ACCOUNTABILITY DESIGN PROPOSAL FOR ELEMENTARY SCHOOLS FOR THE THOMAS B. FORDHAM INSTITUTE ACCOUNTABILIY DESIGN COMPETITION

Ronald F. Ferguson, PhD

Harvard University and Tripod Education Partners, Inc.

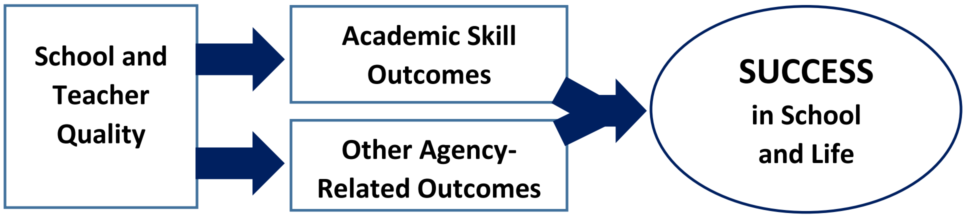
January 24, 2016

DESIGN OBJECTIVES

Schools prepare children for citizenship, economic productivity, parenthood, and self-realization. For each of these, foundations for success include basic academic skills in reading, math, and reasoning, on the one hand, and factors associated with personal agency, on the other hand (Figure 1). By personal agency, we mean the capacity and propensity to take purposeful initiative—the disposition *to actually do* the things that success in life requires. Agency-related factors include personal conduct, growth mindset (the understanding that effort can make one smarter), conscientiousness (the propensity to strive for quality work), and future orientation (anticipatory behaviors based upon recognizing that current actions shape future options). In addition, agency-related factors include social emotional skills required for managing feelings and behaviors in social contexts.

Priority Outcomes. The state accountability system that we propose for elementary schools aims for excellence with equity. It prioritizes the ***academic skills and knowledge*** that standardized tests measure, supplemented by factors associated with ***personal agency***. All measures distinguish discrete performance brackets in addition to whole-school composites. All are reported for major subgroups by race/ethnicity, disability status, prior achievement level, and (where appropriate) English Language learner (ELL) status.

Figure 1


School and Teacher Quality
Academic Skill Outcomes
Other Agency-Related Outcomes
SUCCESS in School and Life

Observational and Survey-Based Metrics. An accountability system should do more than simply measure and reward tested outcomes. Educators need tools and incentives to monitor and manage multiple processes for achieving intended results. Therefore, the state should require the use of ***valid and reliable observational and survey-based assessment tools***.[1] These can provide feedback from students to teachers, and from teachers to administrators, on school climate, teaching quality, and student engagement in learning, as well as the development of agency-related skills and mindsets. For

[1] For example: Ronald Ferguson with Charlotte Danielson (2014). "How Framework for Teaching and Tripod 7Cs Evidence Distinguish Key Components of Effective Teaching." In Thomas Kane, Kerri Kerr and Robert Pianta, eds, *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*. Hoboken, NJ: Jossey-Bass.

these observational and survey-based metrics, schools should NOT be graded on the measured scores. Instead, they should be **rated on the quality of their efforts to use such measures formatively for the improvement of teaching and learning**. Ratings should be provided by officials who supervise principals, contributing 10 percent of a school's composite accountability score. They fit within the ESSA category for *indicators of student success or school quality.*

KEY QUALITY FEATURES OF THE ACCOUNTABILITY SYSTEM

- **Multiple Criteria for High Stakes.** Because every metric is measured with random error and also because of non-random but nonetheless temporary aberrations in performance, high stakes decisions should be based on multiple measures, gathered multiple times, over multiple years.
- **Levels and Growth.** For most indicators, levels (for achievement status) and either growth or value-added measures (VAM) (as measures of learning) should be reported. This applies overall as well as for subgroups. Levels should include at least a four-way set of distinctions (e.g., below basic, basic, proficient, advanced). In addition, school-level mean and median scores should be reported using normal curve equivalent (NCE) or similarly appropriate metrics. Unlike percentiles, the latter do not exaggerate differences in the middle of the distribution.
- **Due Diligence.** Quality control should include accuracy checks in which different measures are used as cross-checks on others—unusually high or low performance on one metric (e.g., value-added test performance) not matched by similar performance on other metrics (e.g., expert observations or student survey assessments) should be cause for additional scrutiny by state or local authorities.
- **Teacher-Level Variation.** This proposal concerns school accountability. Still, because there is more between-teacher than between-school variation in teaching quality, within-school variation in reading and math gains should be reported in order to direct attention *from the district level* to teachers in need of major improvement. School-level means can hide the presence of low-performers in need of special coaching or other supports.
- **Interpretation of Between-Grade Comparisons.** It is sometimes the case that cut points between performance categories (e.g., below basic, basic, proficient, advanced) are not comparable from one grade to the next. Cut points should be calibrated in ways that make between-grade proficiency comparisons meaningful. Schools should be able to correctly interpret—not be misled by—a shift from one grade to the next in the percentage of their students in particular performance categories.
- **State-Wide, Not School-Level Benchmarks for Racial/Ethnic Achievement Gaps.** Each racial or ethnic achievement gap for accountability purposes should be the difference between a school-level subgroup score and an external benchmark. The most logical external benchmark is the statewide average for the state's highest performing group (typically whites) among the three largest racial or ethnic groups in a state's student population. Within-school or within-district racial comparisons *should not* be the focus for two major reasons. First, students from lower achieving groups need to be competitive with others beyond their classmates and neighbors. Second, within-school or within-district benchmarking makes improvement for the high-achieving benchmark group (typically whites) a setback for gap narrowing. Perceptions that gap narrowing requires lower ambitions for white students has undermined progress in a number of communities.
- **Adjusting for Student Characteristics.** In cases where student demographic backgrounds are significant predictors of accountability metrics, both adjusted and unadjusted values should be

reported. ***To be fair to school officials,*** accountability decisions should be based on values that adjust for student background characteristics. ***To be fair to students,*** aspirational goal setting should use values that are not adjusted. These two purposes—accountability and aspirations—should be clearly distinguished in order to avoid unnecessary debates about whether or not to make such adjustments.

- **Transparency versus Complexity.** The design should be simple enough that most people can understand it, but not so simple that it fails to make appropriate adjustments for student background and contextual features. Simple but misleading is not helpful.

SPECIFIC MEASURES

A. Weighting of Performance Categories

There are benefits to distinguishing student-level performance categories, but also to having a single school composite. Weighting enables policy makers to form school composites that reflect priorities across student-level performance categories. Here, I suggest a family of weighting schemes. Consider a four-way classification of performance—e.g., below basic (BB), basic (B), proficient (P), advanced (A).

Weightings can penalize poor performance and reward strong performance to roughly the same degree. For example, where BB, B, P, and A are percentages of students at the respective performance levels, the school-level score (SS) could be:

Option 1: $SS = 100 - 1.0*BB – 0.5*B + 0.5*P + 1.0*A$

Scores would range from zero (for 100% below basic) to 200 (for 100% advanced), while 25% in each category would produce a midpoint score of 100.

Or, to more strongly incentivize moving students out of BB:

Option 2: $SS = 100 - 2.0*BB – 0.5*B + 0.5*P + 1.0*A$

To more strongly encourage both equity and excellence, increase the weight on A as well:

Option 2: $SS = 100 - 2.0*BB – 0.5*B + 0.5*P + 2.0*A$

Here, payoffs would be especially high for moving students out of BB or into A.

The choice between these (or other) options is a policy decision.

I recommend applying selected formulas equally to reading and math for grades 3, 4, and 5. Results for 3rd and 4th grades should be used formatively. High stakes consequences should apply only to 5th grade, thereby rewarding the school's cumulative contribution. Each 5th grader's score should be weighted by the length of their attendance at the school.

B. Achievement Gaps

To measure achievement gaps, the weighting option selected above should be computed separately for each subgroup of a threshold size. Then, the achievement gap for a particular school, grade, subgroup, and subject is the difference between the school-level score and the relevant state benchmark. Subgroups for this purpose are the major racial and ethnic groups and students with disabilities (the state disability benchmark to be determined).

C. Progress Toward English Language Proficiency

Districts should select approaches to ELL instruction that include appropriate formative measures of English acquisition. However, accountability should not be based on the formative measures.

Accountability should be based on the percentage of ELLs scoring at least Basic on reading and math on state accountability exams within 2 years of arrival at the school, Proficient within 4 years, and Advanced within 6. Contributions to composite accountability scores should be scaled using the percentage of the upper-elementary population comprising ELLs who have attended at least two years.

D. **Growth or Value Added**

Growth or VAM scores should be computed for grades 4 and 5. For both grades, they should be calculated within prior-year performance levels (BB, B, etc.). Results within levels should be normed (e.g., z scores), then averaged at the school level. A school's composite should be a weighted average across the levels, treating each student equally. Alternatively, weights can be adjusted to prioritize, for example, low achievers.

E. (Non-Test) Indicators of Student Success and School Quality

This author is the creator of Tripod surveys and co-founder of Tripod Education Partners, Inc.[2] Tripod teacher survey indices for leadership and academic press (or reasonable alternatives) are recommended here. Both predict between-school VAM differences.[3]

From the student survey, the Gates Foundation Measures of Effective Teaching (MET) project showed academic press components strongly predict VAM,[4] while academic support components predict happiness in school and the inspiration to attend college.[5] Press and support both predict agency related factors such as growth mindset and conscientiousness.[6] In addition, current research by this author indicates unacceptably large disparities in access to orderly upper-elementary learning environments.

[2] See: www.tripoded.com

[3] Liu, Keke, et. al. (2014). "The utility of teacher and student surveys in principal evaluations: An empirical investigation." (REL 2015-047). Washington, DC: IES. http://ies.ed.gov/ncee/edlabs.

[4] Thomas Kane, Daniel McCaffrey, & Douglas Staiger (2010), "Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project." Bill & Melinda Gates Foundation. www.metproject.org

[5] Ferguson with Danielson (2014), op. cit.

[6] R. Ferguson et. al., (2015). *The Impact of Teaching* www.agi.harvard.edu/projects/TeachingandAgency.pdf

These teacher and student survey measures can be used formatively and their use can be rated summatively as described above in the third paragraph of this document.

SUMMATIVE SCHOOL GRADES

I have discussed measurements in the following categories.

1. Achievement levels (with only 5th grade counting for accountability)
2. Achievement growth or VAM (for 4th and 5th grades)
3. Achievement gaps by subgroup (*relative to state benchmarks*)
4. Progress of English language learners
5. Progress among students with disabilities
6. Metrics for school leadership and academic press, based on teacher surveys
7. Measures of teaching quality and agency-related factors, based on student surveys

Below, references to categories pertain to this list.

Categories 6 and 7 should be treated as described in the third paragraph of this proposal. The associated rating from the principal's supervisor should contribute 10 percent of the summative school grade.

Performance in categories 1 through 5 will be partly predicable by the demographic compositions of student bodies and community types. To be fair to school administrators, accountability metrics should statistically remove variation predicted by demographics and community type. Differences remaining can be standardized and weighted to form summative school grades. If policy makers prefer to compare schools only within specific types of locations—e.g., inner-city, suburban, rural—scores can be produced separately, by the type of location.

At least 50% of a school's grade should be based on categories 1 and 2—achievement levels and growth (or VAM). If exams and metrics limit growth measures for high achievers—and they may or may not—then weighting can favor growth for schools with low scores and levels for schools with high scores. Related to this, officials should strive to develop an informed public discourse that distinguishes continuous scores (e.g., NCE scores) from discrete performance categories from growth or VAM scores.

Up to 40% and not less than 20% of the summative grade should focus on narrowing racial and ethnic achievement gaps, making progress for ELLs, and for students with disabilities. Details will need to vary in potentially complex ways, depending upon percentages of students in each respective subgroup.