RILS 2018 Looking Back, Looking Forward: Contributions and Challenges in Validation and Evaluation of Assessment and Accountability Systems

Erika Landl, Juan D'Brot, Leslie Keng, Brian Gong

Revised September 13, 2018

A Focus Throughout on Validation and Evaluation

The Center for Assessment's founding purpose was to foster greater student achievement through better assessment and accountability. Over the past 20 years the Center has helped states and others pursue that purpose through providing technical assistance directly (e.g., contracted work with over 30 states currently, including many state technical advisory committees) or indirectly through professional presentations (e.g., RILS, NCME, AERA, and CCSSO NCSA conferences), publications (e.g., CCSSO monographs, EM:IP journal articles), and work supporting other agencies and entities (e.g., advising USED on Peer Review guidance, support of CCSSO convenings, collaboration with NCEO, KnowledgeWorks). A central and increasingly important focus of that assistance has been on **validation** and **evaluation**.

One illustration of that evolving focus throughout the Center's history is the topics of the annual **RILS** (Reidy Interactive Lecture Series) conferences sponsored by the Center since its founding. (See Figure 1.)

Figure 1: RILS topics since inception

RILS 1999: Improving Assessment Practice

RILS 2000: Technical Issues Affecting State Accountability Systems

RILS 2001: Implementing an Accountability System that Improves Schools

RILS 2002: Implementing the No Child Left Behind Act: Alignment, Reliability, and Rationality

RILS 2003: Improving the Validity of States' Standards-Based Assessment and Accountability Systems

RILS 2004: Incorporating Measures of Student Growth Into State Accountability Systems

RILS 2005: Wrestling With High School Assessment and Accountability

RILS 2006: Comprehensive Assessment Systems to Improve Student Learning: Critical Design and Implementation Decisions

RILS 2007: English Language Learner Assessment and Accountability- Critical Considerations for Design and Implementation

RILS 2008: Validating Assessment and Accountability Programs

RILS 2009: Next Generation Education, Assessment, and Accountability Systems

RILS 2010: Next Generation Balanced Assessment Systems: Expanding Our Notion of Technical Quality

RILS 2011: Multiple Measures for Assessment and Accountability

RILS 2012: Evaluating the Evaluators: Evaluating Educator Evaluation Systems

RILS 2013: Assessing College- and Career-Readiness: 2015 and Beyond

RILS 2014: Assessment in the Classroom – Bringing it all Together

RILS 2015: Comprehensive Assessment Systems to Support Learning and Accountability

RILS 2016: Assessment Literacy: Key skills to effectively use assessment information

RILS 2017: Assessing Student Learning of the Next Generation Science Standards

Those topics portray an abiding interest in improving assessment and one of its key contemporary uses: accountability. This concern for effective accountability systems was intimately tied with assessments that would support valid interpretations, and RILS reflects that on-going effort for validity applied to **assessments** in new areas and uses, including standards-based assessments (2003), student growth (2004), high school assessment (2005), English language learner assessment (2007, college- and career-readiness (2013), and next generation science standards (2017).

Similarly, the Center has been concerned with evaluating and improving **accountability** systems through attending to the main purpose of improving schools and student learning (2001), alignment, reliability, and rationality of accountability systems (as distinct from assessment systems' alignment, reliability, and validation) (2002), and explicitly addressing the validity/validation of assessment and accountability systems taken together (2003, 2008).

One theme appearing consistently through the RILS topics is the Center's view that current assessment and current accountability practices are limited—and that a larger view is needed to achieve the desired purposes. Thus there is attention early and often to larger **systems**—such as an accountability system that not only measures and evaluates school performance, but also improves student learning (2006), "next generation" balanced assessment systems (2010), and comprehensive assessment systems to support learning and accountability (2015).

A more detailed portrayal of the Center's attention to promoting validation and evaluation of assessment and accountability systems may be seen in the activities—documented by technical assistance activities, publications, and presentations on promoting and evaluating the **quality** of assessment and accountability systems, which are included in Table 1 below.

Table 1. Evaluating the quality of assessment and accountability systems.

Assessment Quality	Accountability Quality
• Assisting states in implementing, monitoring, and improving their state assessment systems through changes in policy, technology, and operational constraints and opportunities, including computer-based testing • Improving the validity of assessment interpretations by assessing constructs better, especially for diverse students. Areas of focus include ○ Performance assessment ○ Alternate assessments for students with severe cognitive disabilities ○ Competency-based assessments ○ Interim assessments ○ Comprehensive assessment systems • Improving the validity and usefulness of assessment interpretations by promoting better content-referenced interpretations, which include standard-setting and learning progressions • Promoting validation in projects and with clients in AA-AAS projects • Providing tools for evaluating the quality of assessments with a focus on alignment, content, and technical quality evaluation tools • Developing sound theories of action • Supporting consequential validity and the evaluation of consequences	• Evaluating different types of performance, such as status, improvement, growth, and acceleration • Evaluating normative growth • Examining the reliability of accountability decisions • Developing methods for evaluating the reliability of accountability decisions • Assisting states in developing and monitoring their accountability systems • Providing assistance in meeting federal requirements (typically through self-evaluation) under ESEA, NCLB, and ESSA • Contributing to and developing review/evaluation guidelines and participating in Peer Review • Supporting clients in improving the comprehensiveness and coherence of their accountability systems • Extending accountability beyond identification and into support systems • Developing and promoting tools for self-evaluation of the design, development, and implementation of accountability systems

Some Significant Trends & Implications for Validation/Evaluation

We see several trends emerging as a result of the Center's work over the past 20 years. We anticipate these trends will become increasingly important and will involve movement from:

- attending to pieces to attending to systems
- assessment and accountability issues primarily as measurement to include issues of use
- design to evaluation
- advocacy to technical detail
- specialized, one-time studies to more emphasis on embedded, continuous validation and self-evaluation capacities
- state to local in terms of evaluation and validation activities.

These trends will be challenged by new situations and old issues in new contexts. These include evolving federal requirements, state requirements, and state-articulated goals and priorities. Additionally, new assessment technologies (e.g., increasing movement to digital-based assessment), process data, merging of diagnostic and summative assessment from multiple sources, and new assessment models (e.g., data mining, learning analytics, and artificial intelligence) will necessitate a shift in how we validate and evaluate designs and uses. One of the largest shifts will be a merging of traditional **assessment validation** framework to one that combines **validation and educational program evaluation**. This will require an emphasis on mixed-methods and combined formative-summative evaluation efforts that can focus on context-dependent needs and strategies. Some examples of the changes in emphasis are provided in the table below.

Table 2. Emphasis shifts that will need to be supported through validation and evaluation efforts.

Prior Emphasis	Shifted Emphasis
Validation of summative assessments	Defining interim assessments and articulating the relationships between formative, interim, and summative assessments in comprehensive assessment systems
Attention to accountability systems' "audience and purposes"	Developing tools to help define theories of action
Attention on design of accountability systems for identification of low-performing schools	Design accountability systems that include systems of support and articulate a theory of action beyond "accountability ratings will motive schools to improve" or "schools will figure out how to improve"
Isolated technical advice or "one-off" reactive assistance	Developing tools that capture the underlying intelligence of the technical advice and make it possible for end-users to explore accountability systems by applying what-if reasoning with the tools
Expertise applied at the design stage of system development	Applying expertise and technical assistance to evaluation definitions, criteria, procedures, tools, and recognized, trusted authorities to support capacity building
Technical support on limited domains and measures	Supporting the examination, implementation, and validation of expanded domains and less traditional measures across systems
The development of separate assessment, accountability, and support systems	Developing cohesive assessment and accountability programs that include support, curriculum, instruction, and other programmatic systems (e.g., career and technical education programs)

Illustrating the Case for Validation and Evaluation during the RILS Validation/Evaluation Session

The Center staff will provide content-based frameworks or foundations to ground operational discussions in the areas of system coherence, accountability, and assessment. All RILS participants will be encouraged to participate in an opening plenary session around trends in assessment and accountability and the need for validation/evaluation. RILS participants will then be invited to join the Center staff in one of three breakout sessions that will explore operational examples highlighting components of each framework or foundation in one of the following areas:

- Designing for overall system coherence within a Statewide Educational Agency (SEA) using ESSA and Perkins as an example;
- Opportunities and constraints for SEAs to address when evaluating accountability systems, using an accountability evaluation/validation framework as an example; and
- Considerations of assessment quality and how SEAs should think about validation efforts to support high quality assessment systems under different conditions, using the transition from an old to a new assessment as an example.

Session Structure and Additional Reading Materials

The opening session will provide background to participants (building off this document). Subsequent sessions will couple a theoretical opening for each section as described above with an operational example illustrating relevant issues. Participants are invited to read one of the following documents depending on their area of interest and intended breakout group selection. Each paper provides greater detail or an example aligned to one of the three topics identified in the previous section—system coherence, accountability, and assessment—and provides a common foundation to support group discussion in each of the three breakout groups. In addition to the papers discussed below, each topic will be elaborated by a different SEA representative who will discuss how some aspect of the issue has been addressed within his/her state.

System Coherence

Over the last five years, the U.S .Department of Education (ED) has required states to provide more and better evidence supporting the quality and validity of their assessment and accountability systems. Specifically, ED has recently pushed for improved coherence across state plans. This is evident both in ESSA and the recent reauthorization of the *Perkins Act – the Strengthening Career and Technical Education for the 21st Century Act* (i.e., *Perkins V*), which both necessitate and support improved alignment between these pieces of legislation and the *Workforce Innovation and Opportunity Act* (WIOA). ED's appeal for alignment is acknowledgement of the fact that these laws share common goals and objectives that will not be met if a state's response to each is addressed in isolation. While federal efforts to improve alignment are a good start, they are not enough to help those charged with designing and implementing these programs establish accountability provisions that work in a coordinated manner to effectively and efficiency meet the state's goals. That requires not only an understanding of what it means for a system to be coherent (i.e., the core characteristics and features), but a reconceptualization of ESSA, Perkins, and WIOA as complementary elements of a larger state system of accountability.

To that end, this brief discusses the characteristics and features of a coherent system and outlines nine recommendations to support those charged with developing, evaluating or modifying state plans under ESSA, Perkins and WIOA.

Accountability Evaluation

The passage of the Every Student Succeeds Act (ESSA) marked the beginning of a new development cycle for accountability systems. State leaders once again have an opportunity to redesign their accountability systems based on the provisions included in ESSA and to ensure that systems improve outcomes for all students. As states begin implementing and monitoring their accountability systems created under ESSA requirements, the number of stress points across a system becomes more evident. Additionally, effective accountability implementation extends beyond identifying the right schools or obtaining approval for a system that can then be treated as "set it and forget it." The correct identification of schools is a necessary but insufficient condition to build capacity and deliver support to local systems. Systems of accountability, support, and continuous improvement contain a series of feedback loops and information hand-offs that offer opportunities to collect evidence that systems are working as intended. There is a need for states to develop a validity argument for their accountability systems, which require identifying activities and their relevant evidence throughout the design, development, and implementation of accountability systems. This paper leverages a framework[1] that can support a systematic examination of the design, development, and implementation stages of accountability identification that helps practitioners establish validity arguments for their accountability systems.

Assessment Validation

Assessment transitions seem to be happening on a more frequent basis and at a more rapid pace in recent years. Many of these changes have been motivated by the public's demand for shorter tests, faster score reporting, and assessment results that serve multiple purposes (e.g., informing instruction, measuring student progress, determining readiness for college and careers, evaluating teacher effectiveness, and being used in school accountability). Despite the transitions in assessment programs, there still exists a desire or mandate for the new program to maintain performance trendlines. From a technical perspective, this means that the inferences drawn from the benchmark or cut scores (e.g., percentage of students attaining proficiency in ELA or mathematics) are comparable between the old and new assessments or that the reported scores (e.g., scale scores on vertical scales for ELA or mathematics) can be compared across the assessment programs. To support the validity of these types of comparability claims, a validation process that evaluates and compares key aspects of the old and new programs is needed. This paper introduces an alternative approach for evaluating the validity of comparable claims. It describes a framework that PARCC developed known as the *Quality Testing Standards and Criteria for Comparability Claims* (QTS) to support the anticipated shift in its state participation model. The goal of the QTS is to provide guidance to states transitioning from the consortium developed and administered "flagship" forms to an assessment program that continues to include PARCC content and still intends to report results on the PARCC score scale or with the PARCC performance levels. The comparability review process can serve as a more efficient and economical alternative to the more elaborate standards validation approach.

[1] D'Brot (2018). *A framework to monitor and evaluate accountability system efforts*. Dover, NH: Center for Assessment.