Issues to be Considered Before Adopting Local Assessments or a Menu of State Assessments In the A – F Accountability System

Joseph O’Reilly, Ph.D., Executive Director for Student Achievement Support, Mesa Public Schools
Maja Aleksic, Ph.D., Director of Assessment, Accountability & Research, Tempe Union High School District
Dr. Anna McCauley, Director of Research and Assessment, Higley Unified School District
Dr. John Wilson, Director of Research, Assessment & Evaluation, Tempe Elementary School District
Barbara Wright, Ed.D., Director of Curriculum & Instruction, Casa Grande Elementary School District
S. Jay Midyett, Ph.D., Director of Assessment & Accountability, Amphitheater Public Schools
Dr. Kyle Ross, Director of Assessment and instructional Technology, Scottsdale Unified.
Ildiko Laczko-Kerr, Ph.D., Vice-president Academics, Arizona Charter School Association
Steve Jeras, Director of Assessment, Paradise Valley Schools
Dr. Matt Strom, Assistant Superintendent for K-12 Educational Services, Chandler Unified School District
Wendy Davy, Director of Research & Planning, Peoria Unified School District
Ed Sloat, Ph.D., Director of Research & Accountability, Dysart Unified School District
Anju Kuriakose, Ph.D., Senior Research Analyst, Mesa Public Schools & former ADE Director of Accountability
Kathi Marston, Director of Assessment, Phoenix Elementary School District

The Arizona State Board of Education has issued a call for reports on “the feasibility of including a menu of local and/or statewide assessments in an A-F accountability system.” In response to this call, we, a group of experienced LEA-level assessment and curriculum specialists, would like to share our insights into what is needed to practically implement such a system successfully.

It is our assumption that other groups will submit reports detailing the scope of psychometric issues associated with local assessments. Our focus, instead, is on the practical issues of implementing local assessments or a menu of state assessments to be used in an accountability model, as this is what we know best. We have extensive experience implementing assessment and accountability systems for many years in schools, districts and the state.

Have Other States Used Local Assessments or a Menu of State Tests?

Many states use multiple assessments in their accountability systems (http://ecs.force.com/mbdata/mbquestRTL?Rep=AR03). Typical assessments used in such systems include ACT or SAT (21 states), Advanced Placement (11 states, more use participation only), end-of-course assessments (3 states) and ACT Workkeys (2 states). However, data derived from these assessments are applied to all LEAs rather than provided as a menu of choices from which an LEA can
choose. Moreover, these additional assessments do not replace the state assessment, but instead, are “in addition” to the state assessment.

Overall, no State Education Agency appears to allow for the inclusion of local assessments or selections from a menu of assessments in its statewide accountability system. However, there are two current state accountability pilot programs with a limited number of districts that have some of the features that a system with local or a menu of assessments might share.

The first is New Hampshire’s Performance Assessment for Competency Education (PACE) pilot, which had four districts involved last year and eight involved this year. The aim of the PACE program is to reduce the amount of time students spend being assessed by measuring students’ “competency” skills. According to the program’s website, this system “offers a reduced level of standardized testing together with locally developed common performance assessments. These assessments are designed to support deeper learning through competency education, and to be more integrated into students’ day-to-day work than current standardized tests.”

Formal assessment in the PACE program begins with students in grade three, who take the Smarter Balanced Assessments (SBAC). At other grade levels, students participate in PACE or in other local performance assessments or in SBAC (2 other grade levels). PACE requires the demonstration of high technical quality for all assessments, regional scoring, peer review audits, high inter-rater reliability across districts in scoring, a regional support network and extensive teacher training. The state also collaborates with CCSSO’s Innovation Lab Network to develop their system and state policies and practices (http://www.ccsso.org/What_We_Do/Innovation_Lab_Network.html).

Following the 2015-16 school year, the New Hampshire Department of Education will submit results of the PACE pilot for federal peer review, and if the result is undesirable, the New Hampshire Department of Education has agreed that all students will take the SBAC assessments in future years.

Those interested in developing a local or menu-based assessment system can peruse the United States Department of Education’s (USDOE) Assessment Peer Review process, located at: http://www2.ed.gov/admins/lead/account/peerreview/index.html. This document reflects the many issues that must be addressed in creating a high stakes assessment system.


The other state piloting a program investigating decreasing the impact of results on a state assessment is California, where a consortium of ten LEAs make up the CORE (California Office for Reforming Education) districts which have a waiver from USDOE to have a different accountability system (http://coredistricts.org). Sixty percent of their accountability system is based on academic measures such as the SBAC results. The remaining metrics include school climate and student non-academic skills. Although this pilot is not utilizing local assessments,
there are lessons to be learned from the model regarding the levels of support needed to develop measures, the validation process for inclusion in the accountability system, and the scaling model needed to incorporate multiple measures into one scale.

To date, the only example of a large-scale use of local assessments in a state accountability occurred in Nebraska during the early years of NCLB. Some in Nebraska felt it was a successful approach because it required extensive professional development for educators, and it built capacity in districts to develop and score assessments. However, the accountability model could not meet the requirements of peer review in areas such as showing comparability across tests, validity, evidence, etc., and it was replaced by the Nebraska State Assessment or NeSA. (http://www.edweek.org/ew/articles/2007/02/21/24nebraska.h26.html?tkn=YNZFE79gENrDBK12JKoqOvaej79sEoStmrXo&intc=es; http://www.edweek.org/ew/articles/2008/03/26/29nebraska.h27.html?tkn=ZMQFzujf11jrpqaSoccr4iX4cmDO8j%2IFG0w&intc=es; STARS Archive http://www.education.ne.gov/assessment/STARS_Archives.htm

Several lessons can be taken from the examples we have provided. First, no state education agency is including local assessments or a menu of state-approved assessments in lieu of the state assessment in its accountability system. Second, any new assessment approach will require significant new resources, effort and time from the state and LEAs if this to be done properly. Developing assessments on such a scale will necessitate additional staff with specialized skills in assessment item development, psychometrics, advanced statistical analyses, programmatic coding/scripting, etc. Finally, the design of the system requires great care to ensure that it does not systematically disadvantage or advantage particular schools.

**What criteria would multiple assessments have to meet?**

To be used for a high stakes purpose like school letter grades, any assessment would have to meet professional standards for assessment use. The following are some of the key issues from an LEA perspective that would have to be addressed with any additional assessment:

**Validity.** As a practical matter, test validity is the extent to which a test accurately measures what it claims to measure. Validity is specific to the type of use and inferences made from the data. While a locally developed test might be valid for assessing student mastery of content, it may not be valid for evaluating the quality of a school. For a locally designed test to be used in accountability, validity evidence for this new use of the test would need to be defined and collected.

A big threat to validity is when tests really reflect something other than achievement. We see this with current assessments where there is a high correlation with test results and the family income of the students attending the school (http://www.brookings.edu/blogs/brown-center-chalkboard/posts/2016/01/04-implementing-essa-accountable-more-than-a-test-hansen). Whatever (multiple) measures are used in the new system must genuinely align to "... the quality of the school." Perhaps greater use of change or growth indicators may be required.
However, calculating growth using multiple assessments across a variety of schools raises a number of very challenging issues.

**Alignment with State Standards.** Local or menu assessments would need to be aligned to state standards in content and cognitive demand. This would need to be documented through test blueprints, expert review, and other efforts.

The new ESSA legislation raises the possibility of having tests that are aligned with the College and Career Ready goal but not necessarily the state academic standards. The SBE may consider including the ACT or SAT tests in a menu of tests, as this is permitted under ESSA. This would raise interesting questions because the ACT and SAT are not aligned to Arizona’s standards but are consistent with a College and Career focus. Any decisions to choose non-aligned tests should be done explicitly and the rationale clearly stated.  
(http://www.edweek.org/ew/articles/2016/01/06/will-states-swap-standards-based-tests-for-sat.html).

**Reliability.** A reliable assessment would give you comparable results from time to time and across similar students. Demonstrating that the results of the test are consistent will need to be documented.

**Scoring.** Whatever assessment is chosen, the scoring must be accurate and applied uniformly. A school scoring their own tests knowing it would contribute to its letter grade would be very suspect. That is why in the state pilots mentioned earlier there are state or regional scoring centers with rigorous scorer training and monitoring of the scoring process. Scorer training, ongoing monitoring of scorers to ensure they are not drifting, and auditing scores through rescoring a sample of tests are all techniques we and the state have used in the past to ensure accuracy. Any new assessments need to meet this same standard.

**Security.** Once tests are high stakes we have a responsibility to ensure that the tests are secure and cheating has not occurred. The USED provided guidance on expected test security measures that reflects the expectations in this area (http://www2.ed.gov/policy/elsec/guid/secletter/110624.html). Security measures to ensure items and tests have not been compromised, such as erasure analysis and other techniques to identify anomalies, and other efforts will need to be demonstrated for any new assessments chosen.

**Comparability.** If different assessments are used by LEAs and schools, there must be some way to ensure that everyone is being held to a similar level of rigor. If not, there will be a race to the bottom to find the easiest test that leads to the highest letter grade.

Tests must also be comparable because there is a great deal of mobility across schools because of school choice and family relocations. How schools and students will be impacted by students changing schools during the school year to schools that have different local or menu chosen assessments needs to be addressed.
Comparability is calculated on large scale assessments by equating tests psychometrically. With almost 500 LEAs and over 1,800 schools possibly using different local assessments, equating them all may not be feasible. Perhaps an independent expert review could help accomplish this. However, comparability must be addressed.

**Equitability and Fairness.** In addition to being statistically comparable, the use of different assessments must result in an accountability system that is equitable and fair. The design of the system requires great care to ensure that the use of multiple assessments or a menu of assessment options does not systematically disadvantage or advantage particular schools so that the overall A-F rating is a result of the use of these additional assessments rather than a measure of true school quality.

The quality and number of available local assessments differs greatly among LEAs. To ensure that the accountability system is equitable and fair, this difference in the assessments available to include in an accountability model applied across districts must be taken into consideration.

**Testing Time.** This proposal appears to be in addition to existing state required tests such as AzMERIT, AIMS Science, AZELLA, and DIBELS. We already receive complaints about the amount of required testing. Adding additional tests just for the purpose of accountability runs against recent trends in assessment and will receive opposition from parents, teachers and others.

**Clear roles and responsibilities.** New assessment options will require new work by the state and LEAs. In most cases, this will require work that exceeds current capacity at ADE and the districts and charters. What work needs to be done, who needs to do it, and how that will be supported needs to be made clear and explicit. Support for additional ADE and LEA staff to undertake this work will be required.

**A Clear Definition of a “Quality School.”** What the Board means by a ‘quality school,’ if not each letter grade, needs to be operationally defined so that schools know what they must do to reach an “A” grade. It is also needed to allow us to determine if the measurements being used align with the Board’s expectations.

**Peer or Expert Independent Review.** There needs to be an independent review of any assessments developed in order to ensure there is not bias or problems in the assessments or their use. For example, one of our schools created pre- and post-tests to be used in an incentive system. Although the school felt the tests were comparable, an independent review revealed components of the test that made the post-test easier than the pre-test. An independent review can help avoid later problems with assessments.

In addition, it would be useful to have a group like ADE’s expert National Technical Advisory Committee (TAC) for assessments to comment on the structure of the new accountability system and its use of menu or local assessments.
**Going Forward**

We are not opposed to changes in the accountability system or to including new measures. In fact, we think that could be a positive step to broaden the number and type of measures used. However, all measures are not created equal, and simply having more and different assessments does not mean the resulting letter grade metric is somehow improved. We urge the State Board to make sure that any moves to add assessments be done with the effort and resources needed to make such a major new undertaking successful and to be able to withstand the scrutiny it will receive in Arizona and from throughout the country.

Arizona does not currently have the capacity at the state or LEA level to implement local tests or a menu of state assessments as part of an accountability system. If the state moves in that direction it must build the capacity to implement the system to the standards expected by the education and assessment profession (https://en.wikipedia.org/wiki/Standards_for_Educational_and_Psychological_Testing; http://www.apa.org/science/programs/testing/standards.aspx).

Arizona has set high standards and expectations for our schools and students. We cannot reverse course by implementing changes that will, intentionally or not, undermine what we have recently accomplished.