Assessment Validation in the Midst of Change

Leslie Keng, Senior Associate
National Center for the Improvement of Educational Assessment

Paper in support of the Center's *Reidy Interactive Lecture Series (RILS): The Center at 20*

Background: Validation in Transition

As Heraclitus of Ephesus putatively said, "The only thing that is constant is change." This observation certainly seems applicable to K-12 assessment programs. Consider the following statistics from a recent survey of 21 states by the Council of Chief State School Officers about assessment transition in the past few years:

- 16 states changed assessment programs
    - 10 changed from Common Core assessments to state-developed assessments or SAT/ACT;
    - 6 are making changes to their existing assessment programs;
- 12 states changed testing vendors;
- 8 states transitioned from paper-and-pencil to online assessments;
- 5 states shortened their tests;
- Other states implemented new science and/or social studies assessments, removed performance tasks, shifted from untimed to timed tests, changed to 100% machine scoring, added writing tasks, or moved away from an end-of-course model.

Many of these changes are motivated by demand from the field for shorter test time, faster score reporting, and assessment results serving multiple purposes (e.g., informing instruction, measuring student progress, determining readiness for college and careers, evaluating teacher effectiveness, and servicing school accountability). While assessment-program transition is not a new concept, the expected transition rate arguably has increased significantly. In the past, state education agencies (SEAs) usually had at least two years from (a) the ratification of the legislative mandate for a new assessment program to (b) its first operational administration. More recently, however, the expected timeline for transitioning to a new assessment program often is one year or less. This means a SEA has months, not years, to determine its assessed content standards, develop item/test specifications and blueprints, construct new test forms, specify administration policies (including guidelines for accessibility and accommodations), define and implement scoring procedures (including psychometric processes), establish performance standards, design and generate score reports, and develop interpretation guides. Further, this transition often occurs in concert with other changes in the state, such as the procurement of a new testing vendor or the implementation of a new school accountability system.

As if such an aggressive and demanding schedule is not enough, an additional common requirement by stakeholders is the need to maintain trendlines through the transition. From a technical perspective, this means the inferences drawn from benchmarks or cut scores (e.g., percentage of students attaining proficiency in ELA or mathematics) are comparable across the old and new assessments. It also could mean the reported scores (e.g., scale scores on vertical scales for ELA or mathematics) are comparable across the assessment programs. To support the validity of either comparability claim, a validation process that evaluates and compares key aspects of the old and new assessment programs is needed.

A Common Approach: Standards Validation

Many states implement a standards validation process to validate the comparability claims across old and new assessment programs. A standards validation process is similar to a traditional standard setting process in that the former usually involves convening a representative panel of subject matter experts to review tested content and/or student work and, in turn, make recommendations for cut scores on the new assessments. The panelists typically are trained to follow a well-established standard setting method (e.g., Bookmark, Body of Work) involving multiple rounds of judgments, empirical feedback, and committee discussions resulting in cut score recommendations. In contrast to the traditional standard setting, however, a key distinction of the standards validation process is the a priori disclosure of the existing cut scores to panelists. For example, this can be done by pre-marking pages in the ordered item booklets used in a Bookmark procedure or, in a Body of Work procedure, highlighting profiles of student work illustrative of the performance expectation at each cut score. Panelists use these "benchmarks" as the starting points for their content-based judgments. Performance level descriptors (PLDs) for the new assessments usually are written based on PLDs from the old assessments to establish a link in expectations for each performance level between the two programs. Impact data—the percentage of students at each performance level—from the old assessments are also provided to panelists to help evaluate the reasonableness of their recommendations.

A standards validation process generally is considered defensible because it is based on a well-established standard setting method and, further, involves educators and experts who are familiar with the assessed curriculum and have experience working with students in the state. SEAs and their vendors usually present a standards validation plan to their technical advisory committee for review and formative feedback. However, implementing a standards validation process can be time-consuming and costly, particularly if it involves the convening of in-person committees for the standards validation workshop. Also, because the cut score recommendations are based on the judgments of committees of panelists, the process does not always yield consistent outcomes across (or even within) grade levels. For example, one committee may recommend adjusting the Proficiency cut scores for grades 3-5 mathematics, while another committee may recommend retaining the old cut scores for grades 6-8 mathematics. A committee may even make different recommendations for the Proficiency and Advanced cuts in the same grade-level content area test. Such inconsistent recommendations not only can lead to difficulties in interpreting and communicating the cut scores in the new assessments, but inconsistent recommendations also can lead to challenges in scaling if the desire is to maintain the same reporting scale. To avoid these challenges, states must either impose strong restrictions on the types of adjustments a standards validation panel can make during the workshop or overrule the panelists' subsequent recommendations. If not carefully handled, both of these solutions can lead to push-back from panelists and mistrust in the process, the assessment system, and the SEA.

An Alternative Approach: Expert Comparability Review

An alternative approach for evaluating the validity of comparability claims across old and new assessments is to design and implement an expert review process. An example is provided in the framework developed by the Partnership for Assessment of Readiness for College and Careers (PARCC):

The PARCC Quality Testing Standards and Criteria for Comparability Claims (QTS). As the large-scale assessment landscape shifts from a consortium-based model to custom-designed assessment systems, PARCC expects that some of its existing state members may transition to a new state-developed assessment program with a different testing vendor. However, these states could still license PARCC items and tasks to include on their assessments. Table 1 summarizes the potential use cases for licensing PARCC content.

Table 1: Potential Use Cases for Licensing PARCC Content

Use Case	Description of Use Case
State developed "PARCC" assessments	The state licenses PARCC content with test forms designed to match the test specifications and blueprints for the PARCC operational form. The state contracts its own vendor for the other steps in the test development process (such as item and data review, forms construction etc.) as well as for administration, scoring and reporting.
State developed "PARCC" assessments, supplemented with state-developed content	The state licenses PARCC content, but also includes ("non-PARCC") test items from its own item bank. The test forms are designed to match the test specifications and blueprints for the PARCC operational form. The state contracts its own vendor for the other steps in the test development process (such as item and data review, forms construction etc.) as well as for administration, scoring and reporting.
State developed assessments, supplemented with PARCC content	The state develops its own ("non-PARCC") test items but also licenses PARCC content. The test forms are designed to match state-developed test specifications and blueprints. The state contracts its own vendor for test development, administration, scoring and reporting.

An important assumption underlying PARCC's new operating model is that states licensing PARCC content are interested in comparing the results on their new assessment to those on the previous assessment (i.e., the PARCC assessments). To support such comparability claims, states collect and submit evidence demonstrating the defensibility of these comparisons. The evidence is then evaluated by independent expert reviewers to determine if, in their view, the desired comparisons are defensible. If the desired comparisons cannot be defended, the reviewers provide constructive and actionable feedback regarding what the state should do to better support its comparability claims. This is known as the PARCC comparability review process.

The reviewers for this process should be experts, steeped in both technical knowledge and operational assessment experience. Rather than rely on a single expert, multiple independent reviewers are

recommended so different perspectives are considered and then coalesced in the review process. The overarching comparability questions for each expert reviewer are:

- If a student taking the state's assessment with PARCC content took one of the PARCC operational test forms, would he or she obtain the same scale score?
- If a student taking the state's assessment with PARCC content took one of the PARCC operational test forms, would he or she receive the same designation regarding college and career readiness?

If the answer to the first question is "yes," the evidence provides sufficient support for the new assessments to continue using the PARCC score scale. This is called *scale score comparability*. And we have *benchmark comparability* if the answer to the second question is "yes": the evidence provides sufficient support for the new assessments to keep using the PARCC college and career readiness benchmark. Because scale score comparability is more stringent than benchmark comparability, if a state's new assessment is determined to be comparable at the scale score level, the state can also make college and career readiness comparability claims.

The review process focuses on the degree to which a state's new assessment program is comparable to PARCC's standard processes in four areas:

- **Design**: The design of the state's new assessments (e.g., purpose, content representation, item types) and the procedures informing its development are comparable to those for the PARCC operational forms.
- **Administration**: The state's new assessments are administered under comparable conditions (with respect to factors such as testing time, directions, and accommodations allowed) to those of the PARCC operational forms.
- **Scoring**: The procedures used to score the state's new assessments are comparable to those used to score the PARCC operational forms.
- **Reporting**: The results from the state's assessments are communicated in a comparable way to the results from the PARCC operational forms.

The Appendix provides more detailed information about PARCC comparability review process. At the advice of its technical advisory committee, PARCC implemented the comparability review process on its proposed Alternative Blueprinting Options (ABOs), a shortened version of the standard PARCC operational ("flagship") form developed to address testing-burden concerns. This was done in lieu of a traditional standards validation process. The evidence submitted for the ABO comparability review process is similar to the information that would have been presented to a standards validation committee. However, because the expert reviewers have deeper technical knowledge, a more comprehensive set of empirical results (e.g., reliability analyses, classification accuracy and consistency, and test characteristic curves) can be considered in the comparability review process. The expert review process also can be conducted on a significantly reduced timeline—in weeks vs. months—entailing fewer resources and substantially less cost.

Conclusion

The statistics on recent changes to state assessment program at the beginning of this paper is likely a harbinger of more transitions to come. Approaches such as the traditional standards validation and the alternative expert comparability review processes are safeguards that support the technical integrity of assessment outcomes in the midst of such change. By no means are these two approaches the only viable ones. We hope this paper will invite thoughtful dialogue between assessment practitioners and experts so that the important priority of maintaining the validity of test scores can keep pace with the rapidly evolving landscape of educational assessment and accountability.

Appendix – PARCC's Expert Comparability Review Process

Key Aspects of Evaluation

Across the four areas of evaluation, the supporting evidence submitted by a state is organized according to seven key aspects of its testing program:

1. Item and Test Development (Design area)
    a. Test purpose, target population and intended uses;
    b. Assessed content standards, item types, rubrics, blueprints, test formats, eligible content, and time limits, along with the rationale for the test design decisions;
    c. Procedures for review of test items by subject matter experts;
    d. Field testing and data review procedures; and,
    e. Forms construction and review procedures.
2. Fairness and Accessibility (Design area)
    a. Universal design principles;
    b. Accommodations for English learners and students with disabilities; and,
    c. Procedures used to translate forms for students for whom English is a second language.
3. Test Administration (Administration area)
    a. Training and instructions provided to test administrators and coordinators;
    b. Instructions given to test takers;
    c. Information about the modes of administration, such as paper-based vs. computer-based testing, and fixed-form vs. adaptive tests, including rationale for the offering the test in each mode;
    d. Details about test security protocols; and,
    e. Evidence that supports accessibility of the test to all students as part of the test administration.
4. Item Scoring (Scoring area)
    a. Training and qualification procedures for human scorers;
    b. Protocols for both machine and human scoring processes;
    c. Evidence that the scoring process is fair to all students; and,
    d. If used, validation of automated scoring processes.
5. Psychometrics (Scoring area)
    a. Choice of psychometric models;
    b. Scaling and equating design and procedures, including quality control processes;
    c. Analysis of disaggregated student groups;
    d. Sampling, including purpose and methodology; and,
    e. Other psychometric procedures or analyses that support the reliability and validity of test scores.
6. Standard Setting (Scoring area)
    a. Achievement or performance level descriptors (ALDs or PLDs) and how they were established;
    b. Standard setting methodology and procedures; and,

   c. Empirical support for the cut scores.
7. Score Reports (Reporting area)
   a. The reporting of summative scores, subscores, and performance levels;
   b. The reporting of score precision, such as standard errors or probable ranges; and,
   c. Intended use and interpretation of test results, including cautions against misuse.

Available Resources

Table A.1 provides a list of available resources that have been developed for the PARCC comparability review process. These resources will be published soon at the *PARCC QTS Document Repository*.

Table A.1: Summary of PARCC Comparability Review Process Resources

Resource Name	Purpose	Intended Audience
QTS	Provides guidance to states that are interested in including PARCC content and intend to make comparability claims with another PARCC assessment	Any state interested in making PARCC comparability claims
Classification Scheme	Describes a classification scheme for supporting comparability claims under PARCC's new operating model	Any state interested in learning about options for participating in PARCC
PARCC Standard Processes	Describe the PARCC standard process by providing high-level overviews with links or references to additional documents or supporting materials published by PARCC	Expert reviewers in the PARCC comparability review process or anyone interested in PARCC
Evaluation Questionnaire	Collects information about a state's assessments that includes PARCC content	Any state using PARCC content in its assessment program
Evaluation Checklist	Provide a suggested list of evidence that states can provide for the PARCC comparability review process	Any state participating in the PARCC comparability review process
Comparability Review Guidelines	Provide a concrete framework for expert reviewers to follow in their evaluation of the state's evidence	Expert reviewers in the PARCC comparability review process

Figure 1 provides a roadmap illustrating the relationship between the PARCC comparability review resources. The first (red) box captures the resources that provide information about the standards and criteria by which the state's submitted evidence should be evaluated. The second (green) box represents resources that collect information about a state's *evidence* for the review process. The resource in the third (blue) box provides concrete *guidance* on how expert reviewers should compare the evidence in the second box with the standards and criteria in the first box.

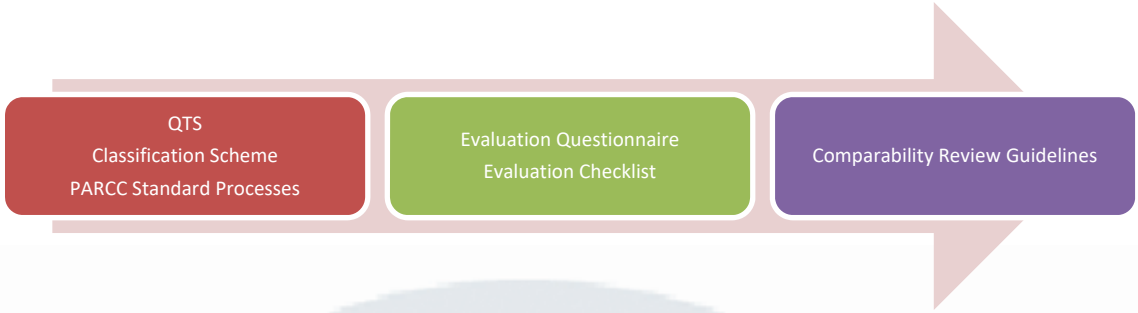

Figure 1: Roadmap for PARCC Comparability Review Resources

Core Questions for Potential Use Cases

In a comparability evaluation, the core question and focus of evaluation differ for each of the potential use cases for licensing PARCC content (see Table 1). Therefore, it is important for expert reviewers to recognize the use case applicable to the state they are reviewing. Table A.2 provides the core question and focus of evaluation for each of the potential uses cases.

Table A.2: Core Question and Focus of Evaluation for Potential Use Cases

Use Case	Core Question	Focus of Evaluation
State developed "PARCC" assessments	Are the procedures, materials, and tools used in the administration, scoring, and reporting of the state developed "PARCC" assessment sufficiently similar to those used by the PARCC operational forms to support the use of the PARCC scale and/or college and career benchmark as if they were equivalent?	• Quality of adherence to the PARCC test specifications and blueprints • Comparability in rigor and quality of procedures used to present, administer, score, and validate the assessment outcomes • Potential sources of construct irrelevant variance that would threaten the comparability of score interpretations and claims between the state's "PARCC" assessments and PARCC operational forms.
State developed "PARCC" assessments, supplemented with state-developed content	Is the construct defined by the test specifications and blueprints, the procedures used to develop and validate content, AND procedures and materials for administering, scoring and reporting of the state developed "PARCC" assessment sufficiently similar to those used by the PARCC operational forms to	Same as for the State developed "PARCC" assessment, with the additional key consideration of being able to support claims that the state developed content measures the Common Core State Standards in the same way as demonstrated on the PARCC operational forms.

Use Case	Core Question	Focus of Evaluation
	support the use of the PARCC scale and/or college and career benchmark as if they were equivalent?	
State developed assessments, supplemented with PARCC content	Is the construct defined by the test blueprint, the procedures used to develop and validate content, AND procedures and materials utilized for administering and scoring PARCC content and reporting test results similar enough to those used by the PARCC operational forms to support the use of the PARCC scale and/or college and career benchmark as if they were equivalent?	Same as the State developed "PARCC" assessment supplemented with state-developed content. The one key difference is rather than evaluating the quality of adherence to the PARCC test specifications and blueprints, a focus of evaluation should be on whether the construct assessed by the state developed assessment is essentially the same as that measured by the PARCC operational forms, even though the blueprints are not the same.