. It would not require new data systems—all states already have the capacity to track and calculate all of the necessary elements—and it does not depend on a state adopting any particular assessments, signing new contracts, or spending any new money.

It starts with existing test score data—used in new ways that reward schools for improving student performance at all levels—as an initial “flag” on school performance. Those scores are used to prioritize subsequent actions, including high-quality, professional, and on-site inspections as a way to observe school quality and provide actionable feedback for school leaders on how to improve.

**Incorporating student achievement**

The system starts with a relatively simple Performance Index. Each school and district would receive a pre-determined number of points based on where students fall on the performance spectrum. Higher performance levels would be worth additional points, and all schools would have an incentive to help all students reach higher and higher levels of achievement. States should place disproportionate weight on the proficiency benchmark, but unlike NCLB, the sole weight of the system won’t be on proficiency. NCLB required all states to define at least three performance levels, but this system works best with more frequent, smaller gradations. Many states already sort students into five levels. In that case, the weighting would be:

0 points: Level 1

15 points: Level 2

35 points: Level 3

70 points: Level 4 (Proficient)

100 points: Level 5 (Advanced)

This sort of point system is simple, clear, and provides an incentive for all schools to care about moving all students along the achievement spectrum. What’s more, this sort of point system could be used at any grade level for any assessment.

A model system for elementary schools would weight all test scores equally. That is, a 3rd grade math score would be worth the same amount as a 4th grade English Language Arts score or a 5th grade science score. If states tested in other subjects, those could be folded in and given equal weight. In building this sort of performance index, states should use the average of three years of data in order to increase rating stability. That stability is particularly important for small schools and rural schools, because one-year results tend to be more susceptible to random fluctuations.

**Adding student growth**

Even after de-emphasizing the proficiency bar, accountability systems must still find a way to incorporate each student’s year-to-year growth. In general, low-income students tend to have lower achievement scores than their higher-income peers. School accountability systems should not merely reflect student demographics but should instead attempt to hold schools responsible for what the school can control. One way to address this is by looking at how much students grow over time.

There are lots of different ways to measure student growth. Some use complex statistical models to “control” for student characteristics and predict student scores in future years. By comparing a student’s actual versus predicted score, the student earns a growth score that can be attributed to his or her teacher or school. While these models have significant appeal, they also have significant downsides. They rely on complex regressions that aren’t easily understandable for parents or teachers, and they often require external vendors to run the numbers.

But there’s an easier ways to measure growth that’s simpler to explain to parents and teachers. It builds on the performance index outlined above, and it produces results that are quite similar even to more complex growth models.[[1]](#footnote-1) Called a “transition matrix,” it gives students points based on whether they advance through various performance thresholds. Unlike under NCLB, where districts focused on students right at the cusp of proficiency—the “bubble” kids—this sort of method creates several, more frequent cutpoints that make it harder to focus on just a small subset of students.

This approach offers several advantages over more complex growth models. Any state could implement a transition matrix without external support, and the calculations could be implemented on any state test. Most importantly, and in contrast to more complex models, the transition matrix provides a clear, pre-determined goal for all students. School leaders and teachers would know exactly where students are and where they need to be to receive growth points.

The Growth Index would look like this:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Year 2 | | | | |
|  |  | Level 1 | Level 2 | Level 3 | Level 4 (Proficient) | Level 5 (Advanced) |
| Year 1 | Level 1 | 0 | 100 | 100 | 100 | 100 |
| Level 2 | 0 | 0 | 100 | 100 | 100 |
| Level 3 | 0 | 0 | 0 | 100 | 100 |
| Level 4 (Proficient) | 0 | 0 | 0 | 0 | 100 |
| Level 5 (Advanced) | 0 | 0 | 0 | 0 | 100 |

States could adjust these weights, but the key concepts are that 1) more frequent cutpoints are better than fewer; and 2) schools should have an external incentive—beyond merely their own good intentions—to care about student growth for all students.[[2]](#footnote-2)

States may be tempted to focus accountability only on student growth. That would be a mistake. First, ESSA clearly requires states to measure and report proficiency rates. Second, growth-only systems require students to take a test that wouldn’t count for accountability purposes (states would have to test all 3rd graders but their 3rd-grade scores wouldn’t count). That also has the effect of lowering sample sizes. Third, growth-only systems do not include mobile students. Depending on the state and year, 5-9 percent of students can’t be matched even one year later, and the mobile students falling into that category tend to be significantly lower-performing.[[3]](#footnote-3) They would be lost in a growth-only system.

**Calculating grades and incorporating subgroup results**

Each elementary school would receive a weighted average of scores on the Performance and Growth Indices (0-100 scale). Next, the state would flag the bottom 5 percent of schools as “comprehensive support” schools. In addition, any school within one standard deviation of the average comprehensive support school would be initially flagged as a “targeted support” school. [[4]](#footnote-4) As a matter of fairness, this system would avoid drawing bright-line distinctions between schools ranked right above and below the 5 percent cut-off line. Those schools are essentially indistinguishable from a statistical perspective, so using standard deviations to identify the next group of schools is an attempt to treat similarly all schools with similar performance.

This method will pick up most “bad” schools, but as a check to ensure that no school ignores subgroups of students, states would repeat the method for each subgroup, as well as for performance on the state’s English Language Proficiency exam.[[5]](#footnote-5) Any school with any subgroup performing at or below the average “comprehensive support” school would also fall into the “targeted support” category.

To be clear, this approach would over-identify schools for improvement. That’s intentional, and in this system no school’s rating is final until it completes the formal inspection process (as outlined in the next section).

Researchers have been unable to pinpoint specific interventions to turn around low-performing schools, but there are positive effects behind the mere act of notifying schools in need of improvement that they faced the *potential* of sanctions. For example, Thomas Ahn and Jacob Vigdor analyzed the impact of NCLB’s accountability sanctions on school performance in North Carolina. They found that the “strongest association between failure to make AYP and subsequent test score performance occurs among those schools not yet exposed to any actual sanctions.” In this case, the failure to meet AYP and the threat of imminent sanctions was a catalyst for schools to improve. Similarly, a Texas study found that students benefitted when their school was at risk of being identified as "Low Performing" under the state’s accountability system. Benefits included short-term gains on test scores as well as higher college-going rates and early-career earnings. In other words, the *threat* of being identified as low-performing caused schools to change their practices in ways that improved long-term student outcomes.

**Indicators of student success and final school ratings**

Under this system, no school’s rating is final until they complete a formal inspection. The inspections would be based off the school inspectorate model used as part of the accountability and school improvement process in England. As Craig Jerald described in a 2012 report, “inspectors observe classroom lessons, analyze student work, speak with students and staff members, examine school records, and scrutinize the results of surveys administered to parents and students.” [[6]](#footnote-6) Although the interviews provide context, the main focus is on observations of classroom teaching, school leadership, and the school’s capacity to improve.

Test scores are used as a screen to identify schools in need of additional, immediate support, and after the initial screening process, all schools are assigned a timeline for a formal school inspection.[[7]](#footnote-7) All comprehensive support schools will receive an inspection in Year 1 (the 2018-19 school year), all targeted support schools will receive an inspection no later than Year 2, and the remainder of schools by the end of Year 3. This would ensure that all schools receive an inspection over a 3-year period, but the state prioritizes attention to comprehensive and targeted support schools.

Like those in England, the inspections would be conducted by professionals trained in rating, evaluating, and providing feedback to schools. No matter how it chose to staff them, states would pay for the inspections out of their 7 percent set-aside in Title I funds plus additional leftover administrative funds for Title II.[[8]](#footnote-8)

Other than comprehensive support schools, which by law must remain as comprehensive support schools for at least three years, all other schools would receive a final summative rating from their inspections (which factor in test scores and achievement gaps as part of the inspections review.) Those ratings would remain until the next inspection. As in England, the inspections should be periodically evaluated to ensure they provide a reasonable amount of differentiation. As of the most recent data, 18 percent of England’s primary schools were rated “Outstanding,” 67 percent were rated “Good,” 14 percent rated “Requires Improvement,” and 1 percent were rated “Inadequate.”[[9]](#footnote-9)

ESSA requires that all measures included in state accountability systems must be disaggregated by subgroup. That would look different under a qualitative system like inspections, but even the British model includes a separate indicator for “outcomes for individuals and groups of pupils.” Applied here, states would expand that indicator and disaggregate school quality as received by subgroups of students. If, for example, the inspectors observed that Hispanic students were disproportionately assigned to classes where low-quality instruction occurred, that would register in the inspection report and the schools would be required to address the issue.

Most importantly, high-quality inspections would serve as a tool both for accountability and improvement. Rather than the low-quality, self-completed improvement plans that were common under NCLB, the inspection reports would be executed by professionals trained to provide an honest review of school quality coupled with guidance on how to improve.

Accountability systems should not merely be an act of punishment. Instead, they should provide clear evidence to parents and the general public about what’s happening in schools, while also providing clear signals to school leaders and teachers about how to improve. The system outlined above would accomplish all of those goals.

1. An evaluation of the federal Growth Model Pilot Program found that a transition matrix approach identified slightly fewer students as meeting growth targets as other, more complicated growth models (“trajectory” and “projection” approaches), but all three models identified broadly similar groups of students. See Chapter IV: U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, Final Report on the Evaluation of the Growth Model Pilot Project, Washington, D.C., 2011. [↑](#footnote-ref-1)
2. High-achieving students may run into ceiling effects on tests, so to earn growth points, students scoring at the “Advanced” level must continue to reach that level in subsequent years. If states revised their testing systems to measure above- and below-grade-level performance—as they’re now allowed to under ESSA—they may want to revise the weighting system or create additional categories of performance. [↑](#footnote-ref-2)
3. See Exhibit 56 from the “Final Report on the Evaluation of the Growth Model Pilot Project,” available at: <http://www2.ed.gov/rschstat/eval/disadv/growth-model-pilot/gmpp-final.pdf>. [↑](#footnote-ref-3)
4. Based on an initial run of data from a medium-size Southern state, I expect this approach would identify roughly 30-40 percent of elementary schools. [↑](#footnote-ref-4)
5. Because not all schools will have a sufficient sample size on the state’s English Language Proficiency exam, it belongs as a check against the system rather than as its own component within an index system. [↑](#footnote-ref-5)
6. For more information, see: <http://educationpolicy.air.org/sites/default/files/publications/UKInspections-RELEASED.pdf>. [↑](#footnote-ref-6)
7. Although ESSA lets LEAs design their own interventions, the inspections would be considered to meet the ESSA requirement for “not less than one indicator of school quality.” Any interventions coming out of the inspections would fall under state authority. [↑](#footnote-ref-7)
8. As of Fiscal Year 2014, Title I funds amounted to just over $15 billion. Seven percent of that amount would be $1.05 billion, well above the low-cost estimate provided by Craig Jerald in “On Her Majesty’s School Inspection Service.” With state administrative funds from Title II, federal funds would surpass even Jerald’s high-cost estimate. [↑](#footnote-ref-8)
9. Source: Ofsted, The Annual Report of Her Majesty’s Chief Inspector of Education, Children’s Services and Skills 2014/15 (London: The Stationery Office Limited, 2015), Figure 25a. [↑](#footnote-ref-9)