CCSSO
Council of Chief State School Officers™

Restart & Recovery: Assessments in Spring 2021

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, Bureau of Indian Education, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

RESTART & RECOVERY: ASSESSMENTS IN SPRING 2021

Jillian Balow (Wyoming), President of the Board

Carissa Moffat Miller, Executive Director, CCSSO

We are grateful to our partners at National Center for the Improvement of Educational Assessment Leslie Keng, Michelle Boyer, and Scott Marion for their help in developing this guide.

Council of Chief State School Officers

One Massachusetts Avenue, NW, Suite 700

Washington, DC 20001-1431

Phone (202) 336-7000 Fax (202) 408-8072

www.ccsso.org

Introduction

The COVID-19 pandemic has had deep and far-reaching disruptive effects on student learning and growth during the 2019-20 academic year. All states and territories canceled or suspended spring 2020 achievement testing, and all federal accountability requirements were waived as well. State assessment leaders and the testing industry have been working feverishly to figure out whether, and if so, how, to assess students this fall (2020) to gauge unfinished learning and changes in the achievement gap. This is an appropriate response that has required considerable attention and effort (Marion, Gong, Lorié, & Kockler, 2020). While there is neither clarity nor consensus regarding the most appropriate fall 2020 assessment response, state assessment leaders have an opportunity to address potential concerns with their spring 2021 assessments now. That may sound like a long time off to the layperson, but to those in the industry it is a mere blink of the eye.

The Council of Chief State School Officers (CCSSO) believes high-quality assessments are one crucial way to measure student learning, identify inequities, and drive the right supports for students. Exactly how those assessments are given may look different in this current environment, and states are working hard to make plans to best meet the needs of students in their state in the 2020-21 school year.

This paper focuses on key considerations for assessment leaders as they plan for their spring 2021 statewide summative assessments in (a) English language arts (ELA) and mathematics, administered in grades 3-8 and high school and (b) science, administered at least once per grade span. There are four major categories of challenges and considerations regarding spring 2021 assessment. It begins with a brief discussion of test design, particularly whether states should consider adjusting their blueprints. It then turns to issues surrounding administration and scoring, including the challenging prospect of at-home test administration and remote proctoring. Third, this paper discusses several psychometric issues, such as field testing, equating, and standard setting. Finally, this paper considers the important matter of interpretation and use.

Let us first consider possible re-entry scenarios for the 2020-2021 school year. It is likely it will not proceed as a "normal" school year—with students and teachers interacting in classroom settings for a full academic year without disruption. Any approach to instruction and school organization—and therefore assessment—should factor in at least three re-entry scenarios:

1. Fully in place. School resumes in-person classes in the fall, and the school year progresses normally.

2. Blended, or partially in place. School resumes in-person classes, but social distancing necessitates some sort of alternative scheduling. For example, to maintain proper spacing in classrooms, half the students attend class in person while the other half attend remotely, alternating weeks.

3. Fully remote. A final possibility is a return to full-time remote schooling, with school buildings remaining closed, as was the case for most U.S. schools in spring 2020.

Possible permutations of these scenarios include earlier or later re-openings and cyclical returns to remote schooling. Some states are preparing to start the school year early (with some opening as early as July), in anticipation of possible COVID-related disruptions throughout 2020-2021. These different scenarios, along with other variations, have implications for the instructional and organizational strategies schools will need to employ to continue fulfilling their many functions. These scenarios undoubtedly will affect the assessment strategies, as well.

Test Design

Blueprints and Opportunity to Learn

We cannot know what classrooms and teaching will look like in the coming school year, but the events of spring 2020 have raised important considerations for spring 2021. It is evident that student learning was affected by the spring 2020 school closures and the concomitant transition to a largely unfamiliar mode of teaching and learning. In the short term, however, we can only speculate on the nature and magnitude of the impact.

With this in mind, we might pause before modifying test blueprints in a well-intentioned effort to avoid assessing students on material which presumably was lost due to school disruptions. It is also important to consider that changing a test's blueprint would affect the stability of the measurement scale, which in turn could break the achievement trend line. In short, changing the blueprint changes what scores mean.

If achievement-level descriptors state that proficient students are expected to demonstrate that they know and are able to do certain things, but those parts are then removed from the test, the description of a proficient student and the corresponding score range are no longer valid.

Based on these considerations, and undoubtedly other issues we cannot foresee, at this time we recommend that states maintain their pre-COVID-19 test designs but are thoughtful about use and policies related to assessment results.

States are considering—based on advice from content experts such as Student Achievement Partners (2020) and others—to focus instruction during the 2020-2021 school year on "priority" content standards. These recommendations are designed to help accelerate learning of the most critical content and skills students need for long-term success in English language arts and mathematics.

If states adopt such recommendations, they will have to consider adjusting their spring 2021 summative assessment blueprints. Otherwise they risk telling educators to teach standards A, B, and C (hypothetically), but test on standards A, B, C, D, and E. In other words, states would be administering a non-aligned test. On the other hand, changing the blueprint carries the risks outlined above, in addition to the financial costs associated with changing the tests. There is no easy solution in this case, and states will face tradeoffs associated with each of the possible choices. One way to lessen the potential negative impact is to eliminate or minimize the consequences specifically associated with test scores.

Use of Previously Developed Tests

Since there was no testing in spring 2020, states will have the option to administer the 2020 tests in 2021 instead. This is a reasonable option, and it is likely to be a common choice across many states. The benefit of this option is obvious: the tests are already constructed. If there are no changes to the summative assessment planned for 2021, this is the logical choice for minimizing effort and cost. However, the content of all tests should be reviewed for items that might be emotionally triggering, or for content that may have been influenced by COVID-19 in ways that make item difficulty anomalous in 2021. Such tems risk trauma for students who have been impacted in profound ways by the pandemic. Psychometricians also risk inaccurate equating results where these items are used in the equating process (for pre- and post-equating alike).

The same cautions apply for states electing to reuse forms from 2019, with the additional consideration of security. The reuse of some test items is common practice in many assessment designs. However, test items from the 2019 assessments have been exposed to examinees in their entirety. States electing to reuse the 2019 tests should review data from prior administrations to ensure there were no breaches of security. It is also important to be alert to any formal item-release decisions (that is, items from old exams which are subsequently made public to ensure that neither teachers nor students have been exposed to test items before administration of the 2021 exams. Should any content be found in the public domain, or otherwise unsecure, the risk of influencing 2021 achievement results may be too great to pursue this option.

An added benefit to using the 2019 tests is that raw scores in 2021 would be directly comparable to 2019, provided no items are added or removed from scoring. This offers the unique opportunity to compare pre- and post-COVID-19 student achievement. It also presents opportunities to examine item parameter stability across the full range of content covered in a test's blueprint. Such data would enable states to better understand scale stability associated with the pandemic.

Administration and Scoring

At-Home Testing and Remote Proctoring

Most commercial-testing platforms are designed to protect the security of test scores and the privacy of students. This works well when all testing devices are under the control of a school or similar entity, which would be the case if schools are still fully or partially in place in 2021. However, what if a resurgence of the pandemic necessitates fully remote schooling during the 2021 testing windows?

Several assessment providers are considering innovative approaches, such as virtual or remote proctoring, to support at-home test administrations. Such approaches are promising not only as a short-term response to the COVID-19 crisis, but also as a long-term solution for test takers with special circumstances, such as home-bound students. If a state or assessment provider is considering at-home testing with remote proctoring for its summative assessments in spring 2021, we recommend that, at a minimum, the state collects evidence to answer these key questions:

- Comparability. Are scores obtained from at-home test administrations and remote proctoring comparable to those from traditional in-school test administrations?
- Technological Accessibility. Do all students have sufficient technological capacity (e.g., Internet access in a secure setting, adequate bandwidth, etc. - not to mention hardware/software that meets minimum specifications)?
- ADA Accessibility. Are students with physical or other disabilities familiar enough with online testing to take the tests at home with remote proctoring? Will they have comparable access to the full range of test-taking accommodations they would have at a conventional test administration site?
- Security. Are safeguards in place to ensure adherence to test-administration procedures and prevent improprieties, such as item-sharing and other forms of cheating (by students or their guardians)?

Any uncertain or negative responses to such questions about could indicate a risk to the validity of scores resulting from at-home testing and remote proctoring. There are also serious equity concerns: students in resource-limited communities are likely to be disadvantaged by this nontraditional mode of test administration. Finally, parents, teachers, students, and other stakeholders may have limited understanding about the logistics and possible complications of at-home testing and remote proctoring. If states end up opting to administer at-home testing in 2021, it will be important for state leaders to implement a clearly and frequently articulated, state-wide communication plan. It also will be important to provide avenues and opportunities for teachers, parents, and students to ask questions and request technical assistance well in advance of the test administration periods.

Traditional Administration Considerations

Even if the spring 2021 summative assessments can be administered in schools, we encourage states to review test-administration procedures to develop contingency plans for various schooling and testing scenarios. In designing the plans, keep in mind such test-administration questions as:

- What procedures or protocols will we incorporate to protect the health and safety of test administrators and students, without creating conditions that might compromise the validity of test scores? For instance, if examinees must be seated in multiple classrooms to facilitate social distancing, will additional proctors be needed?
- 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?
- Are special considerations warranted regarding accommodations and accessibility for students with disabilities or English learners?
- 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 rearranging test items on the same test forms?

In developing an assessment plan for Spring 2021, it will be important to involve district and school testing personnel, such as testing coordinators and administrators, by soliciting input and feedback throughout the planning process. If the state decides to make any adjustments to its 2021 test administration procedures, it will be vital to notify all stakeholders—early, and often, throughout the 2020-21 school year, to give districts and schools sufficient time to prepare.

Scoring

We encourage states to work with their scoring vendors to address the potential impact of COVID-19 on the item performance and test scoring processes. Here, we offer suggestions for both human and automation-based scoring processes.

- Human scoring. Conduct a close examination of previously scored papers used as anchor, borderline, or validity papers. Because we do not know how student responses will be affected by the disruption in schooling caused by the pandemics, we should not assume that prior papers are still representative. Consider selecting responses from spring 2021 to train human scorers.

- Automated scoring. The validity of scores from automated scoring engines rely on the set of papers used for training and calibration. As such, any impact that the COVID-19 disruptions have on the score distribution of papers may affect quality of the automated scoring. Scoring vendors should plan on conducting their existing checks of the training and calibration papers but pay attention to any substantial shifts in score d istributions for specific tasks or prompts.

- Score-related criteria. We recommend that states not modify scorer qualifications, scoring rubrics, or score validity criteria.

Psychometrics

Opportunity to learn (OTL) is widely recognized as a threat to the reliability and comparability of test scores (DePascale & Gong, 2020, Keng & Marion, 2020, Kurz, 2011). In the assessment context, OTL is thought of as the "opportunity to learn what is tested" (Haertel, Moss, Pullin, & Gee, 2008). Disruptions due to COVID-19 are an unprecedented case of OTL loss among all students across a state's schools and districts. Even more concerning, however, is that OTL loss likely will differ based on the student's demographic and socioeconomic conditions. We encourage both states and their assessment providers to closely examine and identify the specific test-development procedures which might be affected by OTL loss and other COVID-related context effects.

In this section, we consider the key psychometric processes of (a) field testing, (b) equating, and (c) standard setting. We also offer additional recommendations for how states can use their spring 2021 assessment results to understand and communicate the impact of COVID-19 on student learning and achievement.

Field Testing

A state's degree of concern about field testing depends on how the field test data are to be used in the test development process. In a post-equating model, field test data are used primarily to determine if an item is eligible for the item bank or operational test forms. Here, the quality of the field test data is less consequential than in a pre-equating model, where field-test item parameters are used to generate the operational score tables. In a post-equating model, special instructions should be given to those involved in data review and test construction for interpreting field test data from the spring 2021 administration. For example, more tolerance could be allowed when selecting or rejecting items based on the corresponding field-test statistics.

If states use a pre-equating model, it will be important to conduct evaluation studies to determine the effects of COVID-19 on the 2021 field test data. One design would be to include, in the spring 2021 embedded field-test slots, items having known statistical properties from an earlier field test or operational form. Comparing the 2021 item statistics with prior item statistics can quantify a "COVID effect," which, in turn, may be used to adjust the statistics for any 2021 field test items on future pre-equated test forms. Of course, making "average" adjustments carries the risk of masking important interactions, which are likely to be manifest in 2021.

Comparing 2021 item statistics with prior item statistics by item type, content area, and student group will help states better understand the impact of COVID-19 on learning loss. However, the limited number of available field-test slots on the 2021 operational exam imposes a practical constraint on this study design—a dilemma

magnified by the fact that, with the suspension of spring 2020 testing, many states' assessment programs now have a backlog of items for field testing. For the sustai ability of the item bank, few states can afford to give up too many field slots for research purposes; a balance should be struck between the needs to replenish the bank and to understand the impact of the pandemic.

This issue speaks to the importance of performing an item-bank analysis to identify gaps in content standards or item types which, in turn, will inform both future test development and field-testing priorities. A rule of thumb could be to allocate enough field-items slots to permit the inclusion of a representative set of items in terms of content and psychometric properties—a "mini-test."

The state's field-test equating process is a psychometric consideration that applies to both pre- and post-equating models. The common-item nonequivalent groups design is a typical field-test equating design. This design uses operational test-form items as a common-item set for placing items from multiple field-test forms onto the base scale. Special attention should be paid to 2021 operational items in the common-item set that may be unduly influenced by COVID-19, eliminating these items from the field-test equating process.

If states are using the common-item nonequivalent groups design equating method, they may want to pay special attention to any items in the 2021 common-item set that could be unduly influenced by COVID-19 (e.g., by possibly triggering a trauma response in some examinees), with an eye to eliminating these items from the field-test equating process.

The question may arise about states which have partial field-test data from spring 2020, because testing was suspended before all assessments were completed. Can the field-test data from those items be salvaged? The answer depends on (a) the representativeness of the sample and (b) how the field test data are to be used.

The representativeness of all students who took the 2020 field tests can be evaluated by examining their demographic characteristics and prior test performance. If the sample is sufficiently similar to what is expected from a full field test, and the number of responses per item meet the minimum n-count thresholds (i.e., to compute differential item functioning statistics for the required student groups), then partial field-test data may be salvageable.

What qualifies as "sufficiently similar" depends on how the field data are to be used. As noted above, the psychometric properties of field-test items are generally more consequential for a pre-equating model than for a post-equating model. In addition, even if schools have insufficient evidence to support using their spring 2020 partial field-test data, these data could be reserved and later combined with data collected from future field tests. If is administratively feasible, for example, items with spring 2020 field-test data could be placed in a separate spiraling of field-test forms having a lower per-item minimum n-count requirement, which would allow for the inclusion of additional field-test forms.

Equating

The primary goal of equating is to adjust for differences in difficulty among operational test forms, so the resulting scaled scores are comparable to those from previous years. In all likelihood, the spring 2021 operational forms will be constructed using items with data that precede the COVID-19 disruptions. This implies that the pre-equated psychometric characteristics of 2021 operational forms, including their item difficulties, should be unaffected by the pandemic. This bodes well for testing programs using a pre-equating model, as the scoring table resulting from the test construction process for spring 2021 forms should still be valid. Moreover, some programs using a pre-equating model incorporate a validation step, or post-equating check, during the operational administration. This step evaluates the stability of certain item types, such as constructed-response tasks, whose psychometric characteristics could change substantially between the field-test and operational test administrations—which in turn could affect the operational-form scoring table.

Given all this, as well as the potential impact of COVID-19 on student performance, we recommend that all states planning to use a pre-equating model for their 2021 tests conduct post-equating checks in spring 2021. It will be important to include an analysis of the stability of item parameters in this process, as well as an evaluation of model and person fit. Results from the post-equating check not only will help validate the pre-equated score table, but also could further the state's understanding of the "COVID effect."

For testing programs using a post-equating model, scoring tables for the operational forms will be estimated from spring 2021 data. It is important, therefore, to identify steps in the post-equating process that may be affected by COVID-19 disruptions and, in turn, devise mitigation strategies to support the comparability of scores resulting from the equating process. We encourage states to work with their assessment providers to evaluate the robustness of their equating process in order to defend the 2021 equating outcomes. The specific investigations to be conducted will depend on the extent to which it's a state's equating design and methodology are influenced by year-to-year population differences. Most programs, however, use an internal anchor design to link the scores on the spring 2021 forms to the base scale. Under this design, the operational items in the anchor set will deserve special attention.

In the Test Design section above, we noted that, if a state's priority is to maintain scale stability and achievement trend lines, we recommend not altering the test design in 2021. It would be preferable to focus on identifying item content that might be emotionally triggering or influenced by the pandemic in ways that render the estimated difficulty anomalous in spring 2021. This recommendation is especially relevant to the equating anchor set: any identified items should be replaced with items that maintain the content representation of the anchor set. This anchor-set review should be completed prior to the 2021 test administration and, further, involve content experts and educators from across the state who have insight into the degree of learning loss and the challenges students face.

Even with a careful review of the anchor set, it is likely that a higher number of items will be flagged by the 2021 post-equating stability check, potentially affecting the viability of the anchor set. If feasible, states should consider embedding some external anchor items in the field test slots as a backup plan. This would offer some flexibility for content and psychometric teams to swap in items with more stable statistical characteristics, as needed, to maintain blueprint coverage of the anchor set. If including external anchor items is not feasible, the state might consider increasing the number of operational items in the anchor set in anticipation of greater instability in 2021. These recommendations about the equating anchor set also apply to programs using a pre-equating model that choose to include a post-equating check in spring 2021; in that case, an appropriate anchor set would need to be identified for the post-equating check.

Regardless of the equating model they use, we recommend that states and their vendors add steps to the quality control (QC) process. The QC process does not end once independent replications of equating results match; a "reasonableness check" is also an essential component. In a reasonableness check, an independent reviewer attempts to answer the question "Do the equating results make sense?" by taking a macro view of the equating results and considering the meaning and implications of the outcomes, looking for unusual patterns. If irregularities are found, the reviewer attempts to find a reasonable explanation, which may require additional information and analysis. Such additional steps to the QC process may affect the score-reporting timeline. States and their vendors therefore should strike a balance between returning test results promptly and performing due diligence to validate and explain the results—especially in spring 2021.

Table 1 summarizes our recommendations for field testing and equating, which we distinguish by equating model. There likely will be a more immediate impact of COVID-19 on post-equated operational forms in 2021. For programs using a pre-equating model, the impact likely will be more prevalent on future operational forms which use items field tested in 2021.

Table 1. Summary of recommendations for field testing and equating

Equating Model	Pre-Equating	Post-Equating
2021 Field Test Data	Design study to estimate and adjust for COVID-19 effects	Address in data review and test construction
Field Test Equating	Exclude operational items affected by COVID-19	
Equating of Operational Form	Conduct post-equating check; identify appropriate anchor set for post-equating check	Collect evidence about the robustness of equating design and method; review and potentially adjust anchor set
Quality Control	Add steps to the equating QC process such a reasonableness check to help explain equating outcomes	
Main Impact of COVID-19	Future operational forms that use 2021 field test items	2021 Operational Forms

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, COVID-19 effects probably are nonrandom, differentially affecting items and students alike. These potential issues give rise to questions such as:

- 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?
- 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?

Importantly, the standard-setting process is a content-driven activity, informed by data. The standard-setting committee members, who are experts in the assessed curriculum and content areas, are charged with using their expertise and experience with students to render judgments about a set of test items or student profiles, given the expectations specified in the performance-level descriptors (PLDs). If the state agrees that the grade-level content expectations should not be lowered because of COVID-19 disruptions, then the PLDs should be unaffected. This is not intended to discount the role of empirical data in the standard-setting process. However, standard setting often takes place at the start of an assessment program, before all content standards have been fully implemented in instruction; consequently, there is differential OTL in this regard. In such cases, it is not unusual to observe depressed impact data or oddities in the item sets or student profiles. The instructions usually given to standard setters is to take a holistic view of the data (with certain caveats), but to base their judgments on the assessed content and PLDs. With this in mind, we offer the following recommendations for states and their vendors planning 2021 standard-setting events:

- Identify content that might be emotionally triggering or unduly influenced by COVID-19. If it can be done without adversely affecting content representation, avoid using such items in the standard-setting item sets, such as ordered item booklets.

[1] For brevity, we will only refer to "standard setting" in this section. However, any issues and recommendations also apply to a standards-validation process in 2021.

- Present impact data as late as possible in the standard-setting process. Consider withholding these data until after the second or third round of standard-setter judgments, which likely will yield "purer" judgments—informed more by the assessed content and PLDs and less by the impact data.

- Compute an additional set of "filtered" impact data, based only on items that are less influenced by COVID-19. Such items could be identified from the typical instructional scope and sequence (i.e., content with prerequisite knowledge taught before school closures) and psychometric characteristics (i.e., item statistics do not differ significantly between 2021 and previous administrations). The filtered impact data could be presented along with the unfiltered impact data as a contrasting data point, with important caveats highlighted (e.g., lack of content representation associated with filtered data).

- With input from the standard-setting committee, establish criteria for reasonable impact data in subsequent administrations as the effects of COVID-19 learning loss gradually subside. Plan to monitor and evaluate the impact data in subsequent years. If the impact data do not meet the established criteria, be prepared to reconvene stakeholder committees to revisit the cut scores.

Many states will need to set cut scores for tests they administer operationally for the first time in spring 2021. It is not necessary to avoid this activity, but it will be important to build in additional time for states' technical advisory committees and other stakeholders to review standard-setting plans and results while considering the recommendations we enumerated above.

Additional Considerations

We encourage states and their vendors to take a close look at the 2021 assessment results to better understand and communicate about the impact of COVID-19. Our recommendations include:

- Rethink student groupings. Most assessment analyses use conventional student groupings based on gender, ethnicity, socioeconomic status, and special education status. These groupings are useful for reporting purposes, and they are the disaggregations to which stakeholders are accustomed. However, they are unlikely to fully capture the differential impact of COVID-19 on students. To better understand the impact, we encourage states to consider defining additional student-group variables for their analysis of 2021 assessment outcomes (e.g., digital literacy, access to high-speed Internet, parental support for at-home learning, etc.). Note that some variables are likely to be state-specific and may be best informed by administering surveys to districts and schools.

- Develop a 2021 research agenda. It will be important for states to develop research agendas for the upcoming academic year, both to better understand COVID-19 effects and to

inform educational policy decisions. A research agenda would specify (a) assessment-related questions regarding the possible consequences of COVID-19 disruptions and (b) the design of the corresponding studies to address these questions. Research-agenda development can be guided by such documents as the Standards for Educational and Psychological Testing Standards (AERA, APA & NCME, 2014), and the Operational Best Practices for Statewide Large-Scale Assessment Programs (CCSSO & ATP, 2010).

Some studies could be extensions of analyses that are performed annually for the program, such as scale score descriptive statistics, impact data, reliability and classification consistency, and comparisons of test characteristic curves and test information functions—disaggregated based on any new student-group variables. Other studies might include surveys of district or school administrators, teachers, parents, and students regarding their experience with online learning and at-home testing; interviews and focus groups involving teachers and students about their respective experiences; and external validity studies examining the relationship of 2021 summative assessment scores with other achievement measures, such as scores from prior years, interim/benchmark tests, and the ACT or SAT.

- Solicit technical advice early—and often. Because of ongoing uncertainties and concerns about the pandemic, many states are now meeting virtually, rather than in person, with their technical advisors and assessment stakeholders. While it is difficult to match the quality of in-person interactions and discussions using a virtual format, the latter provides the opportunity for states to turn what conventionally is a 2- to 3-day, in-person meeting into multiple, much shorter sessions over a period of time. Consequently, the state will have more frequent touch points with its advisory groups, to both inform advisors and collect timely feedback on its plans for 2021. We endorse this approach, recommending that states meet early and often with their advisory and stakeholder groups.

- Document, document, document. A critical part of the annual psychometric work associated with statewide assessments is the clear and comprehensive documentation of goals, frameworks, procedures, outcomes, score interpretation, and intended use of these scores. Because of the unprecedented nature of COVID-19 disruptions, documenting what the state will do to understand and address this crisis with its 2021 assessments results is especially important. Depending on how a state structures its technical documentation, we recommend that any summary of what was different about the 2021 statewide assessments emphasize that this testing cycle was not "business as usual."

Interpretation and Use of Test Results

In the preceding discussion, we offered recommendations to states regarding flexible planning for and heightened scrutiny of test design, scoring, psychometric analyses, and standard setting in 2021. These recommendations are offered in support of the likely need for states to conduct reporting and accountability in a business-as-usual mode—even though the conditions leading up to the test doubtless will be highly unusual.

Contextualizing Assessment Results

The many unknowns associated with remote learning conditions and OTL require this heightened scrutiny; it will be useful to present student scores in the context of how students learned during the pandemic. In addition to understanding the accuracy and fairness of the scores produced, we recommend that states collect additional, nontraditional data to help explain spring 2021 summative assessment results and promote responsible and fair use of test scores. Examples include collecting data related to students' OTL, such as attendance, student and teacher engagement, motivation, availability of remote learning tools, and facility with these tools. Other examples might include collecting data related to students' basic needs, such as whether they have sufficient food and physical or financial security. Some of these data are more difficult to collect and interpret than others; some require collection of sensitive personal information, which may prove prohibitive. Also, some may require expertise in survey methodology, sampling, analysis, and interpretation. Considering reduced budgets, this type of data collection may be difficult for some states to collect reliably. However, the more states can understand the context in which students learn between March 2020 and the spring 2021 testing period, the more they will be able to understand assessment results. This information can be summarized alongside scores and trends, in various forms of reporting test results. This additional information will assist states in providing clear guidance to students, parents, teachers, and the public about how to appropriately interpret and use assessment results in 2021. This could be particularly useful for the study of achievement gaps for disadvantaged students, as well as for student groups that are newly defined in light of COVID-19 impacts.

Decisions Based on Assessment Outcomes

Our final recommendation for the interpretation and uses of 2021 summative assessment scores is that all decisions should be considered considering evidence resulting from the design, scoring, psychometric, and standard-setting approaches chosen by the state. We may conclude from the evidence that score quality is negatively affected in ways that cannot be statistically corrected. A business-as-usual interpretation in this case could be particularly troublesome. For example, if analyses show the presence of high levels of differential item functioning or, for example, a violation of the assumption of score invariance across student groups, certain decisions with potentially negative consequences for individuals, such as graduation or grade promotion, may not be supported.

Summary and Concluding Thoughts

Table 2 summarizes the key issues and the corresponding recommendations.

Table 2. Summary of issues and recommendations.

Issue	Recommendations
Test Design	
Modify test blueprint based on anticipated gaps in student learning.	To maintain scale stability and score comparability, states retain their existing test designs.
Reusing previously developed test forms (e.g., spring 2020 or before) in spring 2021.	States reuse previously developed test forms, but review items for content that might be emotionally triggering or influenced by COVID-19 disruptions.
Administration and Scoring	
Evaluating the validity of at-home testing and remote proctoring	States collect validity evidence to support the comparability, accessibility, and security of at-home testing and remote proctoring. They also address equity concerns and consider communication strategies.
Considerations for traditional (in-school) test administrations	States develop a comprehensive assessment administration plan with contingencies for different schooling and testing scenarios. The plan includes provisions for health and safety, testing time and test window, ADA accommodations and accessibility, and test security.
Maintaining the validity of the performance scoring process	States look closely at previously scored student papers used as anchor, borderline, or validity papers for human scoring, or as training and calibration papers for automated scoring. They pay attention to any substantial shifts in score distributions for specific tasks or prompts.
Psychometrics	
Field testing and equating in spring 2021	See Table 1 for a summary of recommendations based on the equating model (i.e., pre- or post-equated tests)
Standard setting in 2021	States closely examine items used in standard setting; reconsider the role that empirical data play in the standard-setting meeting; compute "filtered" impact data; and generate plans to monitor and possibly revisit cut scores in subsequent years.
Additional psychometric considerations for 2021	States rethink student groupings, develop a 2021 research agenda, solicit technical advice early and often, and detail what is different about assessments in 2021 in their technical documentaton.

Interpretation and Use of Test Results	
Contextualizing 2021 assessment outcomes	States plan to collect, where possible, non-traditional data to help explain spring 2021 assessment results to promote responsible and fair use of test scores.
Decisions based on 2021 assessment outcomes	States consider all decisions based on the evidence that is supplied through the spring 2021 test design, administration, scoring, psychometric, and standard-setting processes. They identify any caveats about the 2021 assessment outcomes that are consequential to their use.

As the saying goes, "We don't know what we don't know." At the time of this writing, most states still have not decided what the 2020-2021 school year will look like for students, teachers, schools, and districts. Some states are only starting to examine how student learning to date has been affected by COVID-19 disruptions. Educators are still evaluating the feasibility and effectiveness of at home learning and remote proctoring.

With all the unknowns, it would behoove states to begin working with their assessment providers, advisors, and stakeholders to identify research studies, develop contingency plans, and discuss communication strategies. If planned and implemented well, results from the 2021 summative assessments can serve as one of several tools that states can use to understand and communicate how the COVID-19 pandemic affected student learning and achievement.

References

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

Council of Chief State School Officers, & Association of Test Publishers (2010). Operational Best Practices for Statewide Large-Scale Assessment Programs. Washington, DC: CCSSO & ATP.

DePasquale, C., & Gong, B. (2020). Comparability of individual students' scores on the "same test". In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of Large-Scale Educational Assessments: Issues and Recommendations,* pp. 25-48. Washington, DC: National Academy of Education.

Haertel, E., Moss, P., Pullin, D., & Gee, J. (2008). Introduction. In P. Moss, D. Pullin, J. Gee, E. Haertel, & L. Young (Eds.), *Assessment, Equity, and Opportunity to Learn* (Learning in Doing: Social, Cognitive and Computational Perspectives, pp. 1-16). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511802157.003

Keng, L., & Marion, S. (2020). Comparability of aggregated groups scores on the "same test". In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of Large-Scale Educational Assessments: Issues and Recommendations*, pp. 49-74. Washington, DC: National Academy of Education.

Kurz, A. (2011). Access to what should be taught and will be tested: Students' opportunity to learn the intended curriculum. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *The handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy.* New York, NY: Springer.

Marion, S. F., Gong, B., Lorié, W., & Kockler, R. (in press). Assessment considerations for fall 2020. Washington, DC: CCSSO.

Student Achievement Partners (2020). *2020–21 Priority Instructional Content in English Language Arts/literacy and Mathematics.* Retrieved from: http://www.achievethecore.org/2020-21_PriorityInstructionalContent