Evidence-Centered Design

Contributors: Louise Yarnall and Geneva Haertel
Questions, or want to add to this topic or to a new topic? Contact CIRCL.

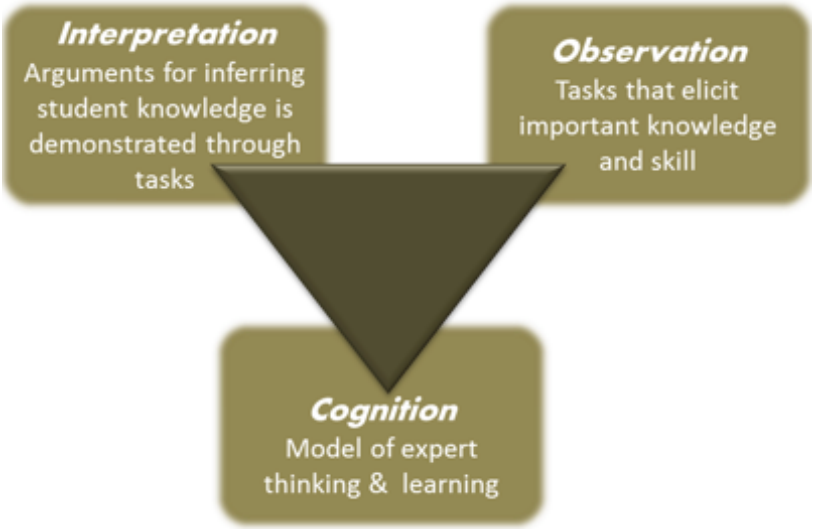
Overview

Evidence-centered design, or ECD for short, takes the art of test design and turns it into a science. Test development usually involves assembling various items into a test and then using statistical techniques and expert review to check for technical quality. ECD reduces the trial-and-error process, and leads to better measures. ECD involves justifying many design decisions *before* the first test item is selected or developed. Through ECD process, test developers create a list of the types of evidence known to accurately reflect what someone knows and can do. Creating this list is particularly valuable when designing tests of hard-to-measure knowledge, skills, and psychological states. ECD is becoming a testing industry standard and provides the kind of documentation that is often imperative in legal situations when evidence of a test's validity is required. This primer will provide a quick overview of this assessment design practice and both its benefits and costs.

To begin the ECD process, test developers study relevant learning sciences research, gather input from subject matter experts, and review previous tests, assessment tasks, scoring rubrics, and scales. Such initial groundwork is important because test designers rarely develop an accurate and reliable measure on their first try. All this upfront work is carefully documented so that later, after the test is administered, ECD test designers may refer back to this rich documentation to systematically review and revise items and tasks to increase comprehensibility, precision and reliability. Moreover, the ECD documentation means that the designers have **evidence** on which to base each subsequent revision to the test questions, media representations, or scoring rules. They also can reuse these documents to efficiently create tests of similar knowledge and skills.

While large-scale testing companies may engage professional item designers and measurement experts in using ECD, the basic ECD principles can serve as a useful guide to teachers and human resources professionals when selecting testing materials for use. These principles may also be helpful to parents and students when considering the fairness of tests administered in schools and elsewhere, especially when high stakes decisions are being made (e.g., admission to a university, a certification examination, for use as evidence of an instructor's competency). This primer summarizes the three core concepts that must be

considered when of designing assessments—Cognition, Observation, and Interpretation (presented in the assessment triangle below), and how they align with the ECD process to contribute to better test design and test product selection.



The Assessment Triangle.
(Pellegrino, Chudowsky, & Glaser, 2001, p. 44)

Key Lessons

The ECD process is a dance around this triangle. In the first move of the dance, the assessment designer jumps to the cognition point at the bottom of the assessment triangle by consulting **research evidence** to understand expert thinking and core problems of novice learning in a subject domain. The assessment designer then steps to the upper right point of the triangle by selecting or creating tasks that create **observable evidence** of the desired, research-based knowledge and skills. Finally, the assessment designer steps to the upper left of the triangle by analyzing how well the test tasks produce **measurable, valid and consistent evidence** of a learner's proficiency. Using the design documentation produced during the ECD process, the assessment designer then goes back to each of these three steps and adjusts the specifics--the model of learning, the observable evidence being collected, and the interpretation of the scores. In practice, the assessment designer is moving among all three points of the assessment triangle in a fluid, iterative way.

Cognition: ECD test developers first refer to learning science research so they can better understand the concepts, reasoning and skills required for individuals to perform successfully in subject area domains. Over the past 40 years, learning scientists have documented the knowledge, skills, and reasoning of domain experts and compared these to novices' knowledge structures and reasoning procedures. They have studied how experts and novices play chess, construct geometry proofs, write computer programs, simplify algebraic equations, evaluate historical evidence, read school textbooks, write argumentative essays, and reason using non-intuitive system models in biology, chemistry, and economics. Learning science has also posited various motivational factors that influence learning, such as effort and confidence. Taken as a group, these cognitive and motivational processes provide greater explanatory power about why some students attain better learning outcomes than others. In the ECD view, tests based on such learning science research can better flag when students are successful in engaging in such learning processes, and when they are engaging in counterproductive practices. Tests that successfully make such distinctions offer a powerful starting point for instructional intervention. In ECD, all these psychological elements associated with learning a subject are documented for future reference in the test design process. The review of this background research is referred to as **domain analysis** and is the initial step in the ECD process.

Observation: In the next step in the ECD process, test developers select or create items that are intended to elicit observable evidence of the underlying cognitive or motivational forms of knowledge and skill from the examinee. At this point the assessment designers have identified task features that they believe will elicit observable behavioral evidence of these knowledge, skills and psychological **constructs**, as psychologists call them. After gathering items relevant to the constructs to be measured, the test developers may find that these items share common task features, such as ways of wording a test question or directing learners to interact with media, and they may find some item scoring criteria are likely to differentiate learners along a useful scoring scale. If no such items or scales are found among the existing tests, ECD assessment designers may construct new items and scales so that the desired knowledge, skills, and psychological states may be observed and measured. Often the first items that assessment developers design around a hard-to-measure construct are embedded within a scenario which presents a short storyline and the learner is asked to reason about the content and provide a short narrative response. In other cases assessment designers may create "stand-alone" items to inform the design of shorter and easier-to-score items. The observations provide data that developers use to score and make inferences about student performance.

Interpretation: After ECD test designers have developed a test, it is time to administer it to learners and check the test for its technical quality (reliability and validity). By having experts in the field review the test items, the ECD test designer can document the **content validity** of the test, meaning that the critical types of knowledge and skills are being tested in the assessment. Through initial pilot testing, the designer observes and interviews learners as they engage in responding to these new test items to determine whether they elicit the targeted knowledge and skill to be measured, or have **construct validity**. Pilot tests can also be conducted to tell test designers how similarly the same results will occur if the test is administered multiple times to the same students—in other words, how consistently will the test measure the construct for the same students, what test statisticians call **test-retest reliability**. Other technical qualities include examining how much overlap there is among certain items designed to measure similar or related underlying psychological states, or **inter item correlation coefficient**; which items are likely to be difficult for most students or correctly answered by students of high- or low-ability levels, or **item difficulty estimates**. These and other technical qualities of the assessment may be studied and used to inform the revision of the test. Through ECD, test developers may then adjust task features to improve the scoring logic or refine their definitions of what psychological constructs are being measured. As a result of this evidence-based approach, the interpretation of the scores produced by the test are strengthened and the test designer has greater confidence that the inference made about what a student knows and can do is valid.

Issues

ECD involves **greater upfront costs** than traditional test development.

Developing careful and explicit design documentation before creating items and tasks **formalizes a step in the test design process that may seem burdensome to some test developers**. Although most test developers are familiar with the production of a test blueprint, fewer regularly engage in creating design patterns that specify the psychological construct being measured. Subject matter experts may resist applying such a principled approach because they believe their content expertise alone is sufficient for creating good test items. In addition, the increased use of technology-enhanced items and tests puts an additional cognitive load on students as they navigate different browsers and use new interfaces and item types (e.g., simulations, drag-and-drop, dynamic graphing). The use of ECD with its emphasis on thoughtful documentation may reduce the number of iterative development cycles needed to produce valid and reliable computer-enhanced tasks

However, there is often **limited guidance** available about how to link higher order psychological constructs, including subject matter content and the steps in complex reasoning processes to the design of assessment tasks and scoring systems. This results in some trial and error in the test development process. Routinely we find an absence of data on the technical quality of many assessments that instructors use from textbooks, item banks, or their own self-designed tests. ECD creates templates that designers may use to create items and tasks that are more likely to have adequate technical quality and support for the valid interpretation of scores. Such documentation may help drive down costs for computer-based tasks, such as interactive simulations.

Projects

Examples of NSF Cyberlearning projects that overlap with topics discussed in this primer.

- EXP: Understanding Computational Thinking Process and Practices in Open-Ended Programming Environments
- EXP: Learning Lens: An Evidence-Centered Tool for 21st Century Assessment
- BCC-SBE/EHR: Developing Community & Capacity to Measure Noncognitive Factors in Digital Learning Environments

Other projects

Principled Assessment Designs for Inquiry – Providing a practical, theory-based approach to developing quality assessments of science inquiry by combining developments in cognitive psychology and research on science inquiry with advances in measurement theory and technology.

Principled Assessment of Computational Thinking – Applying the ECD approach to create assessments that support valid inferences about computational thinking practices, and is using the assessments and other measures to investigate how CS curriculum implementation impacts students' computational thinking practices

Large-scale science assessment – Applying ECD to advance assessment design for large groups of examinees, typically numbering in the thousands, and often administered to make high-stakes decisions.

Next Generation Science Assessment – Developing NGSS-aligned assessments and curricula for the next generation of K-12 students.

Resources

Principled Assessment Designs for Inquiry (PADI) – Advancing Evidence-Centered Assessment design work atpadi.sri.com and ecd.sri.com

Next Generation Science Assessments developed with ECD

In the ETS database of technical papers, the following titles pertain to ECD:

- A Brief Introduction to Evidence-Centered Design
- Monitoring and Fostering Learning through Games and Embedded Assessments
- Designing Adaptive, Diagnostic Math Assessments for Individuals with and without Visual Disabilities
- Supporting Efficient, Evidence-Centered Item Development for the GRE Verbal Measure
- Evidence-Centered Assessment Design for Reasoning about Accommodations for Individuals with Disabilities in NAEP Reading and Mathematics

ECD-developed assessments for learners with disabilities:

- Cameto, R., Haertel, G., & Morrison, K. (2011). Technical Report 5: Synergistic Use of Evidence-Centered Design and Universal Design for Learning for Improved Assessment Design
- Cameto, R., Haertel, G., Haydel-DeBarger, A., & Morrison, K. (2011). Technical Report 1: Project Overview: Applying Evidence-Centered Design to Alternate Assessments in English Language Arts/Reading for Students with Significant Cognitive Disabilities

Readings

References and key readings documenting the thinking behind the concept, important milestones in the work, foundational examples to build from, and summaries along the way.

Baxter, G., & Mislevy, R. (2005). The case for an integrated design framework for assessing science inquiry(PADI Technical Report 5). Menlo Park, CA: SRI International.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). How people learn: Brain, mind, experience, and school. Washington, DC: National Academy Press.

Cheng, B. H., Ructtinger, L., Fujii, R., & Mislevy, R. (2010). Assessing Systems Thinking and Complexity in Science(Large-Scale Assessment Technical Report 7). Menlo Park, CA: SRI International.

Colker, A. M., Liu, M., Mislevy, R., Haertel, G., Fried, R., & Zalles, D. (2010). A Design Pattern for Experimental Investigation (Large-Scale Assessment Technical Report 8). Menlo Park, CA: SRI International.

Mislevy, R., & Riconscente, M. (2005). Evidence-centered assessment design: Layers, structures, and terminology (PADI Technical Report 9). Menlo Park, CA: SRI International.

Mislevy, R., Hamel, L., Fried, R., G., Gaffney, T., Haertel, G., Hafter, A., Murphy, R., Quellmalz, E., Rosenquist, A., Schank, P., Draney, K., Kennedy, C., Long, K., Wilson, M., Chudowsky, N., Morrison, A., Pena, P., Songer, N., Wenk, A. (2003). Design patterns for assessing science inquiry (PADI Technical Report 1). Menlo Park, CA: SRI International. Also presented at American Education Research Association (AERA) in April, 2003.

Mislevy, R., Steinberg, L. S., & Almond, R. G. (1999). Evidence-Centered Assessment Design. Educational Testing Service.

Vendlinski, T., Haertel, G., Chang, B., DeBarger, A., Rutstein, D. Fried, R., Snow, E., Zalles, D., Mislevy, R., Cho, Y., Fulkerson, D., McCarthey, K., & Finkelstein, D. (2013). Using the Principled Assessment Design in Inquiry (PADI) System: Some Frequently Asked Questions (Large-Scale Assessment Technical Report 12). Menlo Park, CA: SRI International.

Cost argument for ECD:

Bennett, R. E. (1999). Using new technology to improve assessment. Educational measurement: Issues and practice, 18(3), 5-12.

Dickison, P., Luo, X., Kim, D., Woo, A., Muntean, W., & Bergstrom, B. (2016). Assessing Higher-Order Cognitive Constructs by Using an Information-Processing Framework. Journal of Applied Testing Technology, 17(1), 1-19.

Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. Association of Test Publishers, 1(1), 1-38.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. Educational Measurement: Issues and Practice, 25(4), 6-20. (An earlier draft)