



# New Mexico's State Assessment System

*Recommendations from the New Mexico  
Task Force for Student Success*

New Mexico//New Measures of Student Success  
OCTOBER 11, 2019



**Michelle Lujan Grisham**  
Governor

**Ryan Stewart, Ed.D.**  
Secretary Designate of Education

**Gwendolyn Perea Warniment, Ph.D.**  
Deputy Secretary for Teaching, Learning, and Assessment

**Lynn Vásquez**  
Director of Assessment of Student Learning

**Report prepared by:**  
Juan D'Brot, Ph.D.  
Scott Marion, Ph.D.  
National Center for the Improvement of Educational Assessment



## STUDENT SUCCESS TASK FORCE MEMBERS

**Felicitas Adame-Reyes,**  
Teacher, Rio Rancho Public Schools

**Laura Adkins,**  
Elementary Principal, Clovis Municipal Schools

**Amanda Allen,**  
High School Science Teacher, Magdalena Municipal Schools

**Eleanor Andrews, EdD,**  
Director of Assessment, Albuquerque Public Schools

**Jaqlyn Baldwin,**  
Executive Director, Siembra Leadership High School, Albuquerque

**Elisa Begueria,**  
Superintendent, Lake Arthur Municipal Schools

**Ellen Bernstein, EdD,**  
President, Albuquerque Teachers Federation

**Charles Bowyer,**  
Executive Director, National Education Association of New Mexico

**Valerie Brea,**  
Associate Director, Southwest Regional Education Cooperative

**Julie Bryant, PhD,**  
Elementary Principal, Bernalillo Public Schools

**Rene Cantu,**  
Assistant Superintendent for Data & Assessment, Hobbs Municipal Schools

**Cyrus Dudgeon,**  
English Department Leader and Teacher, Española Public Schools

**Stacey Eberhart,**  
Biology and Chemistry Teacher, Roswell Independent School District

**Jennifer Estrada,**  
Middle School Teacher, Cimarron Municipal Schools

**Alice Fitzgerald,**  
Elementary Computer Teacher & Test Coordinator, Raton Intermediate School

**Brad Furry,**  
Governing Council Member, The Academy of Technology and the Classics Charter School

**Staci Gallaher,**  
Principal, Career Prep Alternative High School, Central Consolidated Schools

**Veronica Garcia, EdD,**  
Superintendent, Santa Fe Public Schools

**Jessica Gilkison,**  
Teacher, Atrisco Heritage Academy, Albuquerque Public School

## STUDENT SUCCESS TASK FORCE MEMBERS - continued

**Tori Gilpin,**

Director for Research, Evaluation and State Testing,  
Gadsden Independent Schools

**Sharon Gordon-Moffett,**

Director, Central New Mexico Community College

**Lisa Harmon-Martinez,**

Teacher, Bilingual Coordinator, & NMCTE Executive  
Chair, Albuquerque Public Schools

**Jennifer Herschberger,**

STARS Coordinator, Belén Consolidated Schools

**Mike Hyatt,**

Superintendent, Gallup McKinley County Schools

**Pierce Jones, PhD,**

Director of Technology, Los Alamos Public Schools

**Danielle Kusmak,**

Educator and Business Owner, Tularosa Pistachio  
Groves/Tularosa Intermediate School

**Mae LaBella, PhD,**

High School Math Teacher, Taos Municipal Schools

**Stephanie Ly,**

President, American Federation of Teachers,  
New Mexico

**Rafael Martinez,**

Executive Director, Albuquerque Sign Language  
Academy

**Hope Morales,**

Director of TeachPlus New Mexico, TeachPlus

**Marissa Naranjo,**

Policy Director, All Pueblo Council of Governors

**Betty Patterson,**

President, National Education Association of  
New Mexico

**Nathan Pierantoni,**

Director of School Improvement and Assessment,  
Farmington Municipal Schools

**Arsenio Romero, PhD,**

Superintendent, Deming Public Schools

**Elizabeth Russom,**

District Testing and Bilingual Coordinator, & Fine Arts  
Coordinator, Chama Valley Schools

**Lauren Sabato,**

Middle School Art Teacher, Pojoaque Valley Middle  
School

**Sherwin Sando,**

Elementary Teacher, Pueblo of Jemez, Bureau of  
Indian Education

**Kevin Shendo,**

Director, Pueblo of Jemez Department of Education

**Joshua Silver,**

Director of Online Teaching and Learning, Las Cruces  
Public Schools

**Nina Smith,**

Continuous School Improvement Director, Santa Fe  
Indian School

**Kurt Steinhaus, PhD,**

Superintendent, Los Alamos Public Schools

**Vanessa Svihla, PhD,**

Associate Professor, University of New Mexico

**Teresa Tenorio,**

Parent and Education Advocate, Las Cruces

**Justin Trager,**

Director of School Networks, Future Focused  
Education, Albuquerque

**Glenna Voigt,**

Commissioner, New Mexico Public Education  
Commission

**Kalvin White, PhD,**

Education Administrator, Navajo Nation



## TABLE OF CONTENTS

|  |           |
|--|-----------|
| <b>Executive Summary</b> .....   | <b>6</b>  |
| • Task Force Recommendations .....   | 6         |
| - Full System Recommendations .....  | 6         |
| - System Summative Assessment for Grades 3–8 .....                               | 7         |
| - Summative Assessment for High School .....                                     | 8         |
| - Interim Assessment.....  | 9         |
| - Cultural Responsiveness and Sustainability.....                                | 9         |
| • Conclusions.....   | 10        |
| <b>Introduction</b> .....  | <b>11</b> |
| • New Mexico’s Guiding Principles .....  | 11        |
| • Background and Process .....   | 12        |
| <b>Types of Assessments and Appropriate Uses</b> .....                           | <b>13</b> |
| <b>Overview of New Mexico’s Assessment System</b> .....                          | <b>14</b> |
| • Assessments Included in the Proposed System .....                              | 14        |
| • Assessments Not Addressed in this Report.....                                  | 15        |
| <b>Recommended Purposes and Uses of Assessment</b> .....                         | <b>16</b> |
| • Community Conversations .....  | 16        |
| - Purposes and Uses of Assessment System Results .....                           | 16        |
| <b>Goals for New Mexico’s Assessment System</b> .....                            | <b>17</b> |
| • Recommendations to Meet the Goals of New Mexico’s System.....                  | 18        |
| • Key Design Considerations and Requirements.....                                | 18        |
| <b>Key Design Recommendations for Phase I</b> .....                              | <b>19</b> |
| • Overall Assessment System Recommendations.....                                 | 19        |
| • Recommendations for the Summative Assessment in Grades 3–8 .....               | 20        |
| - Purpose and Uses of Assessments in Grades 3–8.....                             | 21        |
| - Alignment, Development, and Timing.....  | 21        |
| - Writing Content, Item Types, and Spanish Language Arts .....                   | 22        |
| - Mode of Administration and Accessibility .....                                 | 23        |
| - Adaptivity .....   | 25        |
| - Reporting Results.....   | 26        |
| • Recommendations for the Summative Assessment in High School .....              | 27        |
| - Purpose and Uses of the Summative Assessments in High School .....             | 27        |
| - College Entrance Exams, Survey Tests, and End-of-Course Testing .....          | 27        |
| • Recommendations for the Interim Assessment for New Mexico .....                | 30        |
| • Evaluating the Validity and Technical Qualities of the Assessment System ..... | 31        |
| - The Use of a Technical Advisory Committee.....                                 | 32        |
| • Conclusions for Phase I.....   | 32        |

# TABLE OF CONTENTS

|  |           |
|--|-----------|
| <b>Key Design Recommendations for phase ii .....</b>                                   | <b>33</b> |
| • Supports for Using Interim Assessments .....   | 33        |
| - High Priority Characteristics.....   | 33        |
| - High-Priority Supports .....   | 34        |
| - High-Priority Educator Needs.....  | 35        |
| • Cultural Responsiveness and Sustainability .....                                     | 36        |
| - Background and Context.....  | 36        |
| - Recommendations for Cultural<br>Responsiveness in Assessment .....                   | 37        |
| • Writing and Authentic Assessment.....  | 38        |
| • Developing and Measuring the Whole Child .....                                       | 39        |
| • Supporting Formative Assessment<br>Practices and Project-Based Learning.....         | 41        |
| • Conclusions for Phase II.....  | 41        |
| <b>References/Sources Consulted.....</b>   | <b>43</b> |
| <b>Appendix A: Glossary of Terms.....</b>  | <b>45</b> |
| <b>Appendix B: Introduction to Assessment Systems .....</b>                            | <b>47</b> |
| <b>Appendix C: Community Conversations Summary (Brief #10).....</b>                    | <b>48</b> |
| <b>Appendix D: Introduction to Principled Assessment Design .....</b>                  | <b>51</b> |
| • The Role and Timing of Assessments in<br>Relation to Standards and Instruction ..... | 51        |
| <b>Appendix E: Mini-summative vs.<br/>Modular Interim Assessment Designs.....</b>      | <b>52</b> |

## EXECUTIVE SUMMARY

In March 2019, The New Mexico Public Education Department (NM PED) convened thirteen statewide community engagements to gather public input to reimagine the state assessment system. Following these community input meetings, the NM PED convened the New Mexico Task Force for Student Success (Task Force), which was comprised of key education stakeholders to make recommendations for New Mexico's next state assessment system. To develop these recommendations, the NM PED held a series of in-person and virtual meetings with the Task Force between April 2019 and June 2019 to deliberate over technical, policy, and practical issues associated with implementing an improved assessment system. Additionally, the Task Force was provided with a summary the community engagement feedback to inform their decision-making. The NM PED 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 consensus (or overwhelming majority) recommendations of the Task Force for the design and implementation of a stable, high-quality, assessment system intended to transition toward a more innovative assessment system for New Mexico.

### Task Force Recommendations

As part of the NM PED's deliberation process, the Task Force addressed a series of assessment design and implementation issues. Through these deliberations, the Task Force established a series of recommendations for design and implementation for the New Mexico State Assessment System. These recommendations are organized in the following way:

1. Overall Assessment System Recommendations
2. Summative Assessment Recommendations for Grades 3–8
3. Summative Assessment Recommendations for High School
4. Interim Assessment Recommendations
5. Supports for Using Interim Assessment
6. Cultural Responsiveness and Sustainability
7. Writing and Authentic Assessment
8. Developing and Measuring the Whole Child
9. Supporting Formative Assessment Practices, Performance-Based Assessments, and Project-Based Learning

Recommendations 1–4 and 6 are addressed in the executive summary, and all nine recommendations are addressed in detail in the full report.

### **Full System Recommendations**

Task Force members unanimously agreed that the end-of-year, statewide, summative assessment is a small part of the larger balanced assessment system. The Task Force's overall system recommendations are associated with the RFP (i.e., Phase I) and the future-focused system (i.e., Phase II). We indicate that association in the list below. The Task Force recommended that the full system incorporate the following components:

1. A summative assessment that complies with federal requirements and reflects greater relevance to the state's student and teacher population (Phase I)
2. Optional, on-demand interim assessments that support improved instruction (Phase I)
3. A common platform that provides the summative assessment, supporting resources, and documentation (e.g., interim assessments, item banks, assessment literacy resources) (Phase I)
4. Assessments and resources that support the whole-child across the assessment system, including early learning in grades K–2 (K–2 as a Phase I option and as an NM PED priority in Phase II)
5. Resources and supports that improve teacher and student use of assessment information (e.g., those that help establish ownership of learning, identify learning goals, and identify next steps for teaching and learning) (Phase II)

6. Information that evaluates readiness and progress for students on alternative pathways (Phase II)
7. Rigorous development and review processes that address cultural sustainability and responsiveness (Phase I and Phase II)

In addition, the Task Force also made several recommendations that addressed how components should be incorporated in the full assessment system. These include the following:

1. **Minimizing Change.** To facilitate a smooth transition and evolution of the assessment system, Task Force members recommended that the same vendor and (potentially expandable) platform be maintained for as long as possible.
2. **Minimizing Footprint.** New Mexico's statewide summative assessment for math and ELA in grades 3–8 and high school should be limited to only what is required by State and Federal law, and testing time should only be as long as necessary to ensure adequate coverage of the content standards.
3. **Assessing Writing.** The State summative assessment should assess writing.
4. **Cultural Responsiveness.** The State should establish processes to advocate for and address culturally responsive and sustainable assessments for Phase I, as well as supports for culturally responsive and sustainable educational resources in Phase II.
5. **Decoupling Performance.** The State should decouple performance on the summative assessment from teacher evaluation and graduation requirements.
6. **New Mexico Review.** New Mexico educators should be involved in the vetting and reviewing of test forms and items, as well as being involved in the development of new items on the State summative assessment, where possible.
7. **Authentic and Innovative Assessments.** The state should pursue options for more authentic and innovative assessment models including performance-based and project-based assessments as part of the full system.

### **Minimizing Footprint.**

*New Mexico's statewide summative assessment for math and ELA in grades 3–8 and high school should be limited to only what is required by State and Federal law, and testing time should only be as long as necessary to ensure adequate coverage of the content standards.*

### **System Summative Assessment for Grades 3–8**

The Task Force identified several purposes and uses of the summative assessment. While the level of endorsement by Task Force members varied, the vast majority of members recommended that the summative assessment should be developed to support the following purposes and uses:

- **Evaluate** performance against defined student expectations on the state standards
- **Determine that students are on track** for the next grade level or postsecondary opportunities
- **Monitor** trend data in both norm- and criterion-referenced ways
- **Support the calculation of student growth** by exhibiting sufficient technical quality to do so (note, the Task Force did not discuss specific methodologies to calculate student growth)
- **Provide large grain size information** of aggregated student performance to inform program and curriculum evaluation decisions and educator professional development needs

### **Alignment, Development, and Timing**

The Task Force recommended that the state engage in a reasonable procurement and development cycle that leveraged as many existing resources to support teaching and learning as possible. This procurement recommendation translates into the following recommendations that address alignment and the development of the assessment:

1. **Alignment.** The summative assessment in grades 3–8 should be tightly aligned to the state standards.

2. **Custom Development.** The state should procure a custom summative assessment in grades 3–8.
3. **Timing of Development.** Deploy a stop-gap assessment in SY 2019–2020 and deploy a fully operational New Mexico assessment in SY 2020–2021.

### **Writing, Item Types, and the Spanish Language Arts**

The Task Force recommended the following with regard to writing, item types, and the role of Spanish Language Arts in the assessment system:

1. **Writing.** Assess writing in every grade, with genres of writing matrixed within each grade.
2. **Item Types.** Include innovative item types only if they provide information that goes above and beyond what can be obtained using traditional selected response and constructed response items.
3. **Spanish Language Arts.** The NM PED should continue to assess Spanish language arts to support schools and districts that implement the CCSS en Español, and include this assessment in the RFP.

### **Mode of Administration and Accessibility**

The Task Force made several recommendations regarding the administration and logistics of the summative assessments for grades 3–8.

1. **Mode of Administration.** Continue to implement online testing and support paper backups only as necessary with sufficient comparability to the online forms. The PED must address infrastructure concerns throughout the state, with a particular focus on Indian, rural, and elementary education settings.
2. **Accessibility.** Whether online or on paper, the NM PED should prioritize accessibility needs for all students. The Task Force identified three sub-recommendations to support accessibility:
  - Support the diversity of learners through translations into languages other than English and support transadaptions (translating and adapting language to ensure it makes sense in languages other than English) of the assessment as needed and appropriate
  - Ensure that sufficient supports are included to students possess sufficient technology skills (e.g., keyboarding) to access items fairly
  - Ensure that all necessary tools, supports, and accommodations are available as needed

### **Reporting**

The facilitators asked Task Force members to consider two facets of reporting: (1) what to report and (2) how to report it over the course of the first two meetings and through the Task Force survey. The Task Force recommended the following:

1. Assessment reports should include information relevant to key groups of users, such as educators, students, and families.
2. Reports should include a mix of both criterion-referenced (comparison to a defined standard) and norm-referenced (comparison to a known group) interpretative information.
3. The assessment system should be on a single platform to the extent possible and support single sign-on if feasible.

### **Summative Assessment for High School**

The Task Force identified several purposes and uses for the high school summative assessment, many of which overlapped with the grades 3–8. However, several purposes and uses, such as college-readiness testing, were unique to high school. The Task Force made several recommendations for New Mexico high school assessments:

1. **College Entrance Exams.** Administer a nationally recognized college readiness test in grade 11.
2. **End-of-Course Testing.** Eliminate current end-of-course testing in high school to enable the state to invest more resources into other aspects of the state's assessment system.



## ***Interim Assessment***

The Task Force made several recommendations regarding the interim assessment approach for New Mexico's assessment system and its procurement.

1. **Use of Interims.** The interim assessments should provide enough granularity to inform instructional decisions.
2. **Optional Interims.** Interim assessments should be optional for district and school use. Interim resources should supplement existing assessment and instructional resources and should remain applicable to historical curricular investments (e.g., maintaining the use of evidence statements).
3. **Interim Design.** The Task Force was split evenly between recommending an item bank and assessment modules aligned to high-leverage needs, based on standards and learning targets. We recommend that the NM PED incrementally build interim assessment resources so educators can (1) use assessments to inform their practice and (2) develop local assessments aligned to both the standards and local curriculum.
4. **Interim Alignment.** Optional interim assessments should be adopted that are based on the same content frameworks as the State summative assessment.
5. **Interim Implementation.** Interim assessments should be on the same platform as the summative assessment if possible. The assessment platform should be expandable over time to support authoring assessment content as part of a future-focused assessment system.
6. **Interim Transparency and Security.** The interim assessment system should be fully transparent and promote educator understanding by involving them in item development, item review, and developing and reviewing associated supporting instructional materials (i.e., ensuring interim assessment materials and resources are not secure).

## ***Cultural Responsiveness and Sustainability***

To arrive at concrete recommendations regarding the role that the NM PED should play in creating culturally sustainable and responsive assessments, the facilitators asked the Task Force to consider (1) the role of culture in assessment development, (2) how to address heritage and diversity in assessment, and (3) the needed supports for stakeholders (i.e., policy makers, legislators, administrators, educators, and the public) to understand how assessments fit into the context of equity and culture.

**Assessment Development.** When asked about the role culture should play, the Task Force recommended that those who develop assessments

1. Construct and adhere to a stakeholder-developed framework for cultural responsiveness and sustainability
2. Adhere to best practices in assessment design with regard to accessibility, bias, and sensitivity, with a particular focus on culturally specific terms, symbols, and representations
3. Support flexibility in local design, development, and selection of assessments that are sensitive to cultural needs and nuance
4. Support the shift in the culture of assessment to prioritize the value of assessment systems and the need to support the whole child (this is discussed in more detail in the *Supporting the Whole Child* section of the full report)

**The Role of NM PED.** When considering the role of the NM PED in addressing the cultural heritage and diversity of values, the Task Force made the following recommendations:

1. Establish a comprehensive training and resource development plan around (1) assessment literacy, (2) the way in which assessments can support cultural heritage, and (3) how performance and access to opportunity differs by location, home language, socio-economic conditions, and racial/ethnic groups
2. Identify and facilitate access to literature and resources that celebrate diversity of culture and promote awareness of the diversity of a community's culture
3. Develop reporting systems that help administrators and teachers understand the demographic makeup of their districts, schools, and communities tied to relevant resources

**Supports for Stakeholders.** When asked about the needed supports for stakeholders (policy makers, legislators, administrators, educators, and the public) to understand how assessments fit into the context of equity and culture, the Task Force recommended the following:

1. Decouple the use of summative assessments from graduation requirements and incorporate additional ways to communicate student readiness for transition
2. Provide training to legislators and legislative staff on the intended purposes and uses of assessment data to inform legislative changes
3. Establish strategies that leverage existing expertise across the state (e.g., universities, strategic partners, advocacy groups) to support communications campaigns, professional development, training around growth mindset, and the use of assessment for continuous improvement
4. Develop resources, professional learning communities (e.g., virtual communities of practice), and partnerships for educator training on the identification, selection, and implementation of appropriate assessment strategies that honors the diversity of student backgrounds and experiences (e.g., portfolios, goal-setting, capstone projects, celebrations of achievement).

## Conclusions

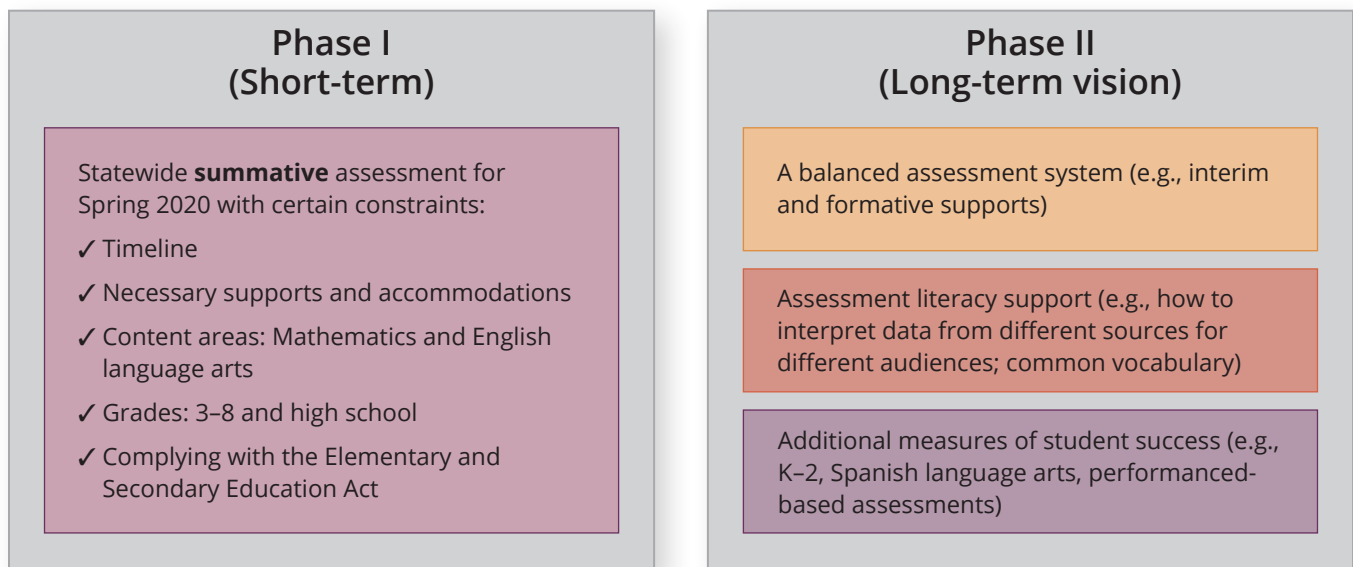
This report presents the work of the New Mexico Task Force for Student Success. The full report includes extensive discussion of the many recommendations associated with the design and implementation of a high-quality, statewide, assessment system. The Task Force included and represented many stakeholders of the New Mexico 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 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 New Mexico's students.

# INTRODUCTION

The New Mexico Public Education Department (NM PED) Assessment staff have sought to evaluate the current state assessment system and make recommendations for its future. The NM PED convened the New Mexico Task Force for Student Success (Task Force) to ensure broad-based stakeholder input into establishing the vision for state assessment in New Mexico. The Task Force was comprised of key education stakeholders in New Mexico and was facilitated by the National Center for the Improvement of Educational Assessment (Center for Assessment), a non-profit, non-partisan consulting firm. The NM PED 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.

The NM PED identified two major phases of recommendations that would be addressed by the Task Force. Phase I addressed the short-term needs for the state and Phase II addressed the longer-term vision for an assessment system (see Figure 1 below).

**FIGURE 1. PHASES I AND II OF NEW MEXICO'S ASSESSMENT SYSTEM**



The Task Force was charged with establishing a set of recommendations to support specifications for a Request for Proposals (RFP) for the NM PED to procure a new assessment system and to help frame NM PED's future assessment efforts. This report presents the results of those deliberations, the subsequent recommendations to the NM PED and the New Mexico Board of Education, as well as considerations for the state's RFP. The contents of this report are based almost exclusively upon consensus decisions of the Task Force. When consensus could not be reached, decisions were based on an overwhelming majority of task force members. This report presents summary of the goals of New Mexico's assessment system as well as the design and implementation recommendations.

## New Mexico's Guiding Principles

In March 2019, Drs. Juan D'Brot and Joseph Martineau, from the Center for Assessment, met with key leadership from the NM PED to discuss the initial perceived assessment needs for the state of New Mexico, how to corroborate those needs through community outreach, the goals for an Assessment Task Force, and the guiding principles that should inform the work. These guiding principles were the foundation for community outreach and a reference point for the Task Force's recommendations moving forward. The four guiding principles for the work are ensuring

1. **Equity.** Students have access to full educational opportunities;
2. **Accessibility.** Assessments are accessible to all students;
3. **Relevance.** Assessments reflect cultural values held by New Mexico students; and
4. **Rigor.** Assessments reflect globally competitive performance expectations.



### *The four guiding principles for the work are ensuring*

- 1. Equity. Students have access to full educational opportunities;*
- 2. Accessibility. Assessments are accessible to all students;*
- 3. Relevance. Assessments reflect cultural values held by New Mexico students; and*
- 4. Rigor. Assessments reflect globally competitive performance expectations.*

## **Background and Process**

The Center for Assessment began working the NM PED leadership in April 2019 to outline the work of the Task Force. Both partners 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. Juan D’Brot, Erika Landl, and Joseph Martineau, in collaboration with NM PED staff, facilitated the first meeting of the Task Force on March 15–16, 2019 to obtain recommendations for crafting an RFP for New Mexico’s next assessment system. The Center for Assessment staff 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 provided important background information that allowed the Task Force and the Center for Assessment facilitators to more quickly address each design consideration. Following the first in-person meeting, the Center for Assessment then solicited feedback from the Task Force through an electronic survey to address key assessment design considerations between April and May 2019.

Upon obtaining survey feedback, Drs. D’Brot and Landl met virtually with Task Force members in May 2019 for a half-day webinar meeting to present survey results and finalize components for the summative assessment portion of the RFP. Following the survey and virtual meeting, Drs. Juan D’Brot and Scott Marion then facilitated a face-to-face Task Force meeting in June 2019 to identify recommendations for future NM PED priorities outside of the RFP. A summary of Task Force recommendations are described throughout this report and summarized in the Executive Summary.

## TYPES OF ASSESSMENTS AND APPROPRIATE USES

Before presenting these recommendations, we first provide a series of definitions and concepts to ensure readers have a shared understanding of assessment types, their purposes, uses, and potential constraints. We define key assessment terminology to avoid misinterpretations as possible. There are several possible categorizations of assessment types, but we focus on the distinction among *summative*, *interim*, and *formative* assessment<sup>1</sup> because of the direct relevance to the Task Force's work. These definitions are critical to understanding what each type of assessment are designed to do, what each can and cannot be expected to provide users, and the categories were helpful for ensuring a shared understanding among Task Force members, as the various design choices were discussed.

**Formative assessment** is inseparable from instruction and can be thought of as a bridge between instruction and classroom assessment (Heritage, 2010, Shepard, in 2019). It has been defined as

*...a planned, ongoing process used by all students and teachers during learning and teaching to elicit and use evidence of student learning to improve student understanding of intended disciplinary learning outcomes and support students to become self-directed learners (CCSSO, 2018, p. 2).*

This definition makes clear that formative assessment is a process better thought of as part of the classroom instructional system, rather than as part of the assessment system (Shepard, 2019). This view follows from the work of Sadler (1989) and Heritage (2010) and makes sense, because for formative assessment to be formative, it must be inseparable from instruction.

**Interim assessments** are defined as

*assessments administered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals in order to inform policymaker or educator decisions at the classroom, school, or district level. The specific interim assessment designs are driven by the purpose and intended uses, but the results of any interim assessment must be aggregable for reporting across students, occasions, or concepts. (Perie, Marion, & Gong, 2009, p. 6)*

**Summative assessments** are designed to support various types of determinations (e.g., proficiency, competency) and administered at the end of a defined instructional period, such as a unit of instruction or a school year, to evaluate students' performance against a set of learning targets for that period. The state summative assessment—because of its prominent role in accountability and reporting—typically plays a disproportionate role in most assessment systems. To be clear, “summative” does not pertain to state-level tests solely; most district and classroom assessment systems include a summative component (e.g., for awarding grades or making competency determinations).


<sup>1</sup> In defining formative, interim, and summative assessment, this section borrows from three sources (Perie, Marion, & Gong, 2009; Michigan Department of Education, 2013; Wiley, 2008).

# OVERVIEW OF NEW MEXICO'S ASSESSMENT SYSTEM

We present an overview of the proposed New Mexico Assessment System in the following section. The Task Force read about and discussed balanced assessments and considered the extent to which the NM PED should try to procure key aspects of a potentially balanced system (see Appendix B for a more complete discussion on assessment systems). 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, but still an important aspect of a state assessment system.

## Assessments Included in the Proposed System

The Task Force spent most of its time discussing the grades 3–8 and high school assessments for English language arts (ELA) and mathematics. As the Task Force continued to meet, they also discussed New Mexico-specific needs, which included a Spanish language arts (SLA) assessment. However, the Task Force did not spend as much time discussing the SLA assessment as they did the primary Phase I (i.e., RFP components) and Phase II (i.e., future assessment process and vision) components. The Task Force spent the majority of its time during the first two meetings discussing the summative assessment design for ELA and mathematics and the majority of its time during the third meeting on high-value priorities for Phase II. We address these Phase II topics later within the *Key Design Recommendations for Phase II* section.



*New Mexico is committed to providing a relevant, competitive, and actionable assessment that communicates that all students have the opportunity to access a college education.*

High school assessment was the focus of considerable attention for the Task Force. New Mexico is committed to providing a relevant, competitive, and actionable assessment that communicates that all students have the opportunity to access a college education. Thus, the Task Force recommended the state procure a college entrance exam as its high school assessment for the *Every Student Succeeds Act* (ESSA) accountability system. That said, the Task Force also explored a limited set of interim assessments that were aligned to the state standards to ensure students and educators had sufficient tools to monitor progress against the state standards throughout high school. Recommendations on interim, end-of-course assessments are presented later in this section of the report and were used to inform optional specifications for the RFP.

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 grain 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 *Recommendations for Summative Assessment for Grades 3–8* section.

The Task Force provided the following high-level recommendations regarding assessments included in the New Mexico assessment system:

- Maintain the current alternate assessment (the New Mexico Alternate Performance Assessment), but as a separate focus and contract
- Recognize that the science assessment is outside of the scope of this Task Force and RFP
- Develop and administer a standards-aligned assessment for mathematics and English language arts in grades 3–8
- Administer a college entrance exam for mathematics and English language arts in high school
- Support an optional interim assessment system to help monitor student progress against the standards
- Invest in instructional and formative resources to support educator's instruction of the standards

## Assessments Not Addressed in this Report

New Mexico has been administering the New Mexico Alternate Performance Assessment (NMAPA), 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, the state of New Mexico plans to continue administering the NMAPA and second, the Task Force did not include enough expertise in alternate assessment in order to make appropriate recommendations. However, the NM PED will continue to engage in improvement efforts around their alternate assessment as part of their support of the overall assessment system. Additional recommendations regarding accessibility for Students with Disabilities (SWD) who are not eligible for the NMAPA are described later in the *Recommendations for Summative Assessments in Grades 3–8* section of this report.

The NM PED 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. The NM PED 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. The *Access for ELLs 2.0* is separate from the Spanish Language Arts assessment, which assesses student mastery of Spanish language arts. Recommendations for Spanish language arts assessments are made in the *Recommendations for Summative Assessments in Grades 3–8* section of this report.

Finally, the Task Force is aware that the state of New Mexico 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 was only able to make general recommendations regarding early childhood assessments.

# RECOMMENDED PURPOSES AND USES OF ASSESSMENT

The first major decisions of the Task Force involved specifying the goals and the intended purposes and uses of New Mexico's 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 based. Before describing the goals specified by the Task Force, we first present findings from the NM PED's Community Conversations, which preceded the Task Force and in which a range of input was collected from a wide breadth of education stakeholders in New Mexico.

## Community Conversations

In March 2019, the NM PED, in partnership with the New Mexico State University and the Center for Assessment, conducted a series of Community Conversations that invited parents, teachers, and community members to provide input on how to best measure student success within New Mexico's future statewide assessment. As a key part of the state's assessment transition plan, it was important that the Task Force consider the themes that emerged as part of the recommendation process, as a *starting point*.

### ***Purposes and Uses of Assessment System Results***

Across the Community Conversations, stakeholders discussed what they believed to be the primary purposes of the state assessment system and the uses the system's information should support. Several of these high-priority purposes and uses emerged repeatedly across locations and stakeholder groups, including the following:

- Clearly identify specific areas of student need and provide detailed feedback that serves to inform instruction and help students improve
- Evaluate student growth/progress over time (i.e., within and across years) on instructed standards
- Inform instruction throughout the school year through the provision of actionable feedback tied to instructional resources
- Inform the development of IEP's (e.g., establish appropriate goals)
- Provide information about the whole child to help educators understand a student's strengths/needs beyond the standards
- Inform decisions about professional development needs at the teacher, school, and district level
- Monitor trends in performance at the student and aggregate level
- Predict performance on the end-of-year summative assessment

These purposes and uses apply to the entire assessment system, since no single assessment can serve all of these needs. In the remainder of this report, we describe the importance of prioritizing and refining these purposes and uses, as they relate to different assessments within the system.


In addition to these desired purposes and uses, the Community Conversations elicited feedback regarding desired assessment features and the supports that the State should consider providing to students, educators, and the community. We provide more detail for the requested assessment features and supports in Appendix C.



# GOALS FOR NEW MEXICO'S ASSESSMENT SYSTEM

The facilitators first engaged the Task Force members in defining their vision and clarifying the big picture goals of New Mexico's assessment system. The vision and goals were grounded in the four guiding principles established by the NM PED. In light of the guiding principles stated above, the Task Force established the following vision statement to help focus their goals and objectives:

*The New Mexico Assessment System should be more than the end-of-year test in math and ELA. The current RFP addresses only part of what is needed to support New Mexico's educators and students. The New Mexico Public Education Department, in addition to complying with federal and state requirements, should expand the assessment system, its supports, and available resources to support educators and students in promoting whole-child success, post-secondary readiness, and universal access.*



*The New Mexico Assessment System should be more than the end-of-year test in math and ELA. The current RFP addresses only part of what is needed to support New Mexico's educators and students. The New Mexico Public Education Department, in addition to complying with federal and state requirements, should expand the assessment system, its supports, and available resources to support educators and students in promoting whole-child success, post-secondary readiness, and universal access.*

While the Task Force members identified many potential goals of the system, they ultimately affirmed the following:

1. Comply with federal requirements that reflect greater relevance to the state, while also meeting State requirements for the State summative assessment
2. Support student and teacher agency in establishing goals and identifying next steps for the assessment system
3. Include optional, on-demand interim assessments to support improved instruction
4. Where possible, utilize a common platform that can be expanded to meet New Mexico's increasing vision and scope for assessment and supports
5. Ensure that system components are reviewed for accessibility, bias, and sensitivity for New Mexico's students, regardless of status, disability, race, ethnicity, gender, and language proficiency, while addressing the need for Spanish language arts

In addition, the Task Force identified three additional needs that go beyond the scope of the current Task Force and RFP. These needs include:

1. Identifying additional assessment components and supports through future Task Force discussions
2. Addressing the whole child across the assessment system, including early learning
3. Evaluating the readiness and progress for students on alternative pathways

The assessment system goals closely mirror and extend those key themes initially raised by Community Conversation participants. Task Force members indicated that, in light of the recent legal decisions in New Mexico<sup>2</sup>, the assessment system must address both cultural and performance issues. Therefore, the Task Force stated that it is critical that their recommendations ensure that (1) the assessment is accessible to all students, (2) the assessment system include instructionally-relevant tools, and (3) performance expectations are both globally competitive and locally accessible. From these three statements alone, it is evident that a single assessment cannot, and will not, meet the goals stated above.

<sup>2</sup> See Yazzie/Martinez vs. State of New Mexico (2018). Retrieved from <http://nmpovertylaw.org/wp-content/uploads/2018/09/Graphic-Yazzie-Martinez-Decision.pdf>.

## Recommendations to Meet the Goals of New Mexico’s System

Assessments and their interpretations are designed and validated to serve a limited number of purposes and uses. As much as policy makers, educational leaders, and other stakeholders want a single assessment 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 purposes and uses identified during the Community Conversations and how the assessment system should be designed to meet those purposes and uses. While the Task Force did not directly specify the intended purposes and uses as a separate activity, they regularly referred to the Community Conversations and described how to meet those purposes and uses through design and operational recommendations.

The remainder of this report outlines the process the Center for Assessment used to facilitate the Task Force and presents their recommendations. These recommendations are organized as follows in the remainder of the report:

- Recommendations for Phase I
- Recommendations for Phase II

## Key Design Considerations and Requirements

After addressing higher-level considerations, the Task Force began to consider many operational decisions in the form of design considerations, requirements for the assessment system, and specifications for the RFP. The Task Force had the opportunity to learn about and discuss the implications of principled approaches for designing assessments. The context for this Task Force was unique in that the timeline to establish a transition plan for the assessment system and release of the RFP was significantly accelerated due to two main factors, including

1. The recent legal decision of *Yazzie/Martinez vs. the State of New Mexico*, and
2. A gubernatorial directive to identify a new assessment vendor and assessment transition plan before the beginning of school year 2019–2020.

Thus, the Center for Assessment utilized a set of briefs as pre-reading for Task Force deliberations. These pre-readings addressed several key design and operational issues one must understand when making defensible assessment development recommendations. These briefs prepared Task Force members with the information needed to help understand principled assessment design. See Appendix D for a brief discussion of *Principled Assessment Design* and the *Role and Timing of Assessment*.

The next section dives into operational considerations that emerged from responses to the types of questions posed above. It addresses topics such as 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 recommendations on adaptive and fixed form tests.

# KEY DESIGN RECOMMENDATIONS FOR PHASE I

When considering how to approach assessment design and understanding where assessments fit into 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 an RFP or summative assessment. This section describes Task Force recommendations for the following topics related to Phase I of the assessment system:

- Overall assessment system recommendations
- Summative assessment recommendations for grades 3–8
- Summative assessment recommendations for high school
- Interim assessment recommendations
- Evaluating the validity and technical qualities of the assessment system

The final section was not discussed by the Task Force, but is critical to managing and improving a statewide assessment system. In that section, we describe high level considerations for the NM PED's consideration.

## Overall Assessment System Recommendations

As part of Task Force deliberations, members discussed those things that should be universally considered across the entire assessment system. These recommendations were established as part of Phase I discussions, but include components of a system that can only be addressed during ongoing Phase II design and implementation. Task Force members were clear they needed to signal the importance of these additional components in light of the desire to minimize the footprint of the summative assessment in Phase I to the greatest extent feasible.

Task Force members, united in their vision for the State's assessment system, unanimously agreed that the end-of-year statewide summative assessment is only a small part of the larger assessment system. As a point of clarity, the Task Force's overall system recommendations can be associated with the RFP (i.e., Phase I), the future-focused system (i.e., Phase II) efforts, or both. We indicate that association in the list below. The Task Force recommended that the full system incorporate the following components:

1. A summative assessment that complies with federal requirements and reflects greater relevance to the State's student and teacher population (Phase I)
2. Optional, on-demand, interim assessments that support improved instruction (Phase I)
3. A common platform that provides both the summative assessment and supporting resources and documentation (e.g., interim assessments, item banks, assessment literacy resources) (Phase I)
4. Assessments and resources that support the whole child across the assessment system, including early learning in grades K–2 (K–2 as a Phase I option and as a Phase II NM PED priority)
5. Resources and supports that improve teacher and student use of assessment information (e.g., those that help establish ownership of learning, identify learning goals, and identify next steps for teaching and learning) (Phase II)
6. Providing information to evaluate readiness and progress for students on alternative pathways (Phase II)
7. Rigorous development and review processes that address cultural sustainability and responsiveness, which consider (Phase I and Phase II)
  - Accessibility, bias, and sensitivity for New Mexico's uniquely diverse student population
  - Student status, disability, race, ethnicity, gender, and language proficiency
  - Spanish language arts needs for students

**Incorporating components into the full assessment system.** The Task Force recognized that even with the rapid-paced timeline for Phase I of the assessment transition, the evolution of the assessment system would take time. Furthermore, members recognized that in order to change the system, there is a certain level of stability that is

necessary to facilitate change within the summative assessment, let alone assessment system. Thus, the Task Force also made several recommendations that addressed the logistics and features that further clarified how components should be incorporated in the full assessment system. These include the following:

1. **Minimizing Change.** Task Force members recognized that in Phase I, the State assessment system will resemble a traditional state assessment system, but the State will investigate, with stakeholders, how to transition to a more authentic assessment system in Phase II. To facilitate a smooth transition and evolution of the assessment system, Task Force members recommended that the same vendor and (potentially expandable) platform be maintained for as long as possible.
2. **Minimizing Footprint.** Task Force members also unanimously recommended that New Mexico's statewide summative assessment for math and ELA in grades 3–8 and high school should be limited to only what is required by State and Federal law, and testing time should only be as long as necessary to ensure adequate coverage of the content standards.
3. **Assessing Writing.** Task Force members overwhelmingly recommended that the State summative assessment include writing. There was less agreement on how writing should be assessed. Specific recommendations regarding writing are provided in the *Recommendations for the Summative Assessment in Grades 3–8* section.
4. **Cultural Responsiveness.** Task Force members specifically recommended that the State establish processes to advocate for and address culturally responsive and sustainable assessments for Phase I *and* supports for culturally responsive and sustainable educational resources in Phase II.
5. **Decoupling Performance and Evaluation.** A major point of discussion among Task Force members centered on the use of summative assessments for high-stakes decisions, like teacher evaluation and requirements for graduation. In-person meetings reflected a near unanimous recommendation to decouple performance on the summative assessment from teacher evaluation and graduation requirements. A follow-up survey, soliciting individual Task Force responses, yielded agreement rates close to 75 percent, recommending decoupling assessment results from teacher evaluation and graduation requirements. Several members cited issues with unintended negative consequences of eliminating performance-based requirements. *We recommend that the NM PED continue to study this issue through other meetings with key stakeholders throughout the state.*
6. **New Mexico Review.** Task Force members recommended that New Mexico educators be involved in the vetting and reviewing of test forms and items, as well as being involved in development of new items on the State summative assessment where possible. This suggestion is obviously dependent on other endorsements, such as when college entrance exams are recommended, as well as considerations of cost and efficiency.

These two sets of recommendations reflect a comprehensive assessment system requiring significant coordination to design and implement well. In the following sections, we describe the detailed Phase I recommendations for the summative assessment in grades 3–8, high school, and the interim assessment.


## Recommendations for the Summative Assessment in Grades 3–8

The Task Force spent a great deal of time during the first two meetings considering the design and operations of the statewide summative assessment in grades 3–8. These deliberations focused on both the general design of, and the specific administration requirements for, the assessment. Topics for deliberation and recommendation included purpose, alignment, development, adaptivity, reporting, and application to future performance.

When thinking about the purpose of summative assessments in grades 3–8, the Task Force was very clear about the interpretation for which the test should be designed: *evaluating performance against the State standards*. 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 concepts 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. On the other hand, the overwhelming majority of the Task Force recommended including writing at every grade level, where reading and mathematics are assessed in grades 3–8.

### **Purpose and Uses of Assessments in Grades 3–8**

The Task Force identified several purposes and uses of the summative assessment. While the level of endorsement by Task Force members varied, the vast majority of members recommended that the summative assessment should be developed to support the following purposes and uses:



*The summative assessment should be developed to support the following purposes and uses:*

- 1. Evaluate performance against defined student expectations on the state standards*
- 2. Support on-track claims about students being on track for the next grade level or postsecondary opportunities*
- 3. Monitor trend data in both norm- and criterion-referenced ways*
- 4. Support student-growth calculations by exhibiting sufficient technical quality to do so (Note: the Task Force did not discuss specific methodologies to calculate student growth)*
- 5. Provide student performance data at a level of granularity that can inform program and curriculum evaluation decisions and indicate educator professional development needs*

- 1. Evaluate** performance against defined student expectations on the state standards
- 2. Support on-track claims** about students being on track for the next grade level or postsecondary opportunities
- 3. Monitor** trend data in both norm- and criterion-referenced ways
- 4. Support student-growth calculations** by exhibiting sufficient technical quality to do so (Note: the Task Force did not discuss specific methodologies to calculate student growth)
- 5. Provide student performance data at a level of granularity** that can inform program and curriculum evaluation decisions and indicate educator professional development needs

In addition to the recommendations above, the Task Force also noted that the NM PED should attend to the following factors as they procure and develop the State summative assessment in grades 3–8. Some of these considerations may be more long-term in nature and will need to be implemented through additional assessment supports but are critical to meeting the stated purposes and uses above. These considerations include the following:

1. The assessment vendor must thoroughly test the system to identify gaps in infrastructure that should be addressed. This cannot be based on simulation alone, but include a statewide, collaborative effort with LEAs to place high demands on the system from every school that will participate in testing.
2. The State should explore the possibility of evaluating student engagement as a component of the system (e.g., through telemetry data or interest inventories). This may be a more future-focused consideration.
3. The State should clearly articulate the appropriate use(s) of summative assessment data and specify the limit of the assessment's purpose through communications and a public relations campaign (i.e., state assessments are a required part of an overall assessment system and are not reflective of the whole child).

### **Alignment, Development, and Timing**

Throughout Task Force deliberations, alignment was a recurring topic of discussion. This was related to whether the state should procure an

off-the-shelf assessment (e.g., consortium-based, vendor-developed assessments) or procure a custom assessment for New Mexico. Task Force members made three recommendations regarding alignment and timeline critical to developing the state summative assessment in grades 3–8. These recommendations are presented at the end of this sub-section.

Critically, the Task Force recognized the amount of time that it takes to develop a high-quality assessment system that is aligned to the New Mexico standards. The assessment development process must begin with a clarification of the uses and purposes of the assessment. In the case of New Mexico'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 under ESSA.

The Task Force recommended that the state engage in a reasonable procurement and development cycle that leveraged as many existing resources that support teaching and learning as possible. To avoid replicating development efforts or wasting resources, the Task Force recommended maintaining a strong alignment to standards-based assessment in grades 3–8 in math and ELA using existing evidence statements. This procurement recommendation translates into the following recommendations that address alignment and the development of the assessment:

1. **Alignment.** Task Force members recommended that the summative assessment in grades 3–8 should be tightly aligned to the State standards. This would allow users of assessment information to evaluate student performance against defined expectations on the standards.
2. **Custom Development.** Task Force members reviewed briefs, detailing development considerations regarding timing and ownership of summative assessments. Based on readings and extensive discussion, the Task Force recommended almost unanimously that the State should procure a custom summative assessment in grades 3–8. Furthermore, the NM PED should capitalize on the already high-quality work NM educators have done with previous multi-state collaboratives to maximize previous investments and minimize disruptions to the timeline.
3. **Timing of Development.** Task Force members recommended that a new assessment vendor is identified in SY 2019–2020; assessment content is developed through SY 2019–2020 and SY 2020–2021; a stopgap assessment is deployed in SY 2019–2020; and a fully operational New Mexico assessment is deployed in SY 2020–2021. This timeline would allow for sufficient training throughout both the 2019–2020 and 2020–2021 school years and allow the NM PED to better manage development, training, operation, and communications risks. This recommendation is also presented in the table below.

**TABLE 1. TIMING OF ASSESSMENT DEVELOPMENT**

| School Year                | 2018–2019                   | 2019–2020                        | 2020–2021                        |
|----------------------------|-----------------------------|----------------------------------|----------------------------------|
| <b>Procurement</b>         | Identify new contractor     | New contractor in place          |                                  |
| <b>Content Development</b> |                             | Rapid development of new content | Continued development of content |
| <b>Administration</b>      | Administer prior assessment | Administer stop-gap assessment   | Administer new custom assessment |

**Writing Content, Item Types, and Spanish Language Arts**

The Task Force engaged a series of discussion that addressed the content of the assessment beyond the generality of math and ELA in grades 3–8. These included writing, how item types are influenced by content, and the role of Spanish language arts in the assessment system.

**Writing Content and Assessing Writing**

A key consideration, associated with content representation, was whether the New Mexico assessment should include writing at every grade level in which math and ELA are assessed. Many states rely on extended 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 necessitate that students frame arguments, based on evidence from reading stimuli rather than the contrived narrative prompts, 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 include writing on the assessment at all. As noted in the overall system recommendations, the Task Force strongly believed that writing should be included on the assessment. Additional deliberations focused on whether the NM PED should prioritize student- or school-level scores. Focusing on school-level information does not mean that the State would give up on student-level scores, but it does mean that the student scores are less descriptive (e.g., providing a writing score, but not a narrative writing score). However, the school may be able to provide robust information on writing performance at the school level that may even be able to support writing subscores (e.g., by genre) using multiple prompts at the school level.

### **Item Types**

Items and tasks are the tools that elicit student responses, which in turn support inferences about what students know and can do. The information elicited 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; the NM PED must focus on obtaining and maintaining item quality as efficiently as possible. For a detailed discussion on item types, please see *Brief #6: Measuring Student Learning: Item Types*. This brief 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

The Task Force reviewed the information in *Brief #6* and considered how item types should best be represented on the assessment. Over the course of the first two meetings, the facilitators were able to obtain more clarity around the content-related recommendations. Specifically, these clarifications were focused on how writing should be assessed, the role of item types in development, and how to support Spanish language arts assessments.

1. **Writing.** The Task Force extended their earlier recommendation to include assessing writing on the summative assessment. The final recommendation, based on survey feedback, was to assess writing in every grade, with genres of writing matrixed within each grade. This allows schools to receive performance information on each type of writing genre, still allowing students to obtain a writing subscore without requiring a student to take three writing prompts.
2. **Item types.** The Task Force recommended that the assessment include innovative item types. Members were clear that these items should only be used when they provide information that goes above and beyond what can be obtained using traditional selected- and constructed-response items. Therefore, the vendor should justify the use of innovative item types (e.g., technology enhanced items) that accurately and appropriately assess the complexity of State’s standards and document how the use of innovative item types move more than nominally beyond what can be done with existing and less expensive item types.
3. **Spanish language arts.** The Task Force also discussed the degree to which Spanish language arts should be assessed in grades 3–8. Members recommended that the NM PED should continue to assess Spanish language arts to support schools and districts that implement the CCSS en Español and that this assessment be included in the RFP.

### **Mode of Administration and Accessibility**

The Task Force made several recommendations regarding the administration and logistics of the summative assessments for grades 3–8. These are focused around the mode (i.e., paper-pencil, online, or both) in which the Task Force believes the assessment should be administered, how its content should be accessed by students, and how educators should interact with reports.

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 the main considerations, the list of concerns for any State deliberating the mode of administration (PPT vs CBT) also includes:

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

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 selection, which were addressed in the Task Force’s pre-reading and throughout the Task Force meetings.



### *Without accessibility*

- *there is no equity in access;*
- *there is no ability to determine whether something is relevant to a student; and*
- *there is no opportunity to demonstrate student ability against a set of competitive performance expectations.*

The Task Force reviewed these mode-related issues and engaged in a discussion focusing on statewide technology readiness. It is important to note that New Mexico’s summative assessments are currently administered fully online, which contributed to the discussions and decisions and helped convince many Task Force members that New Mexico can continue to implement a fully online system. However, several Task Force members indicated that the ability to support a fully online system does not mean that the state is free from online administration issues. In fact, several members cited specific issues related to infrastructure, bandwidth, and device capability that may be concentrated in high-poverty districts and schools, including those served by the Bureau of Indian Education. This is of particular importance to the NM PED, as the department attends to issues of equity and access moving forward.

Extending the topic of mode of administration, the Task Force discussed the need to ensure that students could access the content equally and fairly. These deliberations reference the guiding principles directly. While accessibility—ensuring that all students can access assessments without bias—is directly connected to those principles, ensuring access has direct implications on the remaining three principles. Without accessibility

- there is no equity in access;
- there is no ability to determine whether something is relevant to a student; and
- there is no opportunity to demonstrate student ability against a set of competitive performance expectations.

As a result, the Task Force provided several sub-recommendations under the umbrella of accessibility. These recommendations are detailed below.

Based on conversations regarding mode and accessibility, the Task Force made the following recommendations:

1. **Mode of Administration.** The Task Force recommended that the NM PED continue with online testing but support paper backups as necessary with sufficient comparability to the online forms. The continued implementation of online administration will help facilitate the use of innovative item types. However, the NM PED should help districts address infrastructural concerns throughout the state, with a particular focus on Indian, rural, and elementary education settings.



2. **Accessibility.** Whether online or on paper, the NM PED should prioritize accessibility needs for all students. The Task Force identified three sub-recommendations to support accessibility:

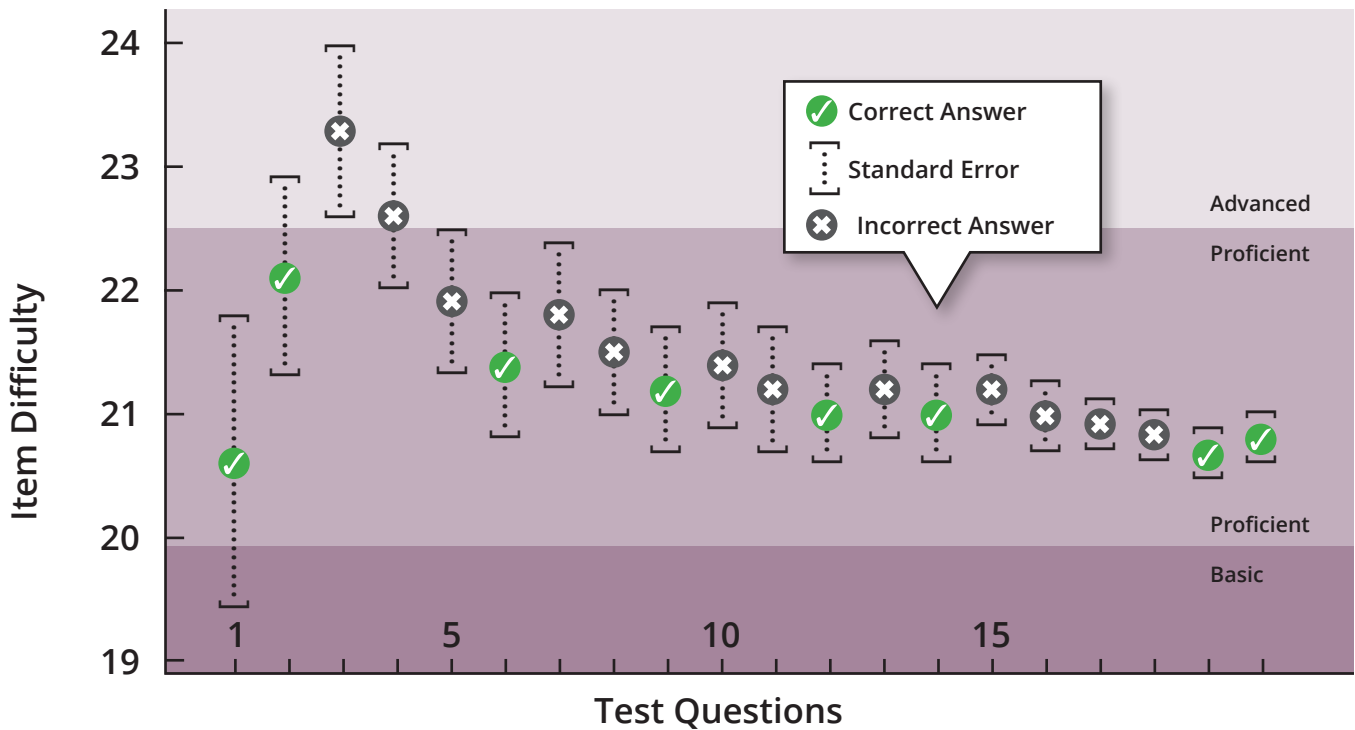
- Support the diversity of learners through translations into languages other than English and support transadaptions (translating and adapting language to ensure it makes sense in languages other than English) of the assessment as needed and appropriate (e.g., decoding vs encoding at different grade levels impact the availability of translations)
- Ensure that sufficient supports are included to ensure students possess sufficient technology skills (e.g., keyboarding) to access items fairly
- Ensure that all necessary tools, supports, and accommodations are available as needed

**Adaptivity**

In addition to deliberations around the mode of assessment, the Task Force reflected about the use of adaptive testing in New Mexico’s new assessment system. Adaptive tests do exactly that, they adapt to an estimate of a student’s achievement by providing more or less difficult items, based the student’s 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 computer-adaptive testing (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, they are given a more difficult item, and conversely an easier item, if answered incorrectly, until a sufficiently accurate judgment about the student’s achievement can be made.

**FIGURE 1. COMMON CONCEPTUALIZATION OF AN ADAPTIVE TEST<sup>3</sup>**




<sup>3</sup> 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. Two common adaptive approaches used in summative state testing include (1) multi-stage testing and (2) computer- adaptive testing. These two approaches are increasingly adaptive in nature; 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 two common types of adaptive testing are described in further detail below.

1. **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 routing 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.
2. **Computer-adaptive testing (CAT).** Also known as item-level CAT, creates fully individualized forms (essentially) for each student by adapting each item, based on answers to the previous items. CAT produces the most precise and potentially the shortest test (but only if there are no content coverage constraints). 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 over a fixed form test.

The benefits of CAT are maximized when CBT is supported fully throughout a state. If a state opts to support dual mode testing, it becomes much more difficult to maintain comparability across PPT and CAT administrations. 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.

Despite a lengthy discussion focusing on the costs, constraints, benefits, and information provided by adaptive tests, there was no clear consensus on how adaptive testing should be implemented in the new assessment system. Task Force members struggled with the tension between the resources required for developing an adaptive test and the desire to minimize the footprint of the summative assessment in order to create the potential for increased allocation of resources to interim assessments and instructional resources. Therefore, the Task Force made the following recommendation for adaptive testing.



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

**Adaptive Testing.** The Task Force recommended that the NM PED include adaptive testing as a cost option in the RFP for both summative and interim assessments. If adaptive testing can only be supported in one of the components of the assessment system, prioritize its use in the interim assessments *unless the design of the interim solution makes the use of adaptive testing irrelevant* (e.g., freely available item banks or focused modular assessments tied to a series of standards with a limited item bank).

### **Reporting Results**

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

Over the course of the first two meetings and through the Task Force survey, 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 exists between the two facets, the Task Force recommendations focus on the user and the interpretation for each user. Due to the accelerated timeline to make recommendations, the Task Force did not provide detailed plans of what reports might resemble. Instead, they provided general recommendations about the types of information that should be provided to different levels of the education system.

Additionally, the Task Force recognized that different groups of people require different levels of detail in the information reported. As a result, members also made recommendations with regard to the logistical considerations for reporting results. In order to support the use of assessment information by educators, students, and families, Task Force members recommended that the assessment report include both criterion-referenced (comparison to a defined standard) and norm-referenced (comparison to a known group) interpretations. Additionally, those reports should be tailored to each stakeholder group, using relevant and instructive information. Information on the reports should, at a minimum, include:

1. Performance levels
2. Comparative performance percentages reported at the school, district, and state levels
3. Distance from a given performance level
4. Comparisons to other students in the class, school, and state
5. Level and degree of improvement between grade levels
6. Historical trend data (e.g., performance and growth trends)
7. Aggregates of the above information at the student group, school, district, and state levels

As noted in the overall system recommendations, the Task Force recommended that the assessment system be on a single platform to the extent possible. If a common platform is unreasonable, the Task Force recommended that the NM PED supports the use of a single-sign-on (SSO) system to minimize the disruption for educators accessing the assessment information. This may also facilitate the integration of additional resources related to other assessment and accountability initiatives. It is important to note that compliance with the *Family Educational Rights and Privacy Act* (FERPA) is necessary but not sufficient to ensure the integrity of data; privacy of student and educator data; or the security of test, student, and educator data.

## Recommendations for the Summative Assessment in High School

ESSA, like the *No Child Left Behind Act* (NCLB) before it, requires states to administer an English language arts, mathematics, and science assessment at least once in high school. Many states have opted to administer survey assessments generally in grade 11. Such survey assessments attempt to sample from all of the high school standards from the respective content areas and are administered to all students in 11th grade. While such assessments provide a general picture of achievement relative to the high school standards, they are a blunt instrument, given the widely varying course-taking patterns of our students. High school assessment design requires the consideration of many difficult tradeoffs. Task Force deliberations focused on exploring several of these tradeoffs, in an effort to determine what high school assessment options best aligned to the needs and goals of the state.

### ***Purpose and Uses of the Summative Assessments in High School***

The Task Force identified several purposes and uses that the summative assessment should support—many of which overlapped with the grades 3–8 purposes and uses, but several, such as college-readiness testing, were unique to high school.

### ***College Entrance Exams, Survey Tests, and End-of-Course Testing***

One of the key constraints operating in New Mexico is that nationally recognized high school assessments (e.g., the SAT and ACT) have rapidly become institutionalized. There is significant political and public support for having all students complete a nationally recognized high school assessment in 11th grade. The Task Force, throughout discussions, kept the needs of students and schools in New Mexico at the forefront, before making recommendations to the State.

Participants in the Community Conversations indicated a strong desire to administer a nationally recognized high school assessment. However, the Task Force was initially unable to reach consensus. In addition to a nationally recognized college entrance exam, two other options were considered. These included the administration of (1) a survey test in grades 9 and/or 10 in ELA and mathematics, or (2) an end-of-course (EOC) exam in a selection of high school courses. We explore each of these in a bit more detail below before detailing the Task Force's recommendations.

### **Grade 9 and/or 10 Survey Tests**

Many states have decided to administer grade-level tests in grades 9 and 10, tied to the state's ELA and mathematics standards. In fact, several states that have adopted this approach are also administering the ACT or SAT in grade 11. This affords the state several opportunities. The state can measure student learning of the state's own standards in these two grades, and it can use these test results as the achievement indicator for high school accountability, while limiting the SAT or ACT to its validated use as a college readiness indicator. It also provides the opportunity to compute student longitudinal growth measures from middle school through grade 11. On the other hand, grade 9 and 10 survey tests suffer from some of the same challenges as a single survey test in 11th grade, in that students participating in either or both the grades 9 and 10 assessments may be in very different courses, leading to motivation and interpretation challenges of these tests. This may be more of an issue in mathematics where tracking is quite common, but it is perhaps less of an issue in early high school ELA, where students often take the same core classes before moving to electives. Other options include administering pre-college readiness assessments (e.g., PSAT or grades 9/10 ACT). However, these latter options face the challenge of alignment when instruction has been tethered to state standards.

### **End-of-Course Tests**

End-of-course (EOC) tests are common in approximately one-half of the states. As the name implies, EOC tests are tied to specific courses (e.g., Algebra 1, Biology) and are tied closely to the expected content of these course. In certain states, the EOC test results are required to be incorporated into the course grade, while in other states, they are prohibited from counting toward student grades. If these are well-aligned assessments, used to increase student motivation to perform well on the tests, and are of high quality, then we would argue that the results should be allowed to count in the student grade, depending on the wishes of the local school leaders. However, challenges begin to amass quickly if end-of-course tests are developed quickly and without technical rigor.

A major challenge with EOC exams is determining which courses to test. Anyone who has looked at a course catalogue of a comprehensive high school knows that there are hundreds of courses available to students. It would be a financial and logistical nightmare to try to have an EOC testing system to cover most courses. Therefore, states have to prioritize which courses they want to include in their EOC testing system. States with EOC testing systems generally test in Algebra 1, Geometry, English 9, English 10 (or some other high-frequency course such as US Literature), Life Science, and perhaps one of the physical sciences. Some states also include EOC exams in commonly required courses like US History, World History, US Government, and perhaps Economics.

There are many benefits to a high-quality EOC exam system, including potentially creating and raising shared expectations across the state and ensuring that students are evaluated by exams that are generally higher quality than those created locally. However, there are some challenges associated with an EOC exam system. The first, discussed already, is prioritizing which courses are tested and determining how the results are used. The second, which is the converse of shared expectations, is that EOC tests, as observed with Advanced Placement (AP) exams, tend to shape course content and instruction and reduce local control. The most serious challenges, though, involve the cost and capacity necessary to maintain a high-quality EOC system. It costs about as much to develop a single 11th-grade survey test as it does to develop only one EOC exam. Therefore, every additional test that is administered has a direct multiplier effect on the cost of high school testing. Every test requires direct supervision by NM PED personnel to ensure that the vendor is providing the assessment as it has been promised, and it is at the level of quality negotiated. Therefore, more testing means more money to hire more NM PED personnel. As money is not unlimited, those dollars spent on high school testing come at the cost of other ways to support student learning—assessments for elementary and middle school or other additional resources associated with the larger New Mexico assessment system.

### **High school recommendations**

After discussing these options, the Task Force immediately recommended the state move away from end-of-course assessments, in large part because of their current use in the teacher evaluation system. This left the two remaining options of a custom survey test or the use of a nationally recognized high school assessment (i.e., college entrance exams). The Task Force further deliberated on a series of characteristics and value propositions. We provide a general overview of these considerations in the table below.

**TABLE 1. CHARACTERISTICS OF A CUSTOM SURVEY TEST VS A COLLEGE ADMISSIONS TEST**

| Characteristics              | Custom Survey Test                               | College Admissions Test                      |
|------------------------------|--|--|
| <b>Alignment</b>             | High: Assuming appropriate test design processes | Moderate: High with appropriate augmentation |
| <b>Interpretation</b>        | Performance on state standards                   | Readiness for college entrance               |
| <b>Control</b>               | High: Custom                                     | Low: Off the shelf                           |
| <b>Administration Timing</b> | More Flexibility: Within Reason                  | Less Flexibility: Windows Defined            |
| <b>Data Ownership</b>        | Easy: Custom end-of-year                         | Restricted: Currently quite difficult        |
| <b>Cost</b>                  | Typical  | Typical: But customizations are costly       |
| <b>Peer Review Support</b>   | High: Custom                                     | Low to moderate: Still evolving              |
| <b>Perceived Value</b>       | Generally lower outside of K-12 environment      | Generally higher for college readiness       |
| <b>Cultural Relevance</b>    | High flexibility for culturally relevant items   | Low flexibility                              |
| <b>Administrative Burden</b> | State-defined administration procedures          | Vendor-defined administration procedures     |
| <b>Issue Resolution</b>      | State-defined and carried out by vendor          | Vendor-defined with some state input         |


These considerations are not exhaustive, but reflect a general summary of the constraints and trade-offs associated with these two options.

There appeared to be an initial slight preference for a custom survey test in high school during the first Task Force meeting. That shifted toward a college admissions test, based on the belief that a college admissions test could provide a signal to students that they could access a college education even if they previously thought they could not (a diamond in the rough). The Task Force believed that having the State pay for college readiness tests for all students made this an attractive option. The Task Force strongly recommended that the NM PED *not* use a college entrance exam in teacher evaluation or as part of a graduation requirement. This sentiment was shared by an overwhelming majority of the Task Force during the first meeting.

We felt that the Task Force needed to reiterate this recommendation, given the impact that it would have on the assessment system design and procurement process. Thus, we surveyed Task Force members as part of a follow up to the Task Force meeting and confirmed that the vast majority of members believed that a college entrance exam in grade 11 was the most appropriate path forward for the State of New Mexico.

Based on the deliberations over two meetings and the survey, the Task Force made the following recommendations for the high school summative assessment:

1. **College entrance exams.** The Task Force recommended administering a nationally recognized college readiness test in grade 11. This assessment will provide a signal of college readiness to all students and can be supplemented by other components of the State's assessment system.
2. **End-of-course testing.** The NM PED should eliminate current end-of-course testing in high school. This will enable the State to invest more resources into other aspects of the State's assessment system. The Task Force



*College entrance exams. The Task Force recommended administering a nationally recognized college readiness test in grade 11. This assessment will provide a signal of college readiness to all students and can be supplemented by other components of the State's assessment system.*

recognized the constraints associated with the use of EOCs under current policy and requirements. Therefore, we recommend that the NM PED continue to study this issue through other meetings with key stakeholders throughout the state.

## Recommendations for the Interim Assessment for New Mexico

The Task Force discussed two major interim assessment designs. The first, the “mini-summative” design, is the most common among commercial interim assessment providers in which each assessment replicates the end-of-year blueprint. 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, is tied to key subdomains within the standards (e.g., Number-Base 10; see Appendix E for a detailed explanation of different types of interim assessment designs)<sup>4</sup>.

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 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.

After discussing these issues, the Task Force deliberated on whether the State should be providing optional interim exams to districts and exactly how those interims should be structured. The Task Force reasoned that while some districts in New Mexico have resources that enable them to procure or develop local assessments that can be used to support instructional or curricular needs, there are many that do not. Thus, the NM PED provided interims could ensure a “lowest common denominator” of resources to the entire state. Task Force members felt that the optional approach would allow districts to retain current interim solutions if they so choose.

There was less clarity around the design recommendation for interim assessments. The Task Force was evenly split between providing (1) a bank of optional assessment modules aligned to high-leverage reporting categories or groups of standards or (2) a bank of items that could be used to allow teachers to create their own assessments. We have rarely seen freely available banks of items aligned to the standards yield the intended impact. This is due to the level of investment necessary on the part of the State (i.e., in time, money, and personnel to develop the needed number of items and tasks), the amount of assessment literacy necessary for educators to use them well, and the level of time and coordination needed for district leaders to develop focused assessments that inform instructional improvement, curriculum development, and professional development selection. Therefore, we recommend that the NM PED consider incrementally building their interim assessment resources by beginning with modular assessments as well as increasing the availability of assessment literacy resources, so educators can (1) use the available resources to inform their practice and (2) develop local assessments aligned to both the standards and their local curriculum.

The Task Force made several recommendations regarding the interim assessment approach for New Mexico's assessment system and procurement. These recommendations include:

1. **Use of Interims.** The Task Force recommended that the interim assessments provide enough granularity that educators could use them to inform instructional decisions. However, they should also be aggregable to facilitate evaluating programs and curriculum, primarily at the school level and, to a degree, at the district level.

<sup>4</sup> 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: [http://www.ccsso.org/Resources/Publications/Using\\_Interim\\_Assessments\\_in\\_Place\\_of\\_Summative\\_Assessments\\_-\\_Consideration\\_of\\_an\\_ESSA\\_Option.html](http://www.ccsso.org/Resources/Publications/Using_Interim_Assessments_in_Place_of_Summative_Assessments_-_Consideration_of_an_ESSA_Option.html)

2. **Optional Interims.** Task Force members also recommended that, regardless of the design of the interim assessments, they should be optional for district and school use. These interim resources should supplement existing assessment and instructional resources and should remain applicable to historical curricular investment (e.g., maintaining the use of evidence statements).
3. **Interim Design.** The Task Force was split evenly between recommending an item bank and assessment “modules” aligned to high-leverage needs, based on standards and learning targets. We recommend that the NM PED incrementally build interim assessment resources, so educators can (1) use assessments to inform their practice and (2) develop local assessments aligned to both the standards and their local curriculum.
4. **Interim Alignment.** The Task Force recommended that the NM PED adopt optional interim assessments that are based on the same content frameworks as the state summative assessment.
5. **Interim Implementation.** The Task Force recommended that the NM PED implement interim assessments that are on the same platform as the summative assessment, if possible. Additionally, members noted that the assessment platform should be able to expand over time to support authoring assessment content as part of a future-focused assessment system.
6. **Interim Transparency and Security.** A critical recommendation that received unanimous support was that the interim assessment system should be fully transparent and promote educator understanding by involving them in item development, item review, and developing and reviewing associated supporting instructional materials. Furthermore, the Task Force recommended that the system should be freely available for teachers for use, review, and reference on demand (i.e., ensuring interim assessment materials and resources are not secure).

## 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), the three standards for validity are

1. establishing intended uses and interpretations;
2. identifying issues regarding samples and settings used in validation; and
3. recognizing 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 New Mexico’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 NM PED 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, the NM PED must collect evidence regarding fairness, accessibility, lack of bias, generalizability, and appropriateness of performance expectations.

The NM PED should include requirements for their prospective vendor to help identify and collect sources of validity evidence, which include those aforementioned. 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, the NM PED 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 New Mexico prepares its peer review submission and engages in continuous evaluation of the assessment system.

### ***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 TACs 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.

### **Conclusions for Phase I**

This section of the report presented a description of the work of the New Mexico Task Force for Student Success and the various issues deliberated by the Task Force. The Task Force included and represented many stakeholders of the New Mexico 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.

This section included extensive discussion of the many recommendations associated with the design and implementation of a high-quality, statewide assessment system that are part of, and related to, Phase I of the work. Adhering, as closely as possible to the recommendations presented herein regarding the New Mexico assessment system, will help ensure the credibility and stability of the system. Establishing stable partnerships within and outside of the state can help promote the ambitious change reflected in Phase I of the work. Such stability is crucial for supporting advances in educational achievement, growth, and attainment for students and schools in New Mexico.



# KEY DESIGN RECOMMENDATIONS FOR PHASE II

A series of topics were identified that could not be addressed as part of the RFP (i.e., Phase I) process. These topics were discussed briefly and, in most cases, tabled until we could attend more directly to these future-focused topics. These became the priorities for the State's efforts for Phase II of the assessment transition.

This section describes Task Force recommendations for the focus for Phase II of the assessment system.

- High-priority supports for using interim assessments
- Cultural responsiveness and sensitivity
- Writing and authentic assessment
- Supporting and assessing the whole child

## Supports for Using Interim Assessments

While the Task Force made recommendations regarding the design and implementation of interim assessments during the first two Task Force meetings, the third Task Force meeting focused specifically on Phase II topics. One such topic was how to best provide supports to use interim assessments. We presented the Task Force with a series of questions that addressed the following:

- What resources would be helpful to interpret summative assessment results?
- What resources would be helpful to interpret interim assessment results?
- What would help educators translate interim assessment results to instructional next steps?
- How should those resources be delivered to districts and schools (e.g., distribution could range from online documentation to training summits)?

Facilitators also asked Task Force members to consider the high-priority needs for educators. Members responded to a series of questions to focus their discussion on the skills and capacities that the NM PED must consider to support educators. These included questions that addressed the following:

- What are teachers' highest needs to help them better understand and interpret assessment data?
- What resources would help facilitate communication between school, families, and the community around progress and performance?
- What resources would help teachers evaluate the alignment of their own classroom assessments to the standards?
- What other tools might the NM PED provide to help educators instruct in support of the state's standards?

Ultimately, Task Force members made recommendations that can be grouped into three primary areas that can help educators leverage information from the state's assessment system. These are categorized into the following groups: (1) High-priority characteristics of interim assessments, (2) high-priority supports for interim assessments, and (3) high-priority needs for educators. The remainder of this section outlines the recommendations made by Task Force members.

### **High Priority Characteristics**

**Granularity of interim assessments.** While the Task Force spent time making recommendations on how interim assessments should be procured for Phase I, we sought to obtain as much feedback as possible to help inform the NM PED in designing and implementing interim tools and resources. With regard to the granularity of interim assessments, Task Force members generally recommended that utility should be the driving factor for granularity. Specifically, recommendations reflected that interim assessments should

- possess sufficient granularity to inform instruction;
- be closer to the standards/skills, so educators can leverage item-level responses to inform progress;
- be built using "mini units" to facilitate actionable feedback; and

- be informed by local curriculum implementation or be developed locally and aligned to curriculum.

**Frequency of administration.** When considering the frequency of administration, Task Force members provided a range of recommendations that, in some cases, contradicted each other. Members recommended that interim assessment timing should

- be left to the discretion of the classroom teacher based on pacing and student learning;
- reflect pacing of the standards using a “reasonable rate of instruction”; and
- occur minimally twice per year, but as frequently as needed, based on teacher discretion.

*We believe the current iteration of the interim design, described in the RFP, is an important first step in supporting the use of interim assessments. However those interims must be aligned to local instructional needs and combined with assessment literacy efforts.*

Reporting. When asked about how performance should be reported on interim assessments, Task Force members recommended that the interim reports include the following information:

- Learning targets or standards to provide information at the question level
- Performance that is reported in way that can be used for teacher reflection, to drive instruction, and to make programmatic decisions by educators
- Descriptions that can be used by students and parents to focus on individual progress
- Data that communicate mastery of the standards, predict future performance, and confirm instructional observations.

The Task Force recommendations reflect the range of characteristics desired of interim assessments. The NM PED will need to consider how to best support the needs of educators and students, while

balancing costs and the local context of New Mexico districts. We believe the current iteration of the interim design, described in the RFP, is an important first step in supporting the use of interim assessments. However those interims must be aligned to local instructional needs and combined with assessment literacy efforts.

### **High-Priority Supports**

Facilitators also asked the Task Force to consider the supports they believed would be most useful to aid educators in accurately and consistently interpreting summative and interim results and then connecting interim results to instructional next steps. States have been providing summative assessment results for some time, but results are still misinterpreted and misused. The Task Force suggested that educators would benefit from more granular, summative assessment information. Facilitators pushed members, so as to better understand the purpose, role, and intended interpretation of summative assessments. The Task Force’s clarification of the summative assessment’s proper role should decrease the expectation of considering it to be a diagnostic tool. Instead, by focusing on the role of other components of an assessment system, efforts can be directed to leverage more useful information to support teaching and learning. These components might include modular interims aligned to the standards and local pacing, locally developed assessments, project-based tasks, and resources to support formative instruction and practices.

**Interpreting interim assessment results.** When asked about the resources that could help educators interpret interim results and link them to instructional next steps, the Task Force made a series of recommendations. With regard to interpreting interim assessment results, the Task Force recommended that the NM PED provide the following resources:

- Suggested data analysis processes and protocols to model disaggregating and examining results from interim assessments
- Item analysis that includes justification for selecting correct and or incorrect items (i.e., distractors) on provided assessments
- Transparent access to questions, grading rubrics, and student answers
- Webinars for different groups (e.g., teachers, administrators, district testing coordinators) that can help them understand how to interpret data, set goals, and align their own assessment practices to scope and sequence

- Aligned formative materials (i.e., instructional delivery and formative assessment practice resources) that can be adapted and revised

**Translating interim assessment results to instructional next steps.** When considering how educators can best use results from interim assessments to actually inform their instruction, Task Force members' responses shifted toward educator needs, resource availability, and long-term outcomes. Task Force members recommended the following when considering how educators can translate interim assessment results to instructional next steps:

- Support districts and schools in self-reflection to identify educator strengths and weaknesses in both content and pedagogy
- Improve assessment literacy throughout the state for educators, students, and the public
- Increase resources that facilitate peer learning, collaboration, intervention periods, and other structured teacher time for professional development, leadership, and data analysis training
- Support structures to improve leadership capacity (e.g., administrators, teacher leaders, school leadership teams) to better support teachers

The Task Force believes interim assessments are a worthwhile investment, but interims reflect the first of many steps in a larger process to build assessment and instructional literacy among educators in New Mexico. New Mexico's RFP sends a signal regarding the value of assessment literacy that, if implemented well, can lay a solid foundation for future capacity-building efforts to help educators efficiently use assessment data.

### **High-Priority Educator Needs**

The facilitators also asked Task Force members about high-priority educator needs as a means to link interim assessments and their use to actual practice. Task Force members considered teacher needs, the resources necessary to facilitate communication to multiple audiences, and the resources that could help educators evaluate the alignment of their instruction and assessments to the state's standards. Member deliberations reflect their responses to a range of general needs and specific activities, but they can generally be categorized into: (1) building safe environments that support teacher learning and (2) exploration of data and resources to facilitate that learning.



*In addition to providing the tools and resources supporting informed use and development of assessments, the Task Force believes that creating the structures to facilitate professional development and capacity building is critical to improving teaching and learning efforts across New Mexico.*

**Four high-priority needs of educators recommendations.** Specific Task Force recommendations that are unique to the high-priority needs of educators include the following:

- Create safe, non-competitive environments that foster self-reflection, self-evaluation, and self-improvement
- Build the capacity of administrators who can serve as trusted and trusting academic leaders
- Establish a communications campaign that conveys the positive aspects about New Mexico education, assessment, and the role of assessment in learning
- Develop communications, assessment literacy, and data literacy toolkits that can be shared with districts, who passes these to school leaders, who integrate them into school personnel's professional development

Task Force comments reflected the need for a systemic approach to supporting the effective use of assessment results. In addition to providing the tools and resources supporting informed use and development of assessments, the Task Force believes that creating the structures to facilitate professional development and capacity building is critical to improving teaching and learning efforts across New Mexico.

## Cultural Responsiveness and Sustainability

State Education Agencies (SEAs) play a vital role in the provision of a fair and equitable education system. *The Elementary and Secondary Education Act* (ESEA) was a foundational component of the “War on Poverty” (McLaughlin, 1975) and reflected the commitment of the Johnson administration to providing a quality education and improve the skill gaps of students from schools with high concentrations of students from low-income families (see Paul, 2016). While the original ESEA has evolved throughout its amendments and reauthorizations, ESEA’s reauthorization under ESSA remains a law focused on equity and student achievement.

States also establish a series of laws and policies that govern and dictate educational oversight. However, they must still comply with federal requirements under ESSA. In the state of New Mexico, the State is required provide equal educational opportunities to all students :

- *A uniform system of free public schools sufficient for the education of, and open to, all the children of school age in the state shall be established and maintained. N.M. Const. art. XII, § 1.*
- *The legislature shall provide for the training of teachers...so that they may become proficient in both the English and Spanish languages, to qualify them to teach Spanish-speaking pupils. N.M. Const., art. XII, § 8.*
- *Children of Spanish descent in the State of New Mexico shall never be denied the right and privilege of admission and attendance in the public schools...and they shall never be classed in separate schools, but shall forever enjoy perfect equality with other children in all public schools and educational institutions in the state. N.M. Const., art. XII, § 10.*

In light of these constitutional requirements, the state of New Mexico has been challenged by plaintiffs in the early 1970s for funding structure, in 1999 for funding school facilities, and in 2014 by *Martinez vs. State* and *Yazzie vs. State* (ELC, 2019). As a primary contextual driver of the New Mexico Task Force for Student Success, it is important to note that the *Yazzie* and *Martinez* cases were consolidated, and in July 2018, Judge Sarah Singleton ruled in favor of *Yazzie/Martinez*—all New Mexico students have a right to be college and career ready. This decision requires the State to ensure that New Mexico schools have the resources necessary, including sufficient funding, to provide all students with a uniform and sufficient education that prepares them for college and career (New Mexico Center on Law and Poverty, 2019).

### Background and Context

In response to this decision, the NM PED engaged the Task Force to consider how to best support cultural sustainability and responsiveness at the state level. This required a discussion of the interconnected role between the State and other educational entities. Furthermore, the topic required a discussion about how assessment and instruction are interconnected and the role that culture should play in learning and teaching. Although the Task Force discussed the role of culture and culturally responsive pedagogy (see Ladson-Billings, 1994), it was important to also understand what we mean by cultural sustainability.

Sustaining culture requires operating at all three levels of culture: local, national, and global (Laine, 2016). Local culture refers to the local environment and traditions. National culture refers to national heritage. Global culture refers to internationality and multiculturalism. All three of these levels must be considered in the design of assessment to promote cultural sustainability that can be used in instruction.

In order to translate frameworks and vague claims about what would promote cultural responsiveness and sustainability in assessment system design, the facilitators asked the Task Force through a series of questions. We presented these questions in the context of existing assessment practices that seek to make assessments maximally accessible to all students, regardless of race, ethnicity, background, and ability. The field of assessment design is guided by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) which, according to Fairness Standard 3.0, requires that

*All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended use for all examinees in the intended population (p.63).*

Furthermore, the fairness standards stipulate several clusters of requirements to promote valid score interpretations. These clusters include the following:

1. **Cluster 1.** Test Design, Development, Administration, and Scoring Procedures that Minimize Barriers to Valid Score Interpretations for the Widest Possible Range of Individuals and Relevant Subgroups

2. **Cluster 2.** Validity of Test Score Interpretations for Intended Uses for the Intended Examinee Population
3. **Cluster 3.** Accommodations to Remove Construct-Irrelevant Barriers and Support Valid Interpretation of Scores for their Intended Uses for the Intended Examinee Population
4. **Cluster 4.** Safeguards Against Inappropriate Score Interpretation for Intended Uses

Clusters 1 and 3 are particularly relevant to work of cultural sustainability and responsiveness, because assessment systems must be accessible and fair but must also provide valid score interpretations. In the case of the state summative assessment, interpretations must inform whether students are making progress toward college and career readiness by the time they exit the New Mexico education system. By developing assessments (and instructional and assessment supports) that are relevant to students regardless of their cultural heritage, we can maintain confidence in the validity of score interpretations for the population of examinees.

### ***Recommendations for Cultural Responsiveness in Assessment***

To arrive at concrete recommendations regarding the role that the NM PED should play in creating culturally sustainable and responsive assessments, the facilitators asked the Task force to consider (1) the role of culture in assessment development, (2) how to address heritage and diversity in assessment, and (3) the needed supports for stakeholders (e.g., policy makers, legislators, administrators, educators, the public) to understand how assessments fit into the context of equity and culture. While we did not expect that members of the Task Force should be experts in culturally responsive and sustainable education, we believed that member experiences could provide guidance to the NM PED, as they develop a process addressing these issues.

After discussing these three focal areas, the Task Force made several recommendations regarding development, the role of the NM PED, and educator supports. The Task Force first discussed what equity meant in educational settings. They repeatedly noted that fairness and equity look different for different students (e.g., “fair is not equal or the same for everyone”) but does require embracing and sharing common values. Members concurred that the ability to support student diversity of culture and experience requires administrator and teacher training to facilitate school climates conducive to supporting culture and that sensitivity may require scaffolding student experiences. However, the Task Force was clear that accessibility to the content and assessments did not mean that performance expectations should be compromised. Discussions among Task Force members reflected this tension between access and expectation and can be seen in their recommendations below.

**Assessment development.** When asked about the role culture should play in assessment development, the Task Force recommended that the NM PED


- Develop and adhere to a stakeholder-developed framework for cultural responsiveness and sustainability;
- Adhere to best practices in assessment design, with regard to bias and sensitivity, and with a particular focus on culturally specific terms, symbols, and representations;
- Support flexibility in local design, development, and selection of assessments that are sensitive to cultural needs and nuance; and
- Support the shift in the culture of assessment to prioritize the value of assessment systems and the need to support the whole child (this is discussed in more detail in the *Supporting the Whole Child* section of this report).

**The role of the NM PED.** When considering the role of the NM PED in addressing the cultural heritage and diversity of values, the Task Force made the following recommendations:

- Establish a comprehensive training and resource development plan around (1) assessment literacy, (2) the way in which assessments can support cultural heritage, and (3) how performance and access to opportunity differs by location, home language, socio-economic conditions, and racial/ethnic groups
- Identify and facilitate access to literature and resources that celebrate diversity of culture and promote awareness of the diversity of community’s culture
- Develop reporting systems that highlight and help administrators and teachers understand the demographic makeup of their districts, schools, and communities; tie resources that are culturally relevant to those students’ needs and specific to their growth.

**Supports for stakeholders.** When asked about the needed supports for stakeholders (e.g., policy makers, legislators, administrators, educators, the public) to understand how assessments fit into the context of equity and culture, the Task Force recommended the following:

- Decouple the use of summative assessments from graduation requirements and incorporate additional ways to communicate student readiness for transition
- Provide training to legislators and legislative staff on the intended purposes and uses of assessment data to inform legislative changes
- Establish strategies that leverage existing expertise across the state (e.g., universities, strategic partners, advocacy groups) to support communications campaigns, professional development, and training around growth mindset and the use of assessment for continuous improvement
- Develop resources, professional learning communities, and partnerships for educator training on the identification, selection, and implementation of appropriate assessment strategies that privilege the diversity of student backgrounds and experiences (e.g., portfolios, goals setting, capstone projects, celebrations of achievement).



*The Task Force recommended that the New Mexico Public Education Department identify and engage with experts in cultural responsiveness and sustainability*

**Recommendations based on the Task Force deliberations.** The Task Force recommended that the New Mexico Public Education Department identify and engage with experts in cultural responsiveness and sustainability to do the following:

- Establish guiding principles, goals, and a systematic process to create a framework supporting culturally responsive and sustainable assessment development practices
  - Identify a culturally representative stakeholder group who can speak to the affect, behavior, and cognitions—whether through personal experience or training—that should be examined to support students and educators
- Establish an action plan to implement these steps and monitor the intended (and tangible) outputs and outcomes
  - Include any other provisions or recommendations made by experts to support New Mexico students and educators, particularly those of Native American or Hispanic heritage

## Writing and Authentic Assessment

As part of the Phase I work, the Task Force recommended that the assessment system include writing on the statewide summative assessment. As a result, the NM PED is asking vendors to propose how they would approach writing on the state summative assessment that includes

- writing at every grade level on the assessment;
- the three genres of writing matrixed at each grade within a school; and
- automated scoring, if efficient and cost effective.

In order to better design, implement, and support writing throughout the state, the NM PED asked the Task Force to deliberate on the potential supports they should provide beyond the state summative assessment.

**Beyond the summative assessment.** To obtain these recommendations, the facilitators asked the Task Force to consider and propose (1) the types of professional learning (broadly) that are necessary to improve writing instruction, (2) whether a formative writing tool would be helpful, and (3) the types of writing tools—if any—the NM PED should invest in for the state of New Mexico students.

**Support for improvement in writing instruction.** With regard to the types of professional learning broadly necessary to support improvement in writing instruction, the Task Force recommended the following:

- Provide professional learning opportunities that improve educators' content and pedagogical knowledge of writing and how writing can be used across content areas

- Build educator and student literacy in Spanish language arts and bilingual programs to help student transfer skills between languages
- Link language resources and professional development to oral traditions and cultural resources to increase the relevance of writing education to students and educators
- Highlight links between academic writing and real-world modes of writing, which might include memos, business letters, technical write-ups, and media publications

**Formative writing resources.** When prompted about the importance of the NM PED procuring interim assessments or formative writing tools to support writing instruction, the Task Force provided mixed recommendations. While some members believed schools and districts would benefit from the State providing a platform that could facilitate faster scoring of student writing, others believed focusing support on writing resources to improve formative practices was more impactful. If the NM PED decided to provide a formative writing resource, participants recommended the following:

- Provide a writing tool that could promote equitable access (which would require infrastructural changes to technology and hardware) that features a typing component, a tutorial available to all schools, accessibility features, and multi-lingual access
- The ability to aggregate results from student, to classroom, to school, to district
- An optional writing tool that uses automated scoring and feedback across multiple genres that can be used independently
- Leverage the power of state-level procurement to provide a common resource for all districts and schools

**Other options.** On the other hand, participants also recommended that the State should not provide a formative writing tool and instead

- Focus support from the State on the summative assessment and allow local control for interim and formative resources;
- Focus only on formative resources to improve writing instruction;
- Prioritize training to improve educator expertise and improve assessment literacy around writing; and
- Allow districts the autonomy to purchase their own writing resources without State oversight.

**Range of recommendations rather than consensus.** It is important to note that the recommendations made by the Task Force represent the range of recommendations, rather than consensus recommendations; Task Force members did not reach consensus around this topic. Member comments may best be interpreted as reflecting the representative needs of a diverse set of stakeholders (e.g., district administrators, school administrators, teachers, students). We recommend that the NM PED take these recommendations as a comprehensive view of desires in districts and schools. It may be most helpful to focus on learning opportunities for educators, resources to support strong writing instruction, and communicating the value of writing in standards, curriculum, instruction, and assessments—both state- and locally supported.

## Developing and Measuring the Whole Child

When facilitating design groups, we often observe participants discussing vague ideas that are difficult to define and specify. These often include topics like balanced assessments, high-quality instruction, social and emotional skills, and developing the whole child. Throughout facilitation exercises, our goal has been to clarify a concept, uncover participant assumptions, establish a shared understanding of the concept, and articulate specific recommendations that result in strategies to affect a specific outcome related to that concept. A relevant concept that emerged during conversations with the Task Force throughout Phases I and II was that of developing the whole child.

In order to better understand what Task Force participants mean when stating the value of developing the whole child, we started with a common definition and solicited recommendations around two key questions. First, what does developing the whole child mean for a particular stakeholder group (i.e., student, parent, teacher, and guardian)?

Second, what are specific ways that we can measure characteristics related to each group’s definition of the whole child? We initially presented the Task Force with a definition of a whole-child approach to education. According to the ASCD (2019), a whole-child approach to education *is defined by policies, practices, and relationships that ensure each child, in each school, in each community, is healthy, safe, engaged, supported, and challenged.*

Upon reviewing the definition of a whole-child approach, participants engaged in deliberation regarding what it means to support and assess the whole child from the perspective of the student, parent, teacher, and guardian. Instead of making recommendations to the NM PED, the Task Force provided a series of responses that could be used to develop a shared understanding of whole-child success. Based on this conceptualization of developing the whole child, the Task Force then considered possible data elements that could be used to assess facets of whole-child education. We grouped participant responses into four key themes:

1. **Affect.** The emotions students are experiencing—feelings about themselves and others
2. **Behavior.** The actions or interactions students have—interacting with others
3. **Cognition.** The active interpretation of events—thoughts
4. **Context.** The environment or conditions for learning

Using this social psychological framework to analyze participant responses, Task Force comments yielded this four-themed definition of developing the whole child. In paying attention to these critical dimensions, children are prepared for transitions throughout their education, and can leverage a growth mindset to overcome adversity.


**TABLE 2. FOUR-THEMED DEFINITION OF DEVELOPING THE WHOLE CHILD**

|   |   |
|---|---|
| <b>AFFECT</b>   | <b>BEHAVIOR</b>   |
| Students, through self-awareness and self-advocacy, have the capacity to learn from their mistakes, are prepared for transitions throughout their education, and can leverage a growth mindset to overcome adversity. | Students, who have the ability to engage in real-world problem solving—both independently and collaboratively, while being socially and civically active—have the capacity to become contributing citizens.     |
| <b>COGNITION</b>  | <b>CONTEXT</b>  |
| Students, who can responsibly meet rigorous academic, cultural, and career challenges through creativity, critical thinking, and problem solving, have the capacity to become life-long learners.                     | Conditions for learning are provided that meet students’ needs for physical and psychological safety, social justice, and appropriate linguistic input. To that end, the school community is equipped to do so. |

In addition to a shared definition of developing the whole child, the Task Force brainstormed potential sources of data aligned to different levels of the educational system. These data elements included the following:

| DATA TO DRIVE THE CHANGES WE SEEK |   |
|-----------------------------------|---|
| Category                          | Potential Data Elements   |
| <b>Student</b>                    | Engagement • Life skills • Academic performance • Healthy student                                       |
| <b>Parent</b>                     | College and career readiness • Safety • Health • Parent involvement                                     |
| <b>Teacher/Teacher Teams</b>      | Curriculum alignment • Professional development availability • Student growth<br>Principal observations |
| <b>Principal</b>                  | Funding and appropriation • Student growth • Discipline   |
| <b>District</b>                   | Funding and appropriations • Student growth   |
| <b>Contextual Characteristics</b> | Integrated data systems • Continuum of assessments • Data dashboards<br>Early warning systems           |





*The Task Force recommended the New Mexico Public Education Department crosswalk the definition of developing the whole child and the suggested data elements to determine the feasibility of using these sources of information to make defensible inferences about whole-child education.*

The Task Force recommended the New Mexico Public Education Department crosswalk the definition of developing the whole child and the suggested data elements to determine the feasibility of using these sources of information to make defensible inferences about whole-child education. Many of the challenges associated with these data elements are similar to those states face when developing accountability systems—especially under ESSA and the school quality/student success indicator. We invite readers to review Marion and Lyons' (2016) discussion on the challenges in selecting other measures of student success, which are briefly described below in our recommendations.

We additionally recommend that the NM PED examine the intended use (low stakes, local accountability, and informational reporting) and the potential level of comparability (i.e., within classroom, within school, within district, and across district) to determine appropriate ranges for each of the following characteristics:

- **Usability.** Do measures reflect lagging (e.g., outcome) or leading (e.g., process) indicators? What information does this indicator provide to stakeholders and for what purpose?
- **Level of Inference.** To what extent does an indicator require a high degree of inference (e.g., school climate based on surveys) vs. a low level of inference, such as counts of data (e.g., days attended or credits earned). While both of these require some inference to their interpretation, the number of logical steps to support that inference can vary significantly.
- **Data burden on districts.** Would this data element place a low level of burden on districts because it uses previously collected data? Does this data element place a high level of burden on districts because new policies, procedures, and mechanisms are needed?
- **Potential corruptibility.** To what extent are data “gameable” in the system? For example, standardized assessments are more difficult to game than self-reported data. However, all data are subject to the risk of corruption if used for high-stakes purposes (see Campbell's law).
- **Comparability.** What is the degree of comparability desired compared to the level of comparability a data element actually affords?

Once data elements are identified, it will be important for the NM PED to evaluate the appropriateness of each measure's intended use and whether it facilitates the intended changes in practice at the district and school levels. Whatever the state decides regarding developing the whole child, it will be necessary to base it on a well-articulated theory of action and ensure the components of that theory of action are measurable.

## Supporting Formative Assessment Practices and Project-Based Learning

One of the key issues the Task Force identified as a near-term priority for NM PED's Phase II work was supporting formative assessment practices and project-based learning. While an important focus, there was insufficient time to address these topics during the Task Force discussions. We recommend that the NM PED continue to work with members of the Task Force, SEA leaders, and other stakeholders across the education system in New Mexico to uncover the topic-relevant perspectives, goals, and needs for students and educators.

## Conclusions for Phase II

This section of the report presented a description of the issues, work, and recommendations of the New Mexico Task Force for Student Success focused on Phase II of the assessment system. Instead of focusing on the operations of assessment systems, this section of the report focused on the long-term needs for and the structures necessary to support the transition to a high-quality assessment system. Many of the recommendations presented in this section represent a

consensus of the Task Force. Where consensus could not be reached, that was explicitly noted. Additionally, there were many areas in which the Task Force was simply unable to make a recommendation. For these, we analyzed Task Force comments and attempted to synthesize them into next steps for the NM PED to consider. While the Phase I components of the assessment system are alone ambitious, the NM PED will need to plan carefully to also include the aspects identified in Phase II. Despite the challenges associated with implementing both Phase I and Phase II features, the findings in this section reflect a promising vision for the future of the New Mexico assessment system. As part of the development process, we suggest continuing to work with members of the existing Task Force and build out the assessment components that we did not have time to address.

## REFERENCES/SOURCES CONSULTED

- AERA, APA, & NCME, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Association for Supervision and Curriculum Development (ASCD, 2019). *A whole child approach to education and the Common Core State Standards initiative*. Retrieved from <http://www.ascd.org/ASCD/pdf/siteASCD/policy/CCSS-and-Whole-Child-one-pager.pdf>.
- Binet, A., & Simon, Th. A. (1905). Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique*, 11, 191–244.
- 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.
- Education Law Center (2019). New Mexico state constitution and major cases. Retrieved from <https://edlawcenter.org/states/newmexico.html>
- Ladson-Billings, G. (1994). *The dreamkeepers*. San Francisco: Jossey-Bass Publishing Co.
- Laine, M. (2016). Culture in sustainability: Defining cultural sustainability in education. *Discourse and Communication for Sustainable Education*, 7, 52–67.
- Marion, S.F. (2018). The opportunities and challenges of a systems approach to assessment. *Educational Measurement: Issues and Practice*, 37, 1, 45–48.
- 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](https://www.nciea.org/sites/default/files/publications/PACE%20Principled%20assessment%20design_092417.pdf)
- Marion, S. F., & Lyons, S. (2016). *In search of unicorns: Conceptualizing and validated the “fifth indicator” in ESSA accountability systems*. National Center for the Improvement of Educational Assessment. [https://www.nciea.org/sites/default/files/pubs-tmp/Marion%20Lyons\\_ESSA%20Accountability\\_5th%20Indicator\\_111416.pdf](https://www.nciea.org/sites/default/files/pubs-tmp/Marion%20Lyons_ESSA%20Accountability_5th%20Indicator_111416.pdf)
- 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](http://www.innovativeassessments.org).
- McLaughlin, M. (1975). *Evaluation and reform: The Elementary and Secondary Education Act of 1965, Title I*. Cambridge, Massachusetts: Ballinger Publishing Company.
- Michigan Department of Education. (2013). *Report on Options for Assessments Aligned with the Common Core State Standards*. Retrieved June 20, 2015, from [http://www.michigan.gov/documents/mde/Common\\_Core\\_Assessment\\_Option\\_Report\\_441322\\_7.pdf](http://www.michigan.gov/documents/mde/Common_Core_Assessment_Option_Report_441322_7.pdf)
- Mislevy, R. J. (1996). Evidence and inference in educational assessment. CRESST Technical Report No. 414. <https://pdfs.semanticscholar.org/5eae/5388283e95a3a8f3e5b291bcae7f3558dd44.pdf>
- Mislevy, R. J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice*, 25: 6–20.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.
- 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, & Alexandra S. Beatty, *Editors*. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

- New Mexico Center on Poverty and Law (2019). *Yazzie/Martinez v. State of New Mexico Decision*. Retrieved from <http://nmpovertylaw.org/wp-content/uploads/2018/09/Graphic-Yazzie-Martinez-Decision.pdf>
- Paul, C. A. (2016). Elementary and Secondary Education Act of 1965. *Social Welfare History Project*. Retrieved from <http://socialwelfare.library.vcu.edu/programs/education/elementary-and-secondary-education-act-of-1965/>.
- 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.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144.
- 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
- Wiley, E. C. (2008). *Formative Assessment: Examples of Practice*. Retrieved August 11, 2015, from [http://ccsso.org/documents/2008/formative\\_assessment\\_examples\\_2008.pdf](http://ccsso.org/documents/2008/formative_assessment_examples_2008.pdf)

## 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 (AI), 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 produces fully individualized tests for each 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) are delivered and captured on an electronic device (e.g., desktop, laptop, tablet).

**Fairness:** Fairness emphasizes that the test must be impartial, 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 items are measuring what they purport to measure and whether the test interpretation validly reflects the claims it intends to make.

**Formative Assessment:** Formative assessment is a planned, ongoing process used by all students and teachers during learning and teaching to elicit and then use evidence of student learning to improve student understanding of intended curricular learning outcomes and support students to become self-directed learners.

**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:** Interim assessments are assessments administered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals, in order to inform educator and policymaker decisions at the classroom, school, and district levels. The specific designs of the interim assessment are driven by the purpose and intended uses, but the results of any interim assessment must be aggregable for reporting across students, occasions, and concepts.

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

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

**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.

**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).

**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, prompt) and response (e.g., score sheet, short answer form).

**Peer Review:** Peer review refers to the process used by the US Department of Education to evaluate the degree to which state assessment systems meet the technical and inclusion requirements spelled out in law and regulations. Peers are individuals with technical and/or operational expertise and experience with state assessment systems.

**Reliability:** Generally, reliability refers to the pieces of information or evidence that help us determine whether a test is precise and provides the same results consistently under the same conditions, sufficient for the intended use of that test.

**Summative Assessment:** Summative assessments are designed to support various types of determinations (e.g., proficiency, competency) given at the end of a defined instructional period, such as a unit of instruction or a school year, to evaluate students' performance against a set of learning targets for that period.

**Survey Test:** A survey test is administered to all students in a given grade, designed to broadly cover the grade level or grade span content standards in that subject area (Marion, 2018).

**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, comprehensive support of multiple purposes and uses, and continuous documentation of 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).

Coherence, utility, and efficiency. 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 then 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) asks 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 is not shared across levels of the system, such loosely-coupled systems are not as efficient as ones in which information from one level (e.g., classroom) is 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 that require 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, in which 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: COMMUNITY CONVERSATIONS SUMMARY (BRIEF #10)

Erika Landl & Juan D’Brot, Center for Assessment

April 5, 2019

As part of an assessment transition plan, the New Mexico Department of Education (NM PED) conducted listening sessions through a series of Community Conversations around the state. The themes that emerged from the Community Conversations are important for the Task Force to consider as part of the deliberation and recommendation process. This brief provides an overview of the themes and some details that emerged from across the Community Conversations that address the (1) prioritized purposes and uses of an assessment system (2) the ideal assessment system features and (3) needs to ensure assessment system success.

## Prioritized Purposes and Uses of Assessment System Results

Across the Community Conversations, several common purposes and uses emerged. These include:

- **Identifying specific areas of student need** and the kinds of feedback that informs instruction and drives improvement
- **Evaluating student growth and progress** over time (within and across years) on instructed standards
- **Informing the development of IEP’s** (e.g., establish appropriate goals)
- **Providing information about the whole child** to understand needs beyond the standards
- **Informing instruction with actionable feedback** tied to instructional resources. This occurs on an on-going basis throughout the year.
- **Informing decisions about professional development** at the teacher, school, and district levels
- **Predicting performance** on the end of year assessment

We would like to reiterate that these purposes and uses are specific to the entire assessment system, as no one assessment can serve all of these purposes and uses. It will be important for us to further prioritize and refine these purposes and uses, as they relate to different assessments within the system.

## Desired Assessment System Features

In addition to the desired purposes and uses, the following desired assessment features emerged from the Community Conversations. In addition to each of the themes, we also present comments representative across each of the Community Conversation meetings.

**TABLE 2. FOUR-THEMED DEFINITION OF DEVELOPING THE WHOLE CHILD**

| THEMES           | SUMMARY OF COMMENTS  |
|------------------|--|
|                  | <b>Assessments within the system</b>   |
| <b>Relevance</b> | <ul style="list-style-type: none"> <li>• are meaningful and have value and purpose for students beyond high school;</li> <li>• measure content/courses that students have already taken (especially in HS);</li> <li>• provide relevant assessments of students with disabilities and IEP’s that produce useful information to inform individualized instruction; and</li> <li>• provide testing consistent with a student’s ability level, even if off-grade (e.g., Computer Adaptive Testing)</li> </ul> |



| THEMES                                  | SUMMARY OF COMMENTS  |
|---|--|
|   | <b>Assessments within the system</b>   |
| <b>Fairness and Accessibility</b>       | <ul style="list-style-type: none"> <li>• are fair (culturally, linguistically, geographically), developmentally appropriate, and relevant to all students</li> <li>• are accessible to all test takers (multiple languages, assistance and accommodations in native language/images, should mimic instruction)</li> <li>• are untimed to reduce stress</li> </ul>  |
| <b>Alignment</b>                        | <ul style="list-style-type: none"> <li>• are aligned to state content standards and the Common Core</li> <li>• are aligned with information provided in text books used for instruction</li> <li>• are aligned to provide ongoing and progressive information about student understanding of the standards (e.g., interim aligned to summative) <ul style="list-style-type: none"> <li>- There is a pool of items that supports development of local assessments aligned to those interim and summative tests</li> </ul> </li> </ul>   |
|   | <b>Assessments within the system</b>   |
| <b>Growth/ Progress</b>                 | <ul style="list-style-type: none"> <li>• support inferences on progress and growth toward proficiency rather than status; and</li> <li>• support testing throughout the year to evaluate student progress.</li> </ul>  |
|   | <b>The assessment system facilitates</b>   |
| <b>Multiple Measures of Performance</b> | <ul style="list-style-type: none"> <li>• collecting data/measures of performance throughout the year;</li> <li>• collecting information about life and work skills that are important after high school (e.g., problem solving, critical thinking, and communication);</li> <li>• addressing foundational skills and standards-based skills (<b>Note:</b> a few comments specifically noted that Istation addresses only foundational content); and</li> <li>• supporting ways of demonstrating understanding (e.g., verbal-, portfolio-, artifact-based evidence).</li> </ul> |
| <b>Comparability</b>                    | <ul style="list-style-type: none"> <li>• a comparison between schools and districts (and potentially compare performance to that of other states and the nation).</li> </ul>   |

Similar to the themes of purposes and uses, many themes regarding ideal assessment features cannot be met by any single assessment. Task Force members should consider these desired features, as we engage in the deliberation and recommendation process.

### Characteristics to ensure the Assessment System is Successful

The following themes emerged from the Community Conversations that were focused on the characteristics that would best support the success of the assessment system. In addition to each of the themes, we also present a summary of the comments that are representative across each of the meetings.

**TABLE 2. DESIRED ASSESSMENT SYSTEM THEMES AND REPRESENTATIVE COMMENTS**

| THEMES   | SUMMARY OF COMMENTS  |
|--|--|
| <b>Clear, Useful, Timely Reporting</b>                                 | <ul style="list-style-type: none"> <li>• Reports provide useful, meaningful, and detailed information that helps teachers, parents, and students understand how and where improvement is needed.</li> <li>• Timely reporting is critical               <ul style="list-style-type: none"> <li>- immediate feedback for anything machine scored</li> <li>- summative assessment results prior to the end of the year</li> </ul> </li> <li>• Reports are clear, understandable, and meaningful for all stakeholders.</li> <li>• Reports are easy for teachers to access on demand and provide detail about specific areas of need. (e.g., standards-based reporting).</li> <li>• Normative information is provided (e.g., national norms).</li> <li>• Data dashboards are used that present relevant information.</li> </ul> |
| <b>A Consistent Platform and Tools that Support the Use of Results</b> | <ul style="list-style-type: none"> <li>• A common platform is used across all assessments in the system to provide for consistency in administration, access, login, accessibility features, and reporting features.</li> <li>• Easy access is provided to detailed longitudinal and trend data for students, schools, districts—overall and disaggregated.</li> <li>• Good tools are available for schools to communicate with parents.</li> <li>• Data-informed resources and supports are available to guide students help themselves.</li> </ul>   |
| <b>Professional Development/ Training</b>                              | <ul style="list-style-type: none"> <li>• Professional development for educators is provided that trains them in using assessment results to improve instruction and communicate with parents (e.g., instructional supports)</li> <li>• Stakeholder training is provided that informs them on the use of assessment data (what it can and what it should do).</li> </ul>  |
| <b>Reduced Testing Time and Frequency</b>                              | <ul style="list-style-type: none"> <li>• Testing is consolidated testing, so there can be a greater focus on learning and instructional formative assessments.</li> </ul>  |
| <b>Transparency</b>  | <ul style="list-style-type: none"> <li>• Greater transparency is provided as to what students will be assessed on (not just standards but skills).</li> <li>• A clear and focused transition plan is developed and communicated for the new assessment system.</li> </ul>  |

The characteristics of a successful assessment system, as described by community members, are relevant to the state summative assessment and the assessment system overall. As the Task Force considers these comments from the Community Conversations, it will be critical to address the tensions among the desired purposes, uses, features, and characteristics, while identifying what aspects of the assessment system address the needs of students in New Mexico.

We, at the Center for Assessment, are excited to engage with Task Force members and the NM PED to establish coherent recommendations for New Mexico’s assessment transition plan. We look forward to hearing your thoughts, concerns, and ideas for the state of New Mexico.

## APPENDIX D: 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 (i.e., two-year and four-year colleges, universities, gainful employment). Assessment developers have had to ensure that 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, 1996, Mislevy & Haertel, 2006) or through the use of the Assessment Triangle (NRC, 2001). The Assessment Triangle draws a connection between 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 when creating tasks (e.g., how to: develop test questions that require that students engage more deeply with a question or prompt, examine evidence of the accuracy of the response, and what is expected of students in their presentation of a coherent response?) that are also germane to the assessment development process. These questions might look like the following:

- 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 make a task harder or easier?
- What characteristics/features 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 New Mexico's next assessment system. Using the pre-reading materials and re-teaching, the facilitators extended the application of this approach to clarify how assessment operations should be defined throughout the Task Force's deliberation.

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

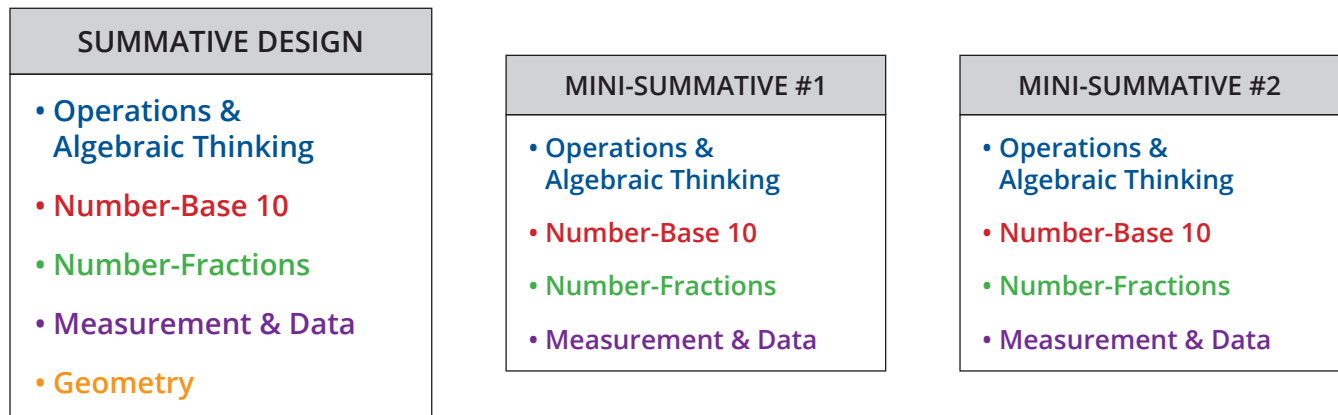
Throughout conversations with the Task Force, we defined and described the assessments types and uses presented here, in order to ensure that members had a shared understanding of assessment. While many of the conversations came back to close-to-the-classroom and in-the-moment instruction (i.e., formative assessment), it was necessary to focus the Task Force on the role of the State (and how the State can support districts) in providing relevant assessment information and communicating those results in meaningful ways. 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—and *interim* assessments to support progress towards meeting requirements described by the standards, which are measured through the state summative assessments.

State-wide summative assessments are, generally 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. Thus, the Task Force spent some time discussing the role and timing and utility of both summative and interim assessments in the educational system. Recommendations for the summative and interim assessments are provided in the *Key Design Recommendations for Phase I* section of this report.

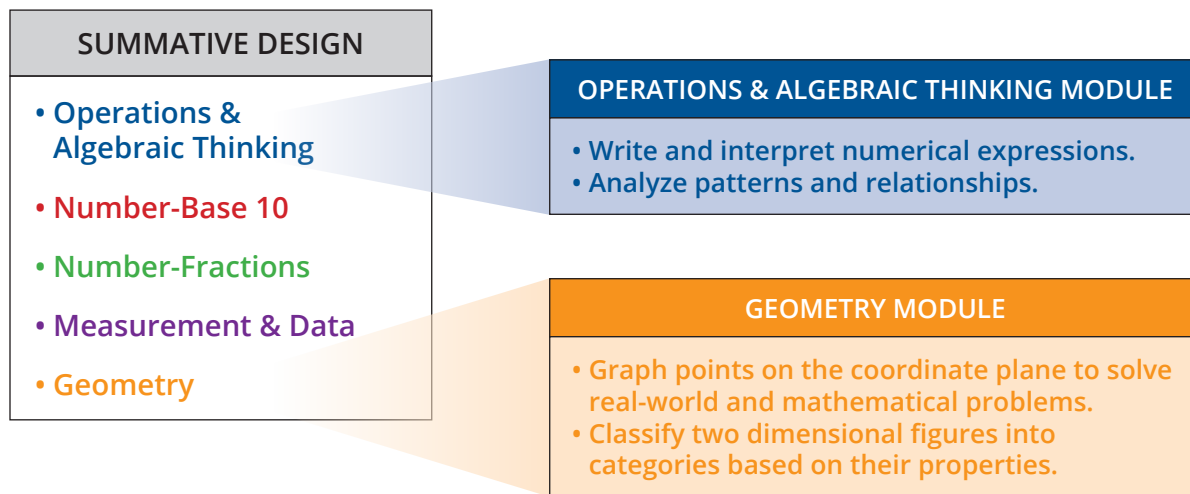
# APPENDIX E: MINI-SUMMATIVE VS. MODULAR INTERIM ASSESSMENT DESIGNS

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 provide 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  |
|---|
| <p><b>Operations &amp; Algebraic Thinking</b></p> <ul style="list-style-type: none"> <li>• Write and interpret numerical expressions               <ul style="list-style-type: none"> <li><i>Use parentheses, brackets, or braces...</i></li> <li><i>Write simple expressions that record calculations...</i></li> </ul> </li> <li>• Analyze patterns and relationships               <ul style="list-style-type: none"> <li><i>Generate...numerical patterns...given rules...</i></li> </ul> </li> </ul>   |
| <p><b>Number &amp; Operations in Base Ten</b></p> <ul style="list-style-type: none"> <li>• Understand the place value system               <ul style="list-style-type: none"> <li><i>Recognize [digit values increase tenfold when one place... left]</i></li> <li><i>Explain patterns in...multiplying by powers of 10...</i></li> <li><i>Read, write, and compare decimals to thousandths</i></li> <li><i>Use place value understanding to round decimals to any place</i></li> </ul> </li> <li>• Perform operations...to hundredths               <ul style="list-style-type: none"> <li><i>Fluently multiply multi-digit whole numbers...</i></li> <li><i>Find whole-number quotients of whole numbers...</i></li> <li><i>Add, subtract, multiply, and divide decimals to hundredths...</i></li> </ul> </li> </ul>    |
| <p><b>Number &amp; Operations—Fractions</b></p> <ul style="list-style-type: none"> <li>• Use equivalent fractions...to add and subtract fractions               <ul style="list-style-type: none"> <li><i>Add and subtract fractions with unlike denominators...</i></li> <li><i>Solve [fraction word problems by comparison...]</i></li> </ul> </li> <li>• Apply and extend...multiplication and division               <ul style="list-style-type: none"> <li><i>Interpret a fraction [as a division problem]...</i></li> <li><i>[Extend whole number] multiplication to...fractions...</i></li> <li><i>Interpret multiplication as scaling (resizing)...</i></li> <li><i>Solve...problems [with] multiplication of fractions...</i></li> <li><i>[Extend division to involve unit fractions]</i></li> </ul> </li> </ul> |
| <p><b>Measurement &amp; Data</b></p> <ul style="list-style-type: none"> <li>• Convert like measurement units [in the same] system               <ul style="list-style-type: none"> <li><i>Convert among different sized measurement units...</i></li> </ul> </li> <li>• Represent and interpret data               <ul style="list-style-type: none"> <li><i>Make a line plot to display [data with fractional units]...</i></li> </ul> </li> <li>• Geometric measurement: understand...volume               <ul style="list-style-type: none"> <li><i>Understand volume as an attribute of solid figures...</i></li> <li><i>Measure volumes by counting unit cubes...</i></li> <li><i>Relate volume to [multiplication and division]...</i></li> </ul> </li> </ul>   |
| <p><b>Geometry</b></p> <ul style="list-style-type: none"> <li>• Graph points on the coordinate plane to solve...               <ul style="list-style-type: none"> <li><i>Use [two] perpendicular lines...to define a coordinate...</i></li> <li><i>Represent... points in the first quadrant...</i></li> </ul> </li> <li>• Classify two-dimensional figures...on...properties               <ul style="list-style-type: none"> <li><i>[Know category] attributes [apply] to all sub-categories...</i></li> <li><i>Classify...figures in a hierarchy based on properties</i></li> </ul> </li> </ul>  |

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 |   |
|--|------------|---|
| <b>Operations &amp; Algebraic Thinking</b> <ul style="list-style-type: none"> <li>• Write and interpret numerical expressions <ul style="list-style-type: none"> <li><i>Use parentheses, brackets, or braces...</i></li> <li><i>Write simple expressions that record calculations...</i></li> </ul> </li> <li>• Analyze patterns and relationships <ul style="list-style-type: none"> <li><i>Generate...numerical patterns...given rules...</i></li> </ul> </li> </ul>   | <b>6</b>   |   |
|  | 4          | 2 |
|  |            | 2 |
|  | 2          | 2 |
| <b>Number &amp; Operations in Base Ten</b> <ul style="list-style-type: none"> <li>• Understand the place value system <ul style="list-style-type: none"> <li><i>Recognize [digit values increase tenfold when one place... left]</i></li> <li><i>Explain patterns in...multiplying by powers of 10...</i></li> <li><i>Read, write, and compare decimals to thousandths</i></li> <li><i>Use place value understanding to round decimals to any place</i></li> </ul> </li> <li>• Perform operations...to hundredths <ul style="list-style-type: none"> <li><i>Fluently multiply multi-digit whole numbers...</i></li> <li><i>Find whole-number quotients of whole numbers...</i></li> <li><i>Add, subtract, multiply, and divide decimals to hundredths...</i></li> </ul> </li> </ul>    | <b>14</b>  |   |
|  | 8          | 2 |
|  |            | 2 |
|  |            | 2 |
|  |            | 2 |
|  | 6          | 2 |
|  |            | 2 |
|  |            | 2 |
| <b>Number &amp; Operations—Fractions</b> <ul style="list-style-type: none"> <li>• Use equivalent fractions...to add and subtract fractions <ul style="list-style-type: none"> <li><i>Add and subtract fractions with unlike denominators...</i></li> <li><i>Solve [fraction word problems by comparison...]</i></li> </ul> </li> <li>• Apply and extend...multiplication and division <ul style="list-style-type: none"> <li><i>Interpret a fraction [as a division problem]...</i></li> <li><i>[Extend whole number] multiplication to...fractions...</i></li> <li><i>Interpret multiplication as scaling (resizing)...</i></li> <li><i>Solve...problems [with] multiplication of fractions...</i></li> <li><i>[Extend division to involve unit fractions]</i></li> </ul> </li> </ul> | <b>14</b>  |   |
|  | 4          | 2 |
|  |            | 2 |
|  | 10         | 2 |
|  |            | 2 |
|  |            | 2 |
|  |            | 2 |
|  |            | 2 |
| <b>Measurement &amp; Data</b> <ul style="list-style-type: none"> <li>• Convert like measurement units [in the same] system <ul style="list-style-type: none"> <li><i>Convert among different sized measurement units...</i></li> </ul> </li> <li>• Represent and interpret data <ul style="list-style-type: none"> <li><i>Make a line plot to display [data with fractional units]...</i></li> </ul> </li> <li>• Geometric measurement: understand...volume <ul style="list-style-type: none"> <li><i>Understand volume as an attribute of solid figures...</i></li> <li><i>Measure volumes by counting unit cubes...</i></li> <li><i>Relate volume to [multiplication and division]...</i></li> </ul> </li> </ul>   | <b>10</b>  |   |
|  | 2          | 2 |
|  | 2          | 2 |
|  | 6          | 2 |
|  |            | 2 |
|  |            | 2 |
|  |            | 2 |
| <b>Geometry</b> <ul style="list-style-type: none"> <li>• Graph points on the coordinate plane to solve... <ul style="list-style-type: none"> <li><i>Use [two] perpendicular lines...to define a coordinate...</i></li> <li><i>Represent... points in the first quadrant...</i></li> </ul> </li> <li>• Classify two-dimensional figures...on...properties <ul style="list-style-type: none"> <li><i>[Know category] attributes [apply] to all sub-categories...</i></li> <li><i>Classify...figures in a hierarchy based on properties</i></li> </ul> </li> </ul>  | <b>8</b>   |   |
|  | 4          | 2 |
|  |            | 2 |
|  | 4          | 2 |
|  |            | 2 |
|  |            | 2 |
| <b>TOTAL</b>   | <b>52</b>  |   |

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         |
| <b>Operations &amp; Algebraic Thinking</b> <ul style="list-style-type: none"> <li>Write and interpret numerical expressions               <ul style="list-style-type: none"> <li><i>Use parentheses, brackets, or braces...</i></li> <li><i>Write simple expressions that record calculations...</i></li> </ul> </li> <li>Analyze patterns and relationships               <ul style="list-style-type: none"> <li><i>Generate...numerical patterns...given rules...</i></li> </ul> </li> </ul>   | 3          | 3         | 3         | 3         | 3         |
|  | 2          | 2         | 2         | 2         | 2         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
| <b>Number &amp; Operations in Base Ten</b> <ul style="list-style-type: none"> <li>Understand the place value system               <ul style="list-style-type: none"> <li><i>Recognize [digit values increase tenfold when one place... left]</i></li> <li><i>Explain patterns in...multiplying by powers of 10...</i></li> <li><i>Read, write, and compare decimals to thousandths</i></li> <li><i>Use place value understanding to round decimals to any place</i></li> </ul> </li> <li>Perform operations...to hundredths               <ul style="list-style-type: none"> <li><i>Fluently multiply multi-digit whole numbers...</i></li> <li><i>Find whole-number quotients of whole numbers...</i></li> <li><i>Add, subtract, multiply, and divide decimals to hundredths...</i></li> </ul> </li> </ul>    | 7          | 7         | 7         | 7         | 7         |
|  | 4          | 4         | 4         | 4         | 4         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 3          | 3         | 3         | 3         | 3         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
| <b>Number &amp; Operations—Fractions</b> <ul style="list-style-type: none"> <li>Use equivalent fractions...to add and subtract fractions               <ul style="list-style-type: none"> <li><i>Add and subtract fractions with unlike denominators...</i></li> <li><i>Solve [fraction word problems by comparison...]</i></li> </ul> </li> <li>Apply and extend...multiplication and division               <ul style="list-style-type: none"> <li><i>Interpret a fraction [as a division problem]...</i></li> <li><i>[Extend whole number] multiplication to...fractions...</i></li> <li><i>Interpret multiplication as scaling (resizing)...</i></li> <li><i>Solve...problems [with] multiplication of fractions...</i></li> <li><i>[Extend division to involve unit fractions]</i></li> </ul> </li> </ul> | 7          | 7         | 7         | 7         | 7         |
|  | 2          | 2         | 2         | 2         | 2         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 5          | 5         | 5         | 5         | 5         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
| <b>Measurement &amp; Data</b> <ul style="list-style-type: none"> <li>Convert like measurement units [in the same] system               <ul style="list-style-type: none"> <li><i>Convert among different sized measurement units...</i></li> </ul> </li> <li>Represent and interpret data               <ul style="list-style-type: none"> <li><i>Make a line plot to display [data with fractional units]...</i></li> </ul> </li> <li>Geometric measurement: understand...volume               <ul style="list-style-type: none"> <li><i>Understand volume as an attribute of solid figures...</i></li> <li><i>Measure volumes by counting unit cubes...</i></li> <li><i>Relate volume to [multiplication and division]...</i></li> </ul> </li> </ul>   | 5          | 5         | 5         | 5         | 5         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 3          | 3         | 3         | 3         | 3         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
| <b>Geometry</b> <ul style="list-style-type: none"> <li>Graph points on the coordinate plane to solve...               <ul style="list-style-type: none"> <li><i>Use [two] perpendicular lines...to define a coordinate...</i></li> <li><i>Represent... points in the first quadrant...</i></li> </ul> </li> <li>Classify two-dimensional figures...on...properties               <ul style="list-style-type: none"> <li><i>[Know category] attributes [apply] to all sub-categories...</i></li> <li><i>Classify...figures in a hierarchy based on properties</i></li> </ul> </li> </ul>  | 4          | 4         | 4         | 4         | 4         |
|  | 2          | 2         | 2         | 2         | 2         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 2          | 2         | 2         | 2         | 2         |
|  | 1          | 1         | 1         | 1         | 1         |
|  | 1          | 1         | 1         | 1         | 1         |
| <b>TOTAL</b>   | <b>26</b>  | <b>26</b> | <b>26</b> | <b>26</b> | <b>26</b> |

Multiple interim assessments built to this design would have different sets of test questions, but they would share the same emphasis on each of the content categories as 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 approximately the same length as the mini-summative designs:

| CONTENT CATEGORY   | # OF ITEMS                   |   |  |  |                                    |
|--|------------------------------|---|--|--|------------------------------------|
|  | 1                            | 2   | 3  | 4                                      | 5                                  |
| <b>Operations &amp; Algebraic Thinking</b><br><ul style="list-style-type: none"> <li>Write and interpret numerical expressions<br/><i>Use parentheses, brackets, or braces...</i><br/><i>Write simple expressions that record calculations...</i></li> <li>Analyze patterns and relationships<br/><i>Generate...numerical patterns...given rules...</i></li> </ul>   | 27<br>18<br>9<br>9<br>9<br>9 |   |  |  |                                    |
| <b>Number &amp; Operations in Base Ten</b><br><ul style="list-style-type: none"> <li>Understand the place value system<br/><i>Recognize [digit values increase tenfold when one place... left]</i><br/><i>Explain patterns in...multiplying by powers of 10...</i><br/><i>Read, write, and compare decimals to thousandths</i><br/><i>Use place value understanding to round decimals to any place</i></li> <li>Perform operations...to hundredths<br/><i>Fluently multiply multi-digit whole numbers...</i><br/><i>Find whole-number quotients of whole numbers...</i><br/><i>Add, subtract, multiply, and divide decimals to hundredths...</i></li> </ul>    |                              | 28<br>16<br>4<br>4<br>4<br>4<br>12<br>4<br>4<br>4 |  |  |                                    |
| <b>Number &amp; Operations—Fractions</b><br><ul style="list-style-type: none"> <li>Use equivalent fractions...to add and subtract fractions<br/><i>Add and subtract fractions with unlike denominators...</i><br/><i>Solve [fraction word problems by comparison...]</i></li> <li>Apply and extend...multiplication and division<br/><i>Interpret a fraction [as a division problem]...</i><br/><i>[Extend whole number] multiplication to...fractions...</i><br/><i>Interpret multiplication as scaling (resizing)...</i><br/><i>Solve...problems [with] multiplication of fractions...</i><br/><i>[Extend division to involve unit fractions]</i></li> </ul> |                              |   | 28<br>8<br>4<br>4<br>20<br>4<br>4<br>4<br>4<br>4 |  |                                    |
| <b>Measurement &amp; Data</b><br><ul style="list-style-type: none"> <li>Convert like measurement units [in the same] system<br/><i>Convert among different sized measurement units...</i></li> <li>Represent and interpret data<br/><i>Make a line plot to display [data with fractional units]...</i></li> <li>Geometric measurement: understand...volume<br/><i>Understand volume as an attribute of solid figures...</i><br/><i>Measure volumes by counting unit cubes...</i><br/><i>Relate volume to [multiplication and division]...</i></li> </ul>   |                              |   |  | 25<br>5<br>5<br>5<br>15<br>5<br>5<br>5 |                                    |
| <b>Geometry</b><br><ul style="list-style-type: none"> <li>Graph points on the coordinate plane to solve...<br/><i>Use [two] perpendicular lines...to define a coordinate...</i><br/><i>Represent... points in the first quadrant...</i></li> <li>Classify two-dimensional figures...on...properties<br/><i>[Know category] attributes [apply] to all sub-categories...</i><br/><i>Classify...figures in a hierarchy based on properties</i></li> </ul>   |                              |   |  |  | 28<br>14<br>7<br>7<br>14<br>7<br>7 |
| <b>TOTAL</b>   | 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.