STATEWIDE SUMMATIVE ASSESSMENT IN SPRING 2021:

A WORKBOOK TO SUPPORT PLANNING AND DECISION-MAKING

Version 1.0 | September 2020

Michelle Boyer, Nathan Dadey, Leslie Keng
The Center for Assessment

Center for Assessment

National Center for the Improvement
of Educational Assessment
Dover, New Hampshire

STATEWIDE SUMMATIVE ASSESSMENT IN SPRING 2021:

A WORKBOOK TO SUPPORT PLANNING AND DECISION-MAKING

Version 1.0 | September 2020

This document draws from a series of webinars, blog posts and papers that emerged in late summer 2020, dealing with anticipated challenges for statewide summative assessments in the Spring of 2021. These resources have been sponsored and produced by a variety of organizations, including the Council of Chief State School Officers, the Center for Assessment, and webinar panelists and participants from many other organizations including states, universities, and education and testing organizations and vendors. Articles from the fall 2020 special issue of *Educational Measurement: Issues and Practice* on the impact of COVID19 on educational measurement are also considered.

The goal of this document is to provide a practical summary of the guidance emerging from these resources by presenting many of the important questions that states can pose and discuss with their TACs, test vendors, and other states to facilitate sound planning for summative assessment in 2021.

We welcome suggestions, comments, or inquiries about the document. Please email Michelle Boyer (mboyer@nciea.org), Nathan Dadey (ndadey@nciea.org), or Leslie Keng (lkeng@nciea.org).

INSIDE

INTRODUCTION

This school year, every state education agency (SEA) is faced with unprecedented, COVID19-related challenges for the implementation of 2021 statewide summative assessments. Two overarching challenges are in how tests will be administered, and how scores will be interpreted and used, with many intervening and related challenges.

Test administrations may necessarily look different in 2021. They may occur at atypical intervals, over longer periods of time, and they may be administered in whole, or in part to students who are logging in and testing remotely from their homes. Second the very nature of learning and assessment in remote or hybrid environments during the year preceding Spring 2021 assessments is widely anticipated to have implications for how we interpret scores.

This workbook presents a non-exhaustive list of discussion topics and questions, grouped by the topic areas discussed in the Center for Assessment's RILS webinar on Spring 2021 assessments which were based on two recent publications, Into the Unknown: Assessments 2021, and *Restart & Recovery: Assessments in Spring 2021*. SEAs can ask and answer these questions to facilitate planning and decision-making for their spring 2021 summative assessment in a way that leverages the collective expertise of those who have contributed their time and thoughtful ideas through manuscripts, presentations, and panel discussions.

Addressing these questions will necessarily involve careful consideration of the needs of multiple stakeholder groups, and the anticipated consequences of their score use choices. We encourage SEAs to work with their summative assessment vendors to address these questions and consider bringing some of the questions to their technical advisory committees (TACs) for advice and guidance as they proceed with their operational planning activities.

TECHNICAL CONSIDERATIONS AND GUIDING QUESTIONS

Following the structure of, *Restart & Recovery: Assessments in Spring 2021*, we have organized the technical considerations and guiding questions for spring 2021 summative assessments into six broad areas:

- Test Design
- Test Administration
- Field Testing
- Equating
- Score Interpretation and Use
- Standard Setting

In each area, we elaborate or extend on the guidance and recommendations given in *Restart & Recovery: Assessments in Spring 2021* by incorporating suggestions and advice from other related recent publications and webinars, including the Center's RILS Webinar. In each discussion, we provide a set of guiding questions for use in SEA planning and decision-making.

Test Design

A primary consideration for many states is whether substantive changes to the test blueprints should be made, and the extent to which schools might be expected to implement a uniform set of "priority" content standards. Decisions made about test design will have a direct impact on whether scores can be equated to existing scales, and ultimately, how they are interpreted and used in Spring 2021. There are no hard and fast criteria for determining when blueprint modifications are large enough that we must consider the possibility of establishing a new or temporary test scale. In the strictest sense, any change to a tests blueprint triggers needs for heightened scrutiny of equivalency of scores. As such, test design discussions should already be underway and should include the following topics:

- Extent of blueprint modification
- Tolerance for changes to the test scale
- Openness of the SEA to holding a possible standards validation

Guiding Questions

- Will the state be advocating or providing guidance for the use of "priority" content standards in any subject areas tested in spring 2021?
  - If so, do those changes affect the content in the blueprints for the state summative assessment?
  - If so, which items (i.e., how many, which item types, etc.) are affected by changes related to priority content standards?
- What are the state's policy priorities related to score use, which may include:
  - Maintaining score trends?
  - Reporting achievement only on the content to which students are taught?
- Given the anticipated changes to instruction and test blueprints, are there substantive gaps in alignment that we might anticipate?
- Does the SEA have, or can it acquire clear information on the extent to which priority content standards can, or have been implemented in districts and schools?

Test Administration

There are two basic administration strategies to consider for spring 2021: in-school and remote. When considering these strategies, it is important to remember that the state will likely be concerned with not only the comparability of the results from 2021 to 2020, but also the comparability of results within 2021 across these approaches, if used.

One way we could frame policy and practice is by considering whether an administration approach like remote administration is more akin to a *mode* of administration (e.g., like pencil and paper or computer based testing) or an *accommodation* (e.g., focused delivery to students with specific needs). Framing remote administration in particular, as akin to a mode or an accommodation may help departments consider strategies to implementation.

Guiding Questions

- Will the administration window and allotted testing time be changed?

- When will the test be available? (For remote administration, the recommendation is to have as narrow a window as possible)
- Will the test be timed? If so, how will timing work (e.g., via proctor or automatically?)
- Will students be able to access the test more than once?

- Will the forms be modified to increase security?
  - Will there be multiple forms?
  - Will items be randomized within form?
  - Will adaptive testing be used?

- Will the tests be administered face-to-face?
  - How will schools implement social distancing while testing all students?
    - Will schools have the capacity and willingness to increase space for testing?
  - Will students and educators feel safe enough to focus on testing?
  - Can sufficient numbers of proctors and test administrators be recruited to conduct testing?
  - Can accommodations be provided following typical practice, or will they need to be adapted?
  - Will test windows be flexible to accommodate any need to test fewer students at a time?
  - What provisions will be made for students who are being instructed in hybrid and remote environments (either individually or at the school level)?

- Will the test be administered remotely online?
  - Have the assessments previously been administered online?
  - How will students be supported during the assessment? In particular, who will act as a "proctor" to help students get to, and stay focused on, the assessment? Specifically, how will students be supported:
    - To be authenticated (i.e., how will we insure that the right students gain access to the right test)?
    - To log on (have the appropriate information to start the assessment) and off the assessment, as well as to resume?
    - To start and end each assessment and section assessment, as well as to resume?
    - If technical problems occur?
  - What requirements will be put in place for a specific technology solution (e.g., will a specific OS, form factor, browser, access to technology)?
    - Will students who have better access to technology have comparable results?
    - How will students know if their technology is assessment ready?
  - How will students **be made familiar and fluent with the testing platform** and the technology they test on (laptop, tablet, etc.)?
  - How will the **student testing environment** be accounted for (computer location, other individuals in room, food and drink, restrictions on technology)? As noted in Camara (2020, p. 4), "it is unrealistic to think that all students have quiet, private spaces at home in which to test. Lower-income students are much more likely to face cramped housing, siblings and parents sharing the same workspace, internet connectivity problems, noisy environments, and less comfortable testing spaces".

        - Device access
        - Internet access
        - Quiet space
        - Family support

- Will the test be remotely proctored? If so:
    - Will the proctoring be done by human proctors, AI proctors, or both?
    - Will exam security software be used to monitor students (e.g., software that allows a test provider to monitor the activity of an examinee's computer)?
    - Can **video proctoring be used** (and include steps like examinees to showing their workspace and in some cases their computer itself; Isbell & Kremmel, 2020). Will the sessions be recorded?
        - If not, can the conditions of administration be changed to increase security?
            - "Steger et al. (2020), in their meta-analysis of 49 studies with over 100,000 test takers, found an effect size of 0.20 standard deviation units favoring those testing in an UIT environment. Despite this finding, they found that effect sizes were reduced to near zero when (a) a test had strict time limits, (b) content was not Internet searchable, and (c) a lockdown browser was employed." (Langenfeld, 2020, p. 1)
    - If a human proctor is used:
        - Who will be proctoring (parent, teacher, district staff, assessment vendor staff)?
        - Will proctoring be done live, or based on recordings?
        - How many students will be assigned to each proctor? Generally, 1 proctor to 16 students is the highest ratio recommended in college admissions testing (Camera, 2020).
        - What procedure or process will the proctors follow?
        - How will irregularities be defined, flagged, reported and handled?
    - If AI proctoring is used:
        - How will students be flagged?
        - What are the follow up steps to AI flagging?
    - How will the flagging procedures be created, or adjusted, to fit 3-8 and high school? Some flagging rules (e.g., another person in the room), may be inappropriate for elementary, middle and high school students who are engaging in virtual learning in shared family spaces.

The preceding questions can summarized within the following six categories:

- **Accessibility:** Do all students have sufficient technological capacity, familiarity with online testing to take the tests at home, and have access to the full range of accommodations as in-person administrations? Are special considerations warranted regarding accommodations and accessibility for students with disabilities or English learners?

- **Equity:** Are certain groups of students, such as those from poor households, systematically are disadvantaged by this nontraditional mode of administration?

- **Communication:** Information about at-home testing should be frequent and clear, addressing any common misconceptions. Provide avenues and opportunities for teachers, parents, and students to ask questions and request technical assistance.

- **Safety:** What procedures or protocols should be incorporated to help protect the health and safety of test administrators and students without compromising the validity of test scores?
- **Test windows:** Should the allowable testing time and/or length of the test windows be adjusted to account for school disruptions during the school year, staggered/rotating school schedules, or social-distancing requirements?
- **Security:** Are there safeguards in place to prevent testing improprieties, such as cheating and test-question sharing, and to ensure adherence to test-administration procedures? Do any adjustments to the test administration processes pose threats to test security? If so, how can such threats be mitigated or minimized? For example, if longer testing windows increase the chance of breached test items or forms, should the state develop additional forms or consider scrambling test items on the same test forms?

Field-testing

There are several challenges that are anticipated for collecting field test data, including:

- Changes to instructional coverage of grade-level content
- Testing conditions (physical setting, level of standardization, and test windows)
- Students (motivation, OTL, other barriers to typical learning conditions such as physical/emotional security, level of support for school activities at home…etc.)

Guiding Questions

- Do item pools need refreshing in 2021?
  - If Yes:
    - Are 2021 field test items needed for operational tests in 2022?
    - Are 2021 field test items needed for use as anchors in 2022?
  - If No:
    - Can I use the field test slots in 2021 for extra equating items to link the 2021 test to the current scale?
    - Can I remove field test slots in 2021 to shorten testing times?

Equating

There are three main features that influence the accuracy of equated scores. They are:

- Test content
- Conditions of measurement
- Examinee populations

Standardized administration procedures are used to minimize the influence of these features, with the goal of isolating examinee ability as the primary feature affecting score differences over examinees and time. Further, because we never trust that test content, conditions of measurement, and examinee populations are held completely constant, equating designs and procedures are used to control the influence of small, and hopefully random variation within these three non-ability related features. And finally, as an extra precaution, we typically evaluate our equating solutions to check for any worrisome influence of non-ability related features.

So, since equating results in 2021 will reflect the confluence of possible Instructional changes, remote access issues, remote admin), and changing patterns of students engagement (content, conditions of measurement, and examinee populations, changes in any of these areas represent risk factors for good equating. If our equating is not good, we move from the realm of equating to one of linking, and our business as usual interpretations will be less defensible.

Guiding Questions

- What are the challenges to standardization in 2021?
  - Content: Are there Content changes that could affect the accuracy of my equating results?
    - Shorter tests?
    - Altered test blueprints?
    - Content matrixed across students (i.e. each student is administered a portion of the content)?
  - Conditions of measurement: What conditions are different and how different are they this year?
    - Changes in testing window?
    - Changes in testing mode?
    - Changes instruction and instructional conditions?
  - Examinee population: How are my student populations different this year than in past years?
    - Changes in access to instruction (differences in devices, connectivity, or bandwidth)?
    - Changes in learning environment (e.g. home conditions vs. school classroom)?
    - Changes in quality, mode, or frequency of instruction?
    - Changes in curriculum coverage or emphasis?

- Based on my states Equating and Field Test Designs and responses to the preceding 3 questions, how will I know if my equating is good:
  - This year?
    - What analyses can be done to evaluating equating quality?
    - How long has my state's test and scale been in place—does it provide sufficient historical context to use in understanding/gauging equating accuracy in 2021? If not, what historical information can be used to judge equating accuracy in 2021? Are there methods that might be used to triangulate a worrisome effect on equating—e.g. administer a previously equated form alongside the 2020 operational form?
    - What are appropriate and useful acceptance criteria to use?
    - Will the typical screens (e.g. robust-Z or other measures of parameter change or invariance) be adequate?
    - For post-equated programs, is there an alternative to avoiding equating in 2021?
  - Future years?
    - What analyses can be done to evaluate field test item parameter quality?
    - What are appropriate and useful acceptance criteria to use?
    - Are there sufficient items in the item bank to support test construction and equating in 2022?

- What are my state's alternate paths to addressing state and federal reporting and use requirements should equating results not adequately meet acceptance criteria?

Score Interpretation and Use

Summative scores have both status and trend uses. Under the current conditions, we are especially worried about trend uses—We expect context differences such as:

- Gaps in access to high quality instruction will likely have consequences for our measures, and how we interpret their results
- Opportunity-to-learn (OTL) data will be useful to contextualize score interpretations
- Clear communication about results, and any limitations for interpretation, will be important to help stakeholders understand their meaning, and use them appropriately

Decisions about use will depend on the quality of the measure, given changes in the context:

- Accountability system(s)
- Results for individual and student groups (e.g., grade promotion, high school graduation, school trend monitoring, etc.)

Guiding Questions

- Can the intended score interpretation(s) from the state summative assessment programs be supported?
  - For the purpose of reporting trends?
  - For accountability? Which levels of accountability?
- What is the timeline for reporting? Does it need to change in 2021?
- What kinds of inferences might be in jeopardy?
- What criteria will be used to evaluate whether this is the case?
- And what should the state do if evaluation criteria are not met for a particular use?
  - Modify accountability indices?
  - Proceed with cautions about trend interpretations?
- What data can be collected to support communication of summative assessment results in 2021 and beyond?
  - Opportunity to learn?
  - Curriculum modifications?
  - Variation in administration conditions (e.g. rotating students into schools in small groups over longer administration windows, remote testing, delayed testing…etc)?
- How important will it be to evaluate the impact of COVID on student learning?
- Are there tests given pre-COVID in Winter 2018/2019, 2019/2020 that can be used in 2020/2021 to examine changes in trends that we might expect to see in operational Spring 2019 to Spring 2021?
- What data can reasonably be collected contextual results and understand the likelihood and nature of OTL gaps?
- As it will be very important to different gaps due to COVID from already existing gaps due to school resource disparities, how can we understand gaps in OTL in 2019-2021, relative to the gaps that preceded COVID?

- How can the state consider and possibly mitigate risks associated with relaxing standards or uses in 2021, anticipating that it could be more difficult to put restrictive rules back in place after they have been relaxed?

Standard Setting

If a state plans to either set new cut scores or validate existing ones, it will be important to consider any 2021 data used in the standard-setting or standards validation process. In general, student performance data are used to select the set of items for item-based methods (e.g., bookmark or Angoff) and the student profiles for student-based methods (e.g., body of work) reviewed by standard setting committees, and to generate impact data showing how students are projected to perform given the recommended cut scores. There is a chance that fewer students will be able to achieve the highest levels of performance in 2021, as compared to previous years. Moreover, COVI D-19 effects probably are nonrandom, differentially affecting items and students alike.

These potential issues give rise to questions that should be asked by SEA's to guide standard setting planning for states planning a 2021 standard setting.

Guiding Questions

- Will as many students as previous years be able to achieve the highest levels of performance in 2021?
- How do we know that the items in an ordered item booklet are ordered properly (for the bookmark method)?
- Is it acceptable to exclude items from certain content strands in the standard-setting item sets or student profiles?
- If we assume overall performance will be depressed in 2021, what is the "real" level of performance we can expect in 2022 and beyond?
- How will the workshop be conducted, remote or in-person? If in-person, what considerations or accommodations might need to be made to gather a group that is fully representative of content area experts?
- How should panels treat impact data?
- If we know that COVID-19 disruptions affect students differentially, how should the standard-setting committee interpret differences in subgroup-level impact data based on 2021 performance?

National Center for the Improvement
of Educational Assessment
Dover, New Hampshire

www.nciea.org