An Interindividual Approach to Reliability

In this Extension we consider—from a consumer’s perspective—some aspects of reliability in greater depth than book space permitted. Although educators who use published tests don’t ordinarily conduct reliability studies, they nonetheless need to know how much confidence they can place in the consistency of test scores.

Therefore, all commercially marketed measures and state assessments should be accompanied by manuals that include information about their reliability in enough detail to enable informed test users to judge whether scores are consistent enough to support the intended use of the test.¹ This Extension is designed to enable readers to make informed judgments concerning the data reported in manuals.

Core Idea Reviewed

Reliability means consistency or capability of being replicated. Consistency of test results is of enormous practical importance for classroom assessments (Gullickson, et al., 2003)². It also has great practical importance in evaluating and using published instruments. Knowledge about a test’s consistency enables educators to exercise appropriate caution in interpreting its scores—a topic to be taken up in Chapter 15. The topic of reliability has a great deal of potential for high-road transfer of learning to other fields because imperfect consistency of scores is common to the (quantitative and qualitative) instruments and methods of all fields—from car appraisals to cholesterol testing to determination of income tax. A universal need exists to enhance the reliability of judgments to the extent that (a) practical constraints enable and (b) the importance of the decision(s) to be made from the test scores merit.

¹ The word “score” in this sentence refers to a test’s total score, to every subscore, and to performance on any item or item cluster that might be interpreted.

² A full citation of this important work can be found in the Suggested Readings for Chapter 14.
A Simple Research Design to Assess Each Major Source of Measurement Error

Recall that four major and common sources of random assessment error in the social sciences are content sampling error, occasion sampling error, examiner error, and scorer error.

Content Sampling Error

Tests, especially Context C tests, sample behavior, and where there is sampling there is usually sampling error. Because of content sampling error, scores often misrepresent examinee status. Content sampling error limits the extent to which useful generalizations can flow from the content sampled to the domain from which the sample was drawn.

Estimating Content Sampling Error. Although the text discusses how to minimize content sampling error, it doesn’t explain how to assess it. Fortunately, a simple research design can estimate its magnitude. To use it, however, one needs two parallel forms of the test under investigation—namely, forms that have been constructed to conform to the same specifications. An appropriate group of examinees is administered both forms, one right after the other. This yields two scores per examinee. These two sets of scores are then correlated. This Pearson $r$ is also a reliability coefficient, more specifically, an alternate-form-immediate reliability coefficient.

If there were no error of measurement, then each person would achieve identical status on the two forms and the correlation would be 1.0. The extent to which equivalent-form reliability coefficients of objective tests are less than 1.0 reflects, for the most part, content sampling error. Research estimates of this and other kinds of measurement error are important for published tests; it is the kind of information we expect to find in test manuals.

---

3 Persons would, that is, get the same score on both perfectly reliable forms if there were no practice effect. Actually, people tend to do a bit better on the second form they take. This doesn’t, however, impact parallel-form-immediate coefficients if everyone gained the same amount (say 1 test point) or if everyone gained proportionally (say achieved a raw score 5% higher) on the second form administered. That is, reliability is impacted only if people’s relative status isn’t consistent.
**Typical Value.** Values for total scores on published tests lasting 40 minutes or so usually fall between about .85 and about .95. Teacher-made objective tests of the same length often have equivalent-form-immediate reliability coefficients in the .5 to .9 range. Of course, teachers who use the test-making procedures described in Part III of your text will, other things being equal, produce much more reliable (and valid) instruments than those who are careless.

**Occasion Sampling Error**

An element of chance is also present when an assessment is administered. Test performance can be influenced by emotional and social factors, illness, fatigue, motivation, weather, etc. Tests administered on a person’s good days tend to overestimate status, while those given on bad days are prone to underestimate it.

**Assessing Occasion Sampling Error.** A way to study occasion sampling error is to give the same form of a test on two occasions. The resulting two sets of scores are then correlated. The $r$ is a test-retest reliability coefficient. If each person performed consistently in comparison with others, then relative status would be identical between the two administrations and $r$ would equal 1.0. The extent by which $r$ is less than 1.0 in objective tests reflects, mainly, inconsistency of examinees from one occasion to the other.

How far apart should the two occasions be? Ideally test-retest reliability of published instruments would be researched for several different time intervals. The interval of greatest interest to a given user may depend largely on (a) the length of time over which the scores are to be used and (b) the extent to which interventions (such as instruction) are changing the attribute being measured.

Scores should be used only while they are current. Some tests serve only short-term purposes. For example, scores from an early unit vocabulary test in Spanish (tía = aunt, etc.) quickly become obsolete. A test-retest interval of overnight or a few days would seem reasonable for such subject matter. Other tests are intended to serve much longer-term purposes. For example, if a state required CPR retesting only every three years for lifeguards, then we would be interested in how consistent the scores were over a three-year period.
Do we hear some readers protesting, “But it isn’t the test’s fault that lifeguards’ real status changes during a period this long.” True; however the point is that we need to know how dependable the use of the scores is. If the results are low, then we might decide to assess more frequently. This is a case of not stick-ing our heads in the sand in an attempt to make a problem disappear.

**Typical Value.** The time interval between test and retest in most published cognitive tests is a week or two, e.g., time enough to get over a common cold or to catch one. Common values for scores based on full-period testing times often range from .85 to about .95.

**Examiner Error**

Scores should reflect examinee status, not who happened to test them. Consequently, differences in scores that occur because of differences in examiners are a source of measurement error.

**Estimating Examiner Error.** An examiner’s performance can be observed while a test is being administered. This performance can then be critiqued. Such practice is common in learning to administer instruments such as individual aptitude tests and individually administered achievement tests of kinds often used in special education. However, there is no direct research design for quantifying the amount by which examiner error reduces reliability.

**Scorer Error**

We know that test results ought to generalize across (a) the form that happens to be used, (b) the occasion on which it chances to be administered, and (c) the particular examiner who administers it. In addition, scores should generalize across scorers; they should reveal examinee status, independent of who happened to assess the work.

**Estimating Scorer Error.** The impact of scorer error can be studied by having a single set of products (e.g., essays) or performances (e.g., tryouts) independently assessed by two raters. This yields two sets of scores that can then be correlated. The $r$ is an interrater reliability coefficient. The extent by which it is less than 1.0 indicates the extent to which inconsistencies in marking detract from the reliability of the scores.
Correlational and Quasi-Correlational Methods of Assessing Reliability

The consistency or reliability of a set of measures can be approached from two somewhat different perspectives. One focuses on the size of measurement error; this is the intra-individual approach used in assessing standard errors of measurement.

The other approach concerns the consistency with which individuals maintain their relative positions in a group on repetition of a measurement procedure (Stanley, 1971). We have just been considering some of these correlational ways to assess inter-individual consistence. In this section this interindividual approach will be explored more fully. Discussion will involve the methods sketched above as well as other important methods.

We saw in the previous section that reliability can be investigated in different ways. It is very important to understand that these ways report different sources of error. Therefore, it isn’t appropriate to say a test’s reliability is so and so (AERA, 1999). Alternative ways of studying reliability are reflected in different kinds of reliability coefficients. That is, conceptualizing replication across different combinations of dimensions—e.g., across different samples of content, on different occasions, with different examiners, and with different scorers—results in different research designs and correspondingly different reliability coefficients (Brennan, 2001). A systematic discussion follows of common means by which the reliability of an instrument or a procedure can be assessed.

Alternate-Form-Immediate

As mentioned earlier, the alternate-form-immediate approach requires that two equivalent forms be administered back-to-back to a group of people. The two resulting sets of scores are correlated. This alternate-form-immediate coefficient reflects the extent to which individuals maintain their same relative positions across the two forms. Therefore, parallel-forms reliability coefficients measure content sampling consistency. Because other sources of random error don’t (significantly) erode them, equivalent-form immediate reliability coefficients tend to be relatively high.
**Test-Retest**

As noted above, the *test-retest* approach requires only one form of a test. It is administered on two occasions with a time interval in between. The two sets of scores are correlated. This test-retest coefficient shows how well individuals maintain their same relative positions across the two occasions. Thus, *test-retest coefficients measure occasion sampling consistency*. Since other sources of random error don’t (markedly) reduce them, they too tend to be relatively high.

**Alternate-Form-with-Interval**

The *alternate-form-with-interval* method requires two parallel forms that are given to a group of people with a time interval in between. The two resulting sets of scores are correlated. This alternate-form-with-interval reliability coefficient reflects the extent to which individuals maintain their same relative positions on the two forms *and* on two occasions. Consequently, they tend to be relatively low.

Hence, *alternate-form-with-interval coefficients are more realistic* as estimates of reliability because *both content sampling error and occasion sampling error really occur when ordinary people are tested*.

Content and occasion sampling error are two major sources of inaccuracy in most tests. Therefore, as practitioners who need an appropriate awareness of tests’ lack of perfect reliability, we prefer manuals to report alternate-form-with-interval coefficients. Typical values of parallel-form-with-interval reliability coefficients for published objective tests 30- to 50-minute length range from about .80 to about .91.

**Less Common Methods**

What about essay tests or the assessment of other products? Their reliability is also decreased by occasion sampling error and content sampling error. In addition, they are vulnerable to scorer error. Realism therefore argues for a research design that allows imperfect reliability of scoring to erode reliability coefficients *in the same way it erodes the test’s reliability in everyday
An appropriate design involves two forms of the test administered on different occasions and scored independently by different persons. The correlation between the resulting sets of scores is low enough to reflect the error from content sampling, occasion, and scorer sampling.

Finally, consider an individually administered performance test, such as an acting audition, an athletic tryout, or an individually administered aptitude test. Its reliability also suffers from examiner error. Therefore, we seek a research design that also permits inconsistency across examiners to be revealed. Research capable of yielding a realistic estimate of the consistency across content samples, time, examiners, and scorers requires having people assessed on two different occasions with different forms of the test or task by different examiners/scorers. The resulting coefficients would, of course, be the lowest, other things being equal, of those we have been considering.

Internal Consistency

The approaches discussed so far are relatively simple conceptually. Unfortunately they present logistic difficulties in execution because they require each examinee either to take a test twice or to take two forms of a test. It is hard for test publishers to obtain cooperation in allowing enough time for two assessments of each participating student. Hence, there is great practical attraction to a group of methods that require only one form of a test to be administered once. Collectively known as internal consistency estimates, these approaches are based on the internal relationships among the parts of tests. Internal consistency estimates are commonly reported for published instruments.

Important Limitations of Internal Consistency Estimates. In addition to being easier and less expensive to secure, internal consistency estimates have another attraction for publishers—one that does not merit consumer sympathy. Internal consistency coefficients tend to be higher than those derived from more rigorous methods such as equivalent-form-with-interval methods because they reflect only one of the major sources of error—content sampling. Naive users look at the numbers and don’t realize that all reliability coefficients are not created equal! Therefore, it isn’t unusual for a test having two or more forms to report only internal consistency reliability. This is not adequate.
Another very important caution applies to all internal consistency reliability estimates. **They are inflated by speed.** If a test’s scores are markedly impacted by speed, none of the internal consistency estimates can appropriately be used (Hopkins, 1998; AERA, 1999) because they exaggerate the test’s consistency. The reliability of such tests should be estimated by another method.

We shall now examine three of the more common kinds of internal consistency estimates.

**Split-Half.** One approach to studying internal consistency is the split-half method in which a test’s scores are split, for research purposes only, into two half-scores. A good way to divide the test into halves is to secure subscores from the test’s odd- and even-numbered items. This common method largely equalizes the impact, if any, of speed and tends to balance content between the halves. Another good method is to divide the test into halves that are as parallel as possible in content.

The two sets of subscores are then correlated. The results are very similar to those of a parallel-form-immediate study. However, the split-half coefficient is the reliability of only a half test, not the full test. Fortunately, there is a formula (known as a special case of the Spearman-Brown Formula) that (given some realistic technical assumptions) enables the half-test reliability to be adjusted to estimate the reliability of the full test. To make this adjustment, the $r$ from the half tests is substituted in the expression:

$$\frac{2r}{1 + r}$$

For example, if the correlation between scores of the odd-numbered items and the even-numbered items on a test is .6, then the adjusted estimate of the full test’s internal consistency reliability is:

$$\frac{2 \times (.6)}{1 + .6} = \frac{1.2}{1.6} = .75$$

**Kuder-Richardson 20.** Kuder and Richardson’s Formula 20 (1937)$^4$ avoids splitting a test for scoring; it yields results that (with minor technical qualifications) are the average of all the split-half coefficients that would result from different splits of the test (Anastasi & Urbina, 1997). Although not a Pearson $r$, it has the relevant features of one; we can be treated as though it were one.

---

Many computerized scoring programs, especially older ones, provide Kuder-Richardson 20 (K-R 20) reliability coefficients. Provided an instrument is not speeded, K-R 20 coefficients can be interpreted somewhat like a parallel-form-immediate coefficient or a corrected odd-even coefficient would be.

A difference, however, is that by virtue of being an average of all ways of splitting a test, **K-R 20 is lowered if the test’s content is not homogeneous.** Thus, if a test consisted of dissimilar kinds of material such as spelling words, arithmetic skill problems, and history items, then K-R 20 would yield much lower results than would the alternate-form or odd-even methods. In the case of more common amounts of dissimilarity among items—such as differences between achievement test items that are assigned to different cells in a table of specifications—the amount that K-R 20 coefficients are reduced by the heterogeneity is relatively minor.

K-R 20 requires that all items count the same amount, that all be scored dichotomously (i.e., no partial credit), and that all be independent (e.g., no blocks of items such as those found in interpretive exercises). These conditions are met in many objective tests, but not in tests containing interpretive exercises, essays (where items may not count the same and where partial credit is common), mathematical proofs, or various kinds of rating forms.

**Coefficient Alpha.** Cronbach (1951)\(^5\) provided a powerful generalization of the K-R 20 that is free of all three limitations specified in the last paragraph. This internal consistency measure, known as **coefficient alpha**, also requires only one administration of one form. This quasi-correlation is routinely provided by many computer programs for scoring and analyzing test data. It can be interpreted much like an odd-even or alternate-form reliability coefficient, except that it, like the K-R 20, (a) is depressed by content heterogeneity and (b) is inflated by speededness.

**To ReCAP**

Several factors contribute to classroom error of measurement. Alternative methods of assessing reliability take different combinations of these everyday factors into account. Table 1 relates various methods of assessing reliability to the kinds of error that impact them. In Table 1, assume the same features for everything not specified. Thus in the first row that specifies test-retest

\(^5\) This work is not cited in your text. Its full citation is: Cronbach, L. J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.
reliability, we should assume that it is the same form of the test being giving on both occasions by the same examiner and that the tests are scored by the same person (or electronic device). Or in the parallel-form-immediate coefficient with the same scorer in the fourth row, assume that the forms are administered by the same person. Such assumptions are appropriate when reading descriptions of reliability studies reported in test manuals.

<table>
<thead>
<tr>
<th>Kind of Reliability Coefficient</th>
<th>Kind(s) of Error Taken into Account</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-retest</td>
<td>Occasion sampling</td>
</tr>
<tr>
<td>Test-retest with different scorers</td>
<td>Occasion sampling</td>
</tr>
<tr>
<td>Test-retest with different examiners</td>
<td>Occasion sampling</td>
</tr>
<tr>
<td>Parallel-form-immediate</td>
<td>Content sampling</td>
</tr>
<tr>
<td>Parallel-form-immediate with different scorers</td>
<td>Content sampling</td>
</tr>
<tr>
<td>Parallel-form-with-interval</td>
<td>Content sampling</td>
</tr>
<tr>
<td>Parallel-form-with-interval and different scorers</td>
<td>Content sampling</td>
</tr>
<tr>
<td>Parallel-form-with-interval and different examiners</td>
<td>Content sampling</td>
</tr>
<tr>
<td>Parallel-form-with-interval and different examiners/scorers</td>
<td>Content sampling</td>
</tr>
<tr>
<td>Split-half</td>
<td>Content (Heterogeneity of content unless halves are parallel)</td>
</tr>
<tr>
<td>Kuder-Richardson 20</td>
<td>Content Heterogeneity of content</td>
</tr>
<tr>
<td>Coefficient Alpha</td>
<td>Content Heterogeneity of content</td>
</tr>
</tbody>
</table>

Thinking CAP Exercise

Following is information for five hypothetical published tests. It is difficult to compare their reliability findings because they report different kinds of studies. Assume each test was not significantly speeded and used suitable random subsamples of students from their norming samples. Also assume these samples to have equal variability. If you need it, use the information in Table 1 to answer the questions.

- Test V: K-R 20 reliability coefficient of .90.
- Test W: odd-even reliability coefficient of .88 (which as been adjusted for full length with the Spearman-Brown formula).
- Test X: test-retest coefficient of .92.
- Test Y: alternate-form-immediate coefficient of .88.

1. Which test had its reliability assessed by a method that always causes heterogeneity of content to erode the coefficient?
2. Which test had its reliability assessed by the most rigorous method?
3. Among Tests V, Y and Z, which test appears to have the least adequate reliability?
4. For which test is no information concerning content sampling error presented?
5. Between Tests Y and Z, which probably has the least content sampling error?
6. Test X’s alternate-form-interval would be most apt to be
   A .94.
   B .92.
   C .91.
   D .85.
7. Test Z’s test-retest coefficient would probably be closest to
   A .98.
   B .92.
   C .88.
   D .84.
Key to Thinking CAP Exercise

1 Test V. K-R 20 and alpha always are reduced when content is heterogeneous. (Unless the two halves of Test W were exactly parallel, its coefficient will also be reduced by content heterogeneity.)

2 Test Z. Alternate-form-with-interval assessment takes account of both content sampling error and occasion sampling error.

3 Test Y. Occasion sampling will surely reduce the consistency of its scores further. Occasion sampling is reflected in the coefficient for Test Z.

4 Test X. Only occasion sampling is reflected in its test-retest study.

5 Test Z. Its coefficient is similar to that of Text Y, but it also reflects occasion sampling error. Therefore its content sampling error is probably smaller.

6 D We already know that occasion sampling alone depressed its reliability to .92. Once content sampling error is also taken into account, the reliability should be considerably lower.

7 B We know that occasion sampling error and content sampling error combined yielded .88. If we use a research design that accounts for only one of these two sources of error, the reliability should go up. (Yet .98 would be “too good to be true.”)

Assessing the Adequacy of Reliability Data Provided in Test Manuals

Let’s now use our understanding of reliability assessment to identify what data must be contained in a test manual if we are to be able to assess the reliability of the test. To clarify, we aren’t now considering how favorable to the test the reliability data are; rather we are identifying the kinds of data needed to facilitate informed evaluation of the instrument’s reliability.
What if No Reliability Data are Reported in a Test Manual?

Sadly, some manuals supply no reliability information. Authors of such deplorable publications usually take one of three approaches.

First, for mastery tests, a way that some authors of manuals try to “weasel out” of going to the trouble of researching a mastery test’s reliability (or to avoid reporting bad news about the instrument’s reliability) is by making a statement along these lines. “Since classical reliability assessment is not relevant to mastery tests, the topic doesn’t apply.” At best, this is a confusing half truth. Although it is correct that correlational methods of assessing reliability are often inappropriate for mastery tests, standard errors of measurement (discussed in Chapter 14) are always relevant to mastery tests. Widely recognized professional standards specify that test manuals of all published tests should provide SEMs for the total scores and for all subscores that are reported (AERA, 1999).

Second, for informal reading inventories and some other informal assessments, the shoddy tactic sometimes used is a statement along the lines, “Since this is not a formal test, formal research concerning reliability is not relevant.” In point of fact, if the scores are to be used to improve decision making, then what the instrument is called is not the issue. (What's in a name? A skunk by any other name would smell the same!) If the results are going to be used, then the consistency of those results needs to be known.

Finally, authors of some manuals of all kinds of published instruments simply ignore the topic. Of course, it is far more dangerous to be unaware of a hazard than to proceed with due cognizance of it. (E.g., if your car’s brakes are inefficient, you need to know it so that you can travel more slowly and more conservatively.) It is foolhearty to use tests as though there were no measurement error; as will be emphasized in Chapter 15, prudent use demands due allowance for the error.

If a manual supplies no information about reliability, the test shouldn’t be used. Sticking our heads in the sand does not make measurement error disappear!
What Reliability Data Should be Reported in Test Manuals?

Information About Reliability Sample(s). A canon of all research reporting is that research samples should be described. This very much applies to the samples used for computing reliability coefficients. The test manual needs to report certain kinds of information for each reliability sample:

- **The Sample Size.** Other things being equal, larger samples yield more stable data than smaller ones. Sample size must always be reported to enable users to know how much credence to give reliability studies.

- **The Mean Score of the Reliability Sample.** Researchers should report the mean score of the sample on every variable used, most especially the test whose reliability is being investigated. This is important in describing the research sample.

- **The Standard Deviation of the Reliability Sample.** Researchers should report the standard deviation of the sample on every variable used. If this information is missing, a study’s findings cannot be interpreted. The reason for this need will be explained after about one more page.

- **Basic Demographic Information.** Judging the relevance of research to one’s own students is facilitated by knowing something about the research sample. This may concern such demographics as location, gender, and ethnicity. For most tests, it is especially important to know the age and/or grade of research subjects.

- **Degree of Speededness If Internal Consistency Estimates Are Provided.** Because internal consistency estimates are inflated by speededness, any test manual that reports an internal consistency coefficient for a timed test should supply evidence regarding the degree to which scores are influenced by speed.

For example, if a manual reported that 96% of the norming sample attempted at least one of the test’s final three items, it would be reasonable to report a measure of internal consistency (although it would be very slightly inflated). On the other hand, if a test is highly speeded (say, less than 70% finish), then its internal
consistency estimates will seriously overrepresent the consistency of scores and should not be reported.

If a manual reports internal consistency estimates, yet fails to report some reasonable measure of degree of test speededness, it is wisest to be skeptical and assume that it was no mere oversight. It is more likely a matter of making obscure the fact that the internal consistency method was inappropriate.

**Standard Error of Measurement.** Test manuals should also report SEMs for scores. In practice, SEMs are computed by a simple formula from reliability coefficients. We know that various kinds of reliability coefficients differ in what sources of error they take into account; it follows that SEMs computed from reliability coefficients will also differ in what sources of error they reflect. Therefore, manuals need to report the kind of reliability coefficient used to compute SEMs.

**Appropriateness of Reliability Sample(s)**

Reliability samples should generally be similar to those with whom a test is going to be used. In terms of demographics, this is a matter of generalizing from a sample to a population. In terms of variability, recall from the Extension on Correlation that variability influences the size of correlation coefficients. Recall also that reliability coefficients are either correlations or quasi-correlations. Uses need to know the sample’s standard deviation on the test if they are to be able to judge whether the sample was appropriately variable. This is often best done by ensuring that the reliability sample’s SD is not materially larger than that of the norming sample.

**Kinds of Measurement Error Taken Into Account**

Ideally, we would like to find in test manuals studies that cast light on all of the major sources of measurement error.

For objective tests, this is content sampling error and occasion sampling error. If two or more forms of the test are available, then we expect alternate-form-with-time-interval reliability coefficients to be reported. If only one form is available, then both
test-retest reliability coefficients and internal consistency coefficients are important. (The latter, however, is only appropriate for tests that are not materially speeded.)

For assessments that involve much latitude for examiners and that involve judgment in scoring, then the major sources of measurement error involve content sampling error, occasion sampling error, examiner error, and scorer error. If two or more forms of the test are available, then we would like to have alternate-form-with-time-interval-different examiners-different scorers reliability coefficients. Sadly, few such tests provide these. At a minimum we would expect to be provided with alternate-form-with-time-interval reliability data. We should then anticipate that the additional measurement error arising from examiners and scorers will substantially lower the rs and substantially raise the SEMs.