International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 12, Number 1, February 2025

DOI 10.7333/2502-1201035

ISSN 2165-6592

Design Considerations and Reporting Solutions for a Multiple Administrations Adaptive Testing System

M. Christina Schneider

Cambium Assessment, Inc.

Seung W. Choi

The University of Texas at Austin

Daniel Lewis

Creative Measurement Solutions LLC

This paper overviews the policy, test development, psychometric, and reporting deliberations that stakeholders engaging in the design of a multiple-administrations adaptive test, otherwise known as a through-year assessment, will have to consider. We use the development of a design prototype for simulations coupled with newly designed score reports to serve as exemplars of the work to be done by states and vendors to develop a through-year assessment with the intended purpose of merging two different assessment systems into a singular system that supports teachers in better understanding where students are in their learning and states in meeting accountability requirements.

Keywords: *computerized adaptive tests, test design, through-year assessments, score reporting*

How does the educational measurement field integrate the intended uses and purposes of interim and summative assessment systems into a single, coherent assessment system that meets the needs of state departments of education, school districts, and teachers? This is a weighty problem, oftentimes with competing goals. In this paper we overview the policy, test development, psychometric, and reporting deliberations that stakeholders engaging in the design of such a system will have to consider. We use the development of a design prototype for simulations coupled with a newly designed mock score report as an example of the work to be done by states and vendors to develop a through-year system. Only through collaboration and innovation in the areas of (1) test design, (2) computerized adaptive test (CAT) algorithm features, and (3) score reporting features, will states and vendors merge two different assessment system purposes into a singular system that supports teachers in better understanding where students are in their learning, and states in meeting accountability requirements.

Multiple administrations adaptive tests (MAAT) in the educational measurement field heretofore have generally been referred to as interim assessments, with Curriculum Associates' *iReady*, Edmentum's *Exact Path*, Renaissance's *Star Assessments*, and NWEA's *Measures of Academic Progress* being examples. State assessments, historically, are administered only one time a year and inform state accountability. However, in 2010 the U.S. Department of Education (USDOE) described a through-course summative assessment (TCSA) in their Race to the Top applications. This has, over 15 years, slowly moved the field to considering MAAT in the context of summative purposes and uses.

Originally the USDOE (2010) encouraged the use of a TCSA.

> [A] through-course summative assessment means an assessment system component or set of assessment system components that is administered periodically during the academic year. A student's results from through-course summative assessments must be combined to produce the student's total summative assessment score for that academic year. (p. 18,178)

While stakeholders often initially like the idea of a TCSA, the research and piloting attempts over the years have shown this design has policy challenges (Gianopulos, this issue; Porter-Magee, 2011). Among the largest challenges are growth interpretations and the production of a summative score (Jerald et al., 2011). Creating a theory of action regarding how TCSA supports adapting instruction for students across the ability distributions is typically not a consideration in such a model.

The USDOE next gave flexibility in the final regulations for *The Every Student Succeeds Act* (ESSA) of 2015 (USDOE, 2016) for different assessment system designs, noting: "States have flexibility to develop new assessment designs, which may include a series of multiple statewide interim assessments during the course of the academic year that result in a single summative assessment score (sometimes described as "modular" assessments, p.3).

These interim assessments that result in a single summative score are currently referred to as through-year assessments. The USDOE (2016) clarified that innovative assessments "… may include items above or below a student's grade level so long as the State measures each student's academic proficiency based on the challenging State academic content standards for the grade in which the student is enrolled (p. 2)." USDOE's latest peer review guidance (2018) stipulated that a state can include additional content from adjacent grades in its assessments to provide additional information to parents and teachers regarding student achievement. There are, however, technical considerations for allowing students to go off grade, both above and below.

A fixed-form assessment measuring on-grade content for a state assessment will have larger measurement error at the tails of the score distribution. The students in the tails of the distribution have ability estimates that are the most imprecise, making it difficult to discern what, specifically, the student knows and can do. This situation influences the types and precision of instructional feedback about learner profiles that are available for teachers if three fixed-form assessments are used across the year. One enhancement is to create three multistage assessments (Texas Education Agency, 2024) or three CATs using only items aligned to the target grade-level standards (Florida Department of Education, 2023). While student proficiency must be assessed using items aligned to the depth and breadth of the on-grade standards, the 2016 USDOE regulations imply that a state may be allowed to measure an outlier student more precisely by identifying where he or she is functioning, after first assessing the student's level of proficiency, so that instruction can be targeted to what the student needs next. This conclusion is derived from text noting assessments "may include items above […] a student's grade level," (USDOE, 2016 p. 2) along with the stringent requirements they have set for states to eliminate the double testing of advanced students.

The USDOE (2016) final regulations for ESSA denoted that students in eighth grade who are enrolled in Algebra may take an end-of-course test if they are taking the equivalent high school course. These students can forego taking the grade-level test of record (double testing) *only* if the state has a mechanism of providing all students "the opportunity to be prepared for and to take advanced mathematics coursework" (USDOE, 2016, p. 3). The removal of the double testing requirement is *only* allowed if the state supports the advancement and instructional supports in the same way for all students, and it sets policy that invokes subsequent changes in instruction for students across its educational system. These criteria should analogously apply to English language arts, specifically because we see states such as South Carolina enroll middle school students in end-of-course subjects in both Mathematics and English. To know which students are ready to exit grade-level standards early and enter more advanced coursework requires three criteria for a test design coupled with policy supports from the state.

1. Students can be moved off-grade with supporting evidence that the student has been measured on the breadth and depth of the on-grade-level standards and has demonstrated proficiency on grade-level standards.
2. Identified students are provided enrichment to prepare them for more advanced coursework.
3. Each year, students must again show they are meeting and exceeding the requirements for grade-level standards until they are prepared for and enrolled in high school coursework.

This policy goal translates to a design requirement for an innovative design for a through-year assessment. The prototype must allow a student to bank an advanced score on a summative blueprint and then access an above-grade-level item bank and blueprint. This would allow advanced students repeated opportunities to demonstrate and sustain advanced skills. This underpinning is related to allowing access to challenging content for all students, and it is consistent with holistic models (e.g., Assouline et al., 2009) for determining if students need acceleration. This also relates to a test design requirement that centers on first supporting a grade-level summative test score interpretation (i.e., the student should first demonstrate they are advanced in the on-grade content) and then providing guidance on where the student is functioning in the standards to support system-level and instructional-level actions through a reporting system. Because stakeholders generally want to engage in such tasks with shorter amounts of testing time, this also includes the design requirement for a CAT.

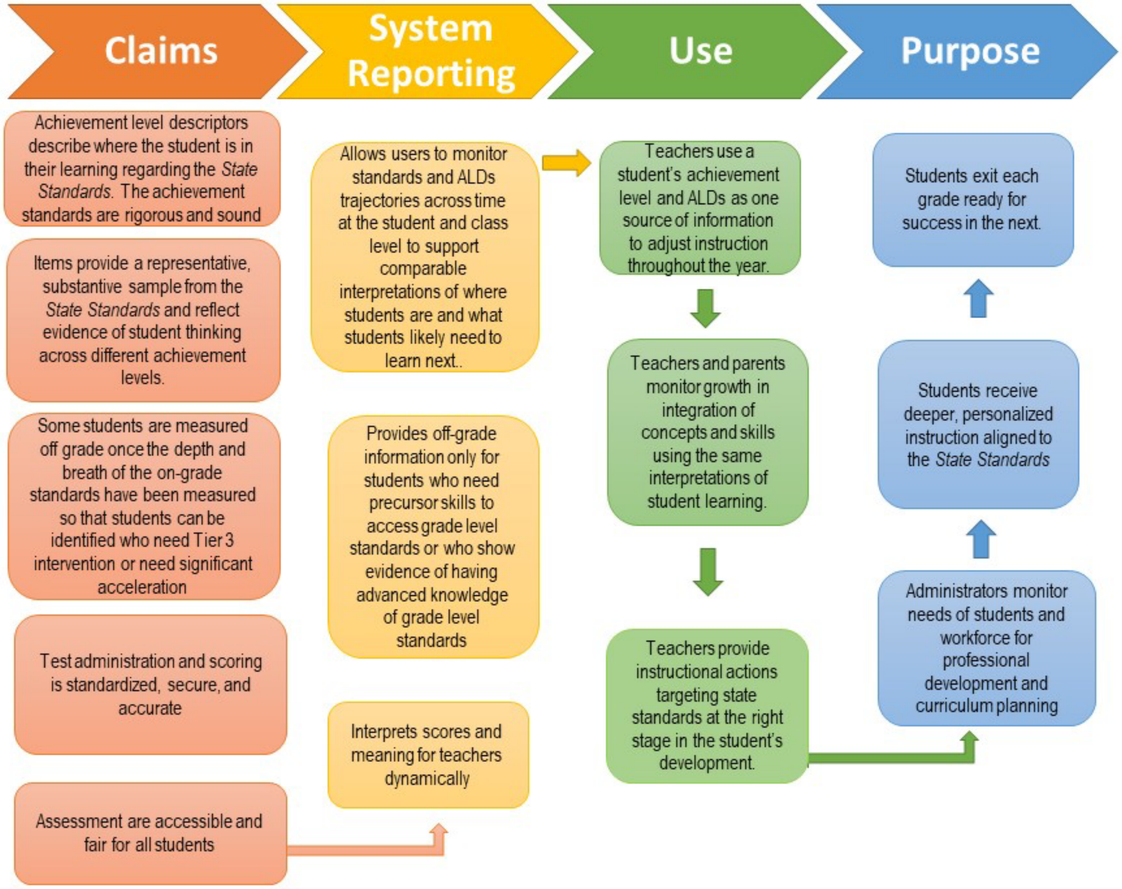
To conform with the USDOE regulations, the identification of what a student can do below-grade should also be equally evidence-based. However, the argument for students in more novice states of learning requires that these students be given access to rigorous on-grade instruction. The USDOE explicitly denoted that proficiency must be established with on-grade items, but it does not preclude students moving to below-grade-level blueprints and item banks to support where instruction needs to begin for these students. District users of interim products frequently use test results to determine which students need to be placed into intensive Tier 3 interventions. Thus, a design requirement for an innovative design for a through-year assessment is that lower-ability students must first be measured on their present level of performance within their grade-level standards at the beginning of each test. If students in the lowest achievement level are not meaningfully accessing the grade-level standards, then these students should be routed to the adjacent below-grade-level bank and blueprint to identify where they are functioning on prerequisite standards to those in their target grade, to assist teachers in efficiently understanding how to scaffold instruction from prerequisite standards to on-grade-level standards to support student growth.

Empirically, Wei and Lin (2015) found that the measurement of students using on-grade content is accurate for most of the student population when a sufficiently large bank of on-grade content is available. However, their results show the highest scoring 10% of students and the lowest scoring 10% of students, might need off-grade content from the adjacent grades to measure their present level of functioning in the state standards. Such interpretations rely heavily on item bank depth and the psychometric qualities of the items. Test design expectations for peer review have consistently noted the need to measure the breadth and depth of the standards, which translate to a design requirement of having enough items within each achievement level bin to measure all students reliably. Because states can expect that students are growing throughout the year, it is critical to have sufficient numbers of items in the lowest and highest achievement levels to measure students on these standards at three different time points in a year with different items. This suggests test design requirements centered in Range achievement level descriptors (ALDs; Egan et al., 2012) as the score interpretation, which are then embedded into item writing, alignment, item bank analysis, standard setting, and reporting processes, along with a commitment to improving score interpretations through iteration (Huff et al., in press; Lewis & Cook, 2020; Leucht, 2020; Schneider et al., 2021). These requirements are necessary to meet the intended theory of action shown in Figure 1.

To support the intended use of monitoring growth over time there is a requirement that the test items be placed on a vertical scale with a common domain blueprint. This is an appropriate design for stakeholders who desire to support students learning at a different pace from one another; that is, they believe there is heterogeneous achievement and growth among students. Further, this design centers the goal of learning in the mastery paradigm. Guskey (2010) wrote that mastery learning (sometimes called standards-based or competency-based grading) is centered in the belief that students earn a grade (or in this context an achievement level) based on achieving mastery, and he advocated those students who need multiple opportunities to master learning targets deserve the same grade (in this case summative achievement level) as those who mastered the learning target faster. Thus, a student's summative score and proficiency is established by the end of the year for most students and can be banked earlier in the year for some students who are advanced.

When comparing the USDOE (2018) *Peer Review Guidelines* with the ESSA regulations (2016), conflicting specificity is found. Whereas the peer review requirements stipulate that the assessment "provides a *score* for the student that is based only on the student's performance on grade-level academic content standards, "(USDOE, 2018, p. 23), the ESSA regulations denote the assessment measure "each student's academic *proficiency* based on the challenging State academic content standards for the grade in which the student is enrolled" (USDOE, 2016, p. 2). The USDOE peer review guidelines also note that each student's score who is measured with off-grade content must be as precise as the score for a student assessed only on grade-level academic content standards. This translates to a criterion that outlier scores have a conditional standard error of measurement that is similar to those students at the edges of the distribution who have tested using only on-grade content. Moreover, the state "may not include off-grade-level content in evidence addressing the critical elements" (USDOE, 2018, p. 25) for peer review, and only student performance based on grade-level academic content and achievement standards will meet accountability and reporting requirements under Title I. Thus, the final requirement for the assessment is that the prototype allow for the easy and clear extraction of data that is on grade versus off grade for accountability and that the on-grade information be sufficiently reliable for its intended purpose.

Figure 1
Theory of Action for a Principally Designed Through-Year Assessment



Achievement level descriptors describe where the student is in their learning regarding the *State Standards.* The achievement standards are rigorous and sound

Items provide a representative, substantive sample from the *State Standards* and reflect evidence of student thinking across different achievement levels.

Some students are measured off grade once the depth and breath of the on-grade standards have been measured so that students can be identified who need Tier 3 intervention or need significant acceleration

Test administration and scoring is standardized, secure, and accurate

Assessment are accessible and fair for all students

Allows users to monitor standards and ALDs trajectories across time at the student and class level to support comparable interpretations of where students are and what students likely need to learn next..

Provides off-grade information only for students who need precursor skills to access grade level standards or who show evidence of having advanced knowledge of grade level standards

Interprets scores and meaning for teachers dynamically

Teachers use a student's achievement level and ALDs as one source of information to adjust instruction throughout the year.

Teachers and parents monitor growth in integration of concepts and skills using the same interpretations of student learning.

Teachers provide instructional actions targeting state standards at the right stage in the student's development.

Students exit each grade ready for success in the next.

Students receive deeper, personalized instruction aligned to the *State Standards*

Administrators monitor needs of students and workforce for professional development and curriculum planning

Translating Design Requirements to System Features for Prototype Software Development

The requirement for CAT was central to the prototype development. It was determined that creating the prototype as an extension of the shadow-test approach would be practical given this approach's ability to fully satisfy complex test blueprint requirements and the availability of an R package implementing the shadow-test approach (Choi et al., 2022) at the time of the prototype development. The goal was to develop smaller adaptive tests to mimic summative blueprints but that strategically permitted access to off-grade items to satisfy the uses of both interim and summative test score users. Therefore, the prototype needed to be underpinned with a large, simulated item bank in sufficient numbers for each achievement level bin to measure students reliably across three assessments. The on-grade section (or module) of the assessment needed to produce a reliable maximum likelihood estimate of student ability which Davey et al. (2016) noted would require more than 15–20 items. In practice, the on-grade module of the test would need to meet a state summative assessment reliability goal of .80 or above.

In summary, there were three high-level policy goals:

1. Summative interpretations of student performance should undergird the score interpretations for each administration, such that a student who was *approaching proficient* in the winter could show comparable ability to a student who was *approaching proficient* in the spring, but who had developed that knowledge and skill faster.
2. Students could pool advanced proficiency and move on when ready.
3. Lower performing students would be allowed a clean slate at each test administration to provide them multiple opportunities to demonstrate on-grade mastery.

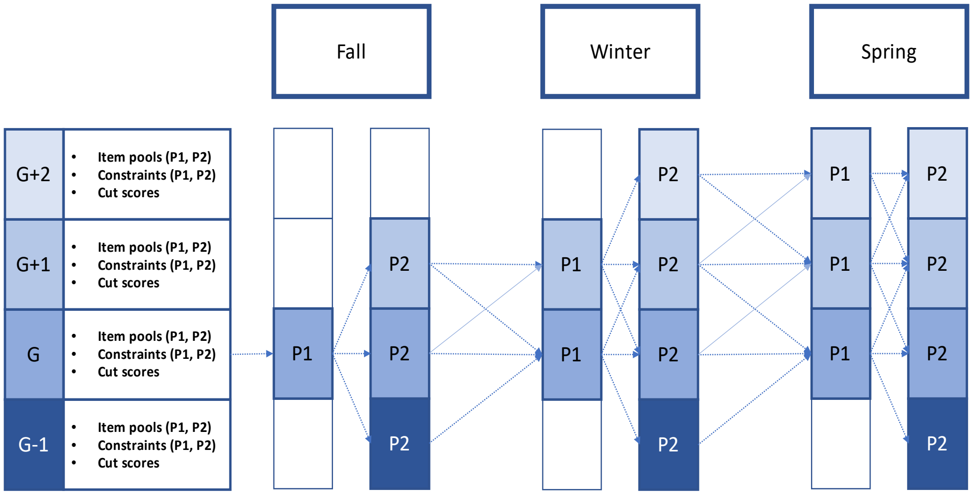
These goals support the intended interim use of test scores to diagnose if students are accessing or exiting on-grade content through the use of configurable routing rules to phases of the assessment, to document if students were growing. They are also intended to support the summative use of the test scores by determining the year-end achievement level of the student. In essence, rather than a multistage fixed form or a multistage assessment that dynamically routes students to a different module for the same blueprint (e.g., Luo & Wang, 2019) that increases or decreases in difficulty, the goal was to create a multistage CAT assessment in which phases shifted item bank content, if needed. The stages were described as phases during the feature development and modules in the actual software build.

Phase Structures

The degree of off-grade adaptivity and routing rules needed to be determined. The USDOE (2016) discussed Grade 8 students taking Algebra, and Assouline et al. (2009) discussed the need to assess students in the actual above-grade content to make acceleration decisions. A system feature decision was to allow advanced students to access content in the next two higher grades sequentially if such a student met the routing criteria for their grade and the next adjacent grade. In this way, the grade-level achievement level and scale score for the student could be reported for accountability and the achievement level descriptors in the grade in which a student was functioning, along with what area of the curriculum the student needed to grow in, could be made available to help teachers prepare students for advanced coursework. For lower-ability students previously routed to the next lower grade level, a system feature decision was to always start the student in the grade-level bank. Each testing event should use the final ability estimate from the previous phase or test to initialize the subsequent phase or test. This led to the conceptualization of the desired functioning of the prototype as shown in Figure 2 (the design discussed here, while funded by NWEA, is different than the design shared in Gianopulos, this issue).

Figure 2 shows three tests, each of which has two CAT phases. In the Fall, the test comprises an on-grade-level phase and a second phase that can move off-grade, if appropriate. The shading is used to depict the grade-level bank and constraints of a phase. In the Fall, the arrows depict the pathways to item banks and constraints that can be followed based on routing rules. The Fall has three possible pathways. One pathway is the on-grade-level phase paired with an above-grade-level phase. A second pathway is the on-grade-level phase paired with an on-grade-level phase. A third pathway is the on-grade-level phase paired with a lower-grade-level phase. The arrows between the Fall and Winter administrations show the possible pathways to the bank and constraints that begin the adaptive phase 1 of the Winter administration, which depends on the student's final ability estimate in the Fall and the routing rules. As can be seen in the figure, the number of possible pathways increase, in particular for high ability students, with each administration.

Figure 2
Routing of Phases
(From https://cran.r-project.org/web/packages/maat/vignettes/maat.html)



To achieve this design the following technical assumptions were required:

1. Within each phase, students can be routed to different grade-level blueprints and banks respectively; however, item parameters along the vertical scale can be aggregated for use in maximum likelihood scoring. This allows the final θ estimate to be comprised of item responses from phase 1 and phase 2.
2. Item banks built to Range ALDs (Egan et al., 2012) with sufficient numbers of items should allow most students in a grade to remain in the grade-level bank and show growth by moving into adjacent, higher achievement levels. This allows the majority of students to demonstrate growth in knowledge and skills while staying in the grade-level bank.
3. Item parameters on the vertical scale need to be vertically articulated such that the minimum item difficulty and the maximum item difficulty of grade G – 1 is lower than that of G. While item difficulties overlap between adjacent grades on a vertical scale, they could not do so at the tails of the distributions. This allows the use of a bank transition rule for moving off grade based on the difficulty percentile approach described below.
4. Vertical content alignment exists across domains and subdomains across grades, such that a single construct is present. This allows the final θ estimate to be comprised of item responses from phase 1 and phase 2.
5. Within and across test administrations, students should not see the same items. This allows

each test event to be comprised of new items for each student so that final θ estimates are not inflated due to item exposure.

These assumptions allow each phase to cover the content blueprint for the grade level of the phase and the corresponding item bank. This means that ultimately there is a high-level blueprint to support growth interpretations and a lower-level blueprint to support grade-level summative proficiency decisions.

Configurable Routing Rules

The design team determined that configurable routing rules should be implemented to give users different methods for determining what accessing or exiting grade-level standards meant, based on state policy. There were two approach variations stipulated for this feature: a student-centered approach and a content-centered approach. The student-centered approach was based on using confidence intervals (CI; Kingsbury & Weiss, 1983; Eggen & Straetmans, 2000) using maximum likelihood scoring (Yang et al., 2006) extended to multiple cut scores (Thompson, 2007). The confidence interval approach was used in comparison to the cutscores for the lowest and highest achievement level in each grade. If the student's θ estimate and CI at the end of phase 1 did not overlap with the cutscore for Level 2 (for lower-performing students) or Level 4 (for advanced students) then students could be routed to an off-grade bank.

For comparison purposes, the content-centered approach was conceptualized as transitioning students when their ability was either lower than or higher than the items in the bank. This was one of the key reasons for vertically articulating the lowest and highest item parameters in the grade-level bank.

Translating System Features to Technical Specifications for Algorithms

The MAAT system is adaptive in multiple levels and built on the optimal test design framework and the shadow-test approach to CAT (van der Linden & Reese, 1998). Each test assembly within the system is performed with a clear optimality criterion and complex test specifications as a constrained combinatorial optimization problem. The test is then adapted to individual examinees through the shadow-test approach to CAT as a sequential simultaneous optimization problem. The current assessment design presumes three tests administered at specific times within a school year (e.g., Fall, Winter, and Spring) and two CAT phases within each test.

The key design features address the needs for (1) satisfying complex test specifications, (2) adapting to examinee ability within phases, (3) tracking individual examinees across test administrations, (4) controlling intra-individual item exposure, and (5) transitioning item banks between phases within a test administration and between test administrations so grade-level feedback can be provided. The system extends the shadow-test approach to CAT to assemble multiple adaptive tests optimally constructed and administered throughout the year using multiple item banks vertically scaled.

Satisfying Complex Test Specifications

A critical design requirement to support the summative use of the test scores is to maintain the same test blueprint for all students at all levels across all test administrations throughout the year. If the test design requires that a separate test blueprint be specified by test administration (or by module within each test), it should also be permissible to specify different test blueprints for the

Fall, Winter, and Spring administrations. Such a requirement might be needed when a single summative assessment is restructured into a sequence of two shorter interim tests administered throughout the year with a slightly longer summative test. For maximum flexibility, the MAAT system also supports a TCSA test design because the software allows (1) separate test blueprints specified for different test administrations (and modules within each test administration), and (2) a common test blueprint enforced for all test administrations. Given that the test assembly is performed via a mixed-integer programing solver, any complex test blueprint constraints can be satisfied while maintaining measurement optimality for individual students so long as items that meet those constraints are represented in the item bank.

Adapting to Student Ability Within Phases

While typical multistage testing presents each test module as a fixed form, the MAAT system presents each module as a fixed-length CAT, fully optimized for each student's ability using the shadow-test approach to CAT. Upon administering each item and obtaining an updated θ estimate, the system re-assembles the module to the updated θ estimate and the same test blueprint constraints. The new module contains all items previously administered within the module and new items optimizing the updated θ estimate. As a result, the new module will fully satisfy the test blueprint constraints while optimized for the updated ability estimate.

Transitioning Item Banks Within and Between Test Administrations

To further enhance the quality of measurement through the increased adaptivity, while meeting the USDOE policy guidelines, each test is designed in two phases with the provision for transitioning from one item bank (and associated test blueprint constraints) to another between the phases as determined necessary according to a prespecified transition policy. Based on an item bank and associated test blueprint constraints, each phase is a CAT assembled optimally using the shadow-test approach. At the completion of the first phase, the θ estimate from the phase determines whether the student should continue with the same item bank or be routed to an off-grade item bank in the second phase.

Figure 3 shows transition rules between phases in Test 1 and between the final θ estimate in Test 1 to the first phase of Test 2. The bank and constraints that begin the adaptive Phase 1 of Test 2 depends on the student's final θ estimate from Test 1 and the routing rules. Students can be routed to, at most, one grade level above or one grade level below between test administrations. Given G denotes the student's enrolled grade of record, any student who was previously routed to a below-grade item bank (G – 1) always starts the subsequent test on-grade, G, as shown between Test 1 and Test 2. This means, with three test administrations, the permissible item banks range from (G − 1) to (G + 2). That is, with the number of tests fixed at three per year, an advanced student can go two grades up and a more novice student in the content area can go one grade down.

As illustrated in Figure 3, transitions to different grade-level item banks can occur between phases and also between tests. In what follows, we present some details on the two approaches to implementing the above-mentioned transition rules: (1) the CI approach, and (2) the difficulty percentile approach. In both approaches, students are routed based on the performance of each phase in a test denoted as $\hat{\theta}$.

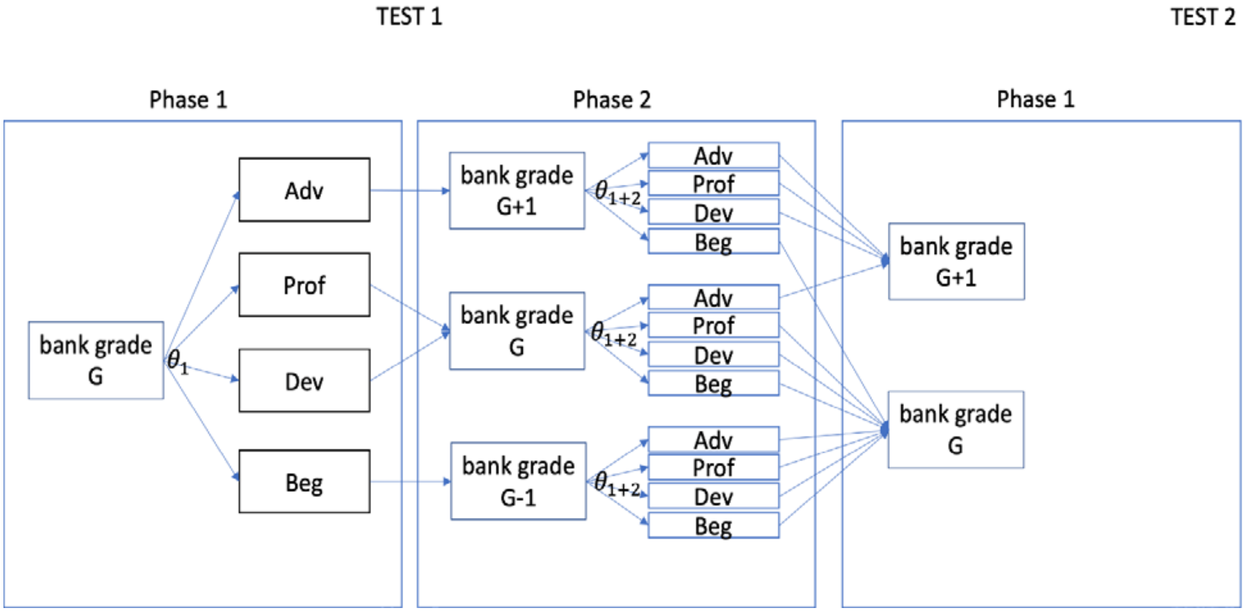
1. The CI approach computes the boundary values for each student's ability estimate:

$$\hat{\theta}_L = \hat{\theta} - z_\alpha \times SE_{\hat{\theta}} \tag{1}$$

$$\hat{\theta}_U = \hat{\theta} + z_\alpha \times SE_{\hat{\theta}} \tag{2}$$

where z_α is the normal deviate corresponding to a $(1 - \alpha)\%$ confidence interval, and $SE_{\hat{\theta}}$ is the standard error of measurement (SEM) associated with a point estimate of $\hat{\theta}$. Using the example with four achievement levels from Figure 3, i.e., Beginning, Developing, Proficient, and Advanced, if the lower boundary value, $\hat{\theta}_L$, falls into Advanced the student is routed to the above-grade item bank. If the upper boundary value, $\hat{\theta}_U$, falls into Beginning, the student is routed to the below-grade item bank. In all other cases, the student will remain in the same item bank.

Figure 3
Transition Rules: Test 1 and Test 2



Note. θ_1 denotes an estimate based on Phase 1 items only; θ_{1+2} denotes an estimate based on a combination of Phase 1 and Phase 2 items.

2. If the lower boundary value, $\hat{\theta}_L$, is higher than the $(1 - \alpha)$th percentile of item difficulty values on the item response theory scale in the current bank, the student is routed to the above-grade item bank. If the upper boundary value, $\hat{\theta}_U$, is lower than the αth percentile of item difficulty values in the current bank, the student is routed to the below-grade item bank. In all other cases, the student will remain in the same item bank.

Adapting Changes in Examinee Ability Across Test Administrations

To enhance continuity and adaptivity across test administrations, each subsequent test is initialized based on the previous test performance with an opportunity to transition to a lower/higher item bank at the end of the first phase of the test. That is, in all tests Phase 1 aims to determine whether the student should be routed to an on- or off-grade bank in Phase 2. The following transition rules are illustrated in Figure 3:

1. Any student who was previously below-grade, i.e., (G – 1), always starts the next test on-

grade, G, with θ_{1+2} as the starting θ.

2. If a student is classified as above-grade after Phase 1, the next phase should be based on an above-grade bank with θ_1 as the starting θ.
3. If a student remains on-grade after Phase 1 and gets classified as above-grade after Phase 2, the next test should begin in above-grade with θ_{1+2} as the starting θ.
4. If a student transitioned into above-grade content in Phase 2, the next test should begin in the same above-grade bank with θ_{1+2} as the starting point unless θ_{1+2} fell into Beginning after Phase 2. Note that these rules are configurable in the MAAT system.
5. If a student transitioned into above-grade content in Phase 2, the next test will not move up again even if θ_{1+2} rose to Advanced.

Controlling Intra-Individual Item Exposure

The system supports options for inter-individual exposure control and intra-individual item overlap control. Exposure control is used to address test security concerns in high-stakes assessment. Overlap control, on the other hand, is used to prevent or reduce the intra-individual overlap in test content across administrations. The primary exposure control method for the shadow-test approach to CAT is the item eligibility probability method (see van der Linden & Choi, 2020). The item eligibility control method can be used to make all items previously seen by the examinee ineligible for the current administration by imposing constraints similarly as

$$\sum_{i \in s_j} x_i = 0\,, \tag{3}$$

where s_j denotes the set of items Examinee j has seen prior to the current administration. Imposing these hard constraints can unduly limit the item bank and potentially affect the quality of measurement. To avoid infeasibility and degradation of measurement we can impose soft constraints in the form of a modification to the maximum information objective function as

$$\text{maximize} \sum_{i=1}^{I} I_i(\theta) x_i - M \sum_{i \in s_j} x_i\,, \tag{4}$$

where M is a penalty for selecting an item from s_j, the subset of items previously administered to Examinee j. This modification to the objective function can effectively deter the selection of previously administered items unless absolutely necessary for feasibility of the model.

Although the same item eligibility constraints for inter-individual exposure control can be used to control intra-individual item exposure, the mechanism for identifying ineligible items for the intra-individual exposure control is quite different. It requires tracking the examinee records across test administrations, which might be months apart. As the number of administrations increases, the ineligible item set (s_j) can grow quickly and adversely affect the quality of measurement progressively. To prevent the ineligible item set from growing quickly, s_j might need to be defined based only on the immediately preceding test administration.

Range ALD-Based Score Reporting to Support Response to Intervention

The dynamic test design implemented in the MAAT R package (Choi et al., 2022) requires a dynamic score report structure centered in Range ALDs to encourage teachers to recognize what students need next to grow. The primary purpose of accountability assessment is to measure students' on-grade achievement and commensurately, the test design and report begin by measuring, and reporting information regarding, the students' on-grade proficiency. Figure 4 illustrates the use of principled reporting features recommended by Lewis (2019) for on-grade reporting. First, to enhance assessment literacy, we provide the most important information in question-and-answer format. That is, if tests are designed to answer questions, we can moderate the need for assessment literacy in several ways.

In particular, we explicitly state the questions and answers so that the teachers and parents do not have to make inferences. In this case, Figure 4 shows the question common for all students: *Where is this student with respect to end-of-year expectations in the grade-level curriculum?* Figure 4 illustrates how reports can answer this question dynamically, depending on each student's test performance. In this case, the answer is: *This student is currently working at Approaches Expectations.*

Another feature of the reporting structure of Figure 4 supporting the principle to enhance assessment literacy is the use of multiple reporting modalities. The question-and-answer format provides the most important information in an optimally accessible format. We also provide the students' test results analytically in two ways to support users of test results with varying degrees of analytic sophistication. We provide the information graphically, to support users capable of comprehending numerical and graphical representations of the results, and we also provide a written annotation of the results that describes in text the information that the graphics reveal.

A modest, but important and often overlooked, feature illustrated in Figure 4 is the reporting and description of the meaning of the SEM in non-technical terms. It is shown as a V-shape spanning the interval of the obtained score plus and minus one SEM, supporting Lewis' (2019) primary principle—"validity first."

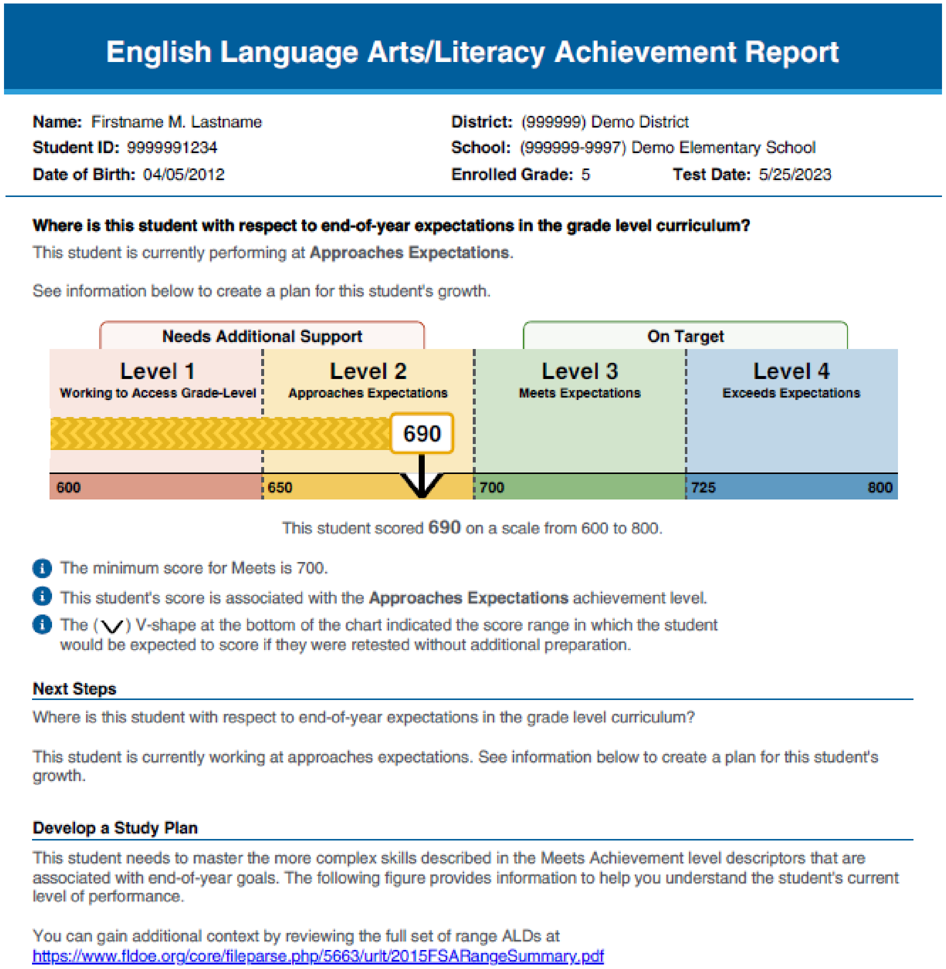
Following this answer to the primary question is a prompt to support another principle suggested by Lewis (2019)—enhance intelligent analytics—let teachers teach: The statement *See information below to create a plan for this student's growth* indicates that more detailed analytics follow and may be used to support student growth. The analytics that follow are also dynamic, depending on whether the student accessed fully on-grade content or on-grade and off-grade content.

Figures 5 and 6 illustrate two mechanisms for conveying more detailed test analytics for on-grade content—providing the percent correct information separately for the sets of items aligned to each performance level (Figure 5) and for the items aligned to the next higher adjacent achievement level (Figure 6) to support teachers eliciting more complex skills that the student needs next.

s

Figure 4

Sample Score Report Showing Principled Reporting Feature



English Language Arts/Literacy Achievement Report

Name: Firstname M. Lastname
Student ID: 9999991234
Date of Birth: 04/05/2012

District: (999999) Demo District
School: (999999-9997) Demo Elementary School
Enrolled Grade: 5 **Test Date:** 5/25/2023

Where is this student with respect to end-of-year expectations in the grade level curriculum?

This student is currently performing at **Approaches Expectations**.

See information below to create a plan for this student's growth.

This student scored **690** on a scale from 600 to 800.

- The minimum score for Meets is 700.
- This student's score is associated with the **Approaches Expectations** achievement level.
- The (V) V-shape at the bottom of the chart indicated the score range in which the student would be expected to score if they were retested without additional preparation.

Next Steps

Where is this student with respect to end-of-year expectations in the grade level curriculum?

This student is currently working at approaches expectations. See information below to create a plan for this student's growth.

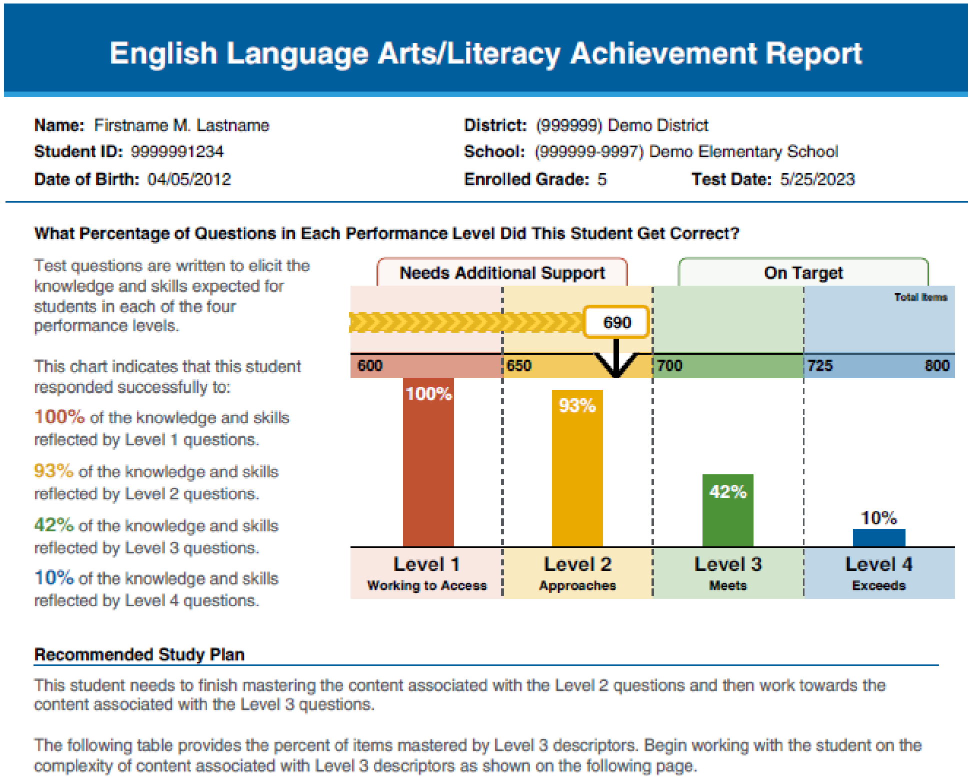
Develop a Study Plan

This student needs to master the more complex skills described in the Meets Achievement level descriptors that are associated with end-of-year goals. The following figure provides information to help you understand the student's current level of performance.

You can gain additional context by reviewing the full set of range ALDs at https://www.fldoe.org/core/fileparse.php/5663/urlt/2015FSARangeSummary.pdf

Figure 5

Sample Score Report Showing Principled Reporting by Achievement Level



English Language Arts/Literacy Achievement Report

Name: Firstname M. Lastname
Student ID: 9999991234
Date of Birth: 04/05/2012

District: (999999) Demo District
School: (999999-9997) Demo Elementary School
Enrolled Grade: 5 **Test Date:** 5/25/2023

What Percentage of Questions in Each Performance Level Did This Student Get Correct?

Test questions are written to elicit the knowledge and skills expected for students in each of the four performance levels.

This chart indicates that this student responded successfully to:

100% of the knowledge and skills reflected by Level 1 questions.

93% of the knowledge and skills reflected by Level 2 questions.

42% of the knowledge and skills reflected by Level 3 questions.

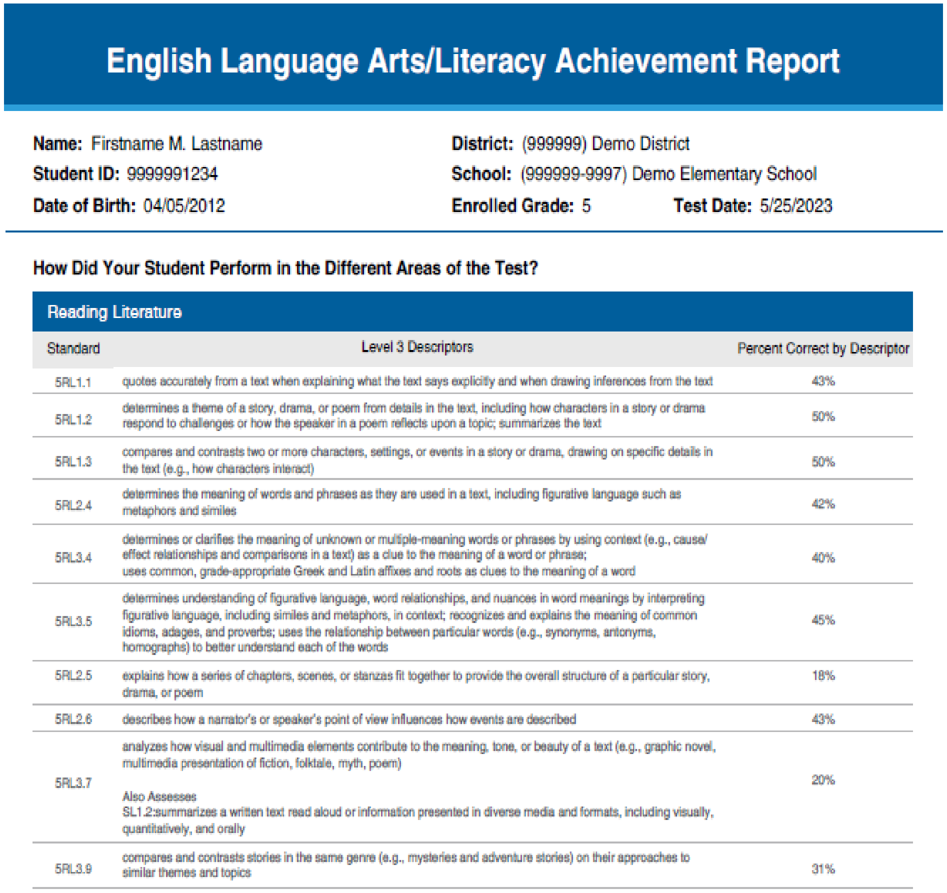
10% of the knowledge and skills reflected by Level 4 questions.

Recommended Study Plan

This student needs to finish mastering the content associated with the Level 2 questions and then work towards the content associated with the Level 3 questions.

The following table provides the percent of items mastered by Level 3 descriptors. Begin working with the student on the complexity of content associated with Level 3 descriptors as shown on the following page.

Figure 6
Sample Score Report Showing Principled Reporting of Percent of Items Aligned to Level 3 ALDs the Student Performing in Level 2 Answered Correctly



English Language Arts/Literacy Achievement Report

Name: Firstname M. Lastname
Student ID: 9999991234
Date of Birth: 04/05/2012

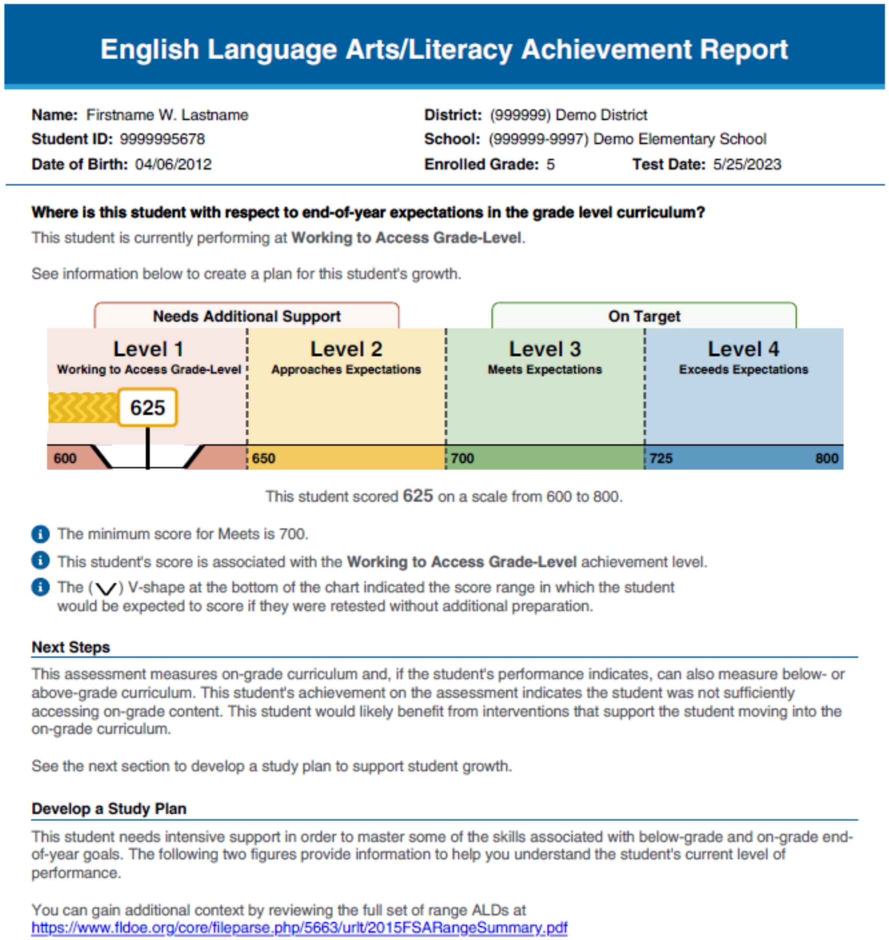
District: (999999) Demo District
School: (999999-9997) Demo Elementary School
Enrolled Grade: 5 **Test Date:** 5/25/2023

How Did Your Student Perform in the Different Areas of the Test?

Reading Literature

Standard	Level 3 Descriptors	Percent Correct by Descriptor
5RL1.1	quotes accurately from a text when explaining what the text says explicitly and when drawing inferences from the text	43%
5RL1.2	determines a theme of a story, drama, or poem from details in the text, including how characters in a story or drama respond to challenges or how the speaker in a poem reflects upon a topic; summarizes the text	50%
5RL1.3	compares and contrasts two or more characters, settings, or events in a story or drama, drawing on specific details in the text (e.g., how characters interact)	50%
5RL2.4	determines the meaning of words and phrases as they are used in a text, including figurative language such as metaphors and similes	42%
5RL3.4	determines or clarifies the meaning of unknown or multiple-meaning words or phrases by using context (e.g., cause/effect relationships and comparisons in a text) as a clue to the meaning of a word or phrase; uses common, grade-appropriate Greek and Latin affixes and roots as clues to the meaning of a word	40%
5RL3.5	determines understanding of figurative language, word relationships, and nuances in word meanings by interpreting figurative language, including similes and metaphors, in context; recognizes and explains the meaning of common idioms, adages, and proverbs; uses the relationship between particular words (e.g., synonyms, antonyms, homographs) to better understand each of the words	45%
5RL2.5	explains how a series of chapters, scenes, or stanzas fit together to provide the overall structure of a particular story, drama, or poem	18%
5RL2.6	describes how a narrator's or speaker's point of view influences how events are described	43%
5RL3.7	analyzes how visual and multimedia elements contribute to the meaning, tone, or beauty of a text (e.g., graphic novel, multimedia presentation of fiction, folktale, myth, poem) Also Assesses SL1.2:summarizes a written text read aloud or information presented in diverse media and formats, including visually, quantitatively, and orally	20%
5RL3.9	compares and contrasts stories in the same genre (e.g., mysteries and adventure stories) on their approaches to similar themes and topics	31%

Figure 7
Sample Score Report Showing Principled Reporting Features for a Student Currently Functioning in Level 1



English Language Arts/Literacy Achievement Report

Name: Firstname W. Lastname
Student ID: 9999995678
Date of Birth: 04/06/2012

District: (999999) Demo District
School: (999999-9997) Demo Elementary School
Enrolled Grade: 5 **Test Date:** 5/25/2023

Where is this student with respect to end-of-year expectations in the grade level curriculum?

This student is currently performing at **Working to Access Grade-Level**.

See information below to create a plan for this student's growth.

This student scored **625** on a scale from 600 to 800.

- The minimum score for Meets is 700.
- This student's score is associated with the **Working to Access Grade-Level** achievement level.
- The (∨) V-shape at the bottom of the chart indicated the score range in which the student would be expected to score if they were retested without additional preparation.

Next Steps

This assessment measures on-grade curriculum and, if the student's performance indicates, can also measure below- or above-grade curriculum. This student's achievement on the assessment indicates the student was not sufficiently accessing on-grade content. This student would likely benefit from interventions that support the student moving into the on-grade curriculum.

See the next section to develop a study plan to support student growth.

Develop a Study Plan

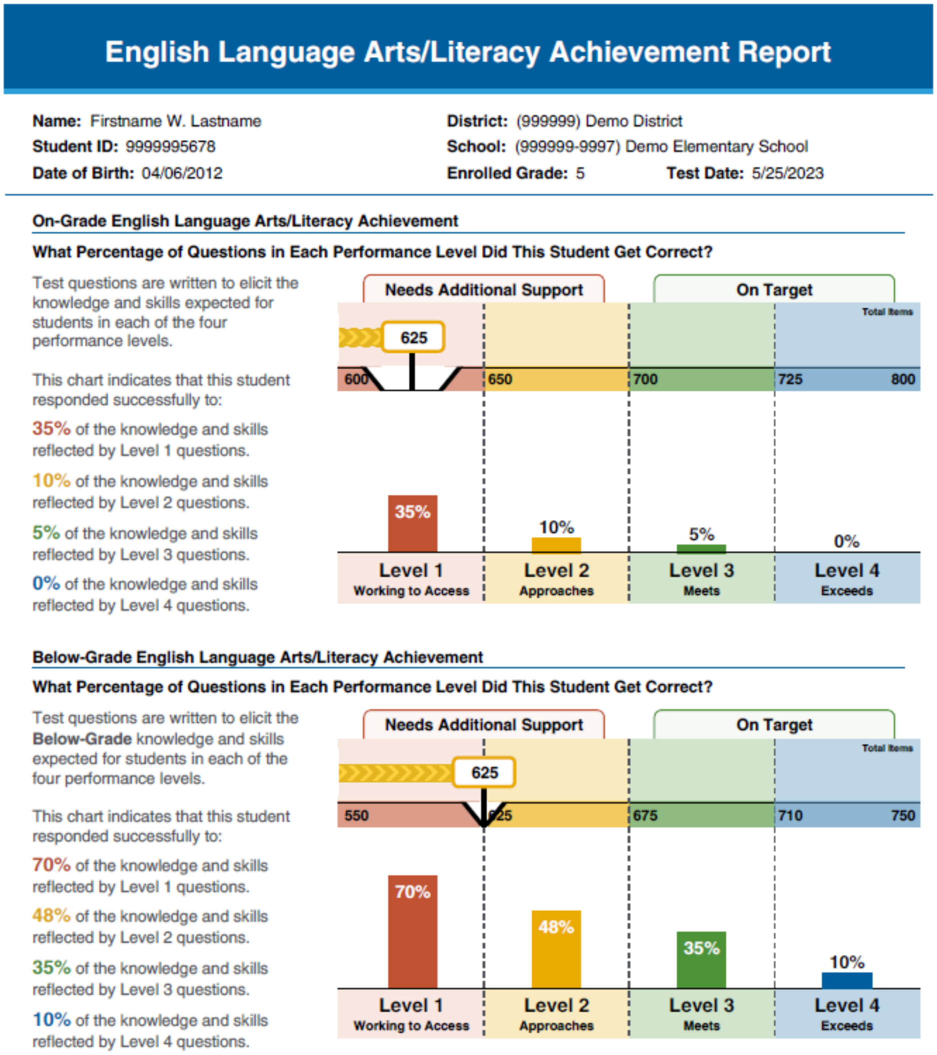
This student needs intensive support in order to master some of the skills associated with below-grade and on-grade end-of-year goals. The following two figures provide information to help you understand the student's current level of performance.

You can gain additional context by reviewing the full set of range ALDs at
https://www.fldoe.org/core/fileparse.php/5663/urlt/2015FSARangeSummary.pdf

When students access on- and below-grade content, the graphic for the on-grade information is again presented; however, the next steps explicitly request the teacher provide more support to the students, as shown in Figure 7.

Depicted in Figure 8 is the breakdown of the student's performance for items by achievement level for both on-grade and below-grade content, to remind the teacher that the student needs support in precursor standards in addition to on-grade standards. This report shows teachers that a student in Level 1 in grade-level standards is likely to also function in or near Level 1 in the lower adjacent grade. Figure 9 shows a policy decision in reporting. Because the goal is to move the student into the on-grade content, we chose to report the student's performance on the Level 2 descriptors in their grade of record to encourage teachers to think about having the student move into more complex content while ensuring precursor content from the lower adjacent level is addressed. The Range ALDs are intended to show the teacher they should have the student use explicit evidence found in texts and do something with it, such as write an explanation in order to

Figure 8
Sample Score Report Showing Principled Reporting of Percent of Items by Achievement Level Both On-and Off-Grade



English Language Arts/Literacy Achievement Report

Name: Firstname W. Lastname
Student ID: 9999995678
Date of Birth: 04/06/2012

District: (999999) Demo District
School: (999999-9997) Demo Elementary School
Enrolled Grade: 5 **Test Date:** 5/25/2023

On-Grade English Language Arts/Literacy Achievement

What Percentage of Questions in Each Performance Level Did This Student Get Correct?

Test questions are written to elicit the knowledge and skills expected for students in each of the four performance levels.

This chart indicates that this student responded successfully to:

35% of the knowledge and skills reflected by Level 1 questions.

10% of the knowledge and skills reflected by Level 2 questions.

5% of the knowledge and skills reflected by Level 3 questions.

0% of the knowledge and skills reflected by Level 4 questions.

Below-Grade English Language Arts/Literacy Achievement

What Percentage of Questions in Each Performance Level Did This Student Get Correct?

Test questions are written to elicit the **Below-Grade** knowledge and skills expected for students in each of the four performance levels.

This chart indicates that this student responded successfully to:

70% of the knowledge and skills reflected by Level 1 questions.

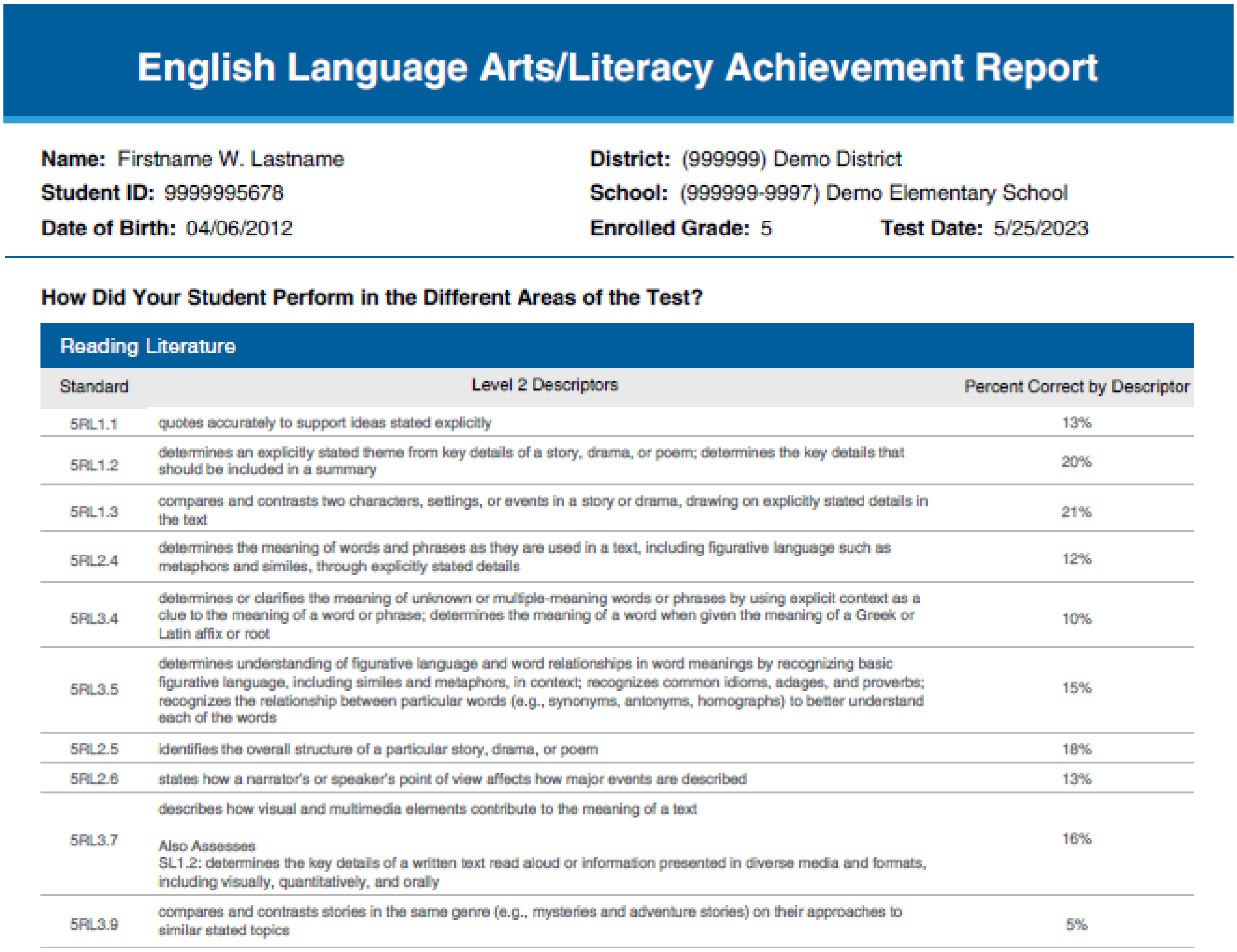
48% of the knowledge and skills reflected by Level 2 questions.

35% of the knowledge and skills reflected by Level 3 questions.

10% of the knowledge and skills reflected by Level 4 questions.

grow. This should help the teacher realize that the student should not just retrieve or locate explicit details. This is in contrast to the Level 3 descriptors for the student currently functioning in Level 2 who needs to grow in the Level 3 skills that are moving into inferencing. However, if these reports were produced for a student who was routed above grade level, we would expect the report to show where the student is functioning in relation to the above-grade cutscores and to show the percent of items aligned to the descriptors in the next higher achievement level in the adjacent grade to support acceleration.

Figure 9
Sample Score Report Showing Principled Reporting of Percent of Items Aligned to Level 2 ALDs the Student Performing in Level 1 Answered Correctly



English Language Arts/Literacy Achievement Report

Name: Firstname W. Lastname
Student ID: 9999995678
Date of Birth: 04/06/2012

District: (999999) Demo District
School: (999999-9997) Demo Elementary School
Enrolled Grade: 5 **Test Date:** 5/25/2023

How Did Your Student Perform in the Different Areas of the Test?

Reading Literature

Standard	Level 2 Descriptors	Percent Correct by Descriptor
5RL1.1	quotes accurately to support ideas stated explicitly	13%
5RL1.2	determines an explicitly stated theme from key details of a story, drama, or poem; determines the key details that should be included in a summary	20%
5RL1.3	compares and contrasts two characters, settings, or events in a story or drama, drawing on explicitly stated details in the text	21%
5RL2.4	determines the meaning of words and phrases as they are used in a text, including figurative language such as metaphors and similes, through explicitly stated details	12%
5RL3.4	determines or clarifies the meaning of unknown or multiple-meaning words or phrases by using explicit context as a clue to the meaning of a word or phrase; determines the meaning of a word when given the meaning of a Greek or Latin affix or root	10%
5RL3.5	determines understanding of figurative language and word relationships in word meanings by recognizing basic figurative language, including similes and metaphors, in context; recognizes common idioms, adages, and proverbs; recognizes the relationship between particular words (e.g., synonyms, antonyms, homographs) to better understand each of the words	15%
5RL2.5	identifies the overall structure of a particular story, drama, or poem	18%
5RL2.6	states how a narrator's or speaker's point of view affects how major events are described	13%
5RL3.7	describes how visual and multimedia elements contribute to the meaning of a text Also Assesses SL1.2: determines the key details of a written text read aloud or information presented in diverse media and formats, including visually, quantitatively, and orally	16%
5RL3.9	compares and contrasts stories in the same genre (e.g., mysteries and adventure stories) on their approaches to similar stated topics	5%

Conclusions

The algorithm and business rule development process described in this paper capitalizes on the USDOE (2016) guidance that innovative assessments are permitted to include off grade items as long as the state determines if the student is proficient with on-grade items reliably. USDOE's latest peer review guidance (2018) stipulated that a state may include additional content from adjacent grades in its assessments to provide additional information to parents and teachers regarding student achievement. The algorithm and business rules described here account for the technical considerations for allowing students to go off grade, both above and below, for only the specific students who are outliers. The algorithms are intended as a tool to be coupled with a test design centered in Range ALDs as the score interpretations that are embedded into item writing, alignment, item bank analysis, standard setting, and validation (Huff et al., in press; Lewis and Cook, 2020; Luecht, 2020; Schneider et al., 2021) to support improved reporting information.

Without innovation in how we (1) design and develop assessments, (2) implement CAT algorithms and business rules to control which students go off grade and when, and (3) dynamically report scores, through-year assessment systems will not meet their potential. To meet the intended uses and purposes of interim and summative assessment systems and solve the problems of bridging these two systems into a single, coherent assessment system requires that the information derived from such assessments is viewed as useful and worthy of educators' and students' time. This is critically important as educational systems work to support students whose education was disrupted by the pandemic. Helping teachers visualize that students respond successfully more often to items in lower achievement levels than to items in higher achievement levels might assist them in making connections that students need more rigor within the standards in order to grow.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Assouline, S., Colangelo, N., Lupkowski-Shoplik, A., Forstadt, L., & Lipscomb, J. (2009). *Iowa Acceleration Scale manual: A guide for whole-grade acceleration K-8* (3rd Ed.). Scottsdale, AZ: Great Potential Press.

Barnard, J. J. (2015). Implementing a CAT: The AMC experience. *Journal of Computerized Adaptive Testing*, *3*, 1–12. CrossRef

Bennett, R. E., Kane, M., & Bridgeman, B. (2011). *Theory of action and validity argument in the context of through-course summative assessment*. Princeton, NJ: Educational Testing Service. WebLink

Choi, S. W., Lim, S., & van der Linden, W. J. (2022). TestDesign: An optimal test design approach to constructing fixed and adaptive tests in R. *Behaviormetrika*, *49*, 191–229. CrossRef

Choi, S. W., Lim, S., Niu, L., Lee, S., Schneider, C. M., Lee, J., & Gianopulos, G. J. (2022). MAAT: An R package for multiple administrations adaptive testing. *Applied Psychological Measurement*, *46*(1),73–74. CrossRef

Choi, S. W., Lim, S., Niu, L., Lee, S. (2022). *MAAT: Multiple administrations adaptive testing*. *R package* (Version 1.0.2.9000) [Computer software]. CrossRef

Davey, T., Pitoniak, M. J., & Slater, S. C. (2016). Designing computerized adaptive tests. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (pp. 483–500).

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice and a proposed framework. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). Routledge.

Eggen, T. J. H. M, & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, *60*, 713–734. CrossRef

Florida Department of Education. (2023). Test design summary and blueprint: FAST ELA reading and B.E.S.T writing. *WebLink*

Jerald, C. D., Doorey, N. A., & Forgione, P. D., Jr. (2011). *Putting the pieces together: Summary report of the invitational research symposium on through-course summative assessments*. Princeton, NJ: Educational Testing Service. *WebLink*

Gianopulos, G. (this issue—2025). A literature review of through-course summative assessment models: The case for an adaptive through-year assessment. *Journal of Computerized Adaptive Testing,12 (1), 4-34. CrossRef*

Guskey. T. (2010). Lessons of mastery learning. *Educational Leadership*, *68*(2), 52-57. *WebLink*

Huff, K., Nichols, P., & Schneider, M. C. (in press). Designing and developing educational assessments. In L. Cook & M. J. Pitoniak (Eds.), *Educational measurement: 5th edition*. NCME.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237–254). Academic Press.

Kolen, M. J., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer.

Lewis, D. (2019, February). A principled approach to score reporting. Invited presentation to the CCSSO winter meeting of the Technical Issues in Large Scale Assessment (TILSA) SCASS, Baltimore, MD.

Lewis, D., & Cook, R. (2020). Embedded standard setting: Aligning standard-setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practice*, *39*(1), 8–21. *CrossRef*

Luecht, R. M. (2020). Generating performance-level descriptors under a principled assessment design paradigm: An example for assessments under the next-generation science standards. *Educational Measurement: Issues and Practice*, *39*(4), 105–115. *CrossRef*

Luo, X., & Wang, X. (2019). Dynamic multistage testing: A highly efficient and regulated adaptive testing method. *International Journal of Testing*, *19*(3), 227–247, *CrossRef*

Porter-Magee, K. (2011). *PARCC eliminates through-course assessments.* Washington, D.C.: Thomas Fordham Institute. *WebLink*

Schneider, M. C., Chen, J., & Nichols, P. (2021). Using principled assessment design and item difficulty modeling to connect hybrid adaptive instructional and assessment systems: Proof of concept. In Sottilare, R. A., & Schwarz, J. (Eds.). *Adaptive Instructional Systems. Adaptation Strategies and Methods. HCII 2021. Lecture Notes in Computer Science*, (12793). Springer. *CrossRef* .

Schneider, M. C., Agrimson, J., & Veazey, M., (2021). Examining alignment of mathematics test score interpretations on a computer adaptive assessment. *Educational Measurement Issues and Practices,41*(2), 12-24. *CrossRef*

Texas Education Agency. (2024, August 1) Texas Through-Year Assessment Pilot (TTAP) Year 1 Pilot Report. *WebLink*

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment, Research, and Evaluation*, *12*(1). *WebLink*

U.S. Department of Education (USDOE). (2010, April 9). Race to the Top Fund Assessment Program; notice inviting applications for new awards for fiscal year (FY) 2010. *Federal Register*, *75*(68), p. 18,178.

U.S. Department of Education (USDOE). (2016, November 29). *Federal Register*, *81*(229). 34 CFR 200 34 CFR 299. *WebLink*

U.S. Department of Education (USDOE). (2018, September 24). *A state's guide to the U.S. Department of Education's peer review process.* WebLink

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22, 259–270. *CrossRef*

van der Linden, W. J., & Choi, S. W. (2020). Improving item-exposure control in adaptive testing. *Journal of Educational Measurement*, 57, 405-422. *CrossRef*

Wei, H., & Lin, J. (2015). Using out-of-level items in computerized adaptive testing. *International Journal of Testing, 15*, 50–70. *CrossRef*

Wise, L. L. (2011). Picking up the pieces: Aggregating results from through-course assessments. Center for K–12 Assessment & Performance Management at ETS. *WebLink*

Yang, X., Poggio, J. C., & Glasnapp, D. R. (2006). Effects of estimation bias on Multiple-category classification with an IRT-based adaptive classification procedure. *Educational and Psychological Measurement*, *66*, 545-564. *CrossRef*

Zwick, R., & Mislevy, R. J. (2011). *Scaling and linking through-course summative assessments.* Center for K–12 Assessment & Performance Management at ETS. *WebLink*

Acknowledgements and Assistance

The prototype development described in this paper was funded by NWEA and is open-source. Thank you to Eli Mintzer at Cambium Assessment for his support in taking an early mock-up draft report and extending the visualizations.

Author's Address

M. Christina Schneider   *Email*: Christina.Schneider@cambiumassessment.com

Citation

Schneider, M. C., Choi, S.W., & Lewis, D. (2025). Design considerations and reporting solutions for a multiple administrations adaptive testing system. *Journal of Computerized Adaptive Testing,12(1)*, 35-53.