5

Extensions

Reliability: Considering Consistency in Inconsistent Containers

Chapter 5′s Assessment-Related Understanding

Assessment Reliability. Assessment reliability, the consistency with which an educational test measures what it's measuring, is indicated by three conceptually different kinds of evidence and can be calculated either for test-taker groups or for individual test-takers.

Better Understanding an Understanding

Two significant yet distinctive notions are embodied in Chapter 5′s assessment-related understanding. After acknowledging that "assessment reliability" equals a test's consistency, it is then explained that a test's degree of consistency can be displayed in three fundamentally different ways. A follow-up notion is that an educational test's consistency can be represented either for *groups* of test-takers or for *individual* test-takers. We will soon take a closer look at the three ways of indicating a test's consistency—as well as the difference between a group versus an individual focus for determining such consistency.

But first, please note that reliability, unlike validity, is ascertained for an educational *test itself*. That is, we determine how consistently a test is measuring whatever it aims to measure. Whereas validity is not an attribute of the test itself, but describes the accuracy of a score-based interpretation and how well a test's results show attainment of an intended purpose. This reliability centers on the consistency with which the *test itself* measures.

However, as indicated in the chapter's assessment-related understanding, three fundamentally different conceptions of consistency have been traditionally employed by educational measurement specialists. It is important that you not only know how these three sorts of reliability evidence are determined but also are constantly on-guard whenever anyone trots out "reliability evidence." Do not assume that the sometimes cavalier designations of a test's reliability coincide with the particular kind of reliability in which you are most interested. More often than you'd suspect, such an assumption is unwarranted.

The three sorts of reliability-designation approaches are based on (1) test-retest evidence, (2) alternate-form evidence, and (3) internal consistency evidence. Because the first two of these require a pair of test-administrations, and the third (internal consistency) necessitates only a single test administration, it should come as no surprise that we bump into internal consistency estimates of reliability much more frequently than its two reliability siblings. Yet, because internal consistency reliability coefficients only indicate the degree to which a test's items are functioning in a similar fashion, it should be evident that such coefficients tell us nothing about whether a test possesses test-retest reliability or alternate-form reliability.

Clearly, there is a profound practical difference between estimates of a test's reliability that centers on groups of students (such as all the fifth-graders in a school district) and a reliability estimate that focuses on a test's reliability for a particular fifth-grader—let's call her Tamara. When we lump together a large number of scores and consider their consistency as a coalesced group, the reliability of those scores will be decisively greater than what we calculate specifically for Tamara. Yet, teachers are often obliged to rely on a test's results to arrive at instructional decisions for Tamara, Tommy, and other *individual* students. Educational tests, when influencing decisions about particular students, usually measure with far less consistency than is widely thought. Those who use educational tests to supply results for individual students must become thoroughly conversant with what's embodied in a standard error of measurement (SEM). Interestingly, because the calculation of an SEM hinges on the type of reliability coefficient one incorporates into the SEM formula, the size of an SEM will be determined by which of the three types of reliability coefficients is chosen.

Collegial Conjecturing

Take a look at the e-mail below. It represents the sort of missive that a reader of *The ABCs* might receive from a close friend or even from a casual acquaintance. If you read what the hypothetical sender of this e-mail has to say, and still have an inclination to do so, please conjure up a response. The activity will help you more fully internalize the two chief elements of the chapter's assessment-related understanding.

TO: THE READER OF A SUPPOSEDLY ENTHRALLING BOOK ABOUT EDUCATIONAL TESTING *FROM:* YOUR LONG-TIME PAL *SUBJECT:* IDEA VETTING

Hello again:

I bet you didn't expect to hear from me again after last week's e-mail exchange, but something came up yesterday that you can make better sense of than I. I say so because of what you were telling me about the new book you were reading about educational testing. And, after all, what are friends for?

There's a guy I work with who is pursuing a master's degree at Tipton University, and he's been taking at least two courses regarding educational assessment. Yesterday, as we were having lunch in the departmental lounge, this guy started talking about the reliability of educational tests, especially for individual students. What he said really shook me up, and I wanted your take on his main point. Does it sound reasonable to you—based on what you've been reading?

His main point was that in many instances, if not most, we are *inappropriately* calculating the reliability of tests we employ to make instructional decisions about individual learners. Apparently, teachers (or sometimes counselors if a school has one) calculate what's called a "standard error of measurement" that's symbolized by the letters SEM. This SEM lets teachers know how consistent a student's score is apt to be on a given test. However, because there are actually *three* different ways of calculating reliability for a test, the wrong kind of reliability estimate is often used in an SEM's computation.

My friend indicates that, in most instances, an SEM is based on how similar a test's items are, but teachers are often more interested in how consistently a test measures students' ability over an extended period of time. As a result of often using the wrong input for an SEM's calculation, teachers are likely to make mistakes in their instructional decisions.

Does this make sense to you, or is the guy misunderstanding something? If you think he is on-target, I'd like to hear some more of his views. If you think he is off, I can find other lunchtime friends.

See you next month at Frank's, and thanks for taking a crack at this topic.

Jamal

Thought-Provocation Queries

Please review the following five questions linked—sometime loosely—to this chapter's assessment-related understanding about reliability, and mentally consider how you might answer such a question

Query 1. Test specialists often assert that, "Assessment validity cannot be present if an educational test is unreliable, yet having a reliable test does not guarantee assessment validity." In practical terms, what does this assertion mean? Do you agree with the statement? And, after agreeing or disagreeing, why?

Query 2. In Chapter 5 of *The ABCs*, a position is taken that busy classroom teachers need not calculate indices of assessment reliability for any but their very most important classroom assessments. Do you agree or disagree with that stance? And, having done so, as usual, what's the reasoning behind your own position on this issue?

Query 3. Busy teachers would often like to know how reliable the scores earned by their students are on significant standardized tests are, and even on significant classroom tests. Yet, in many instances the only sort of evidence about assessment reliability is reported for *groups* rather than for *individual* students. Are there any meaningfully helpful insights provided to teachers by group-focused reliability indices such as, for instance, a standardized exam's test-retest reliability coefficient? How would you defend your point of view regarding this issue?

Query 4. When some measurement-moxie educators chat about a test's standard error of measurement (SEM), they point out that, "All SEMs are not equal." Those educators realize that the size of an SEM is heavily influenced by the magnitude of whatever correlation coefficient has been chosen in the calculation of a specific SEM. What factors influence the choice of a correlation indicator in an SEM computation?

Query 5. Two types of reliability evidence are obvious contributors to score-based decisions about both individual students and groups of students. One of the three kinds of reliability evidence may not be. How might one or more relatively common educational decisions be influenced, if at all, by each of the three types of reliability evidence? Are there any kinds of reliability evidence that don't contribute to educational decisions?

A Real-World Application

Because Chapter 5's assessment-related understanding features three fundamentally distinct kinds of evidence regarding a test's consistency, you need to be thoroughly conversant with those three ways of portraying how consistently a test measures what it is supposed to measure. Distinguishing among these three evidence-types is the focus of the activity described below.

TANGLING WITH RELIABILITY'S BLESSED TRINITY (A SUB-GROUP EXERCISE)

This exercise calls for the creation of sub-groups, perhaps 4–8 members in each. The overall structure of the exercise requires each sub-group to prepare descriptions of evidence chiefly indicative of only one of the three types of reliability, that is, test-retest evidence, alternate-forms evidence, or internal consistency evidence. This phase of the exercise usually take 10–15 minutes. These descriptions are then read aloud, one at a time, to the rest of the group (other than the sub-group that generated the description). After each description has been read aloud, members of the total group indicate—by raised hands—which of the types of reliability evidence was described. Members of a sub-group do not "vote" on their own reliability descriptions.

After each description has been presented and the larger group has had a chance to respond, the presenting sub-group indicates which type of reliability evidence has been described. Ideally the sub-groups' descriptions of reliability evidence should be demonstrably correct, but at the same time should not be so terribly obvious that little, if any, conversance with reliability evidence is really required.

Finally, each participant in the large group votes (by secret ballot, if preferred) to identify the sub-group whose descriptions were accurately labeled without being too blatantly obvious. In other words, a sub-group's efforts are regarded as excellent if the resultant descriptions of reliability evidence are definitively one of the three types of reliability evidence—without embodying an "in your face" obviousness. What's sought in this exercise is a display of subtle discrimination among the three types of reliability evidence one might encounter when using educational tests.