Reliability is such a cherished commodity. We all want our automobiles, washing machines, and spouses to be reliable. The term reliability simply reeks of solid goodness. It conjures up visions of meat loaf, mashed potatoes, and a mother's love. Clearly, reliability is an attribute to be sought.

In the realm of educational assessment, reliability is also a desired attribute. We definitely want our educational assessments to be reliable. In matters related to measurement, however, reliability has a very restricted meaning. When you encounter the term reliability in any assessment context, you should draw a mental equal sign between reliability and consistency, because reliability refers to the consistency with which a test measures whatever it’s measuring:

Reliability = Consistency

From a classroom teacher’s perspective, there are two important ways the concept of reliability can rub up against your day-to-day activities. First, there’s the possibility your own classroom assessments might lack sufficient reliability to be doing a good job for you and your students. Second, if your students are obliged to complete any sort of commercially published standardized tests, you’re apt to find a parent or two who might want to discuss the adequacy of those tests. And reliability, as noted in the previous chapter’s preview, such as for your state’s accountability tests, is an evaluative criterion by which external standardized tests are judged. You may need to know enough about reliability’s wrinkles so you’ll be able to talk sensibly with parents about the way reliability is employed to judge the quality of standardized tests.

As explained in Chapter 2, a particularly influential document in the field of educational assessment is the Standards for Educational and Psychological Testing (2014). Commonly referred to simply as “the Standards,” this important compilation of dos and don’ts regarding educational and psychological measurement was developed and distributed by the American Educational Research Association.
(AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). What's prescribed about educational testing in the Standards are the nuts and bolts of how educational tests should be created, evaluated, and used. Because the 2014 Standards constituted the first revision of this significant AERA-APA-NCME publication since 1999, assessment specialists everywhere will be particularly attentive to its contents and, we can safely predict, most of those specialists will adhere to its mandates. Let's consider, then, what the new Standards say about reliability.

Well, for openers, the architects of the Standards use a different label to describe what has historically been referred to as the “reliability” of tests. The authors of the new Standards point out that the term reliability has been used not only to represent (1) the traditional reliability coefficients so often employed through the years when describing a test's quality, but the term reliability refers more generally to (2) the consistency of students’ scores across replications of a testing procedure regardless of the way such consistency is estimated or reported.

In case you are unfamiliar with the meaning of the technical term coefficient as used in this context, that term typically refers to a correlation coefficient—that is, a numerical indicator of the relationship between the same persons' status on two variables such as students' scores on two different tests. The symbol representing such a coefficient is r. If students score pretty much the same on the two tests, the resulting correlation coefficient will be strong and positive. If the individuals score high on one test and low on the other test, the resulting r will be negative. If there's a high reliability coefficient, this doesn't necessarily signify that students' scores on the two testing occasions are identical. Rather, a high r indicates students' relative performances on the two testing occasions are quite similar. Correlation coefficients can range from a high of +1.00 to a low of -1.00. Thus, what a reliability coefficient really means is that a correlation coefficient has typically been computed for test takers' performances on two sets of variables—as you will soon see.

To illustrate, if the developers of a new achievement test report that when their test was administered to the same students on two occasions—for example, a month apart—and that the resulting test-retest correlation coefficient was .86, this would be an instance involving the sort of traditional reliability coefficient that measurement experts have relied on for roughly a full century. If, however, a test-development company creates a brand new high school graduation test, and supplies evidence about what percentage of students’ scores can consistently classify those test takers into “diploma awardees” and “diploma denials,” this constitutes a useful way of representing a test’s consistency—but in a way that’s clearly not the same as reliance on a traditional reliability coefficient. A variety of different indices of classification consistency are often employed these days to supply test users with indications about the reliability with which a test classifies test takers. Don't be surprised, then, if you encounter an indicator of a test's reliability that's expressed in a manner quite different from oft-encountered reliability coefficients.
Architects of the 2014 Standards wanted to provide a link to traditional conceptions of measurement consistency (in which a single reliability coefficient typically indicated a test's consistency), yet avoid the ambiguity of using the single label reliability to refer to a wide range of reliability indicators such as measures of classification consistency. Accordingly, the 2014 Standards architects employ the term reliability/precision to denote the more general notion of score consistency across instances of the testing procedure. The label reliability coefficient is used to describe more conventionally used coefficients representing different forms of test-takers' consistency.

The descriptor reliability/precision, then, describes not only traditional reliability coefficients but also various indicators of classification consistency and a number of less readily understandable statistical indices of assessment consistency. As the writers of the Standards make clear, the need for precision of measurement increases as the consequences on test-based inferences and resultant decisions become more important.

Whether the label reliability/precision becomes widely employed by those who work with educational tests remains to be seen. In this chapter, you will encounter descriptions of reliability in several contexts. Hopefully, it will be clear to you whether the term applies to a quantitative indicator representing a traditional relationship between two sets of test scores—that is, a reliability coefficient—or, instead, refers to another way of representing a test's consistency of measurement—that is, a reliability/precision procedure.

What does a teacher really need to know about reliability or about reliability/precision? Well, hold off for a bit on that question, and I suspect my answer may surprise you. But one thing you do need to recognize is that the fundamental notion of reliability is downright important for those whose professional lives bump up against educational tests in any meaningful manner. This overriding truth about the significance of assessment consistency is well represented in the very first, most basic expectation set forth in the 2014 Standards:

Appropriate evidence of reliability/precision should be provided for the interpretation of each intended score use. (Standards, 2014, p. 42)

Based on this call for appropriate evidence supporting the reliability/precision of an educational test, fully 20 subordinate standards are presented in the new Standards dealing with various aspects of reliability. Taken together, they spell out how the authors of the 2014 Standards believe the consistency of educational tests ought to be determined.

So, recognizing that we can estimate a test's consistency in a number of ways, and that the expression reliability coefficient refers to particular sorts of evidence that's traditionally served up when educational tests are scrutinized, let's look at the three traditional reliability coefficients presented in Table 3.1: test-retest coefficients, alternate-form coefficients, and internal consistency coefficients. Along the way, while describing how these three sorts of reliability coefficients are obtained,
### Table 3.1: Three Types of Reliability Evidence

<table>
<thead>
<tr>
<th>Type of Reliability Coefficient</th>
<th>Brief Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-Retest</td>
<td>Consistency of results among different testing occasions</td>
</tr>
<tr>
<td>Alternate Form</td>
<td>Consistency of results among two or more different forms of a test</td>
</tr>
<tr>
<td>Internal Consistency</td>
<td>Consistency in the way an assessment instrument's items function</td>
</tr>
</tbody>
</table>

we will consider other approaches to be employed when describing a test's reliability/precision.

If you have only recently been tussling with different conceptions of assessments' consistency (such as in the last few paragraphs), an altogether reasonable question for you to ask about reliability coefficients is: How big must a reliability coefficient be? For instance, looking at the three sorts of reliability coefficients tersely described in Table 3.1, how high do those correlation coefficients need to be for us to regard a test as sufficiently reliable? Regrettably, the answer to this altogether straightforward question is apt to be more murky than you might like.

For openers, if a correlation coefficient is involved, it needs to be positive rather than negative. If test developers from an educational measurement company are trying to measure how consistently their newly created history test does its measurement job regardless of when it is administered, they would typically ask a group of students to take the same test on two separate occasions. Thereupon a test-retest correlation could be computed for the test-takers' two sets of scores. Clearly, the hope would be that students' scores on the two administrations would be decisively positive—that is, the scores would reveal substantial similarity in the way students performed on both of the two test administrations. Okay, the test's developers want a positive **test-retest reliability** coefficient to signify their test's stability in doing its measurement job. But how large does that coefficient need to be before the test's developers start clinking glasses during a champagne toast?

Well, get ready for an answer that's often to be encountered when judging the quality of educational tests. That's right: **It depends.** The size that a correlation coefficient needs to be before we regard a test as sufficiently reliable depends on the context in which a test is being used and, in particular, on the nature of the test-based decision that will be influenced by test-takers' scores.

In general, the higher the stakes involved, the higher should be our expectations for our tests' reliability coefficients. For instance, let's suppose that a teacher is working in a state where the awarding of a high school diploma requires a student to perform above a specific cut-score on both a mathematics test and an English language arts test. Given the considerable importance of the decision that's riding on students' test performances, we should be demanding a much greater indication of test consistency than we would with, say, a teacher's exam covering a one-month unit on the topic of punctuation.
But because the contexts in which educational tests are used will vary so considerably, it is simply impossible to set forth a definitive table presenting minimally acceptable levels for certain kinds of reliability coefficients. Experience shows us that when teachers attempt to collect evidence for their own teacher-constructed tests, it is not uncommon to encounter a test-retest $r$ of .60 plus or minus .10 and an alternate-form $r$ reflecting about the same range. Accordingly, when deciding whether a test's reliability coefficients are sufficiently strong for the test's intended use, this decision boils down to professional judgment based on the correlation coefficients seen over the years in similar settings. Yes, we use historical precedent to help us arrive at realistic expectations for what's possible regarding reliability coefficients. For instance, with significant high-stakes examinations such as nationally standardized achievement tests that have been developed and revised many times at great cost, it is not uncommon to see the test developers report internal-consistency coefficients hovering slightly above or slightly below .90. However, for a district-made test requiring far fewer developmental dollars, the identical sorts of internal consistency coefficients might be closer to .80 or .70.

In general, reliability coefficients representing test-retest or alternate-form consistency tend to be lower than internal consistency coefficients. However, in certain contexts—when a test has been developed in an effort to provide distinguishable subscores—we should expect internal consistency coefficients to be much lower because the overall test is not attempting to measure a single, all-encompassing trait but, rather, a set of related subscales scores. As indicated earlier, it depends. To judge whether a test's reliability coefficients are sufficiently strong, that judgment should hinge on what is a realistic expectation for such coefficients in such situations.

Similarly, when looking at reliability/precision more generally—for instance, when considering the consistency with which a set of test scores allow us to classify test takers' levels of proficiency accurately, once more we need to be guided by historical precedent. That is, based on the experience of others, what expectations about reliability/precision are realistic—irrespective of the quantitative indicator employed? For example, if we look back at recent percentages of identical classifications of students on district-developed end-of-course exams, and find that these decision-consistency percentages almost always reach at least 80 percent (that is, 80 percent of the test takers given the same classifications irrespective of the test form they completed), then educators ought to be wary if a new test yielded only a 60 percent estimate of decision consistency.

**TEST-RETEST RELIABILITY EVIDENCE**

The first kind of reliability evidence we'll be looking at is called test-retest. This conception of reliability often comes to people's minds when someone asserts that reliability equals consistency. Formerly referred to in earlier versions of the Standards as "stability reliability," test-retest evidence refers to consistency of test
results over time. We want our educational assessments of students to yield similar results even if the tests were administered on different occasions. For example, suppose you gave your students a midterm exam on Tuesday, but later in the afternoon a masked thief (1) snatched your briefcase containing the students' test papers, (2) jumped into a waiting armored personnel-carrier, and (3) escaped to an adjacent state or nation. The next day, after describing to your students how their examinations had been purloined by a masked personnel-carrier person, you ask them to retake the midterm exam. Because there have been no intervening events of significance, such as more instruction from you on the topics covered by the examination, you would expect your students' Wednesday examination scores to be fairly similar to their Tuesday examination scores. And that's what the test-retest coefficient conception of test reliability refers to—consistency over time. If the Wednesday scores aren't rather similar to the Tuesday scores, then your midterm exam would be judged to have little test-retest reliability.

To get a fix on how stable an assessment's results are over time, we usually test students on one occasion, wait a week or two, and then retest them with the same instrument. Because measurement specialists typically use the descriptors stability reliability and test-retest reliability interchangeably, you are hereby allowed to do the same thing. Simply choose which of the two labels you prefer. It is important, however, for no significant events that might alter students' performances on the second assessment occasion to have taken place between the two testing occasions. For instance, suppose the test you are administering assessed students' knowledge regarding World War II. If a widely viewed television mini-series about World War II is presented during the interval between the initial test and the retest, it is likely the performances of the students who watched the mini-series will be higher on the second test because of their exposure to test-relevant information in the mini-series. Thus, for test-retest coefficients to be interpreted accurately, it is imperative that no significant performance-influencing events occur during the between-assessments interval.

A reliability/precision procedure for calculating the stability of students' performances on the same test administered on two assessment occasions is to determine the percentage of student classifications that were consistent over time. Such a classification-consistency approach to the determination of a test's reliability might be used, for instance, when a teacher is deciding which students would be exempted from further study about Topic X. To illustrate, let's say that the teacher establishes an 80 percent correct level as the degree of proficiency required in order to exempt students from further Topic X study. Then, on a test-retest basis, the teacher would simply determine the percentage of students who were classified the same way on the two assessment occasions. The focus in such an approach would not be on the specific scores a student earned, but only on whether the same classification was made about the student. Thus, if Jill Jones earned an 84 percent correct score on the first testing occasion and a 99 percent correct score on the second testing occasion, Jill would be exempted from further Topic X study in both cases because she had surpassed the 80 percent correct standard both
DECISION TIME

Quibbling over Quizzes

Wayne Wong's first-year teaching assignment is a group of 28 fifth-grade students in an inner-city elementary school. Because Wayne believes in the importance of frequent assessments as motivational devices for his students, he typically administers one or more surprise quizzes per week to his students. Admittedly, after the first month or so, very few of Wayne's fifth-graders are really "surprised" when Wayne whips out one of his unannounced quizzes. Students' scores on the quizzes are used by Wayne to compute each student's 6-weeks' grades.

Mrs. Halverson, the principal of Wayne's school, has visited his class on numerous occasions. Mrs. Halverson believes that it is her "special responsibility" to see that first-year teachers receive adequate instructional support.

Recently, Mrs. Halverson completed a master's degree from the local branch of the state university. As part of her coursework, she was required to take a class in "educational measurement." She earned an A. Because the professor for that course stressed the importance of "reliability as a crucial ingredient of solid educational tests," Mrs. Halverson has been pressing Wayne to compute some form of reliability evidence for his surprise quizzes. Wayne has been resisting her suggestion because, in his view, he administers so many quizzes that the computation of reliability indices for the quizzes would surely be a time-consuming pain. He believes that if he's forced to fuss with reliability estimates for each quiz, he'll reduce the number of quizzes he uses. And, because he thinks students' perceptions that they may be quizzed really stimulates them to be prepared, he is reluctant to lessen the number of quizzes he gives. Even after hearing Wayne's position, however, Mrs. Halverson seems unwilling to bend.

If you were Wayne Wong and were faced with this problem, what would your decision be?

times. The classifications for Jill would be consistent, so the teacher's decisions about Jill would also be consistent. However, if Harry Harvey received a score of 65 percent correct on the first testing occasion and a score of 82 percent correct on the second testing occasion, different classifications on the two occasions would lead to different decisions being made about Harry's need to keep plugging away at Topic X. To determine the percentage of a test's classification consistency, you would simply make the kinds of calculations seen in Table 3.2.

Whether you use a correlational approach or a classification-consistency approach to the determination of a test's consistency over time, it is apparent you'll need to test students twice in order to determine the test's stability. If a test is yielding rather unstable results between two occasions, it's really difficult to put much confidence in that test's results. Just think about it—if you can't tell whether your students have really performed wonderfully or woefully on a test because their
scores might vary depending on the day you test them, how can you proceed to make defensible test-based instructional decisions about those students?

Realistically, of course, why would a sane, nonsadistic classroom teacher administer the identical test to the same students on two different testing occasions? It's pretty tough to come up with a decent answer to that question. What's most important for teachers to realize is that there is always a meaningful level of instability between students' performances on two different testing occasions, even if the very same test is used. And this realization, of course, should disincline teachers to treat a student's test score as though it is a super-scientific, impeccably precise representation of the student's achievement level.

**Table 3.2: Illustration of How Classification Consistency Is Determined in a Test-Retest Context**

<table>
<thead>
<tr>
<th>Identification</th>
<th>Percent</th>
<th>Classification Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Percent of students identified as exempt from further study on both assessment occasions</td>
<td>42%</td>
<td>100% - 42% - 12% = 88%</td>
</tr>
<tr>
<td>B. Percent of students identified as requiring further study on both assessment occasions</td>
<td>46%</td>
<td>100% - 46% - 12% = 88%</td>
</tr>
<tr>
<td>C. Percent of students classified differently on the two occasions</td>
<td>12%</td>
<td>100% - 42% - 46% - 12% = 88%</td>
</tr>
</tbody>
</table>

**ALTERNATE-FORM RELIABILITY EVIDENCE**

The second of our three kinds of reliability evidence for educational assessment instruments focuses on the consistency between two forms of a test—forms that are supposedly equivalent. *Alternate-form reliability evidence* deals with the question of whether two or more allegedly equivalent test forms are, in fact, equivalent.

In the classroom, teachers rarely have reason to generate two forms of a particular assessment instrument. Multiple forms of educational tests are more commonly encountered in high-stakes assessment situations, such as when high school students must pass graduation tests before receiving diplomas. In such settings, students who fail an examination when it is initially administered are usually given other opportunities to pass the examination. Clearly, to make the assessment process fair, the challenge of the assessment hurdle faced by individuals when they take the initial test must be the same as the challenge of the assessment hurdle faced by individuals when they take the make-up examination. Alternate-form reliability evidence bears on the comparability of two (or more) test forms.

Multiple test forms are apt to be found whenever educators fear that if the same test were simply re-used, students who had access to subsequent administrations of the test would be advantaged because those later test takers would have learned about the test's contents, and thus have an edge over the first-time test
takers. Typically, then, in a variety of high-stakes settings such as (1) those involving high school diploma tests or (2) the certification examinations governing entry to a profession, multiple test forms are employed.

To collect alternate-form consistency evidence, procedural approaches are employed that are in many ways similar to those used for the determination of test-retest reliability evidence. First, the two test forms are administered to the same individuals. Ideally, there would be little or no delay between the administration of the two test forms. For example, suppose you were interested in determining the comparability of two forms of a district-developed language arts examination. Let's say you could round up 100 suitable students. Because the examination requires only 20 to 25 minutes to complete, you could administer both forms of the language arts test (Form A and Form B) to each of the 100 students during a single period. To eliminate the impact of the order in which the two forms were completed by students, you could ask 50 of the students to complete Form A and then Form B. The remaining students would be directed to take Form B first and then Form A.

When you obtain each student's scores on the two forms, you could compute a correlation coefficient reflecting the relationship between students' performances on the two forms. As with test-retest reliability, the closer the alternate-form correlation coefficient is to a positive 1.0, the more agreement there is between students' relative scores on the two forms. Alternatively, you could use the kind of classification-consistency approach for the determination of alternate-form reliability/precision that was described earlier for stability. To illustrate, you could decide on a level of performance that would lead to different classifications for students, and then simply calculate the percentage of identically classified students on the basis of the two test forms. For instance, if a pass/fail cutoff score of 65 percent correct had been chosen, then you would simply add (1) the percent of students who passed both times (scored 65 percent or better) and (2) the percent of students who failed both times (scored 64 percent or lower). The addition of those two percentages yields a classification-consistency estimate of alternate-form reliability for the two test forms under consideration.

As you can see, although both species of reliability evidence we've considered thus far are related—in the sense that both deal with consistency—they represent very different conceptions of consistency evidence. Test-retest reliability evidence deals with consistency over time for a single examination. Alternate-form reliability evidence deals with the consistency inherent in two or more supposedly equivalent forms of the same examination.

Alternate-form reliability is not established by proclamation. Rather, evidence must be gathered regarding the between-form consistency of the test forms under scrutiny. Accordingly, if you're ever reviewing a commercially published or state-developed test that claims to have equivalent forms available, be sure you inspect the evidence supporting those claims of equivalence. Determine how the evidence of alternate-form comparability was gathered—that is, under what circumstances. Make sure that what's described makes sense to you.
Later in the book (Chapter 13), we’ll consider a procedure known as *item response theory* that, whenever large numbers of students are tested, can be employed to adjust students’ scores on different forms of tests that are not equivalent in difficulty. For purposes of our current discussion, however, simply remember that evidence of alternate-form reliability is a special form of consistency evidence dealing with the comparability of two or more test forms.

**INTERNAL CONSISTENCY RELIABILITY EVIDENCE**

The final entrant in our reliability evidence sweepstakes is called *internal consistency reliability evidence*. It really is quite a different creature than stability and alternate-evidence form of reliability. Internal consistency evidence does not focus on the consistency of students’ scores on a test. Rather, internal consistency evidence deals with the extent to which the *items* in an educational assessment instrument are functioning in a consistent fashion.

Whereas evidence of stability and alternate-form reliability require two administrations of a test, internal consistency reliability can be computed on the basis of only a single test administration. It is for this reason, one suspects, we tend to encounter internal consistency estimates of reliability far more frequently than we encounter its two reliability brethren. Yet, as you will see, internal consistency reliability evidence is quite a different commodity than stability and alternate-form reliability evidence.

Internal consistency reliability reflects the degree to which the items on a test are doing their measurement job in a consistent manner—that is, the degree to which the test’s items are functioning *homogeneously*. Many educational tests are designed to measure a single variable, such as students’ “reading achievement” or their “attitude toward school.” If a test’s items are all truly measuring a single variable, then each of the test’s items ought to be doing fundamentally the same assessment job. To the extent the test’s items are tapping the same variable, of course, the responses to those items by students will tend to be quite similar. For example, if all the items in a 20-item test on problem solving do, in fact, measure a student’s problem-solving ability, then students who are skilled problem solvers should get most of the test’s items right, whereas unskilled problem solvers should miss most of the test’s 20 items. The more homogeneous the responses yielded by a test’s items, the higher will be the test’s internal consistency evidence.

There are several different formulae around for computing a test’s internal consistency. Each formula is intended to yield a numerical estimate reflective of the extent to which the assessment procedure’s items are functioning homogeneously. By the way, because internal consistency estimates of a test’s reliability are focused on the homogeneity of the items on a test, not on any classifications of test takers (as you saw with stability and alternate-form reliability), decision-consistency approaches to reliability are not used in this instance.
PARENT TALK

One of your strongest students, Raphael Hobbs, has recently received his scores on a nationally standardized achievement test used in your school district. Raphael’s subtest percentile scores (in comparison to the test’s norm group) were the following:

<table>
<thead>
<tr>
<th>Subject</th>
<th>Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language Arts</td>
<td>85th</td>
</tr>
<tr>
<td>Mathematics</td>
<td>92nd</td>
</tr>
<tr>
<td>Science</td>
<td>91st</td>
</tr>
<tr>
<td>Social Studies</td>
<td>51st</td>
</tr>
</tbody>
</table>

Raphael’s father, a retired U.S. Air Force colonel, has called for an after-school conference with you about the test results. He has used his home computer and the Internet to discover that the internal consistency reliabilities on all four subtests, as published in the test’s technical manual, are higher than .93. When he telephoned you to set up the conference, he said he couldn’t see how the four subtests could all
be reliable when Raphael’s score on the social studies subtest was so “out of whack” with the other three subtests. He wants you to explain how this could happen.

If I were you, here’s how I’d respond to Colonel Hobbs:

“First off, Colonel Hobbs, I’m delighted that you’ve taken the time to look into the standardized test we’re using in the district. Not many parents are willing to expend the energy to do so.

“I’d like to deal immediately with the issue you raised on the phone regarding the reliability of the four subtests and then discuss Raphael’s social studies result. You may already know some of what I’ll be talking about because of your access to the Internet, but here goes.

“Assessment reliability refers to the consistency of measurement. But there are very different ways that test developers look at measurement consistency. The reliability estimates that are supplied for Raphael’s standardized test, as you pointed out, are called internal consistency correlation coefficients. Those correlations tell us whether the items on a particular subtest are performing in the same way—that is, whether they seem to be measuring the same thing.

“The internal consistency reliability for all four subtests is quite good. But this kind of reliability evidence doesn’t tell us anything about how Raphael would score if he took the test again or if he took a different form of the same test. We don’t know, in other words, whether his performance would be stable across time or would be consistent across different test forms.

“What we do see in Raphael’s case is a social studies performance that is decisively different from his performance on the other three subtests. I’ve checked his grades for the past few years and I’ve seen that his grades in social studies are routinely just as high as his other grades. So those grades do cast some doubt on the meaningfulness of his lower test performance in social studies.

“Whatever’s measured on the social studies subtest seems to be measured by a set of homogeneous items. That doesn’t mean, however, that the content of the social studies subtest meshes with the social studies Raphael’s been taught here in our district. To me, Colonel, I think it’s less likely to be a case of measurement unreliability than it is to be a problem of content mismatch between what the standardized examination is testing and what we try to teach Raphael.

“I recommend that you, Mrs. Hobbs, and I carefully monitor Raphael’s performance in social studies during the school year so that we can really see if we’re dealing with a learning problem or with an assessment not aligned to what we’re teaching.”

Now, how would you respond to Colonel Hobbs?

For tests containing items on which a student can be right or wrong, such as a multiple-choice item, the most commonly used internal consistency approaches are the Kuder-Richardson procedures (usually referred to as the K-R formulae). For tests containing items on which students can be given different numbers of
points, such as essay items, the most common internal consistency coefficient is called Cronbach's coefficient alpha after its originator Lee J. Cronbach. Incidentally, if you want to impress your colleagues with your newfound and altogether exotic assessment vocabulary, you might want to know that test items scored right or wrong (such as true–false items) are called dichotomous items, and those that yield multiple scores (such as essay items) are called polytomous items. Try to work polytomous into a casual conversation around the watercooler. Its intimidation power is awesome.

Incidentally, other things being equal, the more items there are in an educational assessment device, the more reliability/precision it will tend to possess. To illustrate, if you set out to measure a student's mathematics achievement with a 100-item test dealing with various aspects of mathematics, you’re apt to get a more precise fix on a student's mathematical prowess than if you asked students to solve only a single, lengthy mathematical word problem. The more times you ladle out a taste from a pot of soup, the more accurate will be your estimate of what the soup's ingredients are. One ladleful might fool you. Twenty ladlesful will give you a much better idea of what's in the pot. In general, then, more items on educational assessment devices will yield higher reliability/precision estimates than with fewer items.

THREE COINS IN THE RELIABILITY/PRECISION FOUNTAIN

You’ve now seen that there are three different ways of conceptualizing the manner in which the consistency of a test's results are described. Consistency of measurement is a requisite for making much sense out of a test's results. If the test yields inconsistent results, of course, how can teachers make sensible decisions based on what appears to be a capricious assessment procedure? Yet, as we have seen, reliability evidence comes in three flavors. It is up to you to make sure the reliability evidence supplied with a test is consonant with the use to which the test's results will be put—that is, the decision linked to the test's results. Although there is surely a relationship among the three kinds of reliability evidence we've been discussing, the following is unarguably true:

\[
\begin{array}{ccc}
\text{Test-Re-test} & \text{Alternate-Form} & \text{Internal Consistency} \\
\text{Reliability} & \neq & \text{Reliability} \\
\text{Evidence} & \neq & \text{Evidence}
\end{array}
\]

To illustrate, suppose you were a teacher in a school district where a high school diploma test had been developed by an assistant superintendent in collaboration with a committee of district teachers. The assistant superintendent has claimed the test's three different forms are essentially interchangeable because each form, when field-tested, yielded a Kuder-Richardson reliability coefficient of
.88 or higher. “The three test forms,” claimed the assistant superintendent at a recent school board meeting, “are reliable and, therefore, equivalent.” You now know better.

If the assistant superintendent really wanted to know about between-form comparability, then the kind of reliability evidence needed would be alternate-form reliability, not internal consistency. (Incidentally, I do not recommend that you arise to your feet at the school board meeting to publicly repudiate an assistant superintendent’s motley mastery of measurement reliability. Simply send the assistant superintendent a copy of this book, designating the pages to be read. And send it anonymously.)

Yet, even those educators who know something about reliability and its importance will sometimes unthinkingly mush the three brands of reliability evidence together. They’ll see a K-R reliability coefficient of .90 and assume the test is not only internally consistent but that it will also produce stable results. That’s not necessarily so.

THE STANDARD ERROR OF MEASUREMENT

Before bidding adieu to reliability and all its raptures, there’s one other thing you need to know about consistency of measurement. So far, the kinds of reliability/precision evidence and reliability coefficient evidence we’ve been considering deal with the reliability of a group of students’ scores. For a few paragraphs, please turn your attention to the consistency with which we measure an individual’s performance. The index used in educational assessment to describe the consistency of a particular person’s performance(s) is referred to as the standard error of measurement (SEM). Often a test’s standard error of measurement is identified as the test’s SEM. A standard error of measurement would be another indicator that writers of the 2014 Standards would classify as a reliability/precision procedure.

You should think of the standard error of measurement as a reflection of the consistency of an individual’s scores if a given assessment procedure were administered to that individual again, and again, and again. However, as a practical matter, it is impossible to re-administer the same test innumerable times to the same students because such students would revolt or, if exceedingly acquiescent, soon swoon from exhaustion. (Swooning, these days, is rarely encountered.) Accordingly, we need to estimate how much variability there would be if we were able to re-administer a given assessment procedure many times to the same individual. The standard error of measurement is much like the plus or minus “sampling errors” or “confidence intervals” so frequently given in the media these days for various sorts of opinion polls. We are told that “89 percent of telephone interviewees indicated they would consider brussels sprouts in brownies to be repugnant” (±3 percent margin of error).

Other things being equal, the higher the reliability of a test, the smaller the standard error of measurement will be. For all commercially published tests, a
technical manual is available that supplies the standard error of measurement for the test. Sometimes, if you ever have an occasion to check out a test's standard error, you'll find it's much larger than you might have suspected. As is true of sampling errors for opinion polls, what you'd prefer to have is small, not large, standard errors of measurement.

In many realms of life, big is better. Most folks like big bank accounts, houses with ample square footage, and basketball players who tower over other folks. But with standard errors of measurement, the reverse is true. Smaller standard errors of measurements signify more accurate assessment.

I don't want you to think that a test's standard error of measurement is computed by employing some type of measurement mysticism. Accordingly, presented here is the formula assessment folks use in order to obtain a standard error of measurement. If, however, you don't really care if the standard error of measurement was spawned on Mysticism Mountain, just skip the formula as well as the explanatory paragraph that follows it.

\[ S_e = S_x \sqrt{1 - r_{xx}} \]

where \( s_e \) = standard error of measurement
\( s_x \) = standard deviation of the test scores
\( r_{xx} \) = the reliability of the test

Take a look at the formula for just a moment (especially if you get a kick out of formula looking), and you'll see that the size of a particular test's standard error of measurement (\( s_e \)) depends on two factors. First, there's the standard deviation (\( s_x \)) of the test's scores—that is, how spread out those scores are. The greater the spread in scores, the higher the scores' standard deviation will be. Second, there's the coefficient representing the test's reliability (\( r_{xx} \)). Now, if you consider what's going on in this formula, you'll see that the larger the standard deviation (score spread), the larger the standard error of measurement. Similarly, the smaller the reliability coefficient, the larger the standard error of measurement. So, in general, you see a test will have a smaller standard error of measurement if the test's scores are not too widely spread out and if the test is more reliable. A smaller standard error of measurement signifies that a student's score is more accurately reflective of the student's "true" performance level.³

The standard error of measurement is an important concept because it reminds teachers about the imprecision of the test scores an individual student receives. Novice teachers often ascribe unwarranted precision to a student's test results. I can remember all too vividly making this mistake myself when I began teaching. While getting ready for my first group of students, I saw as I inspected my students' files that one of my students, Sally Palmer, had taken a group intelligence test. (Such tests were popular in those days.) Sally had earned a score of 126. For the next year, I was absolutely convinced that Sally was not merely above average in her intellectual abilities. Rather, I was certain that her IQ was exactly 126.
I was too ignorant about assessment to realize there may have been a sizeable standard error of measurement associated with the intelligence test Sally had taken. Her “true” IQ score might have been substantially lower or higher. I doubt, if Sally had retaken the same intelligence test 10 different times, she’d ever get another score of precisely 126. But, in my naivete, I blissfully assumed Sally’s intellectual ability was dead-center 126. The standard error of measurement helps remind teachers that the scores earned by students on commercial or classroom tests are not so darned exact.

There’s one place a typical teacher is apt to find standard errors of measurement useful, and it is directly linked to the way students’ performances on a state’s accountability tests are reported. Most states classify a student’s performance in at least one of three levels: basic, proficient, or advanced. A student is given one of these three (or more) labels depending on the student’s test scores. So, for instance, on a 60-item accountability test in mathematics, the following classification scheme might have been adopted:

<table>
<thead>
<tr>
<th>Student’s Mathematics Classification</th>
<th>Student’s Mathematics Accountability Test Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced</td>
<td>54–60</td>
</tr>
<tr>
<td>Proficient</td>
<td>44–53</td>
</tr>
<tr>
<td>Basic</td>
<td>37–43</td>
</tr>
<tr>
<td>Below Basic</td>
<td>36 and below</td>
</tr>
</tbody>
</table>

Now, let’s suppose one of your students has a score near the cutoff for one of this fictitious state’s math classifications. To illustrate, suppose your student had answered 53 of the 60 items correctly, and was therefore classified as a proficient student. Well, if the accountability test has a rather large standard error of measurement, you recognize it’s quite likely your proficient student really might have been an advanced student who, because of the test’s inaccuracy, didn’t earn the necessary extra point. This is the kind of information you can relay to both the student and to the student’s parents. But to do so, of course, you need to possess at least a rudimentary idea about where a standard error of measurement comes from and how it works.

The more students’ scores you find at or near a particular cut-score, the more frequently there will be misclassifications. For example, using the previous table, the cut-score for an “Advanced” classification is 54—that is, a student must get a score of 54 or more to be designated “Advanced.” Well, if there are relatively few students who earned that many points, there will be fewer who might be misclassified. However, for the cut-score between “Basic” and “Proficient,” 44 points, there might be many more students earning scores of approximately 44 points, so the likelihood of misclassifications rises accordingly.

Although teachers who understand that a standard error of measurement can reveal how much confidence we can place in a student’s test performance, and
that this is a useful cautionary mechanism for users of test results. Particularly for important tests employed to classify test takers, SEMs can be tailored so that they allow test users to make more precise interpretations at or near classification cut-scores.

Here’s how this process works: As you probably know already, in a given distribution of students’ scores on a particular test, we will ordinarily find substantial differences in the number of students who earn certain sorts of scores. Usually, for example, there are more students who score toward the middle of what often turns out to be a “bell-shaped curve” than there are students who earn either very high or very low scores. Well, assume for the moment that five classification categories have been established for a state-administered achievement test, and that there are four cut-scores separating students’ performances into the five groups.

Because it will be important to assign test takers to one of the five groups, it is possible to calculate not only an overall standard error for the entire array of students’ test scores, but we can also compute SEMs around each of the four cut-scores established for the test. Yes, these are called “conditional standard errors of measurement,” and conditional SEMs can vary considerably in their sizes when calculated for cut-scores representing different areas of an overall distribution of test scores.

It seems unlikely that classroom teachers will ever want to compute standard errors for their own tests, and almost certain that they’ll never wish to compute conditional SEMs for those tests. However, for particularly high-stakes tests such as students’ admissions to universities or students’ attainment of scholarship support, reliability/precision is even more important than usual. Teachers should recognize that it is possible to obtain both overall and conditional SEMs at the cut-score segments of a score distribution, and that access to this information can prove particularly helpful in estimating how near—or far—a given student’s performance is from the closest cut-score.

WHAT DO CLASSROOM TEACHERS REALLY NEED TO KNOW ABOUT RELIABILITY/PRECISION?

What do you, as a teacher or teacher in preparation, truly need to know about reliability/precision? Do you, for example, need to gather data from your own classroom assessment procedures so you can actually calculate reliability coefficients? If so, do you need to collect all three varieties of reliability evidence? My answers may surprise you. I think you need to know what reliability is, but I don’t think you’ll have much call to use it with your own tests—you won’t, that is, unless certain of your tests are extraordinarily significant. And I haven’t run into classroom tests, even rocko-socko final examinations, that I would consider sufficiently significant to warrant your engaging in a reliability-evidence orgy. In general, if
you construct your own classroom tests with care, those tests will be sufficiently reliable for the decisions you will base on the tests’ results.

You need to know what reliability is because you may be called on to explain to parents the meaning of a student’s important test scores, and you’ll want to know how reliable such tests are. You need to know what a commercial test-manual’s authors are talking about and to be wary of those who secure one type of reliability evidence—for instance, a form of internal consistency evidence (because it’s the easiest to obtain)—and then try to proclaim that this form of reliability evidence indicates the test’s stability or the comparability of its multiple forms. In short, you need to be at least knowledgeable about the fundamental meaning of reliability, but I do not suggest you make your own classroom tests pass any sort of reliability muster.

Reliability is a central concept in measurement. As you’ll see in the next chapter, if an assessment procedure fails to yield consistent results, it is almost impossible to make any accurate inferences about what a student’s score signifies. Inconsistent measurement is, at least much of the time, almost certain to be inaccurate measurement. Thus, you should realize that as the stakes associated with an assessment procedure become higher, there will typically be more attention given to establishing that the assessment procedure is, indeed, reliable. If you’re evaluating an important test developed by others and you see that only skimpy attention has been given to establishing reliability for the test, you should be critical of the test because evidence regarding an essential attribute of an educational test is missing.

I also think you ought to possess at least an intuitive understanding of what a standard error of measurement is. Such an understanding will come in handy when you’re explaining to students, or their parents, how to make sense of a student’s scores on such high-stakes external exams as your state’s accountability tests. It’s also useful to know that for very important tests, conditional SEMs can be determined.

The other thing you should know about reliability evidence is that it comes in three brands—three kinds of not interchangeable evidence about a test’s consistency. Don’t let someone foist a set of internal consistency results on you and suggest that these results tell you anything of importance about test-retest evidence. Don’t let anyone tell you that a stability reliability coefficient indicates anything about the equivalence of a test’s multiple forms. Although the three types of reliability evidence are related, they really are fairly distinctive kinds of creatures, something along the lines of second or third cousins.

What I’m trying to suggest is that classroom teachers, as professionals in the field of education, need to understand that an important attribute of educational assessment procedures is reliability. The higher the stakes associated with a test’s use, the more that educators should attend to the assessment procedure’s reliability/precision. Reliability is such a key criterion by which psychometricians evaluate tests that you really ought to know what it is, even if you don’t use it on a daily basis.
But What Does This Have to Do with Teaching?

Students' questions can sometimes get under a teacher's skin. For example, a student once asked me a question that, in time, forced me to completely rethink what I believed about reliability—at least insofar as it made a difference in a classroom teacher's behavior. "Why," the student asked, "is reliability really important for classroom teachers to know?"

The incident occurred during the early seventies. I can still recall where the student was sitting (the back-right corner of an open-square desk arrangement). I had only recently begun teaching measurement courses at UCLA, and I was "going by the book"; in other words, I was following the course's textbook almost unthinkingly.

You see, as a graduate student myself, I had never taken any coursework dealing with testing. My doctoral studies dealt with curriculum and instruction, not measurement. I wanted to learn ways of instructing prospective teachers about how to whip up winning instructional plans and deliver them with panache. But, after graduate school, I soon began to recognize that what was tested on important tests invariably influenced what was taught by most teachers. I began to read all about educational measurement so I could teach the introductory measurement course in the UCLA Graduate School of Education. Frankly, I was somewhat intimidated by the psychometric shroud that testing experts sometimes employ to surround their playpen. Thus, as a beginner in the field of measurement, I rarely strayed from the "truths" contained in traditional measurement textbooks.

I can't recall my answer to that 1970s student, but I'm sure it must have been somewhat insipid. You see, I was merely mouthing a traditional view of reliability as a key attribute of good tests. In my own mind, at that moment, I really didn't have an answer to her Why question. But her question kept bothering me—actually for several years. Then I finally rethought the realistic value of reliability as a concept for classroom teachers. As you'll see in this chapter's wrap-up, I now downplay reliability's importance for busy teachers. But I suspect I still might be dishing out the same old psychometric party line about reliability—if it hadn't been for this one student's perceptive question. One of those psychometrically sanctioned truths was "reliability is a good thing." Accordingly, I was in the midst of a lecture extolling the extreme goodness of reliability when this memorable student, a woman who was at the time teaching sixth-graders while also working on her master's degree, said, "I can't see any practical reason for teachers to know about reliability or to go to the trouble of computing all these reliability coefficients you've just described. Why should we?"

Teachers do need to make sure they evaluate their students' test performances with consistency, especially if students are supplying essay responses or other sorts of performances that can't be scored objectively. But, as a general rule, classroom teachers need not devote their valuable time to reliability exotica. In reality, reliability has precious little to do with a classroom teacher's teaching. But this doesn't make key notions about reliability totally irrelevant to the concerns of classroom teachers. It's just that reliability is way, way less germane than how to whomp up a winning lesson plan for tomorrow's class!
The situation regarding your knowledge about reliability is somewhat analogous to a health professional’s knowledge about blood pressure and how blood pressure influences one’s health. Even though only a small proportion of health professionals work directly with patients’ blood pressure on a day-by-day basis, there are few health professionals who don’t know at least the fundamentals of how one’s blood pressure can influence a person’s health.

I don’t think you should devote any time to calculating the reliability of your own classroom tests, but I think you should have a general knowledge about what it is and why it’s important. Besides, computing too many reliability coefficients for your own classroom tests might give you high blood pressure.

CHAPTER SUMMARY

This chapter focused on the reliability/precision of educational assessment procedures. Reliability refers to the consistency with which a test measures whatever it’s measuring—that is, the absence of measurement errors that would distort a student’s score.

There are three distinct types of reliability evidence. Test-retest reliability refers to the consistency of students’ scores over time. Such reliability is usually represented by a coefficient of correlation between students’ scores on two occasions, but can be indicated by the degree of classification consistency displayed for students on two measurement occasions. Alternate-form reliability evidence refers to the consistency of results between two or more forms of the same test. Alternate-form reliability evidence is usually represented by the correlation of students’ scores on two different test forms, but can also be reflected by classification consistency percentages. Internal consistency evidence represents the degree of homogeneity in an assessment procedure’s items. Common indices of internal consistency are the Kuder-Richardson formulae as well as Cronbach’s coefficient alpha. The three forms of reliability evidence should not be used interchangeably, but should be sought if relevant to the educational purpose to which an assessment procedure is being put—that is, the kind of educational decision linked to the assessment’s results.

The standard error of measurement supplies an indication of the consistency of an individual’s score by estimating person-score consistency from evidence of group-score consistency. The standard error of measurement is interpreted in a manner similar to the plus or minus sampling-error estimates often provided with national opinion polls. Conditional SEMs can be computed for particular segments of a test’s score scale—such as near any key cut-scores. Classroom teachers are advised to become generally familiar with the key notions of reliability, but not to subject their own classroom tests to reliability analyses unless the tests are extraordinarily important.
Determining Your Outcome Mastery

In this chapter, you encountered one of the three cornerstones of today’s educational assessment—namely, reliability. Along with validity and fairness, reliability represents what the writers of the Standards for Educational and Psychological Testing (2014) regard as the three “foundations” of educational and psychological testing. Clearly, reliability is a concept with which all teachers should be reasonably familiar.

Review the chapter’s chief outcome:

An understanding of commonly employed indicators of a test’s reliability/precision that is sufficient to identify the types of reliability evidence already collected for a test and, if necessary, to select the kinds of reliability evidence needed for particular uses of educational assessments.

Put in other words, the outcome is for readers to understand what’s meant by assessment reliability—understand it well enough to correctly label different sorts of validity evidence and, even more importantly, to know in what sorts of situations a particular type of reliability evidence is needed. The Mastery Checks, then, require identifying what sort of reliability evidence is being described in a test item, choosing what sort of reliability evidence ought to be used in the specific context depicted in a test item, and describing a key reliability indicator that’s useful when making decisions about individual students.

Complete both the Selected-Response and the Constructed-Response quizzes and think about the feedback you receive for each quiz.

MyEdLab Selected-Response Check for Outcome Mastery
MyEdLab Constructed-Response Check for Outcome Mastery

After completing both quizzes, go to the Learning Outcome Mastery Determination, where you will decide whether you’ve mastered the chapter’s learning outcome or whether you need further study.

MyEdLab Learning Outcome Mastery Determination

References


Endnotes

1. A special issue of *Educational Measurement: Issues and Practice* (Volume 33, No. 4, Winter 2014) is devoted to the new *Standards*. The special issue contains a marvelous collection of articles related to the nature of the *Standards* and their likely impact on educational practice. Those who are interested in the 2014 *Standards* are encouraged to consult this informative special issue.

2. Please note the use of the Latin plural for *formula*. Because I once completed two years of Latin in high school and three years in college, I have vowed to use Latin at least once per month to make those five years seem less wasted. Any fool could have said *formulas*.

3. You should recognize that because there are different kinds of reliability coefficients used to reflect a test's consistency, the size of a test's standard error of measurement will depend on the *particular* reliability coefficient that's used when calculating a standard error of measurement.