4

Extensions

Validity: The Crux of the Caper

Chapter 4's Assessment-Related Understanding

Assessment Validation. Assessment validation, the most significant process in all of educational testing, culminates in the creation of a validity argument based on evidence of both inference accuracy and the contribution of a test to the accomplishment of the purpose for which the test is used.

Better Understanding an Understanding

Though surely not as influential as The Ten Commandments—nor as old—the nine assessment-related understandings presented in *The ABCs* do, indeed, constitute an unarguably important collection of truths regarding educational testing. Thus, when Chapter 4's assessment-related understanding characterizes assessment validation as "the most significant process in all of educational testing," we should probably

pay attention to a chapter-understanding that regards itself as focusing on a "most significant process."

This is the first insight regarding the validity chapter's assessment-related understanding. It is such a truly important measurement understanding that a failure to understand it is almost certain to distort a reader's comprehension of *The ABC's'* remaining understandings. Yes, assessment validity is that important! Those who *snooze* in mastering this chapter's assessment-related understanding are definitely destined to *lose.*

The essence of measurement validity is succinctly captured in a single sentence drawn from the 2014 joint *Standards* of AERA, APA, and NCME. "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (page 11). In that definitional statement, assessment validity reflects the extent to which evidence and theory are supportive not only of score-based *interpretations* (or, if you prefer, inferences) but also of the *proposed uses* (or, if you prefer, purposes) of tests. This, then, is the defining essence of educational measurement. We get students to display overt responses to certain sorts of assessment stimuli (such as a test's items) then arrive at hopefully accurate interpretations about those students' covert knowledge and/or skills. But the validation train doesn't stop there, it heads on to the issue of whether the test-based inferences we've drawn about students are truly supportive of whatever purpose the test is attempting to fulfill.

You will note that accuracy determinations of score-based inferences and a test's purpose-support are both *judgmentally* determined. There's not a clever, quantitatively determined number that allows us to definitively assign an A-plus or an F-minus to assessment validity when using an educational test. And this is why the establishment of assessment validity boils down to the use of a *validity argument* whose function is to help a test's users arrive at a defensible judgment about the degree to which a test yields accurate score-based interpretations *and* the extent to which those

interpretations are genuinely supportive of whatever a test's primary purpose is.

The centrality of purpose-support, particularly the presentation of evidence attesting to such support, reminds us of the need to have mastered Chapter 2's assessment-related understanding about the three primary purposes of educational tests. Clearly, whatever evidence is corralled for a test's contribution to decision-making must accurately match a test's dominant purpose. Inference-accuracy, though necessary, takes us only half-way home in assessment's validation sweepstakes. It is also necessary to show that accurate test-based inferences contribute meaningfully to the attainment of an educational test's intended use.

Looking back three paragraphs to the definition of assessment validity in the joint *Standards,* you can see that test-based inferences and purpose-support are confirmed by "evidence and theory." *Please* realize that, in this context, "theory" is simply a ritzy way of explaining how certain variables (as represented by evidence) are related to one another. Consequently, in establishing either the adequacy of a test's interpretive accuracy or its support for an intended use, the emphasis is almost invariably on the amount and quality of *evidence* bearing on both of those issues.

You may recall that in Chapter 4, an attempt was made to dissuade you from ever uttering the phrase "a valid test." This may seem to represent a trifling distinction to you, but in most instances it is far from trivial. If you, and your associates, can *routinely* regard assessment validity as a judgmentally determined attribute of a score-based inference's accuracy *and* the supportiveness of that inference to achieving a test's intended use—and *not* as an attribute of a test itself—then you've surely crossed the assessment-validity finish line with flying colors.

Because assessment validity represents the most significant process in all of educational assessment, this chapter's test-related understanding is, one would think, the most important of all of *The ABCs'* nine understandings. Such thinking is definitely on the mark.

Collegial Conjecturing

After reviewing the contents of the flagrantly phony e-mail below, please generate a response. Then, if you are tangling with *The ABCs* along with other folks, it will be useful to compare different people's responses to the e-mail's writer. As always, the thrust of this Collegial Conjecturing activity is to deepen your understanding of the chapter's assessment-related understanding.

TO: **A READER OF** ***THE ABCS FROM:***
LEE SMITHSON ***SUBJECT:*** **A TEST'S VALIDITY**

Hello:

You may recall meeting a few weeks ago at the Iversons' fund-raising party. As we were chatting, you indicated that you were currently reading "a fascinating and marvelously written" book about the ins and outs of educational testing. I am hoping that you've finished the book by now and can toss a few insights my way regarding educational testing because I am dealing with a remarkably opinionated friend whose opinions about educational testing are making me grind my molars!

What my friend, Tracy, believes—and she typically broadcasts her beliefs for all to hear—is that all high-quality educational tests are valid—regardless of the measurement missions they are attempting to fulfill. I think you said just the opposite.

Here's the event that precipitated my disagreement with her. This year the state has allowed school districts to select their own standardized tests rather than adopt the state-developed tests that annually measure students' mastery of state-approved curricular aims. Our district has chosen a national standardized test that was created as a college entrance exam for high school students—along with a number of lower-grade exams recently developed by the firm distributing the national test. Tracy claims—or should I say *proclaims*—that any test deemed good enough to be used as an admissions test by prestigious colleges is, thereby, good enough to be used for any of the purposes our district's

leaders wish. She says, and I am quoting her accurately, "A properly constructed education test will be valid for just about all assessment needs."

Is Tracy on solid or squishy ground here? If I recall our brief conversation at the Iversons' party, you had been reading that educational tests must be accompanied by evidence regarding their accuracy and their support of a test's intended use. I think you referred to this second factor as "fitness for purpose." What you told me would seem to contradict Tracy's strongly held view.

Could you please straighten me out on this issue? I'll be seeing Tracy again in a few days, and I would love to have some ammunition to contradict her. On the other hand, perhaps she is correct. Help!

Lee

Thought-Provocation Queries

Please look over the following collection of four queries deliberately fashioned with the hope that they would engender a dash of cerebral activity related to this chapter's test-related understanding or, more generally, regarding the chapter's focus on assessment validity. If you find one or more questions that catalyze your cerebration, try to come up with a defensible response to one or more of the queries you selected.

Query 1. Although in previous editions of the AERA-APA-NCME joint *Standards,* the importance of an educational test's purpose was identified, only in the most recent 2014 revision of the joint *Standards* do we see a strong commitment to the assembly of *evidence* supporting a test's contribution to a specific purpose. As a consequence, since the publication of those recent assessment "commandments," we now see markedly increased attention to whether actual evidence—not mere intent—confirms that a test is "fit for purpose." Why do you suppose this shift has occurred?

Query 2. Educational tests are sometimes intended to serve more than one purpose. To illustrate, certain state laws specifically indicate that annual state-sponsored examinations permit evaluations of schools' quality as well as make contributions to improved instruction. How should the *designers* of high-stakes, standardized educational tests deal with such a multi-purpose mandate? How should the *evaluators* of high-stakes, standardized educational tests deal with such a multi-purpose mandate?

Query 3. Why, in Chapter 4, do you think such a big deal was made of the perils flowing from reliance on "a valid test" conception of assessment validity? Is this merely a terminology affectation on the part of measurement devotees as they cavort in educational assessment's vineyards, or do you think it really makes a difference whether we say "a valid test" versus "validity based on evidence of inference-accuracy and purpose-support?"

Query 4. How significant is "most significant?" Should Chapter 4's assertion that assessment validation is "the most significant process in all of educational testing" be taken literally, or is this merely an attempt to have readers pay serious attention to this chapter's understanding? If you were attempting to use an educational test for one of the three major assessment purposes identified in Chapter 2, namely, to compare, to support instruction/learning, or to evaluate instruction, yet insufficient evidence was at hand to support a persuasive validation argument, how would you try to cope with such a situation?

A Real-World Application

As part of the following activity, you will be asked to imagine two sets of evidence for the two twin foci of assessment validation, namely, accuracy of score-based interpretation and

supportiveness of a test's intended use. Although set up as an activity to be carried out by pairs of participants, a solo reader of *The ABCs* can usually profit from undertaking what's asked of participant pairs in this exercise.

WHICH EVIDENCE WINS?

(A PARTICIPANT-PAIRS EXERCISE)

In this exercise, participants are to tackle the assigned activity in pairs. However, if three participants are involved, the exercise can still be carried out by a trio.

The focus of the exercise is the strength of evidence that can be employed to support either (1) the accuracy of a test's score-based inferences or (2) the supportiveness of a test's results in attaining the primary purpose for which a test is being used. Having split into pairs, each participant is allowed up to 10 minutes to arrive individually at two evidence descriptions that can be employed by potential users of the test to reach conclusions about assessment validity as it relates to the test. Although both of these descriptions (again, one emphasizing the accuracy of a score-based inference and one stressing the support supplied by the test for the test's chief purpose) contain positive evidence, one of the two sets of evidence is decisively stronger than the other.

After sufficient time has passed for individuals to isolate fictitious evidence of both kinds, one member of a pair describes (perhaps reads aloud from notes) the two sets of evidence—one dealing with inference accuracy and one dealing with purpose support. The other member of a pair indicates which of the two sets of described evidence is the more compelling—then learns whether this choice coincides with the originator's intent. At that point the roles are reversed, and once more the two sets of evidence are described and a decision is registered regarding the strength of the two sets of evidence.

If time permits, when all pairs (or almost all pairs) have finished their exchange of evidence sets, a general discussion regarding evidentiary strengths and weaknesses should follow.