

[Education Week's blogs](#) > [Assessing the Assessments](#)

Can We Trust Assessment Results?

By Invited Contributor Listed Below on March 20, 2014 6:52 PM | [2 Comments](#)

Today's guest contributor is **Eva L. Baker**, Distinguished Professor of Education, UCLA Graduate School of Education and Information Studies; Director, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

"Oh, great," says the reader "yet another opinion about assessment". Relax. I don't intend to unpack many technical details, mostly because unpack is so East Coast. Instead, I'll point to a few essential elements of assessment quality.

To an average person, the crux of assessment is the scores summarizing their results, scores compared to the results of other students (a percentile), or to the criterion domains, (a percentage). Higher scores are good; lower scores are not; movement up is desirable. That's about it.

For those of us who have focused on the design and meaning of assessments "of learning," or "for learning" (Gordon Commission, 2013), the main story is more than surface results. Rather, it is the legitimacy of the findings as indicators of learning--whether at a particular point, growth over time, connected to instruction, or for next steps. Can we trust assessment results? An assessment of Standards should give correct and sufficient information to make inferences about learning, (including content and cognition), teaching, and schooling. On the downside, it is also critical to estimate whether individuals could game the exam and inflate results.

Pretty obvious, but because of time, cost, and tradition, some large-scale assessments have used severely limited numbers of tasks to measure Standards. As a result, we are left with insufficient information about learning. Shallow surveys can give an overall snapshot of an educational system, barring individual or organizational consequences. But what if some teachers have taught Standards using content that they don't know will be omitted from the survey? Teachers and students' competencies will be underrepresented. To work, assessment design must be exceptionally transparent to those engaged in teaching and learning. Good transparency occurs when test content can be clearly summarized without giving away the specific questions. Logical design and adequate numbers of tasks/items contribute in part to fair and equitable instruction and learning of rich content and challenging thinking. Clouded transparency occurs when tasks don't measure systematically the breadth or depth of intended learning, misleading teachers. Such exams may encourage the learning of the one tested procedure, or worse, the answer to the single test question "measuring" a given Standard. Drill and practice saps everyone's enthusiasm, transforming important goals to trivial outcomes. So results mislead.

Remedies are straightforward, but require breaks with familiar practices.

1. Document how any assessment is apt for its purpose(s).
2. Matrix sample (different learners get different tasks) to allow deeper assessment without extending testing time.
3. Show evidence that assessment scores change with intensities of instruction and that results reflect more than student inputs.
4. Infer validity from trials using instructed students.
5. Circulate transparent specifications including standards-based and transfer tasks, (using applied or new academic situations).

Not hard, but a change. At [CRESST](#), approaches to assessment design and assessment quality come from our community and from recent DARPA and IES research on STEM games. Learn more about assessment futures for all children at CRESST's April 29-30 conference "[Warp speed, Mr. Sulu](#)".

Eva L. Baker,
UCLA Graduate School of Education and Information Studies
National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

References

The Gordon Commission on the Future of Assessment in Education. (E. Gordon, Chairperson). (2013). To assess, to teach, to learn: A vision for the future of assessment. The Gordon Commission Final Report). Princeton, NJ: Author.

IES Institute for Education Sciences

DARPA Defense Advanced Research Projects Agency

Categories: [Validity](#) [High-stakes testing](#) [School accountability](#) [International assessments](#) [Test use](#)

Tags: [accountability policies](#) [high-stakes test](#) [standardized testing](#) [technology-based assessments](#) [validity](#)