MAKING SENSE OF SPRING 2021 ASSESSMENT RESULTS

Nathan Dadey, Leslie Keng, Michelle Boyer, and Scott Marion
National Center for the Improvement of Education Assessment[i]

March 2021

State summative educational assessment is about to begin in earnest. Rightfully, many are raising questions about the quality, meaning, and appropriate use of the assessment results. We have written this document to support state educational agencies (SEAs) and their assessment providers in devising effective and efficient analysis plans. We outline two key strands of analyses for the spring 2021 assessment results: operational and investigative (see Table 1). The operational stand deals with analyses meant to support the *technical quality* of assessment scores and the intended interpretation(s) of the assessment results. The investigative strand deals with analyses meant to aid in understanding the *effects* of pandemic-related disruptions on student performance. Taken together, these two related strands make up a framework that is meant to help SEAs and their assessment providers make sense of student performance on spring 2021 summative assessments.

We have written this document to support state educational agencies (SEAs) and their assessment providers in devising effective and efficient analysis plans. We outline two key strands of analyses for the spring 2021 assessment results: operational and investigative.

This framework, and the recommendations that follow, are informed by emerging work in educational measurement as well as prior publications in this field[1]. For each strand we present a non-exhaustive list of questions and related analyses. Although we present the strands as separate, in application the questions and analyses will likely overlap substantially, requiring an iterative approach. Further, the SEAs may find it helpful to start with the investigative analyses before turning to the operational analyses. Doing so may help inform a process of backwards planning that starts at investigative analyses and leads back to the operational analyses need to support those investigative analyses.

Acknowledgments: This document draws from ideas emerging from the ongoing national conversation around spring 2021 assessments. We thank Damian Betebenner, William Lorie, Carla Evans, and Ted Coladarci in particular, for their valuable comments and insights on the ideas discussed in this document. Any shortcomings or errors are our own.

[1] For example, see *Understanding Pandemic Learning Loss and Learning Recovery: The Role of Student Growth & Statewide Testing* and *Summative State Assessment in Spring 2021: A Workbook to Support Planning and Decision-Making.*

[i] Dadey, N., Keng, L., Boyer, M., & Marion, S. (2021, March). *Making Sense of Spring 2021 Assessment Results*. Dover, NH: The National Center for the Improvement of Educational Progress.

Table 1. Two primary strands for spring 2021 analyses.

	Description	Focus	Responsible Party
Operational	Typical and novel processes and analyses meant to support the *technical quality* of assessment scores and the intended interpretation(s) of the assessment results.	*Validity* of assessment scores: • Comparability • Interpretation	Assessment Provider
Investigative	Prioritized analyses meant to aid in understanding the *effects* of pandemic-related disruptions on student performance.	*Effect* on learning and performance: • Overall effects • Differential effects	Assessment Provider and SEA

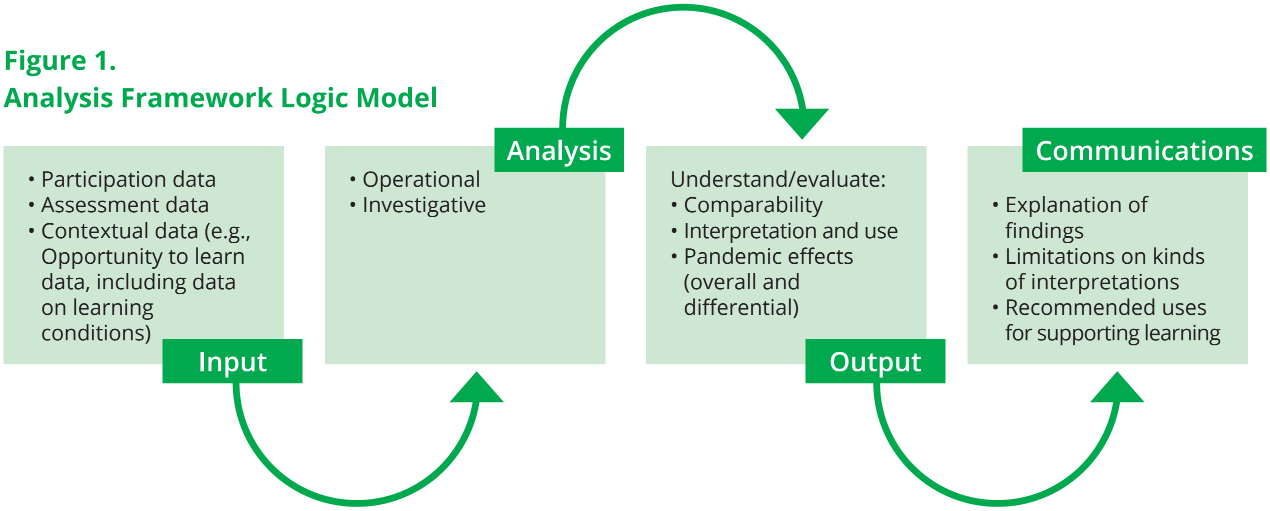
Finally, this framework is meant to support planning *now*, rather than waiting for scores to be returned. Assessment administration this year will be quite different from normal in most cases, requiring additional analyses conducted on accelerated timelines. Standard practice will not be enough. Consequently, this framework should inform the discussions of planned analyses that SEAs have with their technical advisors and assessment providers in preparation for the spring 2021 administration. Below we first present a logic model for structuring analyses based on spring 2021 data, which is followed by a summary of the operational and investigative strands of analyses. We conclude with considerations for implementing these analyses.

Finally, this framework is meant to support planning *now*, rather than waiting for scores to be returned. Assessment administration this year will be quite different from normal in most cases, requiring additional analyses conducted on accelerated timelines.

A LOGIC MODEL

Figure 1 illustrates the logical flow of analyses based on spring 2021 data. For the analyst, this work begins with the inputs and proceeds through to the communication of results. Given their respective contexts, however, states will need to prioritize the research questions and corresponding analyses.

Figure 1.
Analysis Framework Logic Model


• Participation data
• Assessment data
• Contextual data (e.g., Opportunity to learn data, including data on learning conditions)
Input
Analysis
• Operational
• Investigative
Understand/evaluate:
• Comparability
• Interpretation and use
• Pandemic effects (overall and differential)
Output
Communications
• Explanation of findings
• Limitations on kinds of interpretations
• Recommended uses for supporting learning

Since time is short we suggest that SEAs and their assessment partners prioritize those analyses that most likely will inform state leaders' understanding of the effects on learning. Other important elements of planning include establishing reasonable timelines and clearly delineated responsibilities. Table 2 below can be used to structure conversations among state assessment personnel and their partners.

Since time is short we suggest that SEAs and their assessment partners prioritize those analyses that most likely will inform state leaders' understanding of the effects on learning.

Table 2. Supporting Analysis Table: Identification, Prioritization and Responsible Party

Type	Question & Analyses	Description & Outcomes	Priority	Data	Timeline	Actions or Decisions	Responsible Party
Operational							
…							
Investigative							
…							

A Need for Baselines

We do not provide criteria for evaluating the results of suggested analyses. Rather, we hope that each SEA will examine their historical data, in consultation with their technical advisors and assessment providers, to develop baselines that reflect performance in a typical year. Much of the needed data and statistics probably will be available already. Drawing on these data, simulations can determine when problems might arise. For example, at what level of non-participation are aggregate judgments about student performance threatened? These historical baselines can be based on both the individual year (e.g., spring 2017, spring 2018, spring 2019, spring 2020) and between-year changes (e.g., spring 2018 to 2019, spring 2017 to spring 2019); the two-year change likely will better forecast the change between spring 2019 and spring 2021. Ideally, these baselines will inform the degree to spring 2021 data are out of the ordinary.

ANALYTIC STRANDS

The Operational Strand

The operational strand concerns the validity of score interpretations given an assessment's intended purpose. Therefore, the corresponding questions and analyses address the technical quality of the various assessments. These analyses include those typically conducted by assessment providers, as well as novel ones necessitated by concerns regarding the impact of pandemic-related disruptions. Any analysis must begin with a careful examination of descriptive results before moving to more complex analyses. This is particularly true in the present context, where many aspects of the assessment process probably have been affected by the pandemic (e.g., assessed content, measurement conditions, examine population).

The ultimate aim of the operational strand's analyses is to determine if score interpretations from prior years remain valid for spring 2021 scores.

The ultimate aim of the operational strand's analyses is to determine if score interpretations from prior years remain valid for spring 2021 scores. This requires the consideration of multiple factors, such as participation, unit of analysis (e.g., individual student, student subgroup, and school), content area, and the context of assessment (e.g., remote, in-person). Further, analyses can quickly become complex, such as one addressing the question "Is model misfit greater for remotely assessed, economically disadvantaged students compared with other students?"

The following questions, and suggested analyses, will be helpful for determining whether the intended score interpretations can be supported. Some of these questions take on particular importance when considering such factors as remote assessment (e.g., student behavior, model fit) or post-equating designs (e.g., model fit).

- **Has the assessed content changed?** If the assessment blueprints have changed from prior years, confirm that these blueprints are still representative of the prior standards, and determine whether this change nonetheless results in similar scores for students (based on simulations using historical data).
- **Has participation changed?** Determine whether the population of assessed students has changed from prior years, for any change can affect the resulting statistics. With a shift in the student body, as well as the missing data arising from non-participation, item parameters from a post-equated solution may not be stable. Further, judgments about subgroups, schools, and other student groupings may be compromised, possibly precluding judgments about aggregate performance.
- **Has student behavior changed?** Determine whether student assessment-taking behavior has changed from previous years, which can be surfaced by comparing the results of motivation analyses, forensic analyses (e.g., analyses of irregularities), person-fit analyses, and differential item-functioning analyses with earlier results. Shifts in item statistics also could suggest that students are interacting with items in new ways. For example, perhaps items on standards corresponding to little or no instruction have become more difficult.
- **Has measurement precision changed?** Use item response and classical test theories to determine whether measurement precision has changed overall and for targeted subgroups by, for example, inspecting conditional standard errors of measurement.
- **Has the measurement model's fit changed?** Through dimensionality analyses or invariance checks, determine if the fit of the measurement model has changed. Examinations of fit also provide evidence of item-parameter invariance, which is particularly important to consider in a post-equating design.

Considering the Outcomes of Operational Analyses

Again, these analyses ultimately speak to the validity of score interpretations. Practically, such an appraisal informs how scores can be compared both across and within years. There are at least three possible outcomes:

1. Prior score interpretations still hold, so 2021 scores can be compared with 2019 for example.
2. Prior score interpretations are not supported, so 2021 scores can be compared only within the spring 2021 administration.
3. That score interpretations, either for prior years or within 2021 only, are only supported for certain groups of students.

There is no one criterion for determining which outcome applies; rather, an overall judgment must be made after examining the evidence. If the preponderance of evidence suggests that prior-score interpretations are *not* supported, then states should refrain from reporting such comparisons even to the point of limiting the data that is reported for others to make these comparisons on their own. This challenge may involve annotating, flagging, or even suppressing reports for individual students or groups of students. Outcome 3 may be a possibility for aggregate comparisons, because student test participation likely will be non-randomly distributed across factors as subgroup status, type of instruction, or the school or district attended. This kind of nonparticipation means that some comparisons may not be supported for some subgroups, schools, subgroups within schools, and so on.

The Investigative Strand

The investigative strand focuses on understanding the effects of pandemic-related disruptions on student performance. Here, the questions and corresponding analyses build on the results of the operational analyses and, in turn, provide insight into student performance that supports the subsequent actions of stakeholders. SEAs must prioritize their analyses according to potential actions they may employ to support interrupted student learning. They also need to consider what questions need to be answered to support those actions, and how analyses can help inform those actions (e.g., allocating federal relief funding).

The investigative strand focuses on understanding the effects of pandemic-related disruptions on student performance.

Any question regarding the effects of the pandemic on student performance should be sufficiently specific to allow for meaningful analysis. There are a number of factors to consider in this regard:

- **Unit of Analysis.** At what level will the analyses be conducted?
    - Individual students
    - Classrooms, schools, districts, or the state overall
    - Specific subgroups[2] (e.g., race/ethnicity, economically disadvantaged[3], students with disabilities, English language learners)
    - Specific grades (or grade groupings) or content areas
    - Learning conditions (e.g., remote, in-person, hybrid)
- **Outcome Measure.** What measure of student outcome will be examined?
    - Achievement at specific points in time
    - Growth over time
    - Gaps in achievement and/or growth among students groups
- **Comparability.** What comparisons are warranted and among which student groups?
    - To prior years
    - Within the current year
    - Limited to specific groups of students

[2] Some experts suggest that traditional subgroup indicators may need to be augmented to help understand the effects of the pandemic by including data on such variables as instructional approach.

[3] Economically disadvantaged identifications may not be accurate this year for some states. This is because free meals, in many locations, are provided without parents having to sign up for this benefit; consequently, counts may not reflect actual need.

All analyses likely will cross these multiple dimensions in different ways. Analysts will need to maintain the flexibility to uncover unanticipated findings while remaining focused on the key questions and the responsibility to produce timely analyses with the limited resources available.

Below, we provide questions and analyses that can serve as a starting point for SEAs and their partners as they consider the investigative strand. The specific questions likely will evolve as conditions change and new information is produced. Further, as we suggested in Table 2, each key question should be associated with actions to support students. As in the operational strand, historical baseline data are essential for investigating some of these questions.

Key Questions

- **How did state-level performance change from prior years?**
  - Where are the largest decreases and increases from 2019 to 2021? Are they concentrated within specific subgroups, grades, content areas, learning conditions, schools, districts, geographic regions, or perhaps a combination of these contexts? If available, do opportunity-to-learn data throw light on any change in performance?
  - Are the patterns of performance similar to prior years? Have prior trends changed (e.g., regarding specific subgroup gaps)?
- **Which districts, schools, and subgroups have the lowest performance in 2021?** The questions below are similar to those above regarding performance decreases, but they now pertain solely to spring 2021 results.
  - What are the characteristics of the entities with low performance in 2021 (e.g., limited internet access) and what factors might explain this performance? In other words, is there convincing evidence that explain the low performance or are these factors found in high performance entities as well?
- **How does performance relate to key learning conditions during the 2020-2021 school year?**
  - What are the differences in performance between students who experienced different learning conditions, such as in-person, hybrid, or remote learning? Does this difference interact with student subgroup, grade, content area, school, district, or geographic region?
  - Does performance differ by student subgroup? For example, do typical student groups intersect with learning conditions?

Considering Outcomes of the Investigative Strand

Investigative analyses only have value if they inform interventions that support student learning. Ideally, such interventions will have been part of the planning process—planning that both defined the analyses and premised subsequent actions on the analyses' possible outcomes. Interventions are not implemented overnight. And although "we don't know what we don't know," intervention planning must start now (or have started already) based on the developing picture of pandemic effects on student learning. To be sure, doing something will be much better than doing nothing.

Investigative analyses only have value if they inform interventions that support student learning.

IMPLEMENTATION CONSIDERATIONS

Many SEAs, in collaboration with their assessment providers, are in the process of developing their operational and investigative analyses. Conversations about these analyses can be structured by using a table like Table 2 to bring key questions and corresponding analyses into sharp relief, decisions to be made, timelines, and responsible parties. Doing so makes a successful enterprise far more likely.

In addition, the operational and investigative strands do not capture all aspects of score interpretation and use. Performance reporting, for example, could be an entire strand on its own (e.g., Domaleski & Dadey, 2021). Analyses from both strands can help SEAs address questions about whether, and how, to report scores. After all, each SEA will need to decide what to report to the public, and whether reporting entails suppression, caveats, and actions. In other words, depending on participation rates, states may need to alter how if and how they report individual and aggregate results.

Finally, this work is to ensure that the questions and analyses support SEA decision-making. The allocation of federal relief funds is one decision in particular that deserves support by as much information as possible. We now conclude with considerations for prioritizing questions and analyses.

PRIORITIZING ANALYSES

Statewide summative assessment results almost invariably are produced on a tight timeline. SEAs and their assessment providers must ensure they develop timelines that liberally allow for all planned analyses. We list below some considerations for identifying and prioritizing analyses.

- **Internal vs. External Audience.** Are results intended for use outside the SEA? If so, who are the intended stakeholders?
- **Availability of the Data.** How easily accessible and useable are the data?
- **Format of Results.** How will the results be provided (e.g., a simple spreadsheet, posted as part of the state's report card, as an interactive web-based visualization)?
- **Timing.** What is the date by which results are needed?
- **Responsible Party.** Who will conduct the analyses, what is their capacity to do so, and will key stakeholders therefore regard this party's analyses with credibility?

Ideally, SEAs and their assessment providers will have a plan of attack well before assessment administration in the spring[4]. All analyses have been specified and written into procedures and supporting code, with analyses getting underway once student response data are available. One way to make this ideal scenario a reality is to begin the prioritization of analyses as discussed in this paper. Waiting until this summer to begin this essential work will mean that score reporting will occur far too late to inform policies and instructional interventions.

Ideally, SEAs and their assessment providers will have a plan of attack well before assessment administration in the spring[4].

[4] The Center has made progress in developing a hypothetical dataset – at the scale score level – to support some of these analyses in the R environment. See the object sgp_data_covid within the SGPdata package, which is available from GitHub.