

Observations on the ESSA “Menu” Option for High School Statewide Assessment Programs

D. J. McRae, Ph.D.

04/20/18

When ESSA was approved by the feds 2+ years ago, it included statutory language allowing states to implement a “Menu” option for the design and execution of ESSA required statewide assessment programs for grades 3 thru 8 plus one grade level in high school. A “Menu” option suggests that multiple tests might be approved by state authorities for any given grade level, just like a restaurant menu has multiple options for dinner entrees, provided certain requirements are met for tests approved as menu choices. The requirements address both test development characteristics [content specifications as well as test administration features] and test score utilization characteristics [for example, ability to aggregate test results statewide (and disaggregate for statewide subgroups) following good educational measurement practices].

With several exceptions, not many states have seriously considered implementing the ESSA “Menu” option for their statewide assessment program. Those that have taken action to date have limited their considerations to optional use of college entry exams such as the SAT or ACT at a district (or school) level, replacing use of an existing statewide high school test focused on the individual state’s approved academic content standards.

A more general conceptual goal for a “Menu” option program would be to consider *individual student choice* of tests from an approved menu of tests. This consideration is attractive at the high school level simply because many of our high schools offer multiple instructional pathways for their students, both college-bound instructional pathways including specialty emphases [for example, targeted STEM courses] and vocational career-ready pathways including specialty emphases [for example, health-related career choices]. One underlying common-sense requirement for the design of any large-scale K-12 testing program is that test content should follow instruction; for our high schools, with multiple instructional pathways, **no one-size-fits-all test** will ever satisfy all instructional pathways offered to students, regardless whether a single test is designated at the state level or the district/school level. A “Menu” option testing program with choice at the *individual student* level makes a lot of common sense given the variety of instructional pathways offered by comprehensive high schools.

Test Development / Test Administration Observations

One obstacle for a true individual student “Menu” option is the current situation that authorizing legislation (at both state and federal levels) frequently includes the notion that one-size-fits-all high quality academic content standards are required to be the basis for testing all students. When one thinks about the variety of instructional pathways available to students at a comprehensive high school, any

one-size-fits-all set of target content standards will not fit all instructional pathways equally. For example, some states have *Career Technical Education* standards that can serve as the basis for tests for CTE students better than so-called high quality academic content standards for college-bound students. Changes to statutory language and/or regulations and/or guidance documents and/or assessment program peer review protocols may be needed for a true individual student “Menu” option for a high school statewide testing program.

Another obstacle for a true individual student “Menu” option is the practical problem of supplying even a modest quantity of different tests to any given group of students that need to be tested with one or more of those different tests on an “approved” test list. Here the considerable progress toward computer-administered tests in recent years provides real hope – think about any group of high school students taking a statewide test in a specified location within a specified time and add the logistics of different paper/pencil tests and test directions to the mix and one ends up with a nightmare of a test administration situation. But, with computer-administered tests, each student’s log-in ID can be programmed to deliver a specified test that is matched to that student’s chosen instructional pathway. With test directions delivered by computer, and with relatively equal anticipated test administration times for all tests on the menu, the logistics of computer-administered test administration allows for individualized testing for all students with multiple tests without the nightmare of paper/pencil logistics.

Test Score Utilization / Aggregate-Disaggregate Data Observations

Interpretations of scores for multiple approved tests on the menu will add to the complexity of individual student test score interpretation, but personnel in our high schools for each of the different instructional pathways are available to address individual student score interpretations for students and parents. The issue of statewide aggregation-disaggregation is more complex, but I believe it is addressable within the framework of acceptable good educational measurement practice.

There is a need to put scores from approved tests on the menu on a *common scale of measurement* to aggregate these scores to the statewide level, for potential use in a statewide accountability system. There are a number of different methods available to accomplish this technical task. A journal article by highly regarded educational measurement specialist Bob Linn about 25 years ago provides good guidance for what’s acceptable and what’s not acceptable along these lines [Linn, Robert L. (1993) Linking Results of Distinct Assessments. *Applied Measurement in Education*, 6(1), 83-102]. Linn outlined five different methods for *linking scores from distinct tests*, along with properties for each of the five linking methods. In particular, Linn discusses

- Equating methods: These methods require strict comparability of content on the different tests, strict alignment to a common set of target standards, strict conformity of test administration flexibilities such as accommodations and modifications, and strict mathematical rules for establishing scores on a common metric for the tests. These methods are needed for high stakes tests

for individual students, such as college entry tests, and have most notably been routinely used for multiple parallel forms of tests used for any given sitting of SAT and ACT test administrations.

- Equivalency methods: These methods do not require the same degree of strict comparability of content or alignment to a common set of content standards or use of flexibility rules for special groups of students, but do involve a requirement that all tests have good estimates for statewide distributions of scores for each test potentially on the menu. These methods are useful for aggregate data (and disaggregations) such as use for district or school or subgroup accountability use, as long as the specifications do not require individual student comparability of scores across tests. The widely used equi-percentile method for establishing equivalency of scores [the most frequently used method for K-12 national tests from the 1960's through the 1990's] falls in this category for potentially linking scores to a common scale of measurement for approved tests for the Menu option for statewide tests.
- Estimation methods: These methods involve using scores from one test to “estimate” scores on another test. This method of linking scores is widely used for *validity* research to establish the credibility of any given test's results, frequently using regression statistical methods to establish acceptable validity data. This method does not entirely account for the full distribution of scores for tests, rather focusing on averages and variances of scores. One of the disadvantages of this method, according to Linn, is that regression methods frequently do not provide stable estimates over time; new estimates are required from year-to-year to establish comparability of scores from year-to-year.
- Statistical Moderation Methods: These methods require human judgments of comparability across tests with the assistance of available statistical evidence. These methods have not been used widely in the US for comparability of K-12 test scores for data aggregations, but have been more widely used for European elementary-secondary testing programs. I do not think these methods would be sufficiently rigorous for use for aggregations for US statewide accountability system data.
- Social Moderation Methods: These methods are simply human judgments of comparability across tests without statistical evidence at all. Again, these methods are sometimes used in other countries, but have not widely been used in the US. Again, I do not think these methods would be sufficiently rigorous for use for US statewide accountability system use.

Given these five methods, I'd comment that the current federal assessment program peer review requirements for aggregation of scores tend to be based on the Equating methods for comparability of scores. My opinion is that aggregation (and disaggregation) of scores for statewide accountability use can be accomplished using less strict Equivalency methods, and at times by even less strict Estimation methods, for use of multiple tests for the ESSA Menu design option. I would argue that this is a case of a policy trade-off: The potential for use of multiple tests as an ESSA “Menu” option statewide testing design at the high school level is sufficiently attractive to tolerate a less rigorous method used to place the approved tests on the

menu on a common scale of measurement for statewide accountability aggregation (and disaggregation) purposes.

Conclusion

The above set of observations for a potential ESSA “Menu” option use of multiple tests for high school statewide testing needs considerable additional detail before it is ready for implementation. My guess is it will take 2-4 years of planning and preparation work. I would suggest one way to initiate such work would be via a formal Task Force appointed for 6-12 months to meet as a public body to discuss the additional details needed, with a recommendation via a written report to authorizing policy makers at least one year before implementation, well in advance of an operational implementation timeline. This would be a responsible and realistic way to consider the ramifications of an ESSA “Menu” option design for the high school component of a statewide testing program.

One of the considerations for formal Task Force discussion and recommendation would be whether or not to include tests that might be considered “sub-menu” options that would further refine whether approved tests might include specialty emphases within the broad categories of college-ready and vocationally-ready tests. Sub-menu tests [such as Advanced Placement tests, or International Baccalaureate program tests, or Cambridge end-of-course tests for the college-ready category; such as individual career tests for the vocational-ready area] may well be linked to common scale of measurement via one or more of the alternate linking methods described above. In addition, the existing high school content-based test statewide test should also be on the menu (perhaps as the default test should a student not select one of the alternative tests) and is the logical test to provide the underlying common scale of measurement to which other tests would be linked, thus providing continuity for any already established statewide accountability system.

Finally, another topic area for Task Force discussion and recommendation might how best to deal with special testing needs for Students with Disabilities. ESSA and states have already have recognized this area via special tests for severely challenged SWDs, but there is a need for additional considerations for the remaining SWDs [both IEP and 504 Plan students] in the context of a multiple test “Menu” option for high school statewide assessment programs.

Q & A for the ESSA “Menu” Option for High School Statewide Assessment Programs

Q: Can the Menu option also apply to grade 3-8 required ESSA tests?

A: While ESSA statutory language permits this option, in my view it does not need to be pursued. For the most part, grade K-8 curriculum and instruction programs are easily focused on one set of high quality academic content standards for each state, and thus existing one-size-fits-all tests based on those academic content standards do not need a “menu” of optional tests to fit different instructional pathways at the individual student level. [Add on 4/24: *An exception may be optional Math and*

Science end-of-course tests for middle school students taking accelerated high school Math and Science courses.]

Q: How would Menu tests be assigned to individual students?

A: I would advocate assignments would be left to student/parent choice, just like course selections for different instructional pathways are largely left to student/parent choice.

Q: Would individual students be allowed to take more than one test on the Menu?

A: The full answer to this question might be left for Task Force discussion and recommendation, but I would think an allowance for more than one test per student would be a reasonable feature for a high school Menu design. At least two issues would need to be discussed: (1) Who pays for the additional tests? (2) Which score would be used for accountability system use? There might be a variety of ways to resolve these two issues.

Q: How would tests be chosen and approved for the Menu?

A: I would think existing procurement systems could be used for RFI and RFQ processes, with test developers/owners required to provide data showing their tests can generate acceptable representative statewide distributions, data needed for linking to an established common scale of measurement. Developers/owners of prospective tests for the menu may have to fund subsidized test administrations to generate acceptable data for this requirement, but this situation has been common within the K-12 testing industry for many years.