Arkansas Adaptive Assessment System

NWEA
Northwest Evaluation Association
Partnering to help all kids learn®

The NWEA Blended Assessment Program

The Northwest Evaluation Association™ (NWEA™) is pleased to support Arkansas policymakers' vision for the future of their statewide assessments. This vision demonstrates an innovative approach that sets the State apart from all others. While other states have elected to adopt new common standards and assessment programs, Arkansas seeks more innovative ways to assess its students, placing a strong focus on informing and improving instructional practices through the introduction of a comprehensive assessment system.

By strategically blending interim and summative assessments, the State introduces an approach that reduces the amount of testing time and achieves something that summative assessments have traditionally been unable to provide – the provision of data that is relevant and actionable in the classroom. With this solution, the state simultaneously achieves the goals of maintaining high standards, evaluating proficiency relative to on-grade-level expectations, and measuring proficiency and growth for school accountability and educator effectiveness, while still providing instructional feedback to teachers on student learning in an innovative, less burdensome way.

NWEA believes we can support the State's vision by offering a novel solution that provides formative feedback to teachers about student performance and progress, provides student growth information for school accountability and educator evaluation, fulfills state and federal accountability requirements, and offers national comparison data.

We currently support a number of LEAs across Arkansas with our computer-adaptive interim assessments, the Measures of Academic Progress® (MAP®). The State, in turn, has heard these LEAs touting the benefits of the information they receive from MAP test data. They value the information about student learning, achievement, and instructional strengths and weaknesses, and appreciate the sound growth norms we provide. These Arkansas LEAs emphatically express their desire to continue to use MAP tests as part of their strategic improvement plans.

We believe that we can meet the State's needs by augmenting the MAP test Arkansas LEAs are already using as an interim assessment and growth measure to deliver proficiency information about student performance relative to on-grade-level content standards. This model has previously been implemented in both Utah and New Mexico, as well as in partnership with the Bureau of Indian Education.

Our solution effectively and efficiently combines several assessment components to leverage the power of a computer adaptive assessment to provide performance data that is measured in achievement levels, growth data, and instructionally relevant guidance for educators. This is done with alignment to state standards, including performance tasks (constructed-response items), and adherence to a stable, norm-referenced scale. The strength of our solution for Arkansas stems from the design of our test, which will be a truly adaptive assessment that is not merely a patchwork of fixed forms. The MAP test is an item-

level adaptive assessment that meets students where they are. Structured this way, MAP assessments produce greater and more precise data about student achievement and growth than fixed form interim tests.

Our proposed assessment program has three components:

- Component One: Interim MAP assessments, administered in the fall and optionally in the winter
- Component Two: Performance task (constructed-response) assessment, administered in late winter
- Component Three: A two-phase (seamless for students) Summative Blended MAP

assessment, administered in the spring.

The following diagram presents our proposed design for the new Arkansas summative tests.

Figure 1: An illustration of the three components of the NWEA assessment program

The advantage of this model to Arkansas is that the State can provide interim assessments multiple times throughout the year that yield feedback about student achievement, which

allow educators to develop a customized learning path for each individual student. MAP data will help Arkansas educators provide truly differentiated instruction.

Further, the spring Blended MAP test will permit the State to meet federal accountability requirements through a single assessment administration that combines an adaptive on-grade-level assessment to identify proficiency relative to grade-level content standards (as required by federal law), with a traditional adaptive interim assessment to identify the student's achievement and growth for end-of-year performance. Because MAP assessments are item-level adaptive tests, MAP tests provide grade-level proficiency and meet federal requirements more efficiently with a shorter test administration than is traditionally used, thereby reducing the burden to students and schools.

Assessment System

A major change to a state's assessment program always raises questions and concerns among educators, students, and parents. The new assessments in Arkansas will introduce multiple changes, including new item types and ways of administering the tests. To help ease the transition, NWEA provides a reliable source of student achievement and growth information to support instruction. NWEA proposes the assessments be built on the foundation of our MAP tests which report student performance on established and stable Rasch UnIT (RIT) scales using high quality items that have been validated and calibrated for use with millions of students. This streamlines the test development process and eliminates the need – and risk – associated with creating a new scale and recalibrating large pools of items to that scale. Our item pools are strong now, and they will grow stronger as NWEA continues our aggressive item development efforts to measure new, changing, and more rigorous state standards. By using our established scale and our robust and technically sound item banks, Arkansas educators will be provided with stability, certainty, and strong data from the very beginning of a new assessment system.

The RIT Scales

NWEA uses the Rasch Item Response Theory (IRT) model to create the vertical scales, called RIT scales. MAP test results are reported as RIT scores. Using RIT scales to report test results makes it possible to follow a student's educational growth over time.

The strength of our proposed solution for Arkansas stems from the design of our test. Rather than a patchwork of short fixed forms to create a staged adaptive test, NWEA proposes an *item*-adaptive test that adjusts from item to item as the student works through the test. The MAP test is structured this way and produces greater and more precise data about student achievement and growth than fixed form interim tests.

As the State has transitioned to a curriculum which supports its most recently adopted Standards, a challenge facing all stakeholders, including vendors, is the ability to measure those standards that have either not been traditionally taught or, due to their cognitive complexity, have not been measured in a large-scale assessment. One reason this is challenging is that assessment of these standards and skills requires more complex items,

more time to administer, and (in many instances) hand scoring, which makes them difficult to incorporate in an adaptive test.

To improve the efficiency of testing, NWEA proposes that performance tasks (constructed-response items), which can more effectively assess the most complex standards and cognitive levels, be administered separately from the adaptive portion of the State's new tests. The tasks can still be provided in a computer-based administration, and the selection of the performance task can be informed by the student's scores on a computer adaptive test. The performance task assessment will not be scored by teachers. NWEA proposes working with an external scoring partner and Arkansas to determine the scoring model most appropriate to meet the needs of the State.

NWEA proposes a program consisting of three main components:

1. An interim assessment administered in the fall that is fully adaptive and identifies each student's achievement level on a learning continuum, independent of the grade to which the student is assigned. The assessment adapts both above and below the student's current grade level, thereby meeting the student where he or she is. The resulting data informs the teacher regarding the instructional needs for each student and, with resources provided with MAP assessments, allows the teacher to implement appropriate instructional interventions through grouping, re-teaching, remediating, etc. based on individual, small group, and whole class instructional needs in specific areas. Individual student results are delivered instantaneously and class level analysis for the teacher within twenty-four hours. Results also provide information that speaks to whether students are projected to be proficient in on-grade-level content at the end of the school year. The fall administration provides teachers with feedback that permits them to develop a customized learning path for each student. Optionally, the State can administer the interim assessment in the winter at no added cost. The winter administration offers a mid-year check point to allow teachers to evaluate whether students are making progress, to determine whether instructional interventions have been effective, and to make adjustments to each student's learning plan to facilitate ongoing improvement and growth.
2. A late winter assessment using performance tasks (constructed-response items) to measure and classify students' performance on tasks with higher cognitive complexity and depth of knowledge (DOK). This test specifically measures on-grade-level standards and skills and serves both formative and summative purposes. That is, the scoring rubric provides teachers with information useful for instructional planning and the assessment also serves as a portion of the accountability assessment. This test is scored with final results arriving several weeks after the testing has occurred.
3. A spring assessment that is a combination of on-grade-level assessment for the purposes of accountability, along with an adaptive portion that extends above and below grade level to continue to provide information about each student's progress and continued instructional needs. This assessment, described as a Blended MAP test, consists of an initial phase that is adaptive within grade level, providing a proficiency classification for each student, and a second phase that is similar to the fall and winter administrations and adapts outside of grade level for those students who are performing above and below grade level curriculum. Scores and analysis are reported

quickly, as with the fall test, to support instructional planning. Further, utilizing our vertical scale, this provides a measure of academic growth from fall to spring.

This assessment program is intended to gather proficiency data, used primarily as accountability measures, in addition to providing teachers with data and tools for instructional planning. While other summative assessments provide basic data, they miss instructionally relevant opportunities. With the adaptive assessment we propose, we can pinpoint each student's skill level regardless of grade, can identify areas of relative weakness and strength, and can provide an accurate measure of proficiency, as required by federal regulations. All this can be accomplished with greater flexibility and accuracy. Additionally, through the use of an established vertical scale, we can simultaneously provide important information about student growth, which can be used as one component of evaluating educator effectiveness.

During component one of the program, our interim MAP assessments will be administered. MAP assessments are fully adaptive and identify each student's achievement level on a learning continuum, independent of the grade to which the student is assigned.

Our proposed test design for the summative portion program provides proficiency scores and precise evaluation of performance relative to on-grade-level standards. When combined with the second phase of the Blended MAP assessment (seamless for students) which can move outside of grade level for students who are performing significantly above or below grade level, the test provides a highly precise measure of individual student achievement and growth, making it easy for educators, parents, and stakeholders to see what students know and can do, but also whether the individual student is making appropriate progress.

Unlike traditional state growth models, which establish growth parameters based on comparison to the total student population, Blended MAP assessments track individual student growth along a stable vertical scale, tied to standards rather than driven by comparisons to other students. That said, MAP test data can be used in growth models and can enhance their accuracy, because our tests measure with greater precision and less error. We propose traditional interim MAP assessments (component one) be given in the fall (plus an optional winter administration), a performance task assessment (component two) in the winter, and the late spring test would be the Blended MAP assessment (component three).

The on-grade-level summative assessment for accountability is split into two components and is administered in two parts. First, the performance task test (component two) is administered as a standalone assessment. The performance task assessment is delivered electronically and assesses higher DOK levels than does the Computer Adaptive Test (CAT). These tasks allow stakeholders to understand how students apply their skills and knowledge in a new way. Results are reported on a rubric scale that is factored into the overall, summative assessment score.

Second is the administration of the Blended CAT, which is divided into two phases: one with items that assess on-grade-level content standards and a second with items that adapt above and below grade level. To the student, this division of content is not noticeable. Our

intent is to provide breadth of evaluation within the student's grade-level content, as it relates to the Arkansas Core Standards for that grade, as well as depth of understanding about where a student may be advancing or deficient outside of his or her grade-level standards. With this design, the CAT provides a deeper range of insights into each student's knowledge and skills. To ensure accuracy and validity of the scores, the two parts of the summative test event should be administered as closely together as possible. NWEA will work with the State to establish testing windows that work for the state, districts, teachers, and students.

The summative tests produces proficiency scores, which are established through structured standard setting and which utilize a combination of results from the winter performance task assessment and the spring Blended MAP assessment administration. In addition, because the items used for the Blended MAP assessment are calibrated on our vertical RIT scale, the tests also provide growth data from fall to spring. Our proposed total testing time for the spring Blended MAP assessment varies by content area and ranges from approximately 90-115 minutes for the English Language Arts assessment to 60-75 minutes for the mathematics test. Fall (and optional winter) interim assessments are shorter.

Use of our RIT scale permits a wide variety of scoring and reporting options for Arkansas. It is possible for our psychometric teams to use this to create a linking study to the current Criterion Referenced Test (CRT) so that longitudinal data are not lost. The scale also offers the ability to provide a national percentile rank. The RIT scale is applied to the summative and the interim assessments, so it inherently provides the interim assessments with the ability to predict performance on the summative assessment.

Arkansas' assessment system is important not only for educators and students in the State, but also as a model for other states. Arkansas is thinking ahead about how to make the nation's changing educational landscape best benefit the State's students. NWEA is dedicated to that vision and will support the State's leaders with our research, program, technology, and content resources and expertise. Together, we believe we can make this vision a reality that improves student learning across the State.

The NWEA Item Development Process

NWEA content experts continuously develop, field test, and operationalize new machine-scorable items for our assessments in order to provide depth of coverage of a set of standards. As part of this process, our team carefully constructs both the content of the item and the item type to provide the most accurate measurement of each student's knowledge and abilities as it relates to the standard attached to the item.

The item development process begins with NWEA Content Specialists creating item specifications, which derive from an unpacking of the standards. The specifications provide guidance to item writers regarding the content, context, cognitive complexity, item format, item asset (such as passage, graph, or diagram), and any corollary skills or understandings needed to assess the topic or skill. NWEA contracts with highly qualified individuals to write items, who follow the NWEA Item Writing Content Guide for their specific subject area.

Item Specification Creation

The item development process begins with NWEA Content Specialists creating item specifications after a thorough review of both the State's standards and all supporting documents. Item specifications are derived from an unpacking of the standards for the fullest understanding of the intention, scope, and focus of the instruction. From this, they provide specific guidance to item writers regarding the content, context, cognitive complexity, item format, item asset (such as passage, graph, diagram), and any corollary skills or understandings needed to assess the topic or skill. NWEA Content Specialists write item specifications to both the most and least granular levels in the standards.

Basic requirements that guide the development of item specifications are intended to ensure that items follow the best practices for item development and that items align strongly to the standards.

NWEA item specifications provide the following information for each item to be developed:

- The targeted standard.
- Guidance regarding cognitive complexity/Depth of Knowledge (DOK) levels.
- Recommendations concerning passage/item resource/context – for example, a reading item specification provides target Lexile® readability ranges to ensure an appropriate level of challenge
- A discrete statement of the objective of the item.
- The targeted grade or grade range.
- Parameters, examples, definitions, and resources when applicable.
- Suggested language for stems and guidance regarding answer options.

Passage Selection

NWEA Reading tests include sets of items associated with a single stimulus (i.e., an extended passage). Passages are selected from public domain, copyrighted works, or are specifically commissioned from passage writers. Passages created for the Reading assessment will be subject to a review process similar to our current passage review process outlined below.

1. Content Specialists write passage specifications to garner literary, informational, and persuasive passages as well as technical, domain-specific, and historical documents. Specifications detail the desired readability, text complexity, word count, and genre of the passage.
2. Passage specifications are sent for fulfillment to contracted passage writers/passage finders.
3. Contractors send a synopsis of the passage topic to editors for preapproval. Before preapproving a topic, editors ascertain that the topic is age-appropriate; that it does not overlap with the topic of other passages; and that it is unlikely to present bias or sensitivity concerns.
4. If a topic is approved:
    a) Passage writers submit passage files and relevant source documentation to NWEA via a secure FTP site.

b) The passage then undergoes a series of six reviews conducted by NWEA Content Editors and Content Specialists. The reviews include the following tasks:

i. Verifying that the passage is free of plagiarism (if commissioned) and documenting its permissions status (public domain or copyrighted).
ii. Ensuring that the passage does not have copyright, trademark, or rights of publicity issues.
iii. Fact-checking informational passages.
iv. Performing an editorial review to ensure the overall structural and mechanical quality of the passage.
v. Recording Lexile, Flesch-Kincaid, and Coh-Metrix readability analysis scores to help gauge grade-appropriateness and text complexity.
vi. Applying a rubric to analyze the passage qualitatively (i.e., for features such as structure; language conventionality and clarity; knowledge demands; levels of meaning or purpose).
vii. Validating that the passage is free of potential bias and sensitivity issues.
viii. Evaluating the ability of the passage to support a range of item types of varying complexity that will assess a variety of reading standards.
ix. Styling and copyediting the passage.

Item Writing

Although a portion of items are developed in-house by NWEA Content Specialists, the majority of items are written by established vendors or by freelance item writers. The freelance item writers we work with are current educators or professional item writers, who can demonstrate more than two years teaching experience in language arts, mathematics, or science. It is preferred that they also have a master's degree in the content area in which they will contribute and have previous item writing and/or test development experience. In short, we require a demonstrated depth of understanding and expertise with both domain area knowledge and assessment. Writers must also submit sample items, and undergo training with NWEA Item Acquisition Specialists.

Item Quality Review

Each item is subjected to a thorough review schedule, beginning with the item writing phase of development. NWEA items receive an initial quality review performed by Item Quality Reviewers. During this review, the item is evaluated against the following criteria:

- **Bias and sensitivity:** Any potential bias or sensitivity issues found in items are flagged or rejected.
- **Permissions and plagiarism:** We have a strict policy on plagiarism. Editorial staff members verify that the item is an original work. Once these staff members are confident that the item is an original work, the Permissions Editors verify copyright and permissions.
- **Content validity, instructional relevance, and currency:** The content validity evaluation involves ensuring that the item assesses what it is intended to assess and that all components of the item are correct. This includes confirming that any context used in the item is appropriate, the item has only one correct response, and that the distractors are plausible. The instructional relevancy evaluation involves ensuring that the concept

assessed is presented in a way that is consistent with current classroom practices. The currency evaluation involves ensuring that the terminology or information in the item is not dated or likely to become dated.

Content Review One

Next, the item moves into Content Review One. During this review, an NWEA Content Specialist in the appropriate content area reviews the item using a detailed Item Review Checklist to confirm that the item is instructionally relevant, its content is important and aligned, it has clear face validity, it is free of sensitivity and fairness issues, and it is sound in terms of item construction. During Content Review One, items are revised as needed based on the criteria of the Item Review Checklist.

The NWEA Content Specialist also validates the grade appropriateness of the item and assigns a DOK level. Finally, the Content Specialist assigns the item a preliminary difficulty level (called a provisional calibration or provisional RIT) that is needed for field test purposes. The provisional estimate of difficulty is based on the observed difficulty of similar items and the Content Specialist's expertise.

Content Review Two

A second content review is performed by a different NWEA Content Specialist from the same content area. This Content Specialist reviews the item using the Content Review checklist and verifies the alignment, DOK level, and that the item characteristics fields have been set appropriately in the item management system to ensure that the item is ready for field testing.

Copy Edit

After an item successfully passes through Content Review Two, a copy editor checks each item's syntax, grammar usage, spelling, and punctuation to ensure final product quality.

Field Testing and Item Analysis

Items are field tested after passing thorough review by NWEA staff, items are field tested utilizing an embedded field test design. In this design, five item slots of the spring assessment, and three item slots of the fall and winter assessment, are dedicated to field testing new items. This "Common Person" design is an efficient method of ensuring that enough item responses are obtained to accurately calibrate new items to the RIT scale.

Student response data from each new field test item, presented within a set of calibrated items, are used to analyze and to calibrate the difficulty estimate aligned to the existing measurement scale. Successfully calibrated items are added to the operational item banks.