MASSACHUSETTS DEPARTMENT OF
ELEMENTARY AND SECONDARY
EDUCATION

Addendum to Application for Innovative Assessments Demonstration Authority

Massachusetts

April 20, 2020

Table of Contents

Massachusetts's Response to USED and Peer Feedback

Item #1: Alignment with state academic content standards
Regulatory requirement (2)(i)-(ii)

Required information from the SEA

- Evidence that the innovative assessment will align with the challenging State science content standards under section 1111(b)(1) of the Act, including the depth and breadth of such standards, for the grade in which a student is enrolled, specifically:
    - A detailed draft test blueprint that describes how the pilot assessment will assess both the depth and breadth of the science standards.
    - A description of the procedures the State will use to independently evaluate the alignment of the pilot assessment (e.g., an alignment evaluation conducted by persons or entities not involved with the innovative pilot).

Massachusetts Department of Elementary and Secondary Education (DESE) Response

Since the submission of our initial application, DESE has further developed the approach to ensuring sufficient coverage and alignment relative to the breadth and depth of the state's challenging academic standards. This includes more detailed analysis of the tradeoffs for the blueprint for the abbreviated summative, and exploration of possible approaches to blueprint design and standards selection for the innovative performance tasks.

Additionally, DESE has established plans to conduct an alignment study on the innovative assessment (both portions). As part of the existing MCAS technical review process, an in-depth alignment study is conducted on the tests in a specific content area. The review of the MCAS for Science, Technology and Engineering (STE) for grades 5, 8, and high school tests had been planned for this year's test (2020) but due to the cancellation of MCAS in response to the COVID-19 pandemic, this review is being postponed to 2021. This will allow DESE to include the innovative assessment in the same formal alignment study. All items on the abbreviated summative will already be analyzed for alignment to standards and depth of thinking, because they are all included in the statewide MCAS STE being studied. For the new performance tasks, the study will leverage a similar process to analyze alignment to standards and depth of thinking. This alignment study will be conducted by a third-party who is NOT the vendor that created the performance tasks.

Abbreviated Summative Blueprint

In February, DESE staff met with representatives from our existing assessment vendor to explore options for the blueprint of the abbreviated summative portion of the innovative assessment. It was tentatively agreed that it would be achievable to create a blueprint for an abbreviated summative that would maintain comparability, reliability, and validity across reporting categories, keeping the relative proportion of questions in each reporting category in line with the blueprint for the full MCAS STE.

The item types and blueprint for the existing MCAS STE are shown below:

2018 Legacy MCAS: Question Types and Point Values (STE Grades 5 and 8)

Item Type	Number of Items	Number of Points
Selected Response, 1 pt.	26	26
Selected Response, 2 pts.	3	6
Constructed Response, 2 pts	2	4
Constructed Response, 3 pts.	2	6
Module (6 pts = 2 SR-1pt each + 1 CR-3pts)	2	12
Total	41	54

2018 Legacy MCAS: Proportional STE Distribution by Reporting Category and Grade

Reporting Category	***% for Grade 5***	***% for Grade 8***
Earth and Space Science	30	25
Life Science	30	25
Physical Sciences	25	25
Technology/Engineering	15	25
Total	**100**	**100**

For the abbreviated summative portion, the blueprint must balance two goals: shortening the time required for the test, while maintaining enough points aligned to standards in each reporting category. The abbreviated summative will not contain any field test items (as does the MCAS STE), thereby immediately reducing the test length with no reduction in operational questions. This is estimated to be a reduction of 7 items, without any loss of score points.

To determine the minimum number of score points required to maintain integrity of results, DESE is working with the existing MCAS vendor to run simulations on various blueprint lengths using last year's actual student results. For example, for a hypothetical blueprint with 38 points (shown below), the simulation will use a subset of the actual MCAS results from the previous year to create simulated scores for students using that blueprint, and then determine the comparability with the actual reported results using the full test. By testing a number of possible blueprint designs, DESE can weigh the tradeoffs of test length and reliability, and select the shortest blueprint that achieves an acceptable level of reliability. DESE expects the ultimate blueprint will feature a greater proportion of selected response items, to allow for the best possible coverage of the breadth of standards.

Abbreviated Summative: Example Question Types and Point Values (STE Grades 5 and 8)

Item Type	Number of Items	Number of Points
Selected Response, 1 pt.	21	21
Selected Response, 2 pts.	2	4
Constructed Response, 2 pts	2	4
Constructed Response, 3 pts.	1	3
Module (6 pts = 2 SR-1pt each + 1 CR-3pts)	1	6
Total	41	38

Abbreviated Summative: Example Proportional STE Distribution by Reporting Category and Grade

Reporting Category	*% for Grade 5*	*% for Grade 8*
Earth and Space Science	30	25
Life Science	30	25
Physical Sciences	25	25
Technology/Engineering	15	25
Total	**100**	**100**

Innovative Performance Tasks Blueprint

DESE is considering two possible approaches to the blueprint design:

Approach 1: Similar Standards to Abbreviated Summative. In this approach, the set of technology enhanced performance tasks is meant to be reflective of the full MCAS blueprint, when pooled across students. That is, the set of tasks administered in a grade should reflect the depth and breadth of the general blueprint, but in a scaled down fashion in initial years when the performance tasks constitute fewer points than the complete MCAS. Doing so will allow for an examination of the performance task section as a whole, after being scaled concurrently with the mini-summative. This means that we can create a good approximation of a performance-based-only assessment by pooling across the matrix sampled performance items.

Approach 2: Complementary Standards to the Abbreviated Summative. Alternatively, the technology enhanced performance tasks are meant to replace content omitted on the abbreviated MCAS. That is, DESE content experts will identify areas, like combinations of standards, within the current MCAS blueprint that can be replaced with the technology enhanced performance task, and these areas will be measured by the performance tasks.

- A variation on the above would be to intentionally target the performance tasks towards standards that are well suited for performance tasks, due to, for example, their depth. Content experts could select groups of standards that have proven to be difficult to assess on the MCAS.

DESE currently believes that Approach 1 will be the preferred option. This will allow for better direct comparability and allows us to better determine the viability of the performance tasks as a stand-alone test. The ultimate approach to selecting standards for the performance tasks will be decided in partnership with the vendor. We expect that each performance task will encompass multiple standards, as each task will likely contain multiple scored items within it. For some task designs, they may be suited to assessing clusters of related standards within a single reporting category; for other task designs, they may assess standards from multiple reporting categories.

Regardless of the approach ultimately used for the performance tasks, the abbreviated summative blueprint will be designed to ensure that we have sufficient coverage of the breadth and depth of standards to produce reliable scores for students. Over time, we will analyze the results from the performance tasks and adjust as needed to demonstrate that the performance tasks can accomplish the same.

Item #2: Valid, reliable and comparable results for each sub-group
Regulatory requirement (4)(i)-(ii)

Required information from the SEA

- Evidence that the innovative assessment will generate results that are valid, reliable and comparable for each sub-group of students, specifically by describing how the matrix sampling design described in the application will result in comparable sub-group level results.

Massachusetts DESE Response

DESE has a well-established track record of reporting valid and reliable MCAS results for all sub-groups. For official reporting of innovative assessment results, DESE plans to leverage the same processes and approaches to ensure valid and reliable reporting for all sub-groups. In the first year, student scores will be generated entirely from the abbreviated summative portion of the test. As discussed in the response to feedback Item #1, the abbreviated summative will consist entirely of common items with no matrix sampled items. This ensures that the blueprint provides enough information to produce individual students scores that are valid and reliable measurements of student mastery of the state's challenging academic standards.

To establish comparable, reliable and valid results from the performance tasks, in the initial years DESE intends to provide as much information on student performance on the performance tasks as is feasible, in a low-stakes, low-inference manner. Part of the rapid prototyping cycles involves determining what kinds of information can be provided based on the designs of the tasks and how that information can and should be used. DESE will carefully examine what types of information students, parents, teachers, and other stakeholders value and then design student score reports to provide that type of information (and vice versa, ensuring that what can be provided is grounded in what students are actually doing).

The plan for possible matrix sampling of performance tasks will be determined in collaboration with the selected vendor and will depend on the design and time requirements of tasks and their cost to develop. In the RFP, we have required that vendors propose to produce at least 3 performance tasks each for grades 5 and 8. Depending on the design and time requirements for the proposed tasks, we may use all three tasks as common tasks for all students, or we may have one common task and one sampled task. We will evaluate the approach to matrix sampling and the impact on producing valid, reliable results by sub-group in the annual analysis. Because students' official scores will be entirely based on the abbreviated summative in the first year, DESE will have time to establish that the approach to sampling and scoring performance tasks produces valid, reliable results for sub-groups before incorporating them into student scores, thereby ensuring that this requirement is met.

Item #3: Evaluation by an independent, experienced third party
Application selection criteria (e)(1)

Required information from the SEA

- Evidence of criteria used in assessing progress of the pilot assessment throughout the project timeline, specifically detailed descriptions regarding the evaluation of reliability, validity, and comparability that will be conducted on the assessment in its final form.
- Evidence the procedures the State will use to independently evaluate the alignment of the innovative assessment (as noted in regulatory (b)(2) above).

Massachusetts DESE Response

DESE recognizes the importance of high-quality evaluation of this major effort, and plans a multi-faceted approach. Perhaps most important, DESE plans to conduct a thorough, third-party independent evaluation of the final form of the innovative assessment near the end of the IADA period to help inform the decision about whether the innovative assessment is fully ready for statewide use in place of MCAS STE, or whether it makes sense to apply for the 2-year extension. Further, after the 2021 MCAS administration, DESE plans to include the innovative assessment in the planned independent study of alignment of MCAS STE. This is a well-established process for a formal review of the alignment of MCAS test content to standards. DESE will use the same procedure to evaluate the alignment of the innovative assessment by including it in the study to be conducted of 2021 MCAS STE.

In addition to these formal, third-party evaluations, DESE will also work the performance task vendor to conduct annual evaluations of both implementation and the quality of assessment results and findings. While the precise criteria for these evaluations will be established collaboratively, DESE has identified a number of key criteria, below.

For the mini-summative, DESE and its vendor will produce evidence of reliability and validity to mirror the evidence published annually in the MCAS technical reports. For an example, see the 2018 Next-Generation MCAS and MCAS-Alt technical reports. Note that the specifics of these approaches may change as the work evolves and additional insight is provided by key partners, including DESE's technical advisory committee. The current planned analysis includes the below analyses.

- Examination of **reliability** includes both classical and item response theory statistics dealing with measurement precision. The key classical test statistics are (a) Cronbach's alpha, which is calculated overall and for each subgroup with 10 or more students and (b) decision accuracy and consistency calculated following Livingston and Lewis (1995), which is calculated overall and conditional on each achievement level. The key item response theory statistic is the conditional standard error of measurement (CSEM) along the entire scale and in particular at each cut point.
- Examination of **validity** will include evidence of validity from the key sources outlined by the APA, NCME, and AERA Standards for Educational and Psychological Testing (2014), which are also drawn on by the general MCAS. These sources include test content, response processes, internal structure, and relationships to other variables. Since these sources of validity evidence will be collected for the MCAS, examinations of the abbreviated summative will leverage the same data and techniques to essentially reexamine the reduced item pool defining the mini-summative.
- Examinations of **comparability** will explore whether students taking the mini-summative would have received similar scores if they had taken the full summative. Since the mini-

summative is simply a subset of the general MCAS, this question can be answered largely by taking the item responses from the students who took the full MCAS and reducing it to just the items on the mini-summative, then re-estimating scaled scores and achievement levels. Once produced, these mini-summative scores can be compared against those from the general MCAS. While DESE plans to report only achievement levels (and commits to establishing comparability of achievement levels, per the requirements), DESE will also examine the comparability of the scaled scores internally and use this analysis to inform future work.

For the performance task section of the assessment, the focus of work on reliability, validity, and comparability will be aimed at supporting the possible replacement of the general MCAS over a number of years. Again, the performance tasks will *not* be used operationally in the first year, but instead used to support the gradual development of a rich set of performance tasks to meet the department's long-term vision.

- In terms of **reliability**, this work is aimed at exploring what is possible within the context of large-scale assessment, thus the specific measurement model is not known *a priori*. Assuming that the performance tasks result in item response data that is amenable to the classical and item response theory approaches generally taken in large scale assessment, then the vendor will be asked to produce information on measurement precision at the task level. Doing so will provide the department with an idea of how many tasks a student needs to take to produce sufficiently precise scores. Should such models not be appropriate, much of the work on precision will be to determine what statistics appropriately address precision, and then estimating them.
- In terms of **validity**, the work this year will dive deeply into test content and response processes through careful test development and rapid prototyping using a principled approach to assessment design. Drawing on Evidence Centered Design, doing so means deeply articulating the student, evidence and task models that define what is to be measured, what students need to do to provide evidence of that target, and what the features are of tasks that would elicit that evidence. These models will be articulated for each task, and revisited and revised after each round of the rapid prototyping. In addition, cognitive laboratories will be used to understand student response processes and then support task revision. Log files collected from the testing platform may also provide insight into student responses, assuming the vendor is able to create *a priori* structures within the testing platform. Evidence related to internal structure can be investigated within and across tasks using typical methods, e.g. the MCAS currently employs the nonparametric IRT-based methods DIMTEST (Stout, 1987; Stout, Froelich, & Gao, 2001) and DETECT (Zhang & Stout, 1999). Internal structure, however, may also be modeled based on hypotheses about student task interactions based on the student, evidence, and tasks models. Finally, since students in the pilot will be taking both the performance tasks and the mini-summative, a source of external validity evidence is the mini-summative itself, with associations between performance on the mini-summative and performance tasks acting as evidence.

- In terms of **comparability**, the approach for the performance tasks is essentially the same as what has been previously mentioned for validity evidence based on external variables – to compare associations performance on the mini-summative and performance tasks for students participating in the pilot.

Finally, the DESE team for innovative assessments is working closely with the Kaleidoscope team to establish an approach to evaluating the effect on teacher practice. Ultimately, the goal of introducing this innovative assessment is to ensure that more students receive higher quality instruction. Kaleidoscope is currently developing rubrics to assess the presence of deeper learning practices at the classroom and school levels, in addition to a rubric for assessing principal leadership. Kaleidoscope implementation managers will use the rubrics to gather data on each participating Kaleidoscope school during a baseline period and a later period to determine the degree of progress.

Because of the partially overlapping groups for Kaleidoscope and innovative assessments, DESE can use the results to explore questions such as: Do Kaleidoscope schools who used the innovative assessment see a larger change in practice than those that don't? Do Kaleidoscope schools see larger changes in practice in science (with the innovative assessment) than in other subject areas? Do non-Kaleidoscope schools who used the innovative assessment see comparable changes in practice versus Kaleidoscope schools? Seeing the data on these questions will help us target our efforts in future years, and help ensure that the joint efforts of Kaleidoscope and innovative assessment are contributing to a better education for all, especially for students most in need – our ultimate goal.