# EVIDENCE-CENTERED DESIGN: A SUMMARY

APRIL 2015

*Joan L. Herman*
*Robert Linn*

National Center for Research
on Evaluation, Standards, & Student Testing

UCLA | Graduate School of Education & Information Studies

The work reported herein was supported by grant number 2011-7086 from The William and Flora Hewlett Foundation and by grant number S283B050022A between the U.S. Department of Education and WestEd with a subcontract to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

## Evidence-Centered Design: A Summary[1]

Both PARCC and Smarter Balanced adopted Evidence-Centered Design (ECD) as their approach to summative assessment development and validation. Formulated by Robert Mislevy, and colleagues (see, for example, Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 1999), ECD starts with the basic premise that assessment is a process of reasoning from evidence to evaluate specific claims about student capability. In essence, students' responses to assessment items and tasks provide the evidence for the reasoning process, and psychometric and other validity analyses establish the sufficiency of the evidence for evaluating each claim (see also Pellegrino, Chudowsky, & Glaser, 2001).

ECD is a principled approach that proceeds through a series of interrelated stages to support the close correspondence between the claims about student performance that a test is designed to evaluate and the nature and quality of the assessment evidence that is used to draw inferences from student scores. As the following describes further, each of these stages represents a critical link in the argument supporting the validity of score interpretations and a critical leverage point in assuring that an assessment will well represent intended constructs – in the case of PARCC and Smarter Balanced, students' being on-track to and/or ready for college and careers in both English language arts and mathematics.

ECD starts by establishing a clear conceptual foundation for guiding the assessment development. First is the delineation of both the claims about student learning that the test is being designed to evaluate and the specific evidence that is eligible for assessing students' accomplishment of each claim. That is, the specific knowledge and skills that constitute accomplishment of each claim – the domain of possible assessment targets – is laid out, for each subject and grade level to be assessed. Item or task models – also known as specifications – are then developed to guide the actual development of items. The models provide reusable templates for creating items and tasks aligned with each potential assessment target. These item and task models are then used to generate the test items and tasks, which, in turn, are subjected to content and bias reviews, pilot tested and field tested, and revised as necessary to refine psychometric quality and validity. At the

[1] This summary is adapted with permission from a portion of *On the Road to Assessing Deeper Learning: The Status of Smarter Balanced and PARCC Assessment Consortia*, by Joan Herman and Robert Linn (2012).

same time, test blueprints are developed to guide the creation of test forms, the collection of items and tasks to which students respond on a given test form. The blueprints specify how many and what types of items and tasks will be used to assess each claim and, within claim, how many, which and how assessment targets will be sampled and allocated to test forms. The administration of operational test forms then provides additional evidence of psychometric quality and validity of the tests for their intended purpose(s). Standard setting is a final step in this transparent process.

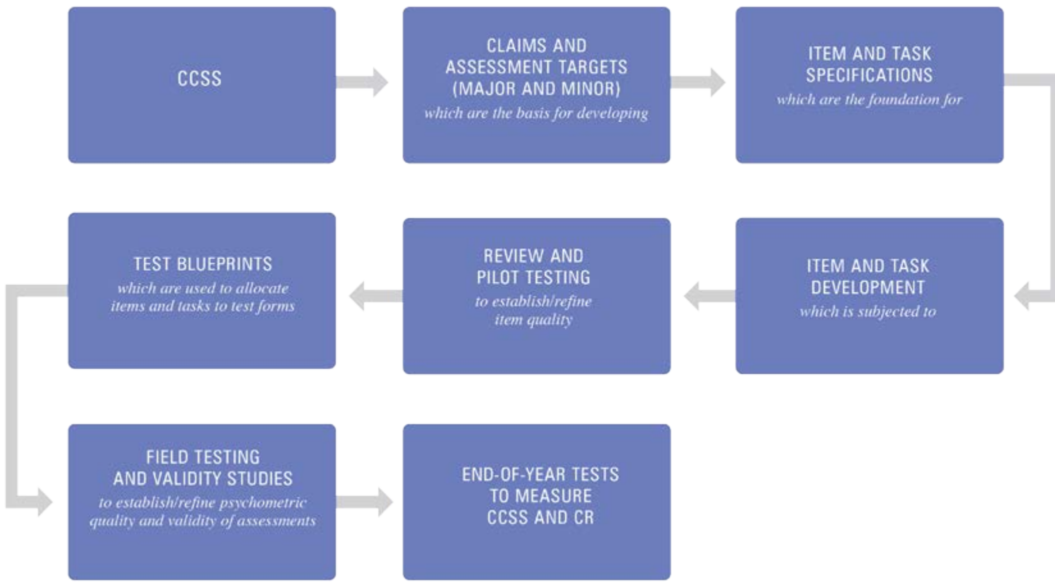


*Figure 1*: ECD General Approach.

Figure 1 lays out these stages in the general context of the PARCC and Smarter Balanced assessment development. Here ECD process here started with the Common Core State Standards themselves, which both consortia reorganized into core claims about student competency in ELA and mathematics that their tests are designed to evaluate. Both consortia start with an overall claim about students becoming college and career ready in ELA and mathematics and then subdivide these overall expectations into more specific sub-claims. Tables 1 and 2 summarize PARCC and SBAC claims for both subject areas.

*Table 1*. PARCC and SBAC Claims for the ELA Summative Assessments

| PARCC | SBAC |
|---|---|
| 1. Reading: Students read and comprehend a range of sufficiently complex texts independently. | 1. Reading: Students can read closely and analytically to comprehend a range of increasingly complex literary and informational texts. |
| 2. Writing: Students write effectively when using and/or analyzing sources. | 2. Writing: Students can produce effective and well-grounded writing for a range of purposes and audiences. |
| 3. Research: Students build and present knowledge through research and the integration, comparison, and synthesis of ideas. | 3. Speaking and Listening: Students can employ effective speaking and listening skills for a range of purposes and audiences. |
| | 4. Research/Inquiry: Students can engage in research and inquiry to investigate topics, and to analyze, integrate, and present information. |

*Table 2*. PARCC and SBAC Claims for the Mathematics Summative Assessments

| PARCC | SBAC |
|---|---|
| 1. Major Concepts and Procedures: Students solve problems involving the major content for grade level with connections to practices. | 1. Concepts and Procedures: Students can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency. |
| 2. Additional and Supporting Concepts and Procedures: Students solve problems involving the additional and supporting content for their grade level with connections to practice. | 2. Problem Solving: Students can solve a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem solving strategies. |
| 3. Expressing Math Reasoning: Students express mathematical reasoning by constructing mathematical arguments and critiques. | 3. Communicating Reasoning: Students can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others. |
| 4. Modeling Real World Problems: Students solve real world problems engaging particularly in the modeling practice. | 4. Modeling and Data Analysis: Students can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems. |
| 5. Fluency: Students demonstrate fluency in areas set forth in the Standards for Content in grades 3-6. | |

Each claim is further defined by specific evidence statements (PARCC) or assessment targets (SBAC) that the claim encapsulates. These statements or targets, in turn, are operationalized relative to particular standards or clusters in the CCSS and specify the DOK or cognitive complexity at which each may be assessed. For example, PARCC subdivides its first reading claim into reading and demonstrating comprehension of grade-level complex literary text and grade-level complex information texts. Within the former claim are standards RL 1, RL 2, RL 3, RL 5, and RL 7. Evidence statements for each standard lay out what student responses are suitable for demonstrating the standards. For example, for RL 1, the student's response:

- Provides questions and/or answers that show understanding of a text, referring explicitly to the text as the basis for the answers; and
- Provides references to details and/or examples in a text when explaining the basis for the answer. (*PARCC ELA evidence statements*, see http://www.parcconline.org/ela-literacy-test-documents)

In essence, the claims and evidence statements or assessment targets circumscribe the domains to be assessed in terms of content and performance expectations, which become the targets for developing item and task specifications. Informed also by each Consortia's performance level (PARCC) or achievement level (Smarter) descriptors, the item and task specifications provide guidance and rules that the item writers will follow in developing items that comport with each assessment target or evidence statement. They circumscribe, among other elements, eligible stimulus materials, prompt content and design, response requirements and scoring criteria, accessibility requirements, administration directions and conditions (e.g., allowable supports), etc. The ideal item or task specification provides sufficient guidance so that two item writers working independently from the same specification would generate essentially comparable items or tasks for a given assessment target or evidence statement – such that students would be expected to perform similarly on both.

Item writers then use the specifications to generate items and tasks, which in turn are subjected to content and bias reviews as well as pilot testing. Items and tasks which survive this process as substantively and psychometrically sound are then assigned to test forms according to blueprints, which provide rules for representing the given claims and assessment targets/evidence statements and for sampling items and tasks. Initial test forms typically are then field tested prior to the tests becoming operational.

The ECD framework thus makes very transparent what is to be assessed and how the CCSS are going to be assessed, and is very different from the “black-box” test development process that has been typical in many state assessments of the past. The black-box process starts with standards and general test blueprints about content coverage, and ends with scores and established proficiency levels, with limited rationale or evidence of the steps in between.

The transparency of the various ECD stages also provides a means for trying to assure that an assessment will represent the depth and breadth of standards and claims it is intended to measure. Each stage influences and constrains subsequent ones, and any omissions in prior stages will result in those same alignment gaps in the actual test. For example, if important standards are not fully represented in the claims and assessment targets/evidence statements, they will not be included on the test, or if some task or item specification do not model suitable depth of knowledge and/or incorporate important practices, the test items appearing on the operational test will be similarly limited. Likewise, the range, balance and depth of knowledge or cognitive complexity of a test could look quite good based on an item and task specifications or even within item and task pools that are developed, but if the test blueprint does not adequately sample higher levels of DOK, the test will under-represent those higher levels.  Analysis of the products of the ECD process during the assessment development process PARCC thus represents an important early warning strategy for supporting the content validity of the tests relative to both CCSS and deeper goals.

## References


Herman, J., & Linn, R. (2012). *On the road to assessing deeper learning: The status of Smarter Balanced and PARCC assessment consortia* (CRESST Report #823). University of California, Los Angeles, CA: CRESST. Retrieved from http://www.cse.ucla.edu/downloads/files/CRESSTReport823.pdf

Mislevy, R. J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice*, *25*, 6-20.

Mislevy, R., Steinberg L., & Almond, R. (1999). *Evidence-centered assessment design*. Princeton, NJ: ETS.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

PARCC. (n.d.). *Reading evidence tables*. Retrieved from http://parcconline.org/ela-literacy-test-documents

PARCC. (n.d.). *Writing evidence tables*. Retrieved from http://parcconline.org/ela-literacy-test-documents