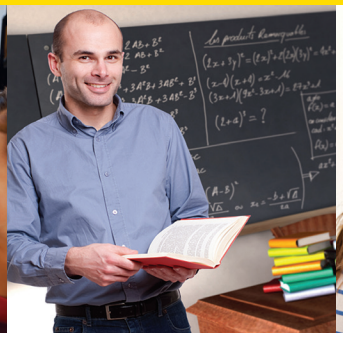
Alabama's State Assessment System: Recommendations from the Assessment Task Force

THE ALABAMA STATE DEPARTMENT OF EDUCATION

Report prepared by:
Juan D'Brot, Ph.D.
Scott Marion, Ph.D.

Center for Assessment

National Center for the Improvement of Educational Assessment

MAY 2018

Task Force Members

Tisha Allred
EL Coach/Parent, Walker County

Nancy Anderson
Associate Director/Attorney, Alabama Disabilities Advocacy Program

Becky Birdsong
Superintendent, Geneva County

Jacqueline Brooks
Superintendent, Macon County

Ashley Chasteen
Teacher, Jefferson County

Jennifer Fernandez
Teacher, Austinville Elementary, Decatur City

Dr. Pamela Fossett
Manager of Ed. Policy & Professional Practice, Alabama Education Association

Carrie Garris
Parent/Speech Pathologist, Clarke County

Khristie Goodwin
Special Education Coordinator, Oxford City Schools

Lisa Heard
Curriculum/Special Ed Coordinator, Tallapoosa County

Dr. Trey Holladay
Superintendent, Athens City Schools

Vickie Holloway
System Test Coordinator, Montgomery City Schools

Jeff Hyche
Principal, Hartselle High School

Heather Johnson
Parent, Tallassee City

Maria Johnson
Principal, Beverlye Magnet School

Kyle Kallhoff
Superintendent, Demopolis City

Mallory Lamb
Teacher, Oneonta City

Josh Laney
ALSDE, Workforce Development

Dr. Eric Mackey
Executive Director, School Superintendents of Alabama

Theresa McCormick
Associate Dean for Academic Affairs and Certification Officer, Auburn University

Jim McLean
Executive Director for the Center of Community-Based Partnerships

Dr. Caroline Novak
President, A+ Education Partnership

Dr. Tonya Perry
Assistant Professor of Secondary Education, University of Alabama, Birmingham (UAB)

Dr. Beth Quick
Dean of College of Education, University of Alabama - Huntsville

Thomas Rains
Vice President of Policy, A+ Education Partnership

Dr. Jimmy Shaw
Superintendent, Florence City

Marcia Smiley
Assistant Superintendent, Perry County

Sally Smith
Executive Director, Alabama School Boards Association

Dr. Shannon Stanley
Superintendent, Boaz City Schools

Keri Southern
Parent, Alabaster City

Lissa Tucker
Alabama School Boards Association

Dr. Vic Wilson
Executive Director, CLAS

Jeana Winter
Executive Director, Alabama Parent Education Center

Kellie Yeager
STC, Jefferson County

ALSDE Representatives

Dr. Joe Morton
Special Consultant

Judy Pugh
State Assessment Program Coordinator

Dr. Tony Thacker
Assistant State Superintendent Office of Evaluation and Innovation

Susan Beard
Student Assessment

Kanetra Germany
NAEP Coordinator

Maggie Hicks
Student Assessment

Dr. Sandy Ledwell
AMSTI/Science Coordinator

Michelle Lee
English Learner Specialist

Kathy Padgett
Student Assessment

Nannette Pence
Student Assessment

Crystal Richardson
Special Education Coordinator

Table of Contents

Alabama's State Assessment System: Recommendations from the Assessment Task Force

Executive Summary

The Alabama State Department of Education (ALSDE) convened an Assessment Task Force comprised of key education stakeholders to make recommendations for Alabama's next state assessment system. To develop these recommendations, the ALSDE held a series of in-person and virtual meetings with the Task Force to deliberate over many technical, policy, and practical issues associated with implementing an improved assessment system. ALSDE contracted with the National Center for the Improvement of Educational Assessment (Center for Assessment), a non-profit, non-partisan consulting organization to facilitate the work of the Task Force and to provide assessment expertise throughout the process. This report presents the recommendations of the Task Force for the design and implementation of a stable, high-quality assessment system in Alabama.

The Task Force held four meetings between December 2017 and April 2018, read numerous design briefs, and reviewed several drafts of this report. The contents of this report are based almost exclusively on consensus decisions of the Task Force. Where consensus could not be reached, decisions were based on an overwhelming majority of task force members. This executive summary presents a summary of the goals of Alabama's Assessment system as well as the design and implementation recommendations.

Goals of Alabama's State Assessment System

The Task Force identified goals for Alabama's assessment system to serve as a touchstone for design decisions and related recommendations. Task Force members identified many potential goals of the system through a series of design exercise, but identified and affirmed the following five goals for Alabama's assessment system.

1. The assessment system must provide a clear and credible measure of student performance on the Alabama state standards in grades 3-8 and high school.

2. The assessment system must provide signals of student readiness informed by external and rigorous criteria as students move through and beyond the Alabama educational system.

3. The assessment system must help improve teaching and learning by providing information useful for evaluating curriculum and instructional programs that promote improved student achievement and growth, and ultimately support the effectiveness of public education efforts.

4. The assessment system must provide stakeholders with varied, informative, and easily interpretable reports that help end-users understand student, local, and statewide educational trends against educational expectations.

5. The statewide assessment system must provide information to support federal and state accountability systems.

Task Force Recommendations

As part of the ALSDE's deliberation process, the Task Force wrestled with a series of assessment design and implementation issues. These discussions were informed by several technical conversations led by experts from the Center for Assessment, Drs. D'Brot and Marion. As a result of the deliberation process, the Task Force made design and implementation recommendations for each of the following assessment considerations:

Design Recommendations
- Assessment Specifications and Content Coverage
- Computer-Based Testing
- Adaptive Testing
- Item Types
- Testing Time and Field Testing
- Interim Assessments and Other System Supports

Implementation Recommendations
- Timing
- Reporting

Prior to enumerating the design and implementation recommendations, the Task Force made clear which components of Alabama's assessment system were addressed by the recommendations and which were considered beyond the scope of the Task Force, at least at this time.

Assessments Included and Not Included

- Develop and administer aligned assessments in grades 3-8 to all students in mathematics and English/language arts, except those very few students eligible for the alternate assessment (no more than 1%).
- Include the science assessments as part of the Alabama assessment system Request for Bids (RFB) requiring an aligned assessment for science at the end of each grade span (grades 3-5, 6-8, and high school). Additionally, the science assessment must be aligned to the state standards and appropriately assesses all three dimensions of science in an integrated manner.
- If the resources are available, the Task Force supported procuring modular interim assessments aligned to key curricular targets for optional use by districts.
- Maintain the alternate assessment for students with the most significant cognitive disabilities as a separate contract and continue to assess these students in the same grades and subjects as the statewide assessments.
- The State is committed to maintaining the use of the ACT for all grade 11 students and in the near term, the Task Force endorsed the use of the "pre-ACT" assessment, offered by ACT, for all grade 10 students in order to maintain the use of the growth indicator for the high school accountability system.

Design Recommendations

✓ **Assessment Specifications and Content Coverage**

- ALSDE should convene Alabama content experts to produce a "test specifications" document to make explicit decisions on content coverage and related issues.
- If more efficient, ALSDE should leverage a prospective vendor or external contractor to facilitate or manage the assessment specification and content coverage process.

✓ **Computer-Based Testing**

- Specify that the assessment will be 100% online in the first year of assessment (i.e., 2020), but support materials (e.g., test coordinator manuals, test administrator manuals, system administrator manuals, ancillaries, scratch paper, etc.) must be provided in both online using the same platform as the assessment and in paper-pencil format.
- Specify that prospective vendors should offer innovative approaches to transitioning Alabama toward 100% online testing. However, the RFB should note the possible need of supporting dual-mode administration early in the contract as a potential cost option.
- Support comparable print versions of the assessment for any emergencies or accommodations that are supported throughout the state that cannot be accomplished online.

- Require that prospective vendors have a comprehensive plan for testing technology infrastructure locally and statewide, training educators and administrators, providing opportunities to engage with the administration interface, on-demand support for test takers and administrators, and contingency plans that are reported to ALSDE well in advance of administration.
- Specify the minimum requirements for a strong online administration platform to ensure students can fairly access the content and have sufficient opportunities to practice in low-stakes settings. Additionally, the Task Force recommended that the ALSDE include performance bonds or liquidated damages requirements if certain key deliverables, milestones, or performance targets are not met in accordance to the design, development, training, and administration proposal.

✓ **Adaptive Testing**

- Specify the use of adaptive testing for the statewide summative component and possibly the interim component in Alabama's RFB. However, the RFB specifications should not determine the type of adaptive testing required (e.g., multi-stage testing vs. fully adaptive testing). Prospective vendors should be given the opportunity to propose innovative solutions that offer the benefits of adaptive testing while balancing the resource requirements for item development and calibration constraints. The specifications should ensure that the vendor's proposed adaptive solution prioritizes within grade content coverage (i.e., alignment), score precision (i.e., minimizing measurement error across the score continuum), and accessibility for all students.
- Require prospective vendors to propose solutions that describe comparison costs associated with different types of testing ranging from fixed form to adaptive. The prospective vendor should be free to suggest approaches that most efficiently address measurement and development issues while identifying opportunities for cost savings.

✓ **Item Types**

- Given the recommendation that the assessments be administered online, ensure that the prospective vendors propose the most appropriate and cost efficient item-types to adequately and accurately assess the construct and domain. That is, if the vendor proposes the use of technology-enhanced or innovative item types, it should be clear why those item types enhance the measurement the relevant grade-level standards.
- For each item type, leverage in-state educators to review the items and perhaps engage in scoring of constructed response items. Depending on an analysis of cost, the RFB may specify leveraging in-state educators to engage in item writing (for either the summative or interim assessments, if applicable).
- In order to more appropriately assess English language arts, ALSDE should include writing in each grade students are assessed (i.e., 3-8 and high school) as part of the ELA assessment.
- To develop the most cost efficient assessment system, the vendor should have the flexibility to propose a blend of human and automated scoring if and when appropriate for the item types and student responses for writing. The prospective vendor should also be required to, in detail, describe the cost/benefit associated with their scoring approach, automated scoring capacity and experience, engine training methods, adjudication rules, risk assessments. Prospective vendors should also provide any additional information that might help the ALSDE evaluate the feasibility and appropriateness of vendor scoring approaches.
- When assessing writing, the RFB should specify that ALSDE prefers to deploy a matrix-sampling approach to ensure that high-quality school-level information can be produced. Vendors should also be required to provide student-level writing subscores for each relevant prompt type.

✓ **Testing Time and Field Testing**

- Where possible and appropriate, leverage embedded field testing to develop and maintain Alabama's new assessment system.
- Target the average testing time to be about 90 minutes for each content area. Assessments that include extended writing responses should be longer with the writing component potentially treated as a stand-alone testing session.
- Explore the ability for a prospective vendor to partner with Alabama's current vendor, Scantron. This could allow the prospective vendor to embed field test items into the operational assessment in the spring of 2019

✓ **Interim Assessments and Other Cost Options**

- Include interim assessments in the RFB as the highest priority cost option. Prospective vendors should propose a set of interim assessments that are modular (see Appendix C), standards-based, and offer information that allows educators to better diagnose student strengths and weaknesses. These interim assessments can be adaptive or fixed form, but the potential vendor should justify their design approach. These assessments should leverage the same test administration platform, have a similar look and feel to the summative assessment, but have the flexibility to be used on-demand by teachers and not be required to be used by the ALSDE.
- Include non-summative support materials and resources in the RFB as the second highest priority support. Non-summative support materials might include resources that assist in the instruction of writing, classroom resources (e.g., units and tasks) for supporting the transition to three-dimensional science standards in a comprehensive way, or model performance tasks to cover multiple mathematics concepts.
- Include science at all grades in the RFB as the third highest priority.
- Include end-of-course (EOC) tests in key courses in high school in the RFB as the fourth highest priority cost-option. The Task Force noted the value of having items ready to be tested to help support this priority, but was aware of the risk of adding testing time in high school. Appropriately leveraging this priority would mean that such EOC assessments be developed to measure student learning of the identified knowledge and skills for the designated Alabama high school courses.

Implementation Recommendations

✓ **Timing**

- Target the new assessment system for a first operational administration in SY 2019-2020. The new assessment would be administered in the spring of 2020 to allow for sufficient training throughout the school year.
- Approach the deployment of a new assessment carefully to better manage development, training, operational, and political risk. The Task Force was concerned that the recent, rapid changes in state assessments has weakened the credibility of the state assessment program. Adopting the timeline described above (lower risk) can help the state and its prospective vendor address these issues and work to proactively avoid previous mistakes.
- Use of a lower-risk timeline can support intensive training efforts for a new assessment administration system. This should be coupled with a communications and professional development effort on the role of summative assessments and their intended purposes, uses, and interpretations.

✓ **Reporting**

- Require several distribution approaches for assessment reports to support different stakeholders and identify what groups of people should have access to similar reports. These reports should be made available through paper-based reports for students and their guardians, and electronically for other groups. Groupings of stakeholders should include at least the following:
    - Student/parent/guardian level reports,
    - Teacher/classroom level access (e.g., principal, educator coach, IEP teams),
    - Local administrator reports (e.g., Principle, district staff, local School Board),
    - Policy maker and legislative reporting, and
    - Public reporting (e.g., Realtors, business, media, community).
- Individual reporting should include accurate performance information displayed visually. This information should include performance from the current assessment, any prior assessments, and any additional information that can help parents or educators interpret performance trends (e.g., scores, performance levels, Achievement Level Descriptors, subscores, comparison data, externally linked information, target performance).
- Support aggregated reporting of student performance where appropriate (e.g., subgroups, growth groupings, performance level groupings, similar schools, grades, classroom, etc.).
- Ensure that assessment reporting mirrors the wider view of accountability reporting under the Every Student Succeeds Act (ESSA) and does not encroach upon interim or diagnostic testing available to schools and districts.

- Specify that the prospective vendor work with ALSDE to develop coherent performance expectations (i.e. Achievement Level Descriptors) that are communicated through intuitive reports to the public, educators, and students.

✓ **Minimum Request For Bid Requirements**

- The ALSDE should continue to work with the Center for Assessment to clearly define the "non-negotiables" that are associated with a new assessment RFB and to ensure that the evaluation process requires prospective vendors to guarantee those minimum requirements will be met.
- ALSDE should include requirements for their prospective vendor to help identify and collect sources of validity evidence, which include the evidence specified in the *Standards for Educational and Psychological Testing* (2014) and the U.S. Department of Education's Standards and Assessment Peer Review requirements.

Conclusions

This report presented a description of the work of the Alabama Assessment Task Force and the various issues deliberated by the Task Force. The report included extensive discussion of the many recommendations associated with the design and implementation of a high-quality statewide assessment system. The Task Force included and represented many stakeholders of the Alabama educational system. They spent considerable time reading, studying, and discussing critical assessment issues. They deliberated respectfully and, in almost all cases, the recommendations presented throughout this report represented a consensus of the Task Force. Adhering as closely as possible to the recommendations presented herein regarding the new Alabama assessment system will help ensure the credibility and stability of the system. Such stability is crucial for supporting advances in achievement, growth, and attainment for all of Alabama's students.

Introduction

The Alabama State Department of Education (ALSDE) sought to evaluate the current state assessment system and make recommendations for its future. To support this evaluation, ALSDE convened an Assessment Task Force comprised of key education stakeholders in Alabama and contracted with the National Center for the Improvement of Educational Assessment (Center for Assessment), a non-profit, non-partisan consulting firm to facilitate the Task Force and provide assessment expertise throughout the process. ALSDE held a series of in-person and virtual meetings with the Alabama Assessment Task Force to deliberate over many technical, policy, and practical issues associated with implementing an improved assessment system. The goal of these meetings was to establish a set of recommendations that could be used to draft specifications for a new Request for Proposals (RFP) or Request for Bids (RFB) that would be led by the ALSDE. This report presents the results of those deliberations, the subsequent recommendations to ALSDE and the Alabama State Board of Education, as well as considerations for the state's RFB. The contents of this report are based almost exclusively on consensus decisions of the Task Force. Where consensus could not be reached, decisions were based on an overwhelming majority of task force members' views. This report presents a summary of the goals of Alabama's Assessment system as well as the design and implementation recommendations.

To support this evaluation, ALSDE convened an Assessment Task Force comprised of key education stakeholders in Alabama.

Process

In October 2017, Drs. Scott Marion and Juan D'Brot from the Center for Assessment met with key stakeholders from Alabama's educational system to describe the various considerations and constraints one should consider when designing and developing a statewide assessment system. As a result of that meeting and subsequent discussions, the Center for Assessment was contracted by ALSDE to support Alabama as the state works to design and implement its next statewide assessment system.

The Center for Assessment began working with ALSDE in late November 2017 to outline the work of the State Assessment Task Force. ALSDE and the Center for Assessment agreed that the process should begin by defining the role of the assessment system and its intended uses, outlining design decisions, and considering implementation constraints. Drs. Marion and D'Brot, in collaboration with the ALSDE, facilitated the first meeting of the Alabama State Assessment Task Force on December 7-8, 2017 to obtain recommendations for crafting a RFB for Alabama's next assessment system. The Center for Assessment prepared a set of technical briefs, which are incorporated throughout this report, to help outline the critical issues associated with several key design considerations. These briefs allowed the Task Force and the Center for Assessment facilitators to more quickly to address each design consideration. The Center for Assessment then solicited feedback from the Task Force through an electronic survey on the following key assessment design considerations:

- Test development and timeline
- Paper pencil vs. computer based testing
- Adaptive vs. fixed-form testing Item types

- Writing and content coverage
- Interim assessments and balanced assessment systems
- High school testing

Upon obtaining feedback, the facilitators met with Task Force members virtually in February for a full-day webinar meeting to clarify the recommendations and design decisions proposed in the prior meetings and identified in the survey results. Drs. D'Brot and Marion used these findings and recommendations to draft a report that would serve as a basis for articulating RFB specifications. A draft report was then presented at the face-to-face Task Force meeting on March 27, 2018. During the March meeting, Task Force members further clarified previous recommendations and made suggestions for modifying the report contents. A summary of Task Force recommendations are described throughout this report and summarized in the Executive Summary.

This report presents a summary of the goals of Alabama's Assessment system as well as the design and implementation recommendations.

Types of Assessments and Appropriate Uses

Before getting into the details of the report, we define some key assessment terminology. There are several possible categorizations of assessment types, but we focus on the distinction among *summative, interim*, and *formative* assessment[1] because of the direct relevance to the Task Force's work. We define and outline the appropriate uses of the three types of assessment below. These definitions are critical to understanding what each type of assessment can and cannot do and were helpful for ensuring a shared understanding among Task Force members as the Task Force discussed the various design choices. Appendix C provides an at-a-glance summary of the typical characteristics, appropriate uses, and examples of each type of assessment.

Formative Assessment

Formative assessment, when well-implemented, should actually be called formative instruction because it is a process with the purpose of evaluating student understanding in order to provide specific to feedback to students in order to adjust instruction on a moment-to-moment basis. The Council of Chief State School Officers (CCSSO) and experts on formative assessment developed a widely cited definition (Wiley, 2017):

Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievements of intended instructional outcomes (p. 3).

The core of the formative assessment process is that it takes place during instruction (i.e., "in the moment") and under full control of the teacher to support student learning. Further, unless formative assessment leads to feedback[2] to individual students to improve learning, it is **not** formative! This is done through very frequent diagnosis of where students are in their progress toward learning goals, where gaps in knowledge and skill exist, and how to help students close those gaps. Instruction is not paused when teachers engage in formative assessment. In fact, instruction and formative assessment are inseparable.

Formative assessment is not a product, but an instruction-embedded process tailored to monitoring learning and providing frequent targeted feedback to individual students. In fact, if anything other than professional development is being purchased, it is most certainly not formative assessment. Effective formative assessment occurs frequently, covering small units of instruction, such as part of a class period. If tasks are presented, they may be targeted to individual students or groups. There is a strong view among leading formative assessment scholars that because formative assessment is tailored to a classroom and to individual students that results cannot, and should not, be meaningfully aggregated or compared (Perie, et al., 2009). There is even a strong, although not universal, view that if the assessment is scored, it is not formative. Given this close connection to instruction and locally-enacted curriculum, formative assessment is beyond the reach of the state and therefore, beyond the scope of this Task Force.

Summative Assessment

Summative assessments are generally infrequent and cover major components of instruction such as units, semesters, courses, credits, or grade levels. They are typically given at the end of a defined instructional period to evaluate students' performance against a set of learning targets for that period. The prototypical assessment conjured by the term "summative assessments" is given in a standardized manner statewide (but can also be given district- or

[1] In defining formative, interim, and summative assessment, this section borrows from three sources (Perie, Marion, & Gong, 2009; Wiley, 2017).

[2] See Sadler (1989).

nation-wide) and is typically used for monitoring, evaluation, accountability, and/or to otherwise inform policy. Such summative assessments are typically the least flexible of the various assessment types in that they require a fair amount of standardization in order to serve the intended purposes. Summative assessments also tend to require a pause in instruction for test administration. Summative assessments may also be used for "testing out" of a course, diploma endorsement, graduation, high school equivalency, and college entrance. Summative assessments, however, are not exclusively used for very high stakes purposes. In fact, many classroom assessments that are used for grading are almost exclusively summative, which may represent some missed opportunities.

They may be controlled by a single teacher (for assessments unique to the classroom), groups of teachers working together, a school (e.g., for all sections of a given course or credit), a district (to standardize across schools), a group of districts working together, a state, a group of states, or a test vendor. The level at which test results are comparable depends on who controls the assessment. They may be comparable within a classroom, across a few classrooms, within a school, within a district, across a few districts, within a state, or across multiple states. Assuming they are well-designed, appropriate uses of a summative assessment include:

- student grading in the specific courses for which it was developed,
- evaluating and adjusting curriculum, programming, and instruction the next time the large unit of instruction is taught,
- serving as a post-test measure of student learning, and
- as an indicator for educational accountability.

Interim Assessment

Many periodic standardized assessment products currently in use that are marketed as "formative," "benchmark," "diagnostic," and/or "predictive" actually belong in the interim assessment category. They are neither formative (e.g., they do not facilitate moment-to-moment targeted analysis of and feedback designed to student learning) nor summative (they do not provide a broad summary of course- or grade-level achievement tied to specific learning objectives).

Many interim assessments are commercial products and rely on fairly standardized administration procedures that provide information relative to a specific set of learning targets—although generally not tied to specific state content standards and certainly not to individual districts curricular choices—and are designed to inform decisions at the classroom, school, and/or district level. Although infrequent, interim assessments may be controlled at the classroom level to provide information for the teacher, but unlike formative assessment, the results of interim assessments can be meaningfully aggregated and reported at a broader level. However, the adoption and administration of such interim assessments are likely to be controlled by the school district. The content and format of interim assessments are very likely to be controlled by the test developer. Therefore, these assessments are considerably less instructionally-relevant than formative assessment in that decisions at the classroom level tend to be *ex post facto* regarding post-unit remediation needs and adjustment of instruction the next time the unit is taught.

Common assessments developed by a school or district for the purpose of measuring student achievement multiple times throughout a year may be considered interim assessments. These may include common mid-term exams and other periodic assessments such as quarterly assessments. Many educators refer to "common formative assessments," but these tend to function more like interim assessments.

Assuming they are well-designed, appropriate uses of a summative assessment include:

- student grading in the specific courses for which it was developed,
- evaluating and adjusting curriculum, programming, and instruction the next time the large unit of instruction is taught,
- serving as a post-test measure of student learning, and
- as an indicator for educational accountability.

This is not a negative connotation because there is tremendous transformative power in having educators collaboratively examine student work.

Finally, all of the uses of interim assessments previously described are "backward looking." There are also "forward-looking" uses of interim assessments such as a pre-test before a unit of instruction to gain information about what students already know in order to adjust plans for instruction before beginning the unit. Such forward-looking assessments may be composed of pre-requisite content or the same content as the end-of-unit assessment. A second forward-looking use of interim assessments via a placement exam used to personalize course-taking according to existing knowledge and skills. Finally, a third type of forward-looking interim assessment use is intended to predict how a student will do on a summative assessment before completing the full unit of instruction. The usefulness of this last type of interim assessment is debatable in that it is unlikely to provide much instructionally relevant information and there is often other information available to determine who is likely to need help succeeding on the end of year summative assessment.

Recommended Purposes and Uses of Assessment and Intended Outcomes of Implementation

The first major decisions of the Task Force involved specifying the goals and intended purposes and uses of Alabama's new assessment system. Assessment system design, like many engineering tasks, is a case of optimization under constraints. Therefore, it was critical for the Task Force to identify the goals and purposes to serve as the foundation from which all other recommendations are be based.

Goals for Alabama's Assessment System

In order to direct the remaining recommendations, the facilitators engaged the Task Force members in defining and clarifying the big picture goals of Alabama's assessment system. The state's assessment system must minimally support both state and federal accountability requirements. However, the Task Force members further defined the possible goals for the state's system. The members identified many potential goals of the system, but the Task Force endorsed the following goals for the Alabama Assessment System:

1. Provide a clear and credible measure of student performance on the Alabama state standards in grades 3-8 and high school;
2. Provide signals of student readiness informed by external and rigorous criteria as students move through and beyond the Alabama educational system;
3. Help improve teaching and learning by providing information useful for evaluating curriculum and instructional programs that promote improved student achievement and growth, and ultimately support the effectiveness of public education efforts; and
4. Provide stakeholders with varied, informative, and easily interpretable reports that help end-users understand student, local, and statewide educational trends against educational expectations.
5. As noted earlier, the statewide assessment system must provide information to support federal and state accountability systems.

Additionally, the Task Force members wanted to ensure that:

- The assessment system and the standards remain stable for as long as possible to facilitate monitoring state and local performance trends;
- The assessments are fair and as accessible as possible to all students;
- The assessment results are presented so they are understandable and useful to students, guardians, educators, and the public; and
- The State should consider employing a system of assessments that would be relevant and meaningful to multiple levels of Alabama's educational system.

The assessment system goals closely mirror those key themes initially raised by Task Force members. Having an assessment system that provides a clear and credible measure of state standards (i.e., is aligned) is a requirement rather than a goal, but the Task Force wanted to highlight its importance. Task Force members indicated that given the recent history in Alabama, it is important that the goal of alignment be stated overtly, in part serving as a promissory note to educators that says "if you teach the standards well, your students will have a high likelihood of success on the state assessment system."

Purposes and Uses of an Assessment System

Assessments are designed and validated to serve a limited number of purposes and uses. As much as policy makers, educational leaders, and other stakeholders want assessments to serve multiple and often far-reaching purposes, it simply cannot be done well. Following directly from the discussion of the high-level goals of an assessment system, the Task Force considered the ways in which assessment data could be used to achieve Alabama's goals. Conversations among Task Force members raised common themes within and across goals. The major purposes and uses of Alabama's assessment system are:

- Monitoring Alabama's Educational Trends
- High Quality Reporting
- Providing Information to Improve Teaching and Learning

Monitoring Educational Trends

Task Force members recognized the need for a clear and consistent set of standards and assessments to allow Alabama's educational stakeholders to track performance over time. Once a stable set of assessments is in place, monitoring progress over time should be supported through the use of high quality reports that describe performance at the state, district, school, grade, subgroup, and subject area. Further, Task Force members described the need for a system with multiple measures in order to focus the right stakeholder group on the right level of data (e.g., helping educators focus on modular interim assessments and policymakers on state-level data and trends).

High Quality Reporting

Task Force members noted the importance reporting plays in making sense of the assessment results for the public and in helping to build credibility for the assessment and larger educational systems. In addition to emphasizing the need for understandable and accessible reports, Task Force members also recommended customizing reports for different audiences that include tailored information that clearly explains the information report is trying to communicate and suggests helpful next steps (either in behavior or for additional reports to explore). This theme is described in more detail under the Reporting Recommendations section.

Providing Information to Improve Teaching and Learning

The Task Force noted that in addition to trend monitoring and reporting, the information provided by an assessment system (rather than a singular assessment) must be designed to support improvements in both teaching and learning. That is, the group recognized that the results from summative end-of-year assessments provide verification of achievement, but results from other assessments – for example, something like an on-demand interim module could be used to gauge student progress throughout the year. Furthermore, Task Force member statements reiterated the need to help the public understand the differences among the various types of assessment and how each assessment type would require its own design and development specifications.

The major purposes and uses of Alabama's assessment system are:

- Monitoring Alabama's Educational Trends
- High Quality Reporting
- Providing Information to Improve Teaching and Learning

Overview of Alabama's Assessment System

We present an overview of the proposed Alabama assessment system in the following section. The Task Force read about and discussed balanced assessment systems and considered the extent to which ALSDE should try to procure key aspects of a potentially balanced system (see Appendix B for a more complete discussion). We first discuss the assessments that are included in the proposed system and therefore included in this report. We also briefly indicate the assessments that are not included in this report.

Assessments Included in the Proposed System

The Task Force spent most of its time discussing the grades 3-8 and high school assessment system for English language arts (ELA) and mathematics. The Task Force also discussed the Alabama science assessment, but understandably did not spend as much time discussing it as they the remainder of the RFB. The Task Force spent the majority of its time deliberating the summative assessment design for ELA and mathematics, but also spent time discussing the various options for an optional interim assessment component. We address these topics later within the *Potential Optional Considerations for an Expanded System of Assessments* section.

High school assessment was the focus of considerable attention for the Task Force. The State is committed to maintaining the use of the ACT for all grade 11 students and, for the near term, the Task Force endorsed the use of the "pre-ACT" assessment, offered by ACT, for all grade 10 students in order to maintain the use of the growth indicator for the high school accountability system. That said, the Task Force also explored a limited set of end-of-course tests for ELA and mathematics to ensure that the high school assessment system is most relevant to students and to meet the Task Force's alignment design requirement. Recommendations on end-of-course assessments are presented later in this section of the report and will be used to inform optional specifications for the RFB.

One of the most critical decisions about assessment design is determining the content to be assessed. It sounds intuitive to say that the assessment should just measure the standards, but unfortunately, it is not that simple. There are too many standards to assess in a reasonable amount of time and the standards are generally too large of a grain size to effectively guide assessment design. The Task Force endorsed the idea of a separate process to help specify the scope and grain size of the assessable standards, which is described in *Assessment Specifications and Content Coverage* section.

The Task Force provided the following recommendations regarding assessments included in the Alabama assessment system:

- Maintain the alternate assessment as a separate contract and continue to assess students with significant cognitive disabilities in the same grades that the state administers general statewide assessments.
- Include the science assessment as part of the Alabama assessment system RFP.

High school assessment was the focus of considerable attention for the Task Force. The State is committed to maintaining the use of the ACT for all grade 11 students and, for the near term, the Task Force endorsed the use of the "pre-ACT" assessment, offered by ACT, for all grade 10 students in order to maintain the use of the growth indicator for the high school accountability system.

Additionally, the science assessment must be aligned to the state standards and assessment evidence should document that it appropriately assesses all three dimensions of science—disciplinary core ideas, science and engineering practices, and cross-cutting concepts—in an integrated manner.

- For science, develop and administer an aligned assessment for science at the end of each grade span, which include grades 3-5, grades 6-8, and high school.
- For mathematics and English/language arts, develop and administer an aligned assessment system for the elementary and middle school grades for grades 3-8.
- Continue to use the ACT (or other college entrance exam) in 11th grade and the pre-ACT in grade 10.

Assessments not Addressed in this Report

Alabama has been administering the Alabama Alternate Assessment (AAA), an alternate assessment system based on alternate achievement standards for students with the most significant cognitive disabilities. The Task Force did not include recommendations related to the alternate assessment in this report. First, Alabama stakeholders are satisfied with the AAA and, second, the Task Force did not include enough expertise in alternate assessment in order to make appropriate recommendations. As noted in the recommendations in the previous section, Alabama plans to maintain the AAA for the foreseeable future.

ALSDE requires the administration of the *Access for ELLs 2.0* developed by the World-Class Instructional Design and Assessment (WIDA) to assess English language proficiency achievement and progress for students identified as English language learners. ALSDE plans to maintain its membership in the WIDA consortium and continue to administer the Access for ELLs 2.0 for its ELL students and therefore, this topic is not addressed in this report.

Finally, the Task Force is aware that Alabama is interested in exploring the development and implementation of an assessment system for students in Kindergarten through second grade that connects to the assessments offered in grades 3-8. However, given the compressed timeline for the Task Force to operate, the need to prioritize the statutorily-required assessments, and the lack of early childhood assessment expertise on the Task Force, the Task Force did not address early childhood assessments.

The Task Force provided the following recommendations regarding assessments included in the Alabama assessment system:

- Maintain the alternate assessment as a separate contract.
- Include the science assessment as part of the Alabama assessment system RFP.
- For science, develop and administer an aligned assessment for science at the end of each grade span, which include grades 3-5, grades 6-8, and high school.
- For mathematics and English/language arts, develop and administer an aligned assessment system for the elementary and middle school grades for grades 3-8.
- Continue to use the ACT (or other college entrance exam) in 11th grade and the pre-ACT in grade 10.

Key Design Considerations and Requirements

After addressing higher-level considerations, the Task Force began to address many operational decisions in the form of design considerations, requirements for the assessment system, and specifications for the RFB. The Task Force, in considering assessment design, had the opportunity to learn about and discuss the implications of principled approaches for designing assessments. Therefore, this section begins with a brief discussion of Principled Assessment Design and the Role and Timing of Assessment, before describing the Task Force's key design recommendations for the Alabama Assessment System.

Introduction to Principled Assessment Design

States across the country have focused their standardized, large-scale assessment development efforts on tests that help us understand whether students are on track or ready for post-secondary endeavors (e.g., 2-year colleges, 4-year colleges and universities, gainful employment). Assessment developers have had to ensure they attend to the inclusion of longer-term claims in their design. One way this can be addressed is through using a principled approach to assessment design, such as Evidence Centered Design (ECD; Mislevy & Haertel, 2006) or through the use of the Assessment Triangle (NRC, 2001), which draws a connection among observation (what we ask), interpretation (how we make sense of their response), and cognition (what they should know) through assessment. Marion and Landl (2017) pose several questions that task developers should consider that are also germane to the assessment development endeavor:

- What claims do we want to be able to make about what students know and can do?
- What knowledge and skills comprise the learning target(s) we are intending to measure?
- What evidence is necessary to demonstrate that a student has mastered those knowledge and skills?
- What type of task will serve to elicit that evidence?
- What characteristics/features will make a task harder or easier?
- What characteristics/features will make a task more or less complex?

Although it seems obvious that test developers would consider these questions during design, newer and more complex assessments have made these questions an explicit part of assessment design. Throughout the work of the Task Force, these types of questions were raised in support of larger questions about goals, purposes, uses, and claims associated with the Alabama Assessment System. The facilitators extended the application of this approach to clarify how assessment operations should be defined.

The remainder of this section dives into operational considerations that emerge from response to the types of questions posed above. It addresses topics like the timing of assessments, the assessment development process, the testing administration mode (and how it relates to the depth and complexity of questions that can be posed), item deployment and field testing, and recommendations on adaptive and fixed form tests.

The Role and Timing of Assessments in Relation to Standards and Instruction

Throughout conversations with the Assessment Task Force, the facilitators defined and described the assessments types and uses presented here to ensure members had a shared understanding of assessment. To address the charge of the Task Force, the members primarily focused on the role and uses of summative assessments—specifically, the state summative assessment for accountability. However, the Task Force discussed ways in which interim assessments could be used to support progress towards meeting requirements described by the standards, which are measured through the state summative assessments. Thus, the Task Force spent some time discussing the role and timing of these assessments in the educational system.

State-wide summative assessments are, by design, backward looking so that such assessments are unable to provide instructionally useful information for the students taking the test. On the other hand, well-aligned and well-constructed assessments can provide information to help evaluate programs and monitor academic progress over time. Therefore, summative assessments can provide information useful for improving the education of "next year's" students. This process is described in more detail in the Implementation section of this report (i.e., Timing subsection).

Key Design Recommendations

After considering how to approach assessment design and understanding where assessments fit in the learning process, the facilitators helped the Task Force navigate a series of key design issues. These design issues were presented to help the Task Force understand the technical and practical implications of design specifications and how they may be operationalized in a RFB or summative assessment. **This section describes Task Force recommendations for the topics of content coverage, computer-based testing, item types, field-testing, and adaptive testing.**

Assessment Specifications and Content Coverage

Alignment to state standards was one of the most important goals articulated by the Task Force. Task Force members indicated that it was critical that the new state assessment accurately reflect the standards that teachers are expected to teach and students are expected to learn. However, what does alignment really mean? The standards include such things as listening, speaking, and research, but when asked, Task Force members acknowledged that there was little interest in trying to assess such learning targets with a statewide summative assessment. We use this example to make the point that all assessments require choices about what will and will not be included on any given assessment. The Task Force recommended not including listening or speaking on the state summative assessment even though both domains are represented in the state content standards. This exclusion is quite common among states. In fact, few if any states currently assess listening and speaking on their state assessments. On the other hand, even though a few states are moving away from assessing writing at every grade level, the Task Force strongly recommended including writing at every grade level where reading and mathematics are assessed. We discuss the possibilities for assessing writing later in this report. In addition, the Task Force recommended the following with regard to content coverage recommendations:

- The ALSDE should convene content experts to produce a "test specifications" document to make such design decisions explicit. The Task Force emphasized that they wanted these decisions made by Alabama experts.
- If more efficient, ALSDE should leverage a prospective vendor or external contractor to facilitate or manage the content specification process. The Task Force recognized that it was beyond their scope to make all of the necessary content decisions and that the content specifications are a critical aspect of a technically and practically defensible assessment system.

Alignment to state standards was one of the most important goals articulated by the Task Force. Task Force members indicated that it was critical that the new state assessment accurately reflect the standards that teachers are expected to teach and students are expected to learn.

The Task Force strongly recommended including writing at every grade level where reading and mathematics are assessed.

Computer-based Testing

Considerations for choosing between paper and pencil testing (PPT) and computer-based testing (CBT) are not limited just to administration experience and technology capability in schools and districts. While both user experience and technological capacity are usually at the fore, the list of considerations for any state deliberating the mode of administration (PPT vs. CBT) also includes:

- Administration monitoring

- Test security and analyses
- Field test administration and design
- Scoring
- Comparability between modes

While this list is not exhaustive, these issues will dictate whether a state would support both PPT and CBT or choose a single administration method, which in turn will influence the cost of the assessment system. Generally, states should expect that dual mode administration (i.e., supporting both PPT and CBT) will be considerably more expensive than supporting either mode alone. The facilitators raised a series of issues for the Task Force around mode issues, their complexity, and associated questions that are described below.

Paper and Pencil Testing (PPT)

Paper and Pencil Testing (PPT) uses paper for both the stimulus (e.g., test booklet) and response (e.g., score sheet). PPT offers an opportunity for easier administration, fewer technological considerations, and less perceived stress, the latter of which is typically attributed to those administering or supporting the test. However, it also presents multiple challenges that limit the types of items that can be administered, reduce speed and efficiency of scoring, eliminate flexible approaches to field testing new items, and complicate the logistics of delivery, packing, and shipping. Furthermore, PPT usually eliminates the possibility of online manuals (e.g., test coordinator, test administrator, system administrator, etc.), ancillaries, graphics, and the like. This last point can be seen as either a benefit or a cost, depending on the case. Additionally, the cost associated with PPT administration is recurring and based on printing needs, pack and ship costs, physical scanning, warehouse space, and long-term storage.

Computer Based Testing (CBT)

The application of computer based testing (CBT) can vary, but for our purposes, we define CBT to be where both the stimulus (e.g., test item) and the response (e.g., item response) are delivered and captured on an electronic device (e.g., desktop, laptop, tablet, etc.). While CBT has greater startup costs (e.g., infrastructure, hardware, and software), these onetime costs are defrayed across the lifespan of the assessment program. In addition, proponents of CBT argue that the tools used for testing, such as the online platform, should be used for instructional delivery and student learning, mitigating some of the initial investment.

CBT offers several opportunities for efficiencies in delivery, administration, field testing, scoring, security, and reporting. Notably, CBT allows for adaptive testing (i.e., adjusting test difficulty to the ability of the student). However, the primary challenges with CBT revolve around the number of available devices, scheduling time on those devices, system readiness (e.g., test administration system installation, system stress tests), and relevant training for educators (e.g., technology surveys, site readiness, administrator training). Like PPT, there are recurring costs. These costs are typically associated with help desk support and any annual printing of support materials. If, however, everything were supported through online administration and documentation, the initial cost would be associated with development of print-ready publication, potentially decreasing the overall costs for the life of a contract. It is important to note that these costs are usually shifted to districts and schools if they desire to have print-based resources.

Level of Complexity

As described in the two previous sub-sections, PPT and CBT each come with their own issues that vary in the challenges posed. We should keep in mind that supporting both modes can potentially negate the benefits of providing only one type of administration, as well as increasing the complexity of managing the program and assessment. The reality is that many states that have implemented CBT have supported dual mode assessment and addressed issues as they emerge. The following table outlines some of the issues and their corresponding complexity for each mode.

TABLE 1. COMPLEXITY OF TESTING ISSUES BY MODE.

TOPIC	PAPER	ONLINE	DUAL MODE
Administration	**Low.** A content area can typically be administered in a single day, limiting the size of test windows.	**Medium.** Limited to the number of devices, often requiring larger windows (and potentially full keyboards).	**High.** Requires managing both types of administration.
Design and Field Testing	**Medium.** Relatively restrictive field test designs, unless there is a heavy investment in administration time.	**Low.** Much easier to embed items to support comparability studies and linking item sets during field-testing.	**High.** Potentially eliminates the flexibility of online field test administration and could over-/under-represent certain schools or districts.
Scoring	**Medium.** Turn-around based on collection and pack/ship dates.	**Medium.** Turn-around based on size of testing windows.	**High.** Requires managing both physical and digital administration conditions and introduces examining differences in mode.
Item Types	**Low.** Limited to non-technology enhanced items.	**Medium.** Provides access to many more item types, but requires a strong rationale for their use (i.e., not using enhancements simply because they are available).	**High.** PPT operates as the driver for item type selection. Any dual mode items have to be examined for administration differences.
Comparability	**Medium.** Issues are limited to accommodated forms, but it is difficult to track administration conditions without a strong policy.	**Medium.** Should be studied by device type, screen size, and peripherals. Administration choices (e.g., tools or accommodations) are easier to monitor.	**High.** Requires careful analysis of mode effects and how it may affect student scores.
Test Security	**High.** Test security can be more challenging due to physical test copies and fewer available analyses to monitor irregularities.	**Medium.** Test security can be less problematic due to electronically-mediated methods of delivery and more robust monitoring and analyses.	**High.** A strong rationale is required to select the appropriate analyses and monitoring techniques, especially if they differ by mode.
Cost	**Medium.** Typically front-loaded in delivery and spread over the cost of the contract; annually based on printing.	**Medium.** Similar to PPT, but also based on the level of support necessary and need for scalability of support (e.g., help-desk support).	**High.** Supporting both PPT and CBT will require factoring the challenges from both modes.

Computer-based Testing Recommendations

The Task Force reviewed these mode-related issues and engaged in a discussion focusing on statewide technology readiness and the claims an assessment could support if administered online or on paper. The fact that the State is testing fully online during this interim period helped convince many Task Force members that Alabama was indeed capable of implementing a fully online system. Based on those conversations, the Task Force made the following recommendations:

- **Specify that the assessment will be 100% online in the first year of assessment (i.e., 2020), but provide support materials (e.g., test coordinator manuals, test administrator manuals, system administrator manuals, ancillaries, scratch paper, etc.) both digitally and via paper.**
- Specify that prospective vendors must offer innovative approaches for moving Alabama toward 100% online testing as a cost option to help the ALSDE understand possibilities if there is a need to support dual-mode administration early in the contract.
- Support comparable print-versions of the assessment for any emergencies or accommodations that are supported throughout the state.
- Require that prospective vendors have a comprehensive plan for testing technology infrastructure locally and statewide, training educators and administrators, providing opportunities to engage with the administration interface, on-demand support for test takers and administrators, and contingency plans that are shared with the state well in advance of administration.
- Specify the minimum requirements for a strong online administration platform to ensure students can fairly access the content and have sufficient opportunities to practice in low-stakes settings. Additionally, the Task Force recommended that the ALSDE include performance bonds or liquidated damages if certain key deliverables, milestones, or performance targets are not met in accordance to the design, development, training, and administration proposal.

There was not consensus initially regarding the online administration recommendation. However, the Task Force opinions converged toward supporting 100% online administration (with appropriate paper-based accommodations as needed) for the first year of the assessment. A key theme that emerged was that a more reasonable implementation date (i.e., 2020) for the new assessment would support better and more comprehensive training and stress testing on a new assessment administration platform. This recommendation is described in more detail under the Implementation section of this report, and in particular the, *The Role and Timing of Assessments in Relation to Standards and Instruction* subsection.

Adaptive Testing

With a recommendation to include online testing, the Task Force had the opportunity to consider adaptive testing. Adaptive tests may help support the Task Force's desire to have a test that is accessible, responsive to student to needs, reduces issues related to motivation, minimizes testing time, and maximizes both precision.

Test developers, practitioners, and educators are often excited when faced with the possibilities associated with computer based testing (CBT). One such possibility includes that of large-scale computer adaptive testing (CAT). Some might think that adaptive testing is a new invention, but we know they have been used for at least 100 years. The Binet IQ test[3] (now known as the Stanford-Binet) is a well-known fully adaptive test (i.e., examinees receive different questions based on their prior correct or incorrect responses). Through the use of modern testing methods and CBT, adaptive tests can be administered simultaneously to students across a state (and in some cases, across a country).

[3] Binet, A., & Simon, Th. A. (1905). Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. L'Année Psychologique, 11, 191-244.

Features of Adaptive Tests

Adaptive tests differentiate themselves from fixed-form tests by providing examinees with different questions depending on how they respond to test items or sets of items. However, all tests—fixed or adaptive—must adhere to certain technical requirements. These include reliability, fairness, and validity[4]. These concepts are described in more detail in relation to summative testing below.

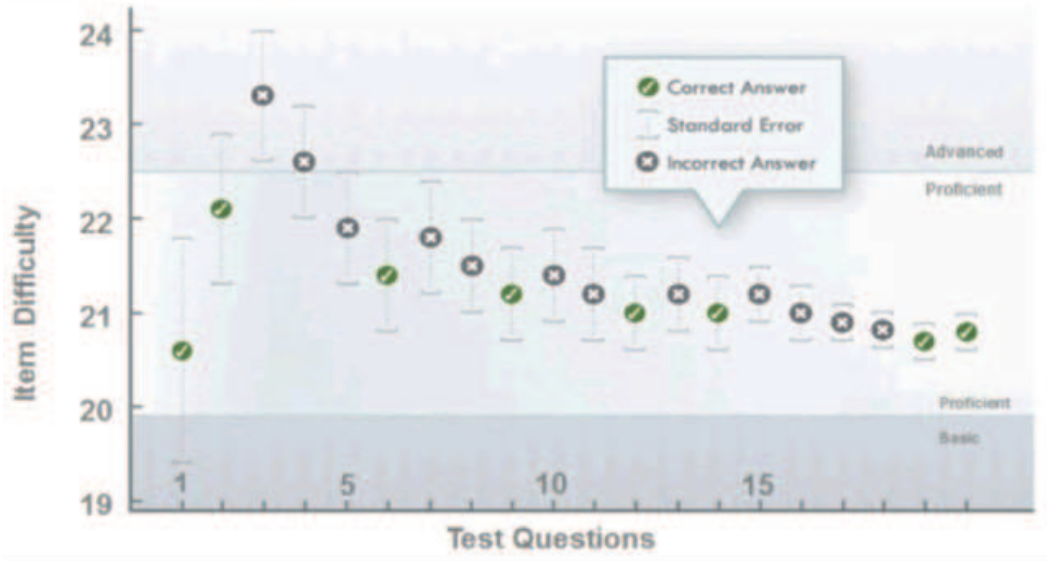
- **Reliability:** Reliability is an acknowledgement that a single test represents a sample of test questions from all possible questions that could be asked, scorers (or open-response questions) from all possible scorers, and so on. A reliability coefficient is a quantification of the consistency of a test score and must be interpreted in light of *the intended use of that test.*
- **Fairness:** Fairness emphasizes that the test must be fair, accessible, and appropriate for all individuals in *the intended population for the intended use of that test.*
- **Validity:** Validity refers to the degree to which evidence and theory support the interpretations of test scores *for the intended use of that test.*

As we can see in the descriptions provided above, all arguments about reliability, fairness, and validity are based on the intended use of the test. Fixed-form tests for accountability prioritize content coverage, sufficient reliability and precision to make claims about proficiency, fairness about students taking the test, and generalizability of claims across students. Adaptive tests for accountability seek to make the same claims, but have a greater ability to be more precise, more efficient, and more targeted with appropriate sets of items for higher and lower performing students.

Types of Adaptive Tests

Adaptive tests "adapt" to an estimate of a student's achievement by providing more or less difficult items based on his or her responses. Furthermore, adaptive tests tend to reduce barriers to motivation associated with test takers receiving items that are too difficult or too easy. However, the degree of adaptivity offered by CAT differs based on the resources dedicated to development, and in particular how many items are available for use (i.e., the size of the item pool). A general conceptualization of CAT is provided in the figure below. In this figure, if a student answers an item correctly, he or she is given a more difficult item, and vice versa, until a sufficiently accurate judgment about the student's achievement can be made.

FIGURE 1. COMMON CONCEPTUALIZATION OF AN ADAPTIVE TEST[5]



[4] AERA, APA, & NCME, & Joint Committee on Standards for Educational and Psychological Testing. (2014). Standards for educational and psychological testing. Washington, DC: AERA.

[5] From http://www.ascd.org/publications/educational-leadership/mar14/vol71/num06/The-Potential-of-Adaptive-Assessment.aspx

In reality, adaptive tests can vary in the level of adaptivity significantly. Three common adaptive approaches include (1) linear on the fly testing, (2) multi-stage testing, and (3) computer adaptive testing. These three approaches are increasingly adaptive in nature - and as adaptivity increases, so does the amount of required resources. These resources include, but are not limited to an increased item pool, immediately scoreable items, increased research capacity to simulate CAT administrations, a CAT delivery system, and appropriate software to account for additional analyses associated with CAT. The three sample types of adaptive testing are described in further detail below.

1. **Linear on the fly testing (LOFT):** Equivalent test forms, based on pre-determined constraints (e.g., content categories) are selected for each student from a large item pool. LOFT forms are not technically adaptive, but offer many benefits over fixed forms, especially in terms of security.

2. **Multi-stage testing (MST):** Pre-determined forms are adapted to the student at pre-determined stages (e.g., after 15 items or after a cluster of topic-specific items). After an initial stage, students are routed to forms (nodes) of varying difficulty in subsequent stages based on performance in the previous stage. Typically no more than three stages are employed.

3. **Computer adaptive testing (CAT):** Also known as item-level CAT, such an approach creates fully individualized forms (essentially) for each student by presenting each item based on answers to the previous items. CAT produces the most precise and potentially the shortest test. If done well, it minimizes the exposure of items more than other types of adaptive testing. However, it requires the most investment and the largest pool of items with appropriate ranges of difficulty and complexity. Further, if alignment requirements must be strictly met, item-level CAT loses much of its test length efficiency over MST and even fixed form.

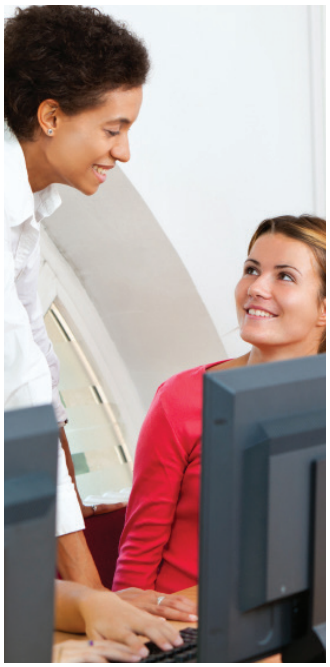
The benefits of CAT are maximized when CBT is supported fully throughout a state. It is possible to support comparable PPT and CAT scores through certain field test designs and administration approaches; however, the constraints are numerous and costly. Supporting a dual mode assessment system with CAT requires an in-depth field test design, a robust research agenda, longer administration windows, a larger budget, and an extensive support plan for training and help-desk access.

Recommendations on Adaptive Testing

The Task Force reviewed these adaptive testing-issues and engaged in discussions focusing on technology readiness, student motivation, adaptivity, and resource requirements to support different types of adaptive approaches. Based on their review of readings and discussions, the Task Force members made the following recommendations regarding the use of adaptive tests:

- The Request for Bids (RFB) should require the use of adaptive testing in Alabama. However, the RFB specifications should not determine the type of adaptive testing required (e.g., multi-stage testing vs. fully adaptive testing). Prospective vendors should be given the opportunity to propose innovative solutions that offer the benefits of adaptive testing while balancing the resource requirements for item development and calibration constraints. The specifications should require the proposed adaptive solution to prioritize within-grade content coverage (i.e., alignment), score precision (i.e., minimizing measurement error across the score continuum), and accessibility for all students.

The Request for Bids (RFB) should require the use of adaptive testing in Alabama. However, the RFB specifications should not determine the type of adaptive testing required (e.g., multi-stage testing vs. fully adaptive testing).



- The RFB should require prospective vendors to propose solutions that describe and compare costs associated with different types of testing ranging from fixed form to adaptive. The prospective vendor should be free to suggest approaches that most efficiently address measurement and development issues while identifying opportunities for cost savings.

- If the state of Alabama includes an interim assessment, any mini-summative versions should be adaptive (this is described in greater detail in the interim assessment section of this report). If possible, modular diagnostic interims may be adaptive if prospective vendors can offer innovative solutions to provide increased amounts of information to educators on student progress against the State's standards.

Item Types

If you want to know what a test measures, look at the items! While this oft-repeated axiom in education is a bit of an over-simplification, it is more true than not. Items and tasks are the tools that elicit student responses which in turn support inferences about what students know and can do. The information produced from test items is the foundation of a validity argument – the argument that organizes the evidence and theory supporting the interpretation(s) of test scores. Therefore, the quality of test items and tasks builds or detracts from the credibility of the assessment system in the eyes of students, educators, parents, and the public. Importantly, test item development is one of the major cost drivers of a state testing program; so in addition to the primary focus of item/task quality; ALSDE must focus on obtaining and maintaining item quality as efficiently as possible. This section of the report discusses the following:

- An overview of the types of items and tasks that can be included on a summative test
- The opportunities and challenges associated with each of the commonly used item types
- Considerations for how to balance the tradeoffs

Overview of items and tasks

Large-scale test items historically have been classified into two very broad categories of items: selected- and constructed- response. Selected-response includes the ubiquitous multiple-choice item, but can also include a variety of related item types or arrangements such as item clusters and evidence-based selected-responses (two-part multiple-choice items). Constructed-response items or tasks can range from very short responses of a few words to multi-hour or even multi-day activities. These "extended" constructed-response tasks share many features with performance-based tasks, but only a few states include extended performance tasks on end-of-year state assessments due to time requirements and cost. With the advent of computer-based testing, we have seen a new class of items, often referred to as technology-enhanced items. The common feature of such items is that they rely on the digital environment to support interactions among students and the content in ways that are not possible on paper. This is an area that is rapidly developing and holds tremendous promise for improving the ways in which we measure student learning.

Opportunities and challenges

Test design is an exercise in optimization under constraints. The same is true for item development. Every choice involves considerable tradeoffs. The name or class of item means less than what the responses to the item tell us about student performance. We must keep in mind the following questions as we consider our choices:

- What are we trying to measure?
- How will this type of item help us to measure these learning targets well?
- What is a close enough approximation to what we really want to measure?
- What resources are available?
- Will the assessment be given solely on computer or split between computer and paper/pencil?
- What are the potential intended positive and unintended negative consequences associated with our choice of items?

In the table that follows, we highlight the opportunities and potential shortcomings with the various item types.

ITEM TYPE	OPPORTUNITIES	CHALLENGES
Multiple-choice	Multiple-choice items have a long track record of success and efficient use. The student is presented with a prompt and is asked to select from among 4-5 response options (generally). The field has developed robust measurement models for scoring, scaling, and evaluating multiple-choice items and they are able to generate a considerable amount of "measurement information" quite efficiently. Multiple-choice items are often presented to students as being independent of one another, but they are also grouped as clusters or testlets around a scenario or reading passage.	The major challenge with multiple-choice items is that they are limited in the complexity of thinking they can elicit from students. While the field has generally advanced beyond populating tests with items that call on rote memory, many multiple-choice items still rely on factual and procedural information. Consequentially, many are concerned that if the accountability test is populated with multiple-choice items, teachers may try to mimic such approaches in their classrooms at a cost to deeper learning.
Evidence-based selected-response	These items are essentially two-part multiple-choice items where the student answers a multiple-choice item, but then answers a second item in the pair to "explain" their original answer. Such items have considerable promise for going beyond the generally lower levels of thinking called for in typical multiple-choice items.	These items have been used on PARCC and Smarter Balanced and once some of the kinks were worked out, they have been somewhat effective. The scoring rules associated with the item type are still tricky (e.g., does the student have to get the first item right in order to get the second one correct?) and the scoring and scaling (creating and maintaining score scales) is less straightforward.
Short constructed-response	Short constructed-response items often ask students to generate a written response that is generally a paragraph or less or to solve a fairly straightforward problem in mathematics. When designed well, such items are very effective at generating complex thinking from students that goes beyond multiple-choice items.	These items can cost considerably more than selected response items to score if they have to be scored by a human rater. They require more testing time than multiple-choice items and tend to generate fewer points (test information) per minute than multiple-choice items. Some short constructed-response items can be scored effectively by computer, but most cannot at this point, especially if they call for the student to generate content-specific responses or use specific evidence from text or other sources.
Extended constructed-response[6]	Extended constructed response tasks are best at probing strategic and deep thinking by students. They generally require between 30-90 minutes each. They are often the most authentic types of items because they can better draw on real-world scenarios or problems than other types of items and tasks. Importantly, such extended-response tasks send a powerful signal for the types of activities that we would like to see teachers use in their classrooms.	Extended-response items are expensive to score, except in cases where writing responses can be scored efficiently by computer (becoming more prevalent). As the name suggests, such items require considerable time and including such tasks on a test can greatly increase testing time. Finally, because such tasks can be memorable and time consuming, they pose challenges for field-testing and year to year comparability.

[6] We discuss "short constructed" and "extended constructed" response items/tasks as if there is a dichotomy. There is not! There are many item types that would fall somewhere in the middle, but we focus on the short and extended for simplicity.

ITEM TYPE	OPPORTUNITIES	CHALLENGES
Technology-enhanced items	As the name implies, these items rely on the technology platform to enhance the interactions between the student and the content. The field is still new, but progressing rapidly. Early versions of TEIs were not much more than video clips embedded in multiple-choice items, but newer items allow for sophisticated simulations that require students to think deeply in order to respond to the question. These items offer considerable promise for advancing our measurement capacity in a cost-efficient manner.	Technology-enhanced items require students to be testing in a digital environment. The more "enhanced" the item, the harder it is to create a "paper clone" and, therefore, the greater the threats to comparability if paper testing is still used. Further, the field is still learning about the scoring and scaling of innovative item types and how such items contribute to our understanding of what students know and can do. Another risk with TEIs is that many schools do not have enough of a digital footprint to allow students the opportunity to learn in a digital environment. Therefore, the only time many students experience such procedures and approaches in on the state test.

Wrestling with tradeoffs

The Task Force recommended using a balance of item types so the state can capitalize on the advantages of each type while trying to minimize the unintended negative consequence or other risks of the item type. The Task Force recognizes that finding that balance is the real challenge.

The Task Force reviewed the various item types that are currently used in large-scale testing. They also discussed the differences between each item type when presented in a CBT vs. a PPT format. The task force offered the following recommendations regarding item types:

- The Task Force recommended including some proportion of open-response item types on the summative assessment.
- The Task Force encouraged the State to consider exploring technology-enhanced items to the degree that digital capacity and digital opportunities for learning can be expanded.
- Given the CBT recommendation, ensure that the prospective vendors propose the highest quality and measurement appropriate item-type to adequately and accurately assess the construct and domain while focusing on cost efficiency. That is, if the vendor proposes the use of enhanced-technology or innovative item types, it should be clear why those item types enhance the measurement of the relevant grade-level standards.
- For each item type, leverage in-state educators to at least review the items and potentially engage in scoring of constructed response items. Depending on a cost analysis, the RFB may specify leveraging in-state educators to engage in item writing (for either the summative or interim assessments, if applicable).
- Specify that the prospective vendor has the flexibility to propose blends of human- and automated-scoring for relevant item types and indicate how these approaches would lead to cost savings for the state of Alabama while maintaining scoring quality. This recommendation is described in more detail in the next section of this report addressing the assessment of writing.

The Task Force recommended including some proportion of open-response item types on the summative assessment.

The Task Force encouraged the State to consider exploring technology-enhanced items to the degree that digital capacity and digital opportunities for learning can be expanded.

Assessing Writing

Many states rely on extended constructed-response tasks as the primary way to assess writing. If we want to measure writing achievement, it makes sense to have students write. Including direct writing on state assessments has been shown to increase the amount of writing that students do in classrooms, at least in the grades where writing is assessed. Further, newer approaches to writing tasks that call on students to frame arguments based on evidence from reading stimuli, rather than the "imaginary narrative" prompts of the past, can help incentivize such practices in classrooms.

However, there are measurement challenges associated with the use of a single writing prompt. Even though student response times can range from 30-90 minutes, the score from the single writing task contributes very little "test information" to an overall English language arts score. Additionally, there are known challenges with the generalizability of the results from a single writing prompt. In other words, since prompts are often not directly comparable (e.g., address different topics, reference different sources of evidence) and students perform differentially on various prompts, it is hard to support valid inferences about individual student writing achievement based on a single prompt. The solution to this problem—administering two or more writing tasks to each student—is not often practically feasible due to the increased testing time required.

Given these challenges, the Task Force wrestled with several options for assessing writing in a meaningful way. The Task Force first discussed whether to prioritize student- or school- level scores. Focusing on school-level information does not mean that the state is giving up on student-level scores, but it might mean that it will tolerate somewhat lower levels of student comparability in order to get more information at the school level. For example, the state could administer multiple writing prompt (at least three and probably not more than eight) at the school level with each student completing only one prompts. This means that students would be completing different prompts, but with writing as just one component of ELA, student-level score comparability will not be compromised to any great degree. On the other hand, the multiple prompts at the school level will produce robust information on writing performance at the school level that may support writing subscores (e.g., by genre).

When student-level scores are the priority, there are ways to gather more "information" than can be gathered through a single writing response. The state can choose to include a few short constructed-response writing tasks throughout the test in addition to the extended-response task to noticeably increase the amount of writing information generated by the test. Some states and consortia have explored measuring both reading and writing with these shorter writing tasks and while this sounds intuitively sensible, it has proved challenging because of phenomena such as "halo effects" (e.g., students getting the same score on both reading and writing).

Given these considerations, the Task Force offered the following recommendations regarding the assessment of writing:

- In order to more appropriately assess the construct of ELA, ALSDE should include writing in each grade students are assessed (i.e., 3-8 and high school) as part of the ELA assessment.
- To develop the most cost efficient assessment system, the vendor should have the flexibility to propose a blend of human and automated scoring if and when appropriate for the item types and student responses for writing. The prospective vendor should also be required to, in detail, describe the

In order to more appropriately assess English language arts, ALSDE should include writing in each grade students are assessed (i.e., 3-8 and high school) as part of the ELA assessment.

To develop the most cost efficient assessment system, the vendor should have the flexibility to propose a blend of human and automated scoring if and when appropriate for the item types and student responses for writing.

costs and benefits associated with their scoring approach, automated scoring capacity and experience, engine training methods, adjudication rules, risk assessments, and ways in which to support educator training on any writing rubrics that are used to score student responses. Prospective vendors should also provide any additional information that might help ALSDE evaluate the feasibility and appropriateness of vendor scoring approaches.

- When assessing writing, the RFB should specify that ALSDE prefers to deploy a matrix-sampled (i.e., multiple prompts at the school level, but each student completing only one prompt) set of writing prompts to better collect school-level information. Vendors should also be required to provide student-level writing subscores against each relevant prompt type.

Testing Time and Field Testing

Testing time is a hotly debated topic in large-scale testing. What is the maximum amount of time that a student should spend taking a standardized testing to reflect their grade-level mastery? Arguments for and against longer tests are prevalent and include authenticity (i.e., requiring longer tests) and minimizing interruptions to instruction (i.e., shorter tests). While testing time is an important factor to consider for an administration, it is important not to confuse the timing and length of a single end-of-year summative assessment (typically a very small percentage of available instructional time) and the timing and length of the assessments required by schools and districts. Depending on the dual state and district requirements, the number of tests can be much larger than expected. Thus, test length is a key consideration in test design.

Factors that Influence Test Length

There are several factors that influence test length, but not all are addressed in this report. As with most aspects of test development, there are tradeoffs associated with each of the following factors:

- Content coverage
- Item types*
- Desired reliability
- Subscore reporting
- Adaptivity
- Field test design*
- Public perception

For the purposes of this section, only those concepts with asterisks will be discussed. The remaining factors are covered throughout this report.

Item Types

We will only cover item types briefly, as they are a primary focus in the previous sub-section of this report. However, item types are a major driver of test time. The test will be longer if an assessment calls for inquiry-based tasks or performance tasks that seek to measure knowledge construction, synthesis, and problem-solving. The trade-off to consider is whether the additional information gained from the assessment justifies the increased amount of time students spend on that particular task. Often, an in-depth task will include multiple pieces of evidence, multiple questions, and multiple responses. Thus, an in-depth task should yield greater amounts of information. However, a very well-specified task could take multiple sessions or days to fully answer. These tasks have been used successfully in the past but require buy-in across the spectrum from the classroom to the state office. While the preceding sub-section includes Task Force recommendations on item types, item types are a key design consideration in field testing and test length.

Field Testing

Field testing, or the act of trying out items in actual situations reflecting their intended use, seeks to provide an initial view of item quality. Traditionally, items can be field tested in one of two ways: through (1) stand-alone field tests or (2) embedded field tests.

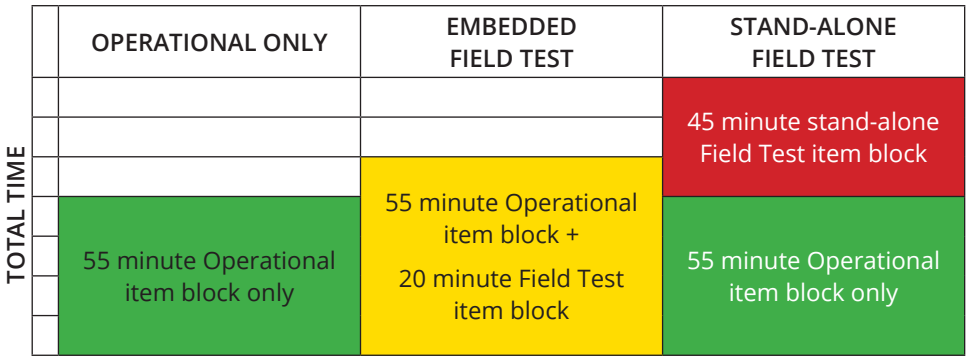
Stand-alone field tests are those instances when items are tried out in an independent administration. Stand-alone field tests are traditionally optional for districts and schools. In some states, mandatory field tests have

been deployed, which can lead to public resistance. The benefit of stand-alone field tests is that they can accelerate test development timeframes[7] but student motivation during the field test can be quite low. That is, stand-alone field tests must be supported through a separate administration meaning that educators and students are aware that they are experiencing a field test. Even if test length itself is not a concern, the presence of a separate "summative" testing event could be.

Embedded field tests place the field test items into the administration of another operational assessment. The primary benefit of this approach is that educators and students do not know what items are operational and what items are for the field test. This mitigates issues regarding decreased student motivation on field test items. However, it does increase the length of the assessment. Depending on the number of test items necessary to field test, "blocks" of items are usually administered to different students in the same grade (e.g., any one student would receive 10 additional field test items in addition to the typical set of items on a form, but those blocks would differ by student). Embedded field tests can also be used to create vertical scales or to link different assessments.

Understanding the role and impact of field test designs is critical to determining their value in the test length conversation. The following figure presents three conditions: (1) an operational (OP) test, (2) an OP test with an embedded field test (FT), and (3) an OP test with a stand-alone FT.

FIGURE 2.
VISUALIZING TEST LENGTH IN EMBEDDED VS. STAND-ALONE FIELD TESTING



One can see, in Figure 2, the differences in the impact of a single administration's testing time using embedded field testing compared to the overall testing time for a stand-alone field test.

Testing Time and Field Testing Recommendations

The Task Force reviewed these test length-related issues and engaged in discussions focusing on field testing, typical testing time, and how testing time relates to item types. While many of the previous recommendations affect test-length recommendations (e.g., online vs. paper-pencil testing), the Task Force was thoughtful in considering the interrelated nature of these issues. Based on their review of readings and subsequent conversations, the Task Force offered the following recommendations:

- Where possible and appropriate, leverage embedded field testing to develop and maintain Alabama's new assessment system.

The Task Force offered the following recommendations:

- Where possible and appropriate, leverage embedded field testing to develop and maintain Alabama's new assessment system.
- Target the average testing time to be about 90 minutes for each content area. Assessments that include extended writing responses should be longer and potentially treated as their own stand-alone testing session.

[7] See section on test timeline: *Timeline for Test Development and Administration.*

- Target the average testing time to be about 90 minutes for each content area. Assessments that include extended writing responses should be longer and potentially treated as their own stand-alone testing session.
- Explore the possibility for a prospective vendor to establish an agreement with Alabama's current vendor, Scantron (facilitated by ALSDE) to allow the prospective vendor to embed field test items into the operational assessment in the spring of 2019.

Potential Optional Considerations for an Expanded System of Assessments

The Task Force discussed several components of a state assessment system that may be considered if funding allows once the base requirements are met:

- Interim assessments
- Science assessments at every grade
- High school end-of-course assessments

The discussion of balanced assessment systems earlier in this report noted that several states are beginning to implement "loosely-coupled" systems of assessment. While the states acknowledge that such approaches are not fully balanced assessment systems, they are designed to eliminate much of the incoherence among the state summative assessment and the various district-purchased commercial interim assessments. Further, having the state support interim assessments tied to the same specifications as the summative assessment helps to ensure that districts have access to higher quality interim assessments than those they purchase on their own.

The Task Force has come to learn that there is no "free lunch" in assessment. Procuring an interim assessment system along with the summative assessment requires additional capacity at the state level (ALSDE) to monitor the quality of the interim assessments in addition to the critical oversight ALSDE must play on the summative assessment. Additionally, districts might be reluctant to give up their current interim assessments, which could mean that the state-sponsored interim assessments go unused or, worse, districts administer both their own interim assessments as well as the state-sponsored interims, leading to over-testing.

The Task Force discussed two major interim assessment designs. The first, the "mini-summative" design, is the most common among commercial interim assessment providers and is where each assessment replicates the end of year design. While such designs may have some use for evaluating within-year student growth, their use for informing instruction is severely limited for a host of well-documented reasons. The second design known as "modular" interim tests are tied to key subdomains within the standards (e.g., Number-Base 10; see Appendix A for a detailed explanation of different types of interim assessment designs)[8].

In addition to interim assessments, the Task Force recognized the value of more general non-summative support materials. These supports could be developed by a prospective vendor and be coherent with the standards and the new assessments. The Task Force distinguished these types of supports from interim assessments and recognized the complementary roles they might serve. Non-summative support

The Task Force recommended including interim assessments in the RFB as the highest priority support. Prospective vendors should propose a set of interim assessments that are modular, standards-based, and offer information for educators to better diagnose student strengths and weaknesses.

[8] See also: Dadey, N. & Gong, B. (2017, April). Using interim assessments in place of summative assessments? Consideration of an ESSA option. Washington, DC: Council of Chief State School Officers (CCSSO). Available online: https://www.nciea.org/sites/default/files/inline-files/ASR%20ESSA%20Interim%20Considerations-April%202017.pdf

materials and resources might include things like exemplar curricular units or lessons, targeted writing instructional materials, sample lessons on evidence-based writing, or phenomena-based lessons on multi-dimensional science standards.

In addition to the interim and non-summative assessment supports, the Task Force also considered the value of two additional areas for the Alabama Assessment System: (1) assessing science at every grade level and (2) end-of-course tests in high school. With regard to science, the Task Force exhibited concern that testing science only once per grade-span sent an implicit signal that science was not important enough to assess each year. Assessing science each grade level would communicate a strong signal to prioritize science instruction in every grade. However, the Task Force recognized the potential cost associated with developing a science assessment for each grade and how that would increase testing time for students.

End-of-course (EOC) tests received considerable interest from the Task Force as an instructionally-relevant way to expand high school testing beyond the ACT. The Task Force believed that these assessments may be more relevant reflections of course-content for students. Additionally, EOC tests may be less prone to motivation issues because they could be used as part of students' course grades and because they are tied directly to the content that students have been learning. The state of Alabama has a stable of items that are aligned to current courses of study that could be used as a basis for developing EOCs, but these items still require field testing and validation. The Task force also recognized that the need to maintain both the ACT and pre-ACT, potentially resulting in over-testing concerns if EOC tests are added to the mix.

As the Task force discussed these four concepts, they offered the following recommendations describing their relative priorities. However, the Task Force was clear that the ALSDE should try to support all four priorities and work to seek additional legislative funding to develop the most effective and appropriate assessment system possible for the state of Alabama.

- The Task Force recommended including interim assessments in the RFB as the highest priority support. Prospective vendors should propose a set of interim assessments that are modular, standards-based, and offer information for educators to better diagnose student strengths and weaknesses. These assessments should leverage the same test administration platform, have a similar look and feel to the summative assessment, but have the flexibility to be used on-demand by teachers and not required by ALSDE.
- The Task Force recommended including non-summative support materials and resources in the RFB as the second highest priority support. Non-summative support materials might include resources that assist in the instruction of writing, considerations for covering three-dimensional science standards in a comprehensive way, or model performance tasks to cover multiple mathematics concepts.
- The Task Force recommended including science at all grades in the RFB as the third highest priority.
- The Task Force recommended including end-of-course tests at all grades in the RFB as the fourth highest support. The Task Force noted the value of having items ready to be tested to help support this priority, but was aware of the risk of adding testing time in high school. Appropriately leveraging this priority would require EOCs to be attached to specific courses of study in the state of Alabama.

The Task Force recommended including non-summative support materials and resources in the RFB as the second highest priority support. Non-summative support materials might include resources that assist in the instruction of writing, considerations for covering three-dimensional science standards in a comprehensive way, or model performance tasks to cover multiple mathematics concepts.

Implementation Recommendations

While many assessment development efforts begin with very tangible questions around timing, costs, and constraints, it is important that the goals, purposes, uses, and claims of an assessment system are first defined. These larger decisions can help steer the practical conversations of development in meaningful and manageable ways. The Task Force appropriately specified these larger concepts to help focus a series of design decisions that have been described in earlier parts of this report. Additionally, several very practical recommendations were made to support RFB development. **The Task Force recommendations for timing, requirements, and evaluating the assessment system are described in this section.**

Timing

One of the most important issues when specifying a RFB for an assessment program is determining the time line for assessment design, field testing, and operational administration. The timeline of an operational administration dictates the timing and pace of development. There are many activities that are sequential in nature (i.e., the first must be completed before the next can be started) when developing an assessment, all of which are reliant on the specified purposes and uses of the assessment. The following sub-section describes two key concepts in assessment design, followed by recommendations made by the Task Force.

The Role and Timing of Assessments in Relation to Standards and Instruction

As noted previously, the Task Force recognized the backwards-looking nature of the information gleaned from the test and its potential uses (e.g., evaluate achievement, monitor progress over time, support accountability). Given these uses, it is important to understand how these types of assessments follow standards and instruction but can still be used to inform practice. That is, how does the statewide summative assessment help us understand how students are making progress against the state standards and grade-level expectations? After-the-fact assessment results can be used to inform broader adjustments to curriculum that may lead to revisions in instruction on a wider scale. This reiterates the notion that large-scale assessment should be dependent on state standards and thus great efforts are taken to determine the facets of the standards that are most appropriate to assess. This process is described in more detail in the next section.

The Assessment Development Process

The assessment development process must begin with a clarification of the uses and purposes of the assessment. In the case of Alabama's state summative assessment, the assessments must provide evidence of student proficiency of grade-level standards, inform progress toward college- and career-readiness (CCR), and support student and school accountability.

In order to appropriately determine the assessment development timeline, we should consider the general steps that are necessary to develop an assessment. Those steps include, but are not necessarily limited to the following[9] – depending on the uses of the assessment:

1. Develop assessment specifications, which are based upon the state's academic standards and provide detailed descriptions about the learning objectives that support the standards and the rules, including prioritization, dictating requirements for test content, format, and accessibility for all students;
2. Develop and review assessment materials, which include item development guides, scoring rubrics, graphic design requirements, a verification of content and standard alignment, and score report requirements;
3. Evaluate existing item banks and develop new items according to numbers 1 and 2 above;

[9] Adapted from DRC|CTB (2016). Designing assessment systems: A primer on the test development process.

4. Conduct pilot tests, cognitive laboratories, usability studies (to ensure ease of use by students and educators), tryout studies (to confirm consistent and accurate scoring if relevant), and bias and sensitivity reviews (to ensure content is validly and fairly represented for all students);
5. Conduct field tests to determine how well items are performing, the effectiveness by which the items represent the content being assessed, and that items can be accessed fairly and appropriately by all students;
6. Produce final assessment materials, which include reports for educators and students and supporting information/data that helps contextualize test results to those consuming reports from the test such as administrative manuals and interpretative guides;
7. Administer, score, and report student performance using the final version of the tests; and
8. Engage in ongoing evaluation of the assessment system to ensure the assessment is meeting the intended goals and to determine if any refinements or revisions to improve its quality and effectiveness are needed.

While these can be considered a general set of steps for assessment development, there may be additional or fewer steps depending on the intended uses and types of the assessment results.

Specifying a Timeline

The aforementioned list above intends to provide a relatively robust view of the steps required to develop an assessment. However, the three main cycles that we will use to describe an assessment development timeline are (1) item development, (2) field testing, and (3) operational administration.

For the purposes of this report, **item development** includes a review of the standards, the creation of item specifications and development guides, the writing of the items and ensuring alignment of the items. Item development is a critical aspect of test development because test claims are dependent on the assessment specifications, item quality, and the ability of the items to represent the content. Incidentally, item development is also where many are trying to reduce timelines and such shortcuts can produce problems later on.

Field testing includes steps 3-5 (see above), all of which are intended to test the test. In other words, are the items measuring what they purport to measure and can we argue that the test will convincingly support the intended claims? Field tests can either be stand-alone as their own administration or embedded into other operational administrations, which increases testing time but yields higher quality item information than stand-alone field tests.

Operational administration includes steps 6-8 to result in reportable scores that, in the case of statewide summative assessments, are used for accountability decisions. The operational administration refers to when students complete the test that "counts."

Two sample timelines are presented below that connect these three cycles with possible dates for the development of the Alabama state assessment system. The first timeline is relatively low risk, while the second timeline is more aggressive and thus poses higher risks to a quality testing experience. ALSDE and the Task Force opted to go with the lower-risk timeline shown below.

FIGURE 3. LOW-RISK ASSESSMENT DEVELOPMENT TIMELINE

	PRE-ASSESSMENT PREPARATION		ASSESSMENT DEVELOPMENT				LIVE ASSESSMENT
Activity	RFP Release	Vendor Award	Item Dev	Embedded Field Test	Form Dev	Form Dev; Training	Operational Adminstration
Semester	Spring 2018	Summer 2018	Fall 2018	Spring 2019	Summer 2019	Fall 2019	Spring 2020
School Year	SY 2017-2018		SY 2018-2019			SY 2019-2020	

FIGURE 4. HIGH-RISK ASSESSMENT DEVELOPMENT TIMELINE

	PRE-ASSESSMENT PREPARATION		ASSESSMENT DEVELOPMENT	LIVE ASSESSMENT
Activity	RFP Release	Vendor Award	Item Dev; Stand-alone Field Test; Training	Training; Form Dev; Operational Administration
Semester	Spring 2018	Summer 2018	Fall 2018	Spring 2020
School Year	SY 2017-2018		SY 2018-2019	

Recommendations for Test Timing

The Task Force examined the two figures above and engaged in conversation around the associated with risks with each timeline. It was evident to the Task Force that the high-risk timeline attempts to compress too many activities into a single year, leading to higher costs and risks and reducing the amount of time the state could address any unforeseen issues. The higher-risk timeline also exacerbated any training and infrastructure test issues that would likely require a full year (i.e., School Year (SY) 2019-2020) to address. Based on a review of risks and a conversation with facilitators, the Task Force made the following recommendations:

- Target a new assessment to be administered in SY 2019-2020. The new assessment would be administered in the Spring of 2020 to allow for sufficient training throughout the school year.
- Approach the deployment of a new assessment carefully to better manage development, training, operational, and political risk. Due to the rapidly changing assessment conditions in the state of Alabama, the current summative assessment systems have lost credibility. Adopting a lower-risk timeline can help the state and its prospective vendor address these issues and work to proactively avoid previous mistakes.
- Use a lower-risk timeline to support intensive training efforts for a new assessment administration system. This should be coupled with a communications and professional development effort on the role of summative assessments and their intended purposes, uses, and interpretation.

As noted in the Task Force recommendations, members recognized that a lower-risk timeline eliminates the possibility of having new operational assessment during SY 2018-2019. However, there was unanimous concern about taking the time to develop, deploy, and support the assessment implementation well.

Target a new assessment to be administered in SY 2019-2020. The new assessment would be administered in the Spring of 2020 to allow for sufficient training throughout the school year.

Reporting

Assessment reporting serves a pivotal role in building credibility with the public and educators. It is the primary—if not the only—point of contact stakeholders have with high-stakes assessment. Thus, reporting should be informative, flexible, understandable, and useful.

The facilitators asked Task Force members to consider two facets of reporting: (1) what to report and (2) how to report it. While some overlap emerged between the two facets, the Task Force recommendations address both facets as a function of the intended audience and the ways in which a system of reports should support coherence and "depth on demand." The Task Force recommended supporting the system's capability to support users' ability to dig into data to answer questions and follow-up questions (or at least point someone in the direction of where they can find additional information to answer their questions).

The Task Force provided the following recommendations around reporting and the roles associated with reports:

- Support several role-specific access points for assessment reports and identify what groups of people should have access to similar reports. These reports should be made available through paper-based reports for students and their guardians, and electronically for other groups. Groupings of stakeholders should include at least the following:
  - Student/guardian level reports
  - Teacher/classroom level access (e.g., principal, educator coach, IEP teams)
  - Local administrator reports (e.g., Principle, district staff, local School Board)
  - Policy maker and legislative reporting
  - Public reporting (e.g., Realtors, business, media, community)

- Individual reporting should include accurate and visually-displayed information performance information. This information should include performance from the current assessment, any prior assessments, and any additional information that can help parents or educators interpret performance trends (e.g., when appropriate, scores, performance levels, Achievement Level Descriptors, subscores, comparison data, externally linked information, target performance).

- Support aggregate reporting for multiple combinations of student performance where appropriate (e.g., subgroups, growth groupings, performance level groupings, similar schools, grades, classroom, etc.).

- Ensure that assessment reporting mirrors the wider view of accountability reporting under the ESSA and does not encroach upon interim or diagnostic testing available to schools and districts.

The Task Force also considered the role of reporting against college-ready expectations. While there was no clear recommendation to support state by state comparisons, members generally agreed that there was significant value in understanding how Alabama students perform against college- and career-ready expectations. Therefore, it will be important that the expectations specified in high school through the college and career readiness standards have a clear and coherent connection to expectations (i.e., performance standards) in earlier grades and that they are well described through accessible reports.

The Task Force offered the following recommendation around performance expectations:

- Specify that the prospective vendor develop coherent performance expectations that are communicated through intuitive reports to the public, educators,
  and students.

Key Minimum Requirements for an Assessment Request For Bids

In addition to the design requirements specified throughout this report, the Task Force discussed how to handle minimum expectations of a prospective vendor under the Alabama's new state assessment system. These conversations included, but were not limited to:

- Collaboration with the State Education Agency
- Planning and timelines
- Cost proposal detail
- Alignment with RFP components

Individual reporting should include accurate and visually-displayed information performance information.

Ensure that assessment reporting mirrors the wider view of accountability reporting under the ESSA and does not encroach upon interim or diagnostic testing available to schools and districts.

- Program/project management
- Data sharing/ownership
- Collection processes
- Vendor qualifications
- Risk management
- Platform technical requirements
- Liquidated damages

The Task Force recommended that ALSDE work with the Center for Assessment to clearly define the "non-negotiables" that are associated with a new assessment RFB and to ensure that the evaluation process requires prospective vendors to guarantee those minimum requirements will be met.

Evaluating the Validity and Technical Qualities of the Assessment System

Throughout conversations with the Task Force, the facilitators continuously raised the notion that we design assessment systems for specific purposes and uses. It is these purposes and uses that help us determine what evidence to collect to establish a validity argument for the assessment system. As noted in the *Standards for Educational and Psychological Testing* (2014), validity evaluations include:

1. Establishing intended uses and interpretations;
2. Issues regarding samples and settings used in validation; and
3. Specific forms of validity evidence.

The purposes and uses of a summative assessment system are quite specific and should support other components of a balanced assessment system. It is incumbent upon the state to collect evidence that supports the interpretations made based on the results of assessment system, as well as evidence on whether the intended goals of the system are being achieved.

In specifying explicit goals, purposes, and uses of Alabama's assessment system, the Task Force essentially suggested the types of validity evidence to which the state should attend. One such piece of evidence includes educators', administrators', and policy makers' ability to interpret and make inferences using summative assessment results. Additionally, the ALSDE will need to attend to the claims made based on the summative assessment results of, for example, student progress toward college- and career-readiness and their mastery of the state standards, by examining external and related data on student performance and preparedness. Furthermore, ALSDE must collect evidence regarding fairness, accessibility, lack of bias, generalizability, and appropriateness of performance expectations.

ALSDE should include requirements for their prospective vendor to help identify and collect sources of validity evidence, which include the aforementioned evidence. Prospective vendors will likely collect and examine content-oriented, cognitive process, construct-related, criteria-based, and consequential sources of evidence throughout the assessment's design, development, field testing, and implementation life cycle. However, ALSDE should work to define what evidence will be collected by the state and what will be collected by the prospective vendor *a priori*, as well as who will be responsible for synthesizing that evidence. Also, it will benefit the state greatly to specify that test developers will need to lead and support the monitoring and continuous improvement of the assessment to ensure it is reliable, fair, and valid for its intended uses. This monitoring and evaluation will be instrumental as the state of Alabama prepares its peer review submission and engages in continuous evaluation of the assessment system.

ALSDE should include requirements for their prospective vendor to help identify and collect sources of validity evidence.

The Use of a Technical Advisory Committee

Employing a high-quality, nationally-reputable technical advisory committee (TAC) is a critical aspect of maintaining the on-going quality of the State assessment system. It can be hard for states to pay for technical advisory committees separately, so many states fold the costs and logistical responsibilities for TAC advising and meetings into the operational assessment contract. It is often helpful to have a separate entity coordinate the TAC because there is potential for a conflict of interest when the test vendor coordinates the TAC.

Employing a high-quality, nationally-reputable technical advisory committee (TAC) is a critical aspect of maintaining the on-going quality of the State assessment system.

Conclusions

This report presented a description of the work of the Alabama Assessment Task Force and the various issues deliberated by the Task Force. The report included extensive discussion of the many recommendations associated with the design and implementation of a high-quality statewide assessment system. The Task Force included and represented many stakeholders of the Alabama educational system. They spent considerable time reading, studying, and discussing critical assessment issues. They deliberated respectfully and, in almost all cases, the recommendations presented throughout this report represented a consensus of the Task Force. Adhering as closely as possible to the recommendations presented herein regarding the new Alabama assessment system will help ensure the credibility and stability of the system. Such stability is crucial for supporting advances in educational achievement, growth, and attainment for students and schools in Alabama.

References/Sources Consulted

1 AERA, APA, & NCME, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

2 Binet, A., & Simon, Th. A. (1905). Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. L'Année Psychologique, 11, 191-244.

3 Chattergoon, R. & Marion, S.F. (2016). Not as easy as it sounds: Designing a balanced assessment system. *The State Education Standard, 16, 1, 6-9*

4 Marion, S.F. (2018). The opportunities and challenges of a systems approach to assessment. *Educational Measurement: Issues and Practice, 37*, 1, 45-48.

5 Marion, S.F. & Landl, E. (2017). Principled Assessment Design for the Performance Assessment of Competency Education (PACE). https://www.nciea.org/sites/default/files/publications/PACE%20Principled%20assessment%20design_092417.pdf.

6 Marion, S.F., Lyons, S., Pace, L., & Williams, M. (2016). A Theory of Action to Guide the Design and Evaluation of States Innovative Assessment and Accountability System Pilots. www.innovativeassessments.org.

7 Michigan Department of Education. (2013). *Report on Options for Assessments Aligned with the Common Core State Standards*. Retrieved June 20, 2015.

8 Mislevy, R. J. (1996). Evidence and inference in educational assessment. CRESST Technical Report No. 414.

9 Mislevy, R. J. and Haertel, G. (2006). Implications for evidence-centered deisgn for educational assessment. *Educational Measurement: Issues and Practice*, 25: 6–20.

10 National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academies Press.

11 National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards.* Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education, James W. Pellegrino, Mark R. Wilson, Judith A. Koenig, and Alexandra S. Beatty, *Editors.* Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

12 Perie, M., Marion, S.F., & Gong, B. (2009). Moving towards a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28, 3*, 5-13.

13 Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144.

14 Shepard, L. A., Penuel, W. R., & Pellegrino, J. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice, 37, 1,* 21-34

15 Wiley, E. C. (2017). *Formative Assessment: Examples of Practice.* Retrieved August 11, 2015, from http://ccsso.org/resource-library/formative-assessment-examples-practice

APPENDIX A: Glossary of Terms

Adaptive testing: Tests that provide examinees with different questions depending on how they respond to test items or sets of items. That is, the test difficulty adapts to the ability of the student.

Automated scoring: Item responses that are evaluated by artificial intelligence, often against a rubric or set of criteria. Automated scoring is typically used to evaluate writing responses and to monitor scoring drift of human scorers.

Computer adaptive testing (CAT): This type of adaptive testing produced fully individualized tests by student and are adaptive at each item.

Computer-based testing: Testing where both the stimulus (e.g., test item) and the response (e.g., item response) would be delivered and captured on an electronic device (e.g., desktop, laptop, tablet, etc.).

Fairness: Fairness emphasizes that the test must be fair, accessible, and appropriate for all individuals in the intended population for the intended use of that test.

Field testing: The activities that are intended to test the test. Activities help determine whether are items measuring what they purport to measure and whether test validly reflects the claims it intends to make.

Formative Assessment: Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievements of intended instructional outcomes (Wiley, 2008).

Human scoring: Item responses that are evaluated by a human scorer, often against a rubric or set of criteria. Human scoring is typically used to evaluate writing responses and to train automated scoring engines.

Interim Assessment: Periodic, semi-standardized assessments that are often referred to as "formative," "benchmark," "diagnostic," and/or "predictive." They are neither formative (e.g., they do not facilitate moment-to-moment targeted analysis of and feedback designed to student learning) nor summative (they do not provide a broad summary of course- or grade-level achievement tied to specific learning objectives), but are intended to provided information relative to a specific set of learning targets.

Item development: The steps in assessment development that include a review of the standards, item specifications, development guides, and item alignment.

Linear on the fly testing (LOFT): This type of adaptive testing allows for all items are selected at the start of the test (i.e., a fixed form) and are adapted based on prior test performance.

Multi-stage on the fly testing (MSOFT): This type of adaptive testing combines LOFT and MST testing. Forms are created on the fly at pre-determined stages (e.g., after 15 items or after a cluster of topic-specific items).

Multi-stage testing (MST): This type of adaptive testing uses pre-determined forms that are adapted to the student at pre-determined stages (e.g., after 15 items or after a cluster of topic-specific items). Students are routed to forms of varying difficulty based on performance in the previous stage.

Operational administration: The activities associated with test administration that produce reportable scores

Paper-pencil testing: Testing that uses paper for both the stimulus (e.g., test booklet) and response (e.g., score sheet).

Reliability: Generally, reliability refers to the pieces of information that help us determine whether a test is precise, reliable, or consistent enough for the intended use of that test.

Summative Assessment: Infrequent and often standardized assessments that cover major components of instruction (e.g., units, semesters, courses, credits, or grade levels).

Validity: Validity refers to the degree to which evidence and theory support the interpretations of test scores for the intended use of that test.

APPENDIX B:
Introduction to Assessment Systems

Balanced and comprehensive assessment systems are receiving a lot of attention these days. Unfortunately, many are realizing that it is easier to talk about assessment systems than actually design them. Assessment systems are balanced when the various assessments in the system are coherently linked - often through clear specification of the learning targets, comprehensively support multiple purposes and uses, and continuously document student progress over time. These properties of coherence, continuity, and comprehensiveness, originally described in *Knowing What Students Know* (NRC, 2001), help create a powerful image of a high-quality system of assessments. Building from NRC, 2001, we have found that coherence, utility, and efficiency are a bit more practitioner-oriented when working with district and state leaders (Chattergoon & Marion, 2016).

Drawing from *Knowing What Students Know* (NRC, 2001), a **coherent** assessment system must be compatible with the ways in which student learning is expected to progress within domains. **Utility** cannot be evaluated in the abstract, but it must follow from a well-articulated theory of action that specifies the various intended outcomes for the system and the processes and mechanisms by which these outcomes will be realized (e.g., Marion, et al., 2017). Further, depending on the explicit purposes and uses, utility must be addressed for each stakeholder group for each intended use. **Efficiency** means getting the most out of assessment resources and eliminating redundant, unused, and untimely assessments. Evaluation of an assessment system, therefore, should identify and reduce assessments that are not serving the stated purposes or are redundant with other, more useful assessments. Unfortunately, many district personnel assume a set of assessments functions as a system if it contains at least summative, interim, and formative components. In particular, there is an implicit and often wrong assumption that simply including interim assessments produces a balanced assessment system; including interim assessments does not magically produce a balanced assessment system.

Moving Into Practice

Defining criteria has been critical for conceptualizing and offering a vision for assessment systems that can advance student learning. Several have argued that districts are the appropriate organizational level for instantiating balanced systems of assessment because of the need for assessment systems to be coherent with the enacted curriculum (and not just standards) in order to be balanced (Shepard, Penuel, & Pellegrino, 2018; Marion, 2018). States, in general, are the wrong entity for the development of balanced assessment systems, but states can play a role in supporting high-quality assessment systems.

The criteria outlined in *Knowing What Students Know* (NRC, 2001) and further developed by Chattergoon and others (Chattergoon & Marion, 2016) are based on visions of "tightly-coupled" systems with information flowing among the various components to maximize efficiency and utility. This is a high bar and, based on the lack of real-world examples, are likely beyond the current reach of most educational systems. Recent work on designing assessments to evaluate student learning of the Next Generation Science Standards (NRC, 2014; Marion, 2018) has us to consider "loosely-coupled" systems. Such systems have multiple levels of assessments tied to the same learning targets and vision of learning science to at least partially address the coherence criterion. However, because the information would not be shared across levels of the system, such loosely-coupled systems would not be as efficient as ones where information from one level (e.g., classroom) could be used to support purposes at another level (e.g., accountability).

Several states are beginning to implement such loosely-coupled systems of assessment by awarding assessment contracts requiring the development of interim assessments explicitly tied to the states' summative assessment in reading and math. The interim assessments in Wyoming and Utah, for example, are based on a "modular" design whereby interim tests are tied to key subdomains within the standards (e.g., Number-Base 10). Many states that have procured interim assessments along with the state test have allowed districts to decide if and when to use the interim assessments. While these are not fully balanced assessment systems, they are designed to eliminate some incoherence between the state summative assessment and the various district-purchased commercial interim assessments. These examples illustrate how states can support coherent assessment approaches.

APPENDIX C: Mini-Summative vs. Modular Interim Assessment Designs

To help illustrate the differences between a mini-summative and modular design, we present an abbreviated pictorial representation of the two designs below. In a mini-summative design, the interim assessments are in essence, just shorter versions of the summative assessment. In a modular design, the interim assessments focus on specific portions of what was covered by the complete summative assessment to give more fine-grained information about student achievement within the content area of the summative assessment. A more detailed explanation of how this might be accomplished is given on the following pages.

FIGURE 5. MINI-SUMMATIVE INTERIM ASSESSMENT DESIGN SCHEMATIC

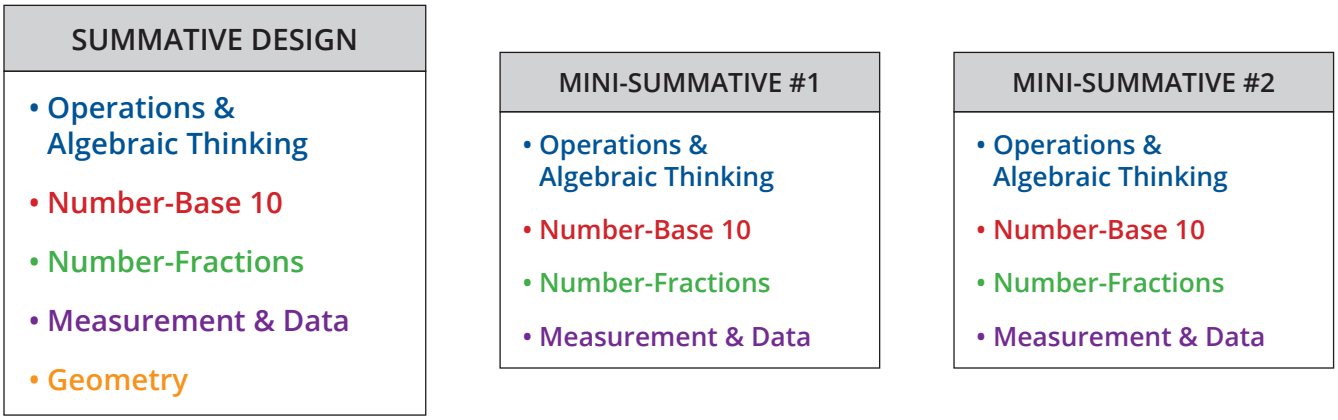
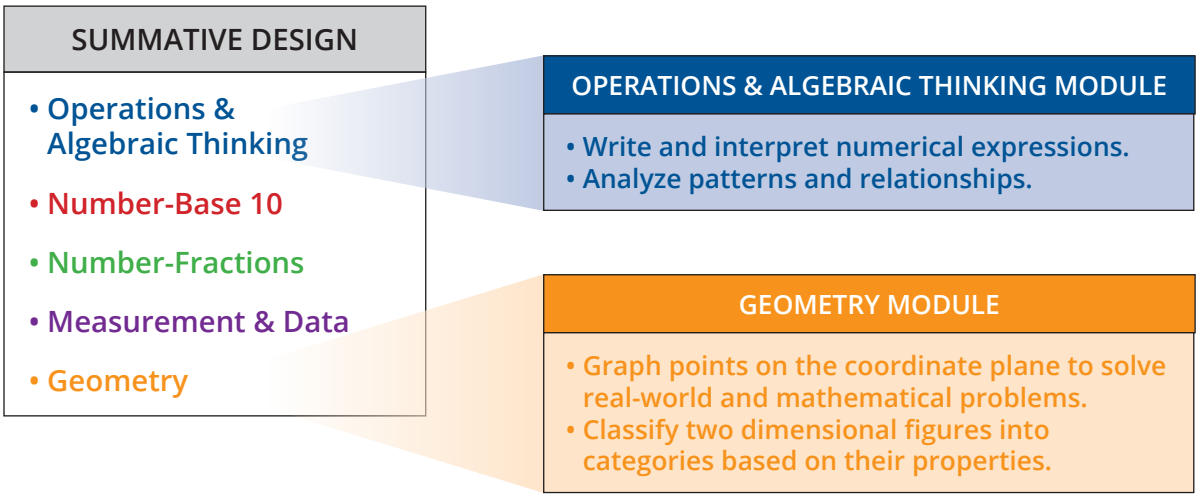


FIGURE 6. MODULAR INTERIM ASSESSMENT DESIGN SCHEMATIC



As an aid in further understanding assessment design, we first describe the general hierarchical format that content standards take by providing an example from grade-5 mathematics:

CONTENT CATEGORY
Operations & Algebraic Thinking • Write and interpret numerical expressions *Use parentheses, brackets, or braces…* *Write simple expressions that record calculations…* • Analyze patterns and relationships *Generate…numerical patterns…given rules…*
Number & Operations in Base Ten • Understand the place value system *Recognize [digit values increase tenfold when one place… left]* *Explain patterns in…multiplying by powers of 10…* *Read, write, and compare decimals to thousandths* *Use place value understanding to round decimals to any place* • Perform operations…to hundredths *Fluently multiple multi-digit whole numbers…* *Find whole-number quotients of whole numbers…* *Add, subtract, multiply, and divide decimals to hundredths…*
Number & Operations—Fractions • Use equivalent fractions…to add and subtract fractions *Add and subtract fractions with unlike denominators…* *Solve [fraction word problems by comparison…]* • Apply and extend…multiplication and division *Interpret a fraction [as a division problem]…* *[Extend whole number] multiplication to…fractions…* *Interpret multiplication as scaling (resizing)…* *Solve…problems [with] multiplication of fractions…* *[Extend division to involve unit fractions]*
Measurement & Data • Convert like measurement units [in the same] system *Convert among different sized measurement units…* • Represent and interpret data *Make a line plot to display [data with fractional units]…* • Geometric measurement: understand…volume *Understand volume as an attribute of solid figures…* *Measure volumes by counting unit cubes…* *Relate volume to [multiplication and division]…*
Geometry • Graph points on the coordinate plane to solve… *Use [two] perpendicular lines…to define a coordinate…* *Represent… points in the first quadrant…* • Classify two-dimensional figures…on…properties *[Know category] attributes [apply] to all sub-categories…* *Classify…figures in a hierarchy based on properties*

To aid in explanation, the broadest content categories (at the top of the hierarchy) are displayed in bold. Sub-categories are indented presented in the same color as the broad category they belong to. Sub-sub-categories are further indented and presented in italics.

In a *highly simplified* version of test design, the number of test questions or score points that come from each sub-sub-category is clearly specified to reflect the relative importance of each category. For example, if every sub-sub-category were considered equally important, a reasonable test design might specify that every sub-sub-category be measured using two test questions, resulting in the following hypothetical summative test design:

CONTENT CATEGORY	# OF ITEMS		
Operations & Algebraic Thinking	**6**		
• Write and interpret numerical expressions		4	
Use parentheses, brackets, or braces…			2
Write simple expressions that record calculations…			2
• Analyze patterns and relationships		2	
Generate…numerical patterns…given rules…			2
Number & Operations in Base Ten	**14**		
• Understand the place value system		8	
Recognize [digit values increase tenfold when one place… left]			2
Explain patterns in…multiplying by powers of 10…			2
Read, write, and compare decimals to thousandths			2
Use place value understanding to round decimals to any place			2
• Perform operations…to hundredths		6	
Fluently multiple multi-digit whole numbers…			2
Find whole-number quotients of whole numbers…			2
Add, subtract, multiply, and divide decimals to hundredths…			2
Number & Operations—Fractions	**14**		
• Use equivalent fractions…to add and subtract fractions		4	
Add and subtract fractions with unlike denominators…			2
Solve [fraction word problems by comparison…]			2
• Apply and extend…multiplication and division		10	
Interpret a fraction [as a division problem]…			2
[Extend whole number] multiplication to…fractions…			2
Interpret multiplication as scaling (resizing)…			2
Solve…problems [with] multiplication of fractions…			2
[Extend division to involve unit fractions]			2
Measurement & Data	**10**		
• Convert like measurement units [in the same] system		2	
Convert among different sized measurement units…			2
• Represent and interpret data		2	
Make a line plot to display [data with fractional units]…			2
• Geometric measurement: understand…volume		6	
Understand volume as an attribute of solid figures…			2
Measure volumes by counting unit cubes…			2
Relate volume to [multiplication and division]…			2
Geometry	**8**		
• Graph points on the coordinate plane to solve…		4	
Use [two] perpendicular lines…to define a coordinate…			2
Represent… points in the first quadrant…			2
• Classify two-dimensional figures…on…properties		4	
[Know category] attributes [apply] to all sub-categories…			2
Classify…figures in a hierarchy based on properties			2
TOTAL	52		

A mini-summative interim assessment design is intended to reasonably replicate the summative assessment experience with the exception of being shorter. For example, on an interim assessment with five testing opportunities, this could be accomplished by measuring each content standard with 1 rather than 2 items, giving the following mini-summative interim assessment design, making each interim assessment half as long as the summative assessment:

CONTENT CATEGORY	# OF ITEMS				
	1	2	3	4	5
Operations & Algebraic Thinking	**3**	**3**	**3**	**3**	**3**
• Write and interpret numerical expressions	2	2	2	2	2
Use parentheses, brackets, or braces…	*1*	*1*	*1*	*1*	*1*
Write simple expressions that record calculations…	*1*	*1*	*1*	*1*	*1*
• Analyze patterns and relationships	1	1	1	1	1
Generate…numerical patterns…given rules…	*1*	*1*	*1*	*1*	*1*
Number & Operations in Base Ten	**7**	**7**	**7**	**7**	**7**
• Understand the place value system	4	4	4	4	4
Recognize [digit values increase tenfold when one place… left]	*1*	*1*	*1*	*1*	*1*
Explain patterns in…multiplying by powers of 10…	*1*	*1*	*1*	*1*	*1*
Read, write, and compare decimals to thousandths	*1*	*1*	*1*	*1*	*1*
Use place value understanding to round decimals to any place	*1*	*1*	*1*	*1*	*1*
• Perform operations…to hundredths	3	3	3	3	3
Fluently multiple multi-digit whole numbers…	*1*	*1*	*1*	*1*	*1*
Find whole-number quotients of whole numbers…	*1*	*1*	*1*	*1*	*1*
Add, subtract, multiply, and divide decimals to hundredths…	*1*	*1*	*1*	*1*	*1*
Number & Operations—Fractions	**7**	**7**	**7**	**7**	**7**
• Use equivalent fractions…to add and subtract fractions	2	2	2	2	2
Add and subtract fractions with unlike denominators…	*1*	*1*	*1*	*1*	*1*
Solve [fraction word problems by comparison…]	*1*	*1*	*1*	*1*	*1*
• Apply and extend…multiplication and division	5	5	5	5	5
Interpret a fraction [as a division problem]…	*1*	*1*	*1*	*1*	*1*
[Extend whole number] multiplication to…fractions…	*1*	*1*	*1*	*1*	*1*
Interpret multiplication as scaling (resizing)…	*1*	*1*	*1*	*1*	*1*
Solve…problems [with] multiplication of fractions…	*1*	*1*	*1*	*1*	*1*
[Extend division to involve unit fractions]	*1*	*1*	*1*	*1*	*1*
Measurement & Data	**5**	**5**	**5**	**5**	**5**
• Convert like measurement units [in the same] system	1	1	1	1	1
Convert among different sized measurement units…	*1*	*1*	*1*	*1*	*1*
• Represent and interpret data	1	1	1	1	1
Make a line plot to display [data with fractional units]…	*1*	*1*	*1*	*1*	*1*
• Geometric measurement: understand…volume	3	3	3	3	3
Understand volume as an attribute of solid figures…	*1*	*1*	*1*	*1*	*1*
Measure volumes by counting unit cubes…	*1*	*1*	*1*	*1*	*1*
Relate volume to [multiplication and division]…	*1*	*1*	*1*	*1*	*1*
Geometry	**4**	**4**	**4**	**4**	**4**
• Graph points on the coordinate plane to solve…	2	2	2	2	2
Use [two] perpendicular lines…to define a coordinate…	*1*	*1*	*1*	*1*	*1*
Represent… points in the first quadrant…	*1*	*1*	*1*	*1*	*1*
• Classify two-dimensional figures…on…properties	2	2	2	2	2
[Know category] attributes [apply] to all sub-categories…	*1*	*1*	*1*	*1*	*1*
Classify…figures in a hierarchy based on properties	*1*	*1*	*1*	*1*	*1*
TOTAL	26	26	26	26	26

Multiple interim assessments built to this design would have different sets of test questions, but with the same emphasis on each of the content categories as on the summative assessment.

Modular interim assessment designs are different, however. Modular designs are intended to focus in on strategically selected subsets of the content standards (typically selected to represent potential moderate-sized units of instruction). Therefore, modular interim assessment designs are not similar to the summative test design. For example, in a highly simplified approach, each of the five broadest content categories could be selected as the focus for each of five interim assessment modules, giving the following modular interim assessment design of approximately the same length as the mini-summative designs:

CONTENT CATEGORY	# OF ITEMS				
	1	2	3	4	5
Operations & Algebraic Thinking	**27**				
• Write and interpret numerical expressions	18				
Use parentheses, brackets, or braces...	*9*				
Write simple expressions that record calculations...	*9*				
• Analyze patterns and relationships	9				
Generate...numerical patterns...given rules...	*9*				
Number & Operations in Base Ten		**28**			
• Understand the place value system		16			
Recognize [digit values increase tenfold when one place... left]		*4*			
Explain patterns in...multiplying by powers of 10...		*4*			
Read, write, and compare decimals to thousandths		*4*			
Use place value understanding to round decimals to any place		*4*			
• Perform operations...to hundredths		12			
Fluently multiple multi-digit whole numbers...		*4*			
Find whole-number quotients of whole numbers...		*4*			
Add, subtract, multiply, and divide decimals to hundredths...		*4*			
Number & Operations—Fractions			**28**		
• Use equivalent fractions...to add and subtract fractions			8		
Add and subtract fractions with unlike denominators...			*4*		
Solve [fraction word problems by comparison...]			*4*		
• Apply and extend...multiplication and division			20		
Interpret a fraction [as a division problem]...			*4*		
[Extend whole number] multiplication to...fractions...			*4*		
Interpret multiplication as scaling (resizing)...			*4*		
Solve...problems [with] multiplication of fractions...			*4*		
[Extend division to involve unit fractions]			*4*		
Measurement & Data				**25**	
• Convert like measurement units [in the same] system				5	
Convert among different sized measurement units...				*5*	
• Represent and interpret data				5	
Make a line plot to display [data with fractional units]...				*5*	
• Geometric measurement: understand...volume				15	
Understand volume as an attribute of solid figures...				*5*	
Measure volumes by counting unit cubes...				*5*	
Relate volume to [multiplication and division]...				*5*	
Geometry					**28**
• Graph points on the coordinate plane to solve...					14
Use [two] perpendicular lines...to define a coordinate...					*7*
Represent... points in the first quadrant...					*7*
• Classify two-dimensional figures...on...properties					14
[Know category] attributes [apply] to all sub-categories...					*7*
Classify...figures in a hierarchy based on properties					*7*
TOTAL	27	28	28	25	28

The benefit of a modular interim assessment design is that it can provide much more granular and instructionally useful information because there are enough items measuring fine-grained categories of content to inform broad (not day-to-day) instructional and/or remedial decisions.

THE ALABAMA STATE DEPARTMENT OF EDUCATION