Measuring the Power of Learning.®

Research Memorandum
ETS RM–18-01

Test Reliability—Basic Concepts

Samuel A. Livingston

January 2018

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Test Reliability—Basic Concepts

Samuel A. Livingston
Educational Testing Service, Princeton, New Jersey

January 2018

Corresponding author: S. A. Livingston, E-mail: slivingston@ets.org

Suggested citation: Livingston, S. A. (2018). *Test reliability—Basic concepts* (Research Memorandum No. RM-18-01). Princeton, NJ: Educational Testing Service.

Action Editor: Gautam Puhan

Reviewers: Shelby Haberman and Marna Golub-Smith

Abstract

The reliability of test scores is the extent to which they are consistent across different occasions of testing, different editions of the test, or different raters scoring the test taker's responses. This guide explains the meaning of several terms associated with the concept of test reliability: "true score," "error of measurement," "alternate-forms reliability," "interrater reliability," "internal consistency," "reliability coefficient," "standard error of measurement," "classification consistency," and "classification accuracy." It also explains the relationship between the number of questions, problems, or tasks in the test and the reliability of the scores.

Key words: reliability, true score, error of measurement, alternate-forms reliability, interrater reliability, internal consistency, reliability coefficient, standard error of measurement, classification consistency, classification accuracy

Preface

This guide grew out of a class that I teach for staff at Educational Testing Service (ETS). The class is a nonmathematical introduction to the topic, emphasizing conceptual understanding and practical applications. The class consists of illustrated lectures, interspersed with written exercises for the participants. I have included the exercises in this guide, at roughly the same points as they occur in the class. The answers are in the appendix at the end of the guide.

In preparing this guide, I have tried to capture as much as possible of the conversational style of the class. I have used the word “we” to refer to myself and most of my colleagues in the testing profession. (We tend to agree on most of the topics discussed in this guide, and I think it will be clear where we do not.)

Table of Contents

Instructional Objectives

Here is a list of things I hope you will be able to do after you have read this guide and done the written exercises:

- List three important ways in which chance can affect a test taker’s score and some things that test makers can do to reduce these effects.
- Give a brief, correct explanation of the concept of test reliability.
- Explain the difference between reliability and validity and how these two concepts are related.
- Explain the meaning of the terms “true score” and “error of measurement” and why it is wise to *avoid* using these terms to communicate with people outside the testing profession.
- Give an example of an unwanted effect on test scores that is *not* considered “error of measurement.”
- Explain what alternate-forms reliability is and why it is important.
- Explain what interrater reliability is, why it is important, and how it is related to alternate-forms reliability.
- Explain what “internal consistency” is, why it is often used to estimate reliability, and when it is likely to be a poor estimate.
- Explain what the reliability coefficient is, what it measures, and what additional information is necessary to make it meaningful.
- Explain what the standard error of measurement is, what it measures, and what additional information is necessary to make it meaningful.
- Explain how the reliability coefficient and the standard error of measurement are related.
- Describe the relationship between the length of a test (the number of questions, problems, or tasks) and its alternate-forms reliability.

- Explain why the length of a constructed-response test (the number of separate tasks) often affects its interrater reliability.
- Explain what "classification consistency" and "classification accuracy" are and how they are related.

Prerequisite Knowledge

This guide emphasizes concepts, not mathematics. However, it does include explanations of some statistics commonly used to describe test reliability. I assume that the reader is familiar with the following basic statistical concepts, at least to the extent of knowing and understanding the definitions given below. These definitions are all expressed in the context of educational testing, although the statistical concepts are more general.

Score distribution: The number (or the percentage) of test takers at each score level.

Mean score: The average score, computed by summing the scores of all test takers and dividing by the number of test takers.

Standard deviation: A measure of the amount of variation in a set of scores. It can be interpreted as the average distance of scores from the mean. (Actually, it is a special kind of average called a "root mean square," computed by squaring the distance of each score from the mean score, averaging the squared distances, and then taking the square root.)

Correlation: A measure of the strength and direction of the relationship between the scores of the same people on two tests.

What Factors Influence a Test Score?

Whenever a person takes a test, several factors influence the test taker's score. The most important factor (and usually the one with the greatest influence) is the extent to which the test taker has the knowledge and skills that the test is supposed to measure. But the test taker's score will often depend to some extent on other kinds of knowledge and skills, that the test is *not* supposed to measure.

Reading ability and writing ability often influence students' scores on tests that are not intended to measure those abilities. Another influence is the collection of skills we call "test-wiseness." One such skill is using testing time efficiently. Another is knowing when and how to guess on a multiple-choice test. A kind of test-wiseness that is often useful on an essay test is knowing how to include relevant knowledge you have, for which the question does not specifically ask.

One factor that can influence a test score is the test taker's alertness and concentration on the day of the test. In test taking, as in many other activities, most people perform better on some days than on others. If you take a test on a day when you are alert and able to concentrate, your score is likely to be higher than it would be if you took it on a day when you were drowsy or distracted.

On most tests, the questions or problems that the test taker is confronted with are not the only ones that could have been included. Different editions of the test include different questions or problems intended to measure the same kinds of knowledge or skill. At some point in your education, you have probably been lucky enough to take a test that just happened to ask about the things you knew. And you have probably had the frustrating experience of taking a test that happened to include questions about several specific things you did not know. Very few test takers (if any) would perform equally well on any set of questions that the test could include. A test taker who is strong in the abilities the test is measuring will perform well on any edition of the test—but not equally well on every edition of the test.

When a classroom teacher gives the students an essay test, typically there is only one rater—the teacher. That rater usually is the only user of the scores and is not concerned about whether the ratings would be consistent with those of another rater. But when an essay test is part of a large-scale testing program, the test takers' essays will not all be scored by the same rater. Raters in those programs are trained to apply a single set of criteria and standards in rating the essays. Still, a test taker's essay might be scored by a rater who especially likes that test taker's writing style or approach to that particular question. Or it might be scored by a rater who particularly dislikes the test taker's style or approach. In either case, the rater's reaction is likely to influence the rating. Therefore, a test taker's score can depend on which raters happened to score that test taker's essays. This factor affects any test that is scored by a process that involves judgment.

The Luck of the Draw

Which of these influences on a test taker's score can reasonably be assumed to be operating effectively at random? Where does chance ("the luck of the draw") enter into the measurement process?

Does chance affect the test taker’s level of the knowledge or skills that the test is intended to measure? In the testing profession, we make a distinction between the test taker’s knowledge of the specific questions on the test and the more general body of knowledge that those questions represent. We believe that each test taker has a general level of knowledge that applies to any set of questions that might have appeared on that person’s test, and that this general level of knowledge is not affected by chance.

What we do consider as chance variation is the test taker’s ability to answer the specific questions or solve the specific problems on the edition of the test that the test taker took. We reason that the test taker could have been presented with a different set of questions or problems that met the specifications for the test. That set of questions or problems might have been somewhat harder (or easier) *for this test taker*, even if they were not harder (or easier) for most other test takers.

Does chance affect the test taker’s level of *other* kinds of knowledge and skills that affect a person’s test score even though the test is *not* intended to measure them? Again, we make a distinction between the test taker’s general level of those skills and the effect of taking the particular edition of the test that the test taker happened to take. We believe that a test taker’s general level of these skills is not affected by chance, but the need for these skills could help to make a particular edition of the test especially hard (or easy) for a particular test taker.

What about the test taker’s alertness or concentration on the day of the test? We generally think of it as a chance factor, because it affects different test takers’ scores differently and unpredictably. (That is, the effect is unpredictable from our point of view!)

When different test takers’ essays are scored by different raters, we generally consider the selection of raters who score a test taker’s essay to be a chance factor. (The same reasoning applies to other kinds of performance tests.) In some testing programs, we make sure that selection of raters is truly a chance factor, by using a random process to assign responses to raters. But even when there is no true randomization, we think that the selection of raters should be considered a chance factor affecting the test taker’s score.

Reducing the Influence of Chance Factors

What can we testing professionals do to reduce the influence of these chance factors on the test takers’ scores? How can we make our testing process yield scores that depend as little as possible on the luck of the draw?

We cannot do much to reduce the effect of day-to-day differences in a test taker's concentration and alertness (beyond advising test takers not to be tired or hungry on the day of the test). We could reduce the effect of these differences if we could give the test in several parts, each part on a different day, but such a testing procedure would not be practical for most tests. On most tests that have important consequences for the test taker, test takers who think their performance was unusually weak can retake the test, usually after waiting a specified time.

There are some things we can do to reduce the effect of the specific selection of questions or problems presented to the test taker. We can create detailed specifications for the content and format of the test questions or problems, so that the questions on different forms will measure the same set of knowledge and skills. We can avoid reporting scores based on only a few multiple-choice questions or problems. And we can adjust the scores to compensate for differences in the overall difficulty of the questions on different editions of the test.[1] But we cannot make the different editions of a test equally difficult for each individual test taker.

There are also some things we can do to reduce the effect of the specific selection of raters who score a test taker's essays, performance samples, or other responses. We can create explicit scoring instructions, so that all the raters will use the same criteria. We can train the raters thoroughly, with carefully chosen examples of responses to score, so that all the raters will use the same standards in deciding what rating to award. We can test the raters by having them rate essays that have previously been rated by expert raters, and we can require a certain level of accuracy before we allow them to rate operationally. We can monitor the raters' performance, comparing ratings they award with ratings awarded to the same responses by expert raters. We can provide additional training for raters whose ratings do not agree closely with the experts' ratings, and we can replace those raters for whom the retraining is not successful. But we cannot get all raters to agree about the appropriate rating for every response. For some kinds of tests, we can get close to this ideal of perfect agreement, but for others, we cannot.

Exercise: Test Scores and Chance

1. Identify three ways in which luck can influence the score of a test taker who does not guess at any of the answers.
2. Suppose you have a young friend who recently failed the driving portion of her driver's test. Her driving skills have not changed much since then, but she thinks they

> are good enough to pass the test. The first time, she took the test at the Greenwood testing station. That is the most convenient place for her to take the test. The Meadowbrook testing station is farther away, but she will go there if her chances of passing the test would be better there.

Where do you think she would have a better chance of passing the test on her second try—Greenwood or Meadowbrook? Why would her chances be better there?

(Answers to the exercises appear in the appendix.)

What Is Reliability?

Reliability is the extent to which test scores are *not* affected by chance factors—by the luck of the draw. It is the extent to which the test taker's score does *not* depend on ...

- the specific day and time of the test (as compared with other possible days and times of testing),
- the specific questions or problems that were on the edition of the test that the test taker took (as compared with those on other editions), and
- the specific raters who rated the test taker's responses (if the scoring process involved any judgment).

Another way to say this is ...

Reliability Is Consistency

Test scores are reliable to the extent that they are consistent over ...

- different occasions of testing,
- different editions of the test, containing different questions or problems designed to measure the same general skills or types of knowledge, and
- different scorings of the test takers' responses, by different raters.

Why is reliability important? To answer this question, ask yourself whether a test score is useful if it does not indicate, at least approximately ...

- how the test taker would have performed on a different day,

- how the test taker would have performed on a different set of questions or problems designed to measure the same general skills or knowledge, and
- how the test taker's responses would have been rated by a different set of raters.

By now, you may realize that there is more than one kind of reliability. Different kinds of reliability refer to different kinds of consistency. Some kinds of reliability have names that indicate the kind of consistency they refer to. "Alternate-forms reliability" is the consistency of test takers' performance on different editions of the test. "Interrater reliability" is the consistency of the scores awarded by different raters to the same responses (essays, performance samples, etc.). Consistency of test takers' performance on different days or times of testing is called "stability" (or sometimes "test-retest reliability").

Reliability and Validity

Reliability and validity are the two most important properties that test scores can have. They are often mentioned together, but they give us different kinds of information.

- Reliability tells us how consistently the test scores measure *something*.
- Validity tells whether the test scores are measuring *the right things* for a particular use of the test.

Figure 1 is an analogy that illustrates this difference. Using a test to measure a test taker's proficiency in a particular set of knowledge or skills is like shooting at a target, with each shot representing one administration of the test.

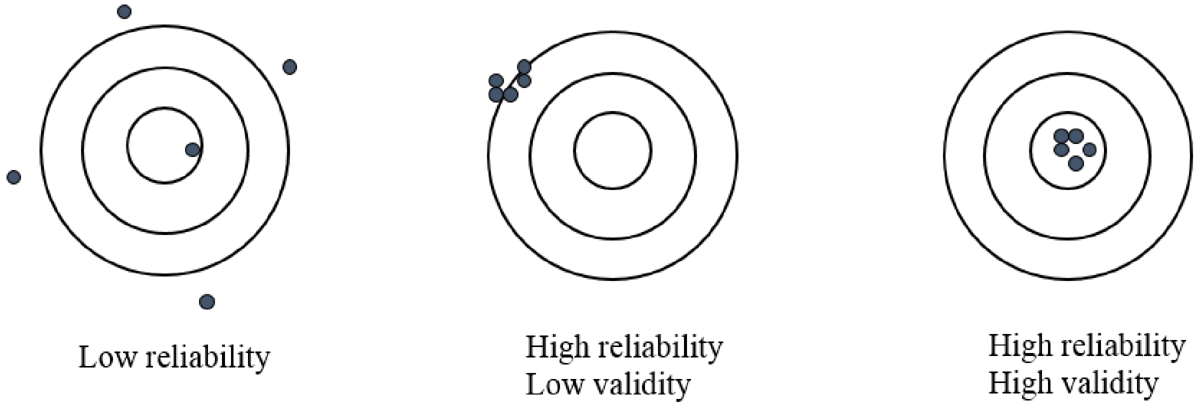

Figure 1. Analogy: Testing a person's proficiency is like shooting at a target.

The concepts of reliability and validity are similar in some important ways and different in others. Reliability refers to a particular source of inconsistency in the scores (or possibly more than one). Validity refers to a particular use of the test. A test can have higher reliability in one group of test takers than in another group; it can also have higher validity in one group of test takers than in another group. But its validity will depend on how the scores are being used. Its reliability will not.

Sometimes people ask, “Which is more important, reliability or validity?” There are two simple ways to answer this question:

- Simple answer #1: Validity is more important. If you are measuring the wrong thing, it doesn’t matter how well you measure it.
- Simple answer #2: Reliability is more important. If the scores depend heavily on chance, you are not measuring anything.

These two answers make it clear that there is really no way to decide whether reliability or validity is more important. Test scores cannot be valid for any purpose unless they are reliable.

Exercise: Reliability and Validity

1. Can a test produce scores that are reliable but not valid? Explain briefly why or why not.
2. Can a test produce scores that are valid but not reliable? Explain briefly why or why not.

(Answers to the exercises appear in the appendix.)

Consistency of What Information?

A test taker’s score provides more than one kind of information. One kind of information a test taker’s score can provide is the test taker’s relative position in some relevant group of test takers. That group might be all the people who took the test in a given time period. Or it could be the much smaller group of people applying for admission to a particular academic department or training program.

A test score also provides information that does not depend on the test taker’s relative position in some group. Saying that a test taker has a *GRE*® Verbal score of 158 provides information that is meaningful to the people in charge of admissions for a graduate program.

They have had the opportunity to see how previous students with GRE Verbal scores near 158 have performed in that program.

Sometimes test takers' scores are used to classify the test takers into groups. For example, students may be classified as "Advanced," "Proficient," or "Not Proficient" in a particular subject. In this case, the most important information the test score provides is which group it places the test taker in.

On some tests, the score is the basis for a decision about the test taker. A test score can determine whether the test taker is awarded a degree, admitted to a training program, or allowed to practice a profession. In these cases, the most important information the test score provides is the decision it leads to. (An old joke may be relevant here. Question: "What do they call the person who gets the lowest passing score on a medical licensing exam?" Answer: "Doctor.")

It is meaningful to talk about the reliability of each of these different kinds of information that a test score can provide and to try to measure their reliability. To do that, we need to use different reliability statistics for the different kinds of information.

"True Score" and "Error of Measurement"

When we testing professionals talk about reliability, there are two terms that we use a lot. The terms are "true score" and "error of measurement" (or "measurement error"). In testing, these words have special meanings. *Do not use these terms when you communicate with people outside the testing profession.* If you do, you are almost sure to be misunderstood.

People outside the testing profession think that your true score on a test is the score you would get if the testing procedure worked perfectly. That is not what people in the testing profession mean by the term "true score." A test taker's "true score" on a test is the *average* of the scores the test taker *would have had*, averaging over *some specified set of conditions*. The test taker's "true score" could be the average over all possible editions of the test. It could be the average over all days and times at which the test taker could possibly have taken the test. If the scoring involves any judgment, it could be the average over all possible raters who could have scored the test taker's responses. It could be the average over two of these sources of inconsistency, or over all three. *The test taker's "true score" depends on which of these factors we are averaging over.* If you use the term "true score," you need to make sure the people you are talking to (or writing for) clearly understand the sources of inconsistency to which you are referring.

Another confusing thing about the “true score” is that it is not a score the test taker can actually get! For example, on most tests, the only possible scores are whole numbers. But the average of many whole numbers is almost never a whole number. Because the “true score” is an average, it is almost never a score that the test taker could get by taking the test once.

Yet another confusing thing about the “true score” is that it may not be a true indication of the skills the test is intended to measure. Anything that affects the test taker’s score *consistently* will affect the test taker’s “true score” in the same way, even if it is not something that the test is intended to measure.

But possibly the most confusing thing about the “true score” is that a test taker’s “true score” can never be known, and yet “true scores” are the basis for the reliability statistics that we report. The test taker’s “true score” can never be known because it is an average of the scores the test taker *would have had* under many circumstances that mostly did not happen. These circumstances include the editions of the test that the test taker did not take, the raters who did not score the test taker’s responses, and so forth. We can report reliability statistics based on “true scores” by using the data from many test takers and making some reasonable assumptions about the data we do not have.

What about “error of measurement”? People outside the testing profession know what an error is—somebody makes a mistake. A driver goes around a curve too fast and the car slides off the road. A baseball player drops the ball. A student subtracts 7 from 16 and gets 8. But that is not how we use the term in testing. When we say “error of measurement,” we do not mean that something is wrong with the test, or that somebody made a mistake in administering or scoring it. In the testing profession, “error of measurement” is the difference between a test taker’s “true score” and the score the test taker actually got. We call the test taker’s actual score the “observed score.” If the test taker’s observed score is higher than his or her “true score,” the “error of measurement” is positive. If the test taker’s observed score is lower than his or her “true score,” the “error of measurement” is negative.

Notice that the definition of “error of measurement” depends on the definition of “true score.” “Error of measurement” includes those sources of inconsistency that are averaged over, in the definition of the test taker’s “true score.”

Reliability and Measurement Error

Reliability is the extent to which test scores are consistent, with respect to one or more sources of inconsistency—the selection of specific questions, the selection of raters, the day and time of testing. Every reliability statistic refers to a particular source of inconsistency (or a particular combination of sources), which it includes as measurement error. *You cannot understand a reliability statistic unless you know what sources of inconsistency it includes as "measurement error."*

The term "measurement error" includes *only* those influences that we think can be assumed to operate randomly. Some kinds of knowledge and skills that affect the test taker's score are not part of what the test is intended to measure. Because they do not operate randomly, we do not consider them to be part of "measurement error." For example, a history test may include an essay intended to measure the students' understanding of some important concepts. If the students type their essays into a computer, their scores will depend, to some extent, on their typing skills. If the students write their essays on paper, their scores will depend, to some extent, on their handwriting skills. In either case, the students will differ on a skill the test is not intended to measure. These differences will affect the students' scores on the test, and the differences will not be random. Differences in typing skill or speed will tend to be consistent over different essay tests taken on the computer. Differences in handwriting skill or speed will tend to be consistent over different essay tests taken on paper. Because the differences in these skills are consistent, they are *not* considered a source of measurement error. Differences in these skills make the test less *valid* (as a measure of the students' understanding of history), but they do not make it less reliable.

Exercise: Measurement Error

For each event described below, indicate whether it is likely to be mainly a result of measurement error.

1. A student in Taiwan wants to apply to an American university that requires a score of at least 20 on each section of the *TOEFL*® test. He takes the test and finds that his speaking score is only 19. He enrolls in a conversational English class that meets 5 hours a week. When he takes the test again, a year later, his speaking score is 22.

( ) Probably measurement error. ( ) Probably not measurement error.

2. A high school student who grew up in Vietnam has been living in the United States for only 1 year. When he takes the state's required test in mathematics, he sees that half the items are word problems; to get the right answer, he must read the question and figure out what mathematical operations to do. His score on the test is lower than the scores of the American-born students who do about as well as he does in math class.

( ) Probably measurement error. ( ) Probably not measurement error.

3. A high school student has a quarrel with his girlfriend on the night before he takes a college admissions test. His scores are lower than those of his classmates whose reading and math skills are similar to his.

( ) Probably measurement error. ( ) Probably not measurement error.

4. A high school student is preparing to take her final exam in U.S. history. She knows the test will include 100 multiple-choice questions and two essay questions. She practices for the test by writing essays on two of the many topics that were covered in class. When she takes the exam, one of the two essay questions is about a topic on which she practiced. Although her performance in the class has been about average, she gets the highest score in the class on the final exam.

( ) Probably measurement error. ( ) Probably not measurement error.

(Answers to the exercises appear in the appendix.)

Reliability and Sampling

When you take a test, the questions or problems you see are only a sample of the questions or problems that could have been included. If you took a different edition of the test, you would see a different sample of questions or problems.

When you take an essay test in class, usually there is only one possible rater—the instructor, who scores all the students' essays. But if the essay test is a large-scale, standardized test, the raters who score your essays are only a sample of the raters who could have been assigned to score them. If your paper were scored again, your essays would be scored by a different sample of raters.

In sampling, other things being equal, more is better.[2] If we increase the number of questions or problems in the test (by adding more questions like those we already have), we will get a better sample of the test taker's performance. The score will be more likely to generalize to other editions of the test containing different questions or problems. If we increase the number of qualified raters who participate—independently—in the scoring of each test taker's responses, we will get a better sample of raters' judgments of the test taker's work. The score will be more likely to generalize to other raters.

The improvement in reliability that results from increasing the number of questions, or the number of raters, can be predicted fairly accurately. Later in this guide, you will see some examples showing the amount of improvement that you can expect from a given increase in the number of questions or the number of raters.

Alternate-Forms Reliability and Internal Consistency

Alternate-forms reliability is the consistency of test takers' scores across different editions of the test, containing different questions or problems testing the same types of knowledge or skills at the same difficulty level. Alternate-forms reliability answers the question, "To what extent do the test takers who perform well on one edition of the test also perform well on another edition?" This is the type of reliability with which test makers are most concerned. It applies to any test that exists in more than one edition. Even if a test exists in only one edition, test users often want the test takers' scores to generalize beyond the specific questions on that edition. Alternate-forms reliability gives the test makers information about a source of inconsistency over which they have some control. By making the test longer, including more questions or problems, they can improve the alternate-forms reliability of the scores.

To measure alternate-forms reliability directly, we need data from test takers who take two different editions of the test without changing in the knowledge and skills the test measures. We don't often have that kind of data. We have to estimate alternate-forms reliability from the data we do have—usually, the test takers' responses to one edition of the test.

Internal consistency is the consistency of test takers' performance on different questions or problems in the same edition of the test. It answers the question, "To what extent do the test takers who perform well on one question also perform well on other questions?" If all the questions on the test measure similar knowledge or skills, the internal consistency will be high. If the questions measure different kinds of knowledge or skills, the internal consistency will not be so high.

Most of the reliability statistics I have seen in technical reports, test manuals, and so forth, have been internal consistency statistics. But does internal consistency really matter? Do we really care whether the same test takers tend to do well on different questions or problems in the same edition of the test? Does it even make sense to refer to internal consistency as a kind of reliability? I think the answer to these questions is No. What we really care about is alternate-forms reliability. We compute internal consistency statistics for two reasons. First, we can compute these statistics from the data we have available—the test takers' responses on a single edition of the test. Second, under an assumption that is usually close to the truth, internal consistency statistics are a good estimate of alternate-forms reliability statistics. [3] That is why we compute internal consistency statistics—not because we care about internal consistency, but because these statistics usually give us a good estimate of alternate-forms reliability.

Sometimes we compute internal-consistency statistics for two or more different editions of a test. When we do, we typically find that the internal-consistency values are not exactly the same for all the different editions. Does that mean that some editions of the test are more reliable than other editions? No—not if the reliability we are concerned about is alternate-forms reliability. Alternate-forms reliability is the extent to which the scores on one edition of the test are consistent with the scores on another edition. It doesn't make sense to say that the alternate-forms reliability of the test is higher on one edition and lower on another. What we have in this case is two or more different *estimates* of the alternate-forms reliability of the test.

When is internal consistency *not* a good estimate of alternate-forms reliability? Here is an example. Suppose we have a writing test consisting of two essays, each measuring a different kind of writing ability. Essay 1 is a descriptive writing task; Essay 2 is a persuasive writing task. And suppose we have two editions of this test; let's call them Form A and Form B. Figure 2 illustrates this situation.

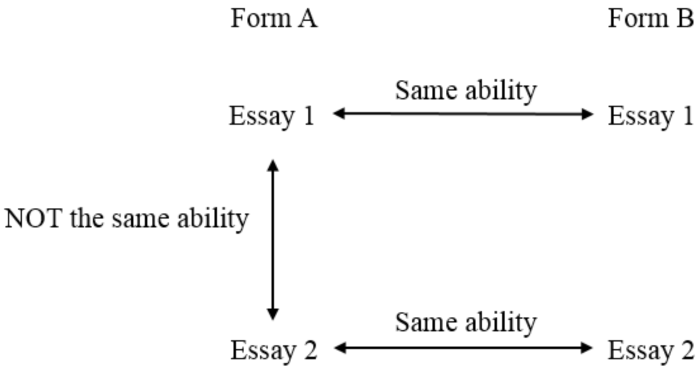

Figure 2. Internal consistency is not always a good estimate of alternate-forms reliability.

Notice that Essay 1 on Form A and Essay 1 on Form B measure the same ability. Similarly, Essay 2 on Form A and Essay 2 on Form B measure the same ability. But Essay 1 and Essay 2 on Form A do not measure the same ability. Therefore, the relationship between Essay 1 and Essay 2 on Form A will not be a good estimate of the relationship between Essay 1 on Form A and Essay 1 on Form B; it will be too low. And it will not be a good estimate of the relationship between Essay 2 on Form A and Essay 2 on Form B; again, it will be too low. As a result, the internal consistency of Form A will not be a good estimate of the alternate-forms reliability of this test; it will be too low.[4]

Interrater Reliability

Interrater reliability is the consistency of the scoring process, on a test for which the scoring involves judgments by raters. It is the agreement between the scores produced by different raters scoring the same responses. It includes only the selection of raters as a source of possible inconsistency. We need to estimate it whenever the scoring involves judgment. If the scoring is a purely mechanical operation, as on a multiple-choice test, the interrater reliability of the scores will be (or should be) perfect.

There are two main approaches to scoring a constructed-response test. In “analytic” scoring, the rater awards points for specific features of the response. For example, in scoring a history question, the rater might award points for correctly identifying up to three causes of a historical event and for correctly identifying up to three results of the event. In “holistic” scoring, the rater makes a single judgment of the whole response and awards a numerical score. Analytic scoring produces scores with high interrater reliability—usually much higher than holistic scoring. Unfortunately, for many kinds of tests, we cannot devise analytic scoring systems that measure the qualities of the response that we think are really important.

On a large-scale test scored by raters, the raters who score a test taker's responses are not the only raters who could have been assigned to score those responses. If another set of raters had scored those responses, it is likely that the test taker's score would have been different. But how likely, and how different? Interrater reliability statistics give us a way to answer this question.

How is interrater reliability related to alternate-forms reliability? If a test taker took two different editions of a test scored by raters, it is unlikely that the same raters would score that person's responses to both editions. Even if the pool of raters were the same for the two editions, it is unlikely that an individual test taker would draw the same raters both times. The alternate-forms reliability of such a test will include *both* the selection of raters *and* the selection of questions or problems as sources of possible inconsistency (i.e., measurement error). For the scores to have high alternate-forms reliability, the effects of both these sources of inconsistency must be small. If the scoring is inconsistent (interrater reliability is low), test takers' *scores* on the two editions of the test will not be consistent, even if their *performance* is consistent. High interrater reliability does not guarantee high alternate-forms reliability, but *low* interrater reliability *does* guarantee *low* alternate-forms reliability.[5]

Test Length and Reliability

"A longer test is a more reliable test." This is a sentence you are likely to hear if you work in the testing profession for any substantial length of time. Is it true of alternate-forms reliability? For most tests, it is. A longer test provides a larger sample of the questions or problems the test consists of. It gives us a better look at the test taker's responses to those kinds of questions or problems. If each question or problem is a separate task, and we increase the number of questions or problems on the test, we will increase the alternate-forms reliability of the scores.

However, on some tests, the questions are grouped into "item sets" consisting of two or more questions based on a common stimulus—a reading passage, a picture, a map, and so forth. If we want to improve the alternate-forms reliability of this kind of test, it may not help much to increase the number of questions in each item set. The questions in an item set all refer to the same stimulus. A test taker who has difficulties with that particular stimulus will have trouble with all the questions about it. To improve alternate-forms reliability, we need to increase the number of item sets.

What about test length and interrater reliability? If the scoring involves judgment, does a longer test produce scores that have higher interrater reliability? The answer to this question depends on how the scoring of the test is organized.

Interrater reliability depends greatly on the number of raters whose judgments contribute independently to each test taker's score. Suppose each test taker's score depended entirely on the judgments of a single rater. If a test taker's responses were scored by a rater who especially liked the test taker's approach, the test taker would tend to get high ratings on all his or her responses. The test taker's score would be much higher than it would have been if a typical rater had scored those responses. If the single rater especially disliked the test taker's approach, the test taker's score would be much lower than if a typical rater had scored those responses. The interrater reliability of the scores would be low. A good way to improve interrater reliability is to increase the number of raters who independently contribute to each test taker's score.

But does increasing the length of the test—the number of questions—increase the number of raters contributing to each test taker's score? If an individual test taker's responses to the different questions are all scored by the same raters, increasing the number of questions will not increase the number of raters. Therefore, it will not increase the interrater reliability of the scores. But if each of a test taker's responses is scored by a different rater (or pair of raters), increasing the number of questions will increase the number of raters. Therefore, it will increase the interrater reliability of the scores.

Exercise: Interrater Reliability and Alternate-Forms Reliability

Three constructed-response tests all measure the **same skill**, using the **same kinds of questions**.

The scoring is holistic.

Test A consists of **one question**. Each response is scored independently by **three** raters.

Test B consists of **two questions**. Each response is scored independently by **two** raters. There are separate panels of raters for the two items.

Test C consists of **five questions**. Each response is scored by **one** rater. There are separate panels of raters for the five items.

Which test will have the highest alternate-forms reliability? Why?

Which test will have the highest interrater reliability? Why?

Reliability and Precision

The 2014 *Standards for Educational and Psychological Testing*, issued jointly by three professional associations,[6] includes a chapter with the title, "Reliability/Precision and Errors of Measurement." The second paragraph of the chapter explains the reasons for this awkward terminology:

> To maintain a link to the traditional notions of reliability while avoiding the ambiguity inherent in using a single, familiar term to refer to a wide range of concepts and indices, we use the term *reliability/precision* to denote the more general notion of consistency of the scores across instances of the testing procedure ... (p. 33).

I think this choice of terms was a really bad idea, not just because it is awkward, but because it is misleading and confusing. Reliability is consistency. Precision is exactness. These two concepts are not the same.

Most standardized tests report "scaled scores," computed by a statistical process that adjusts for the difficulty of the questions on each edition of the test. When we compute scaled scores, we compute them to several decimal places, but we typically report them as whole numbers. For example, if our statistical procedures produce a scaled score of 153.72785126, we will report that score as 154. If we rounded to the nearest 10th, instead of the nearest whole number (e.g., 153.7 instead of 154), the scores would be more precise. But individual test takers' performance would not be more consistent from one day to another. Their performance would not be more consistent from one edition of the test to another. The ratings of their responses would not be more consistent from one set of raters to another. The scores would be more precise, but they would *not* be more reliable.

The authors of this chapter of the *Standards* have chosen to use the term "precision" in a way that is very different from the way the rest of the English-speaking world uses it. Measurement experts many years ago made a similar choice in their use of the terms "true score" and "error," and this usage continues to cause confusion and misunderstanding today.

Reliability Statistics

Reliability is the extent to which test scores are *not* affected by one or more chance factors—the sources of inconsistency that we call “measurement error.” Reliability statistics measure the extent to which test scores (and the information they provide) are consistent over those chance factors. Every reliability statistic refers to a particular set of chance factors—one or more sources of inconsistency in the scores. You don’t know what a reliability statistic is telling you until you know what sources of inconsistency it includes as measurement error. You cannot compare reliability statistics for two different tests unless those statistics include the same sources of inconsistency as measurement error.

A test taker’s score can provide several different kinds of information:

- the score itself, as a meaningful number,
- the test taker’s relative position in a group of test takers, and
- a classification of the test taker or a decision about the test taker, made on the basis of the score.

These different kinds of information require different reliability statistics.

The two most common reliability statistics are the *reliability coefficient* and the *standard error of measurement*. They can (and usually do) refer to the same sources of inconsistency in the scores, but they answer different questions about it.

The Reliability Coefficient

The reliability coefficient is an absolute number that can range from .00 to 1.00. A value of 1.00 indicates perfect consistency. A value of .00 indicates a complete lack of consistency. Scores assigned to the test takers by a completely random process would have a reliability coefficient very close to .00.

The reliability coefficient is actually a correlation. It is the correlation between the test takers’ scores on two applications of the testing process—or, in the case of interrater reliability, two applications of the scoring process. (Although correlations can vary from -1.00 to 1.00, we assume that the scores from two applications of the same testing process would not correlate negatively.)

The reliability coefficient always refers to a group of test takers. It answers the question, "How consistent are the test takers' relative positions in the group, as indicated by their scores?" The answer to this question can differ substantially between one group and another. If the test takers in the group are nearly equal in the abilities the test measures, small changes in their scores can easily change their relative positions in the group. Consequently, the reliability coefficient will be low. The more the test takers differ in their ability, the less likely that small changes in their scores will affect their relative positions in the group, and the higher the reliability coefficient will be. You cannot compare reliability coefficients for two different tests unless they refer to the same population of test takers.

In reading or hearing about a test, you may come across a sentence that says something like this:

> "The reliability coefficient of this test is .90."

To make this statement meaningful, you need more information. You need to know two things:

1. What group of test takers does it refer to?
2. What sources of inconsistency are being included as measurement error?

The Standard Error of Measurement

The standard error of measurement (often referred to as the "SEM") is a number expressed in the same units as the scores it refers to—questions answered correctly, percent correct, scaled-score points, or any other units used to report scores.

Like the reliability coefficient, the SEM refers to a group of test takers. (Often, both the reliability coefficient and the SEM are reported for the same group of test takers.) The SEM answers the question, "On the average, how much do the test takers' scores differ from their 'true scores'?" Although the SEM can differ from one group of test takers to another, it usually does not differ very much. It tends to be nearly the same for different demographic groups, for groups tested at different times, and so forth. It could possibly differ by enough to matter, but it almost never does.[7]

The SEM indicates the consistency of the test takers' scores—not the test takers' relative positions in the group, but the scores themselves. An SEM of 0.00 would indicate that the scores were perfectly consistent—that each individual test taker would get the same score on any application of the testing procedure.

The SEM is actually a standard deviation (or, more precisely, an estimate of a standard deviation). We can never actually compute the error of measurement in any individual test taker's score, no matter what sources of inconsistency we want to include. But if we could compute the error of measurement in each test taker's score, those errors of measurement would have a distribution, just as the scores do. The standard deviation of that distribution would indicate how far, on the average, the test takers' actual scores were from their "true scores." If we could estimate that standard deviation, we could use it to indicate how strongly the errors of measurement are affecting the scores. In fact, we can estimate it, and we do use it that way. Because it is awkward to say "standard deviation of the distribution of errors of measurement," we call it the "standard error of measurement."

In a large group of test takers, the errors of measurement in their test scores tend to have a normal distribution (the familiar bell curve)—even if the scores themselves do not.[8] This fact makes the SEM a useful statistic for describing the reliability of the scores. We can say that about two thirds of the test takers have scores that differ from their "true scores" by less than the SEM. If the SEM is 5 scaled-score points, we can say that about two thirds of the test takers have scaled scores that are within 5 points of their "true scores."

In reading or hearing about a test, you may come across a sentence that says something like this:

"The standard error of measurement of this test is 3.4."

To make this statement meaningful, you need more information. You need to know two things:

1. What is the unit of measurement? 3.4 what? Questions correct? Percentage points? Scaled-score points? Or what?
2. What sources of inconsistency are being included as measurement error?

When we estimate the SEM for interrater reliability, which is the reliability of the scoring process, we often refer to it as the "standard error of scoring." It will always be smaller than the alternate-forms SEM for the same test, which includes both sampling of raters and sampling of questions as sources of measurement error.

How Are the Reliability Coefficient and the Standard Error of Measurement Related?

If the reliability coefficient and the SEM refer to the *same sources of measurement error* and the *same group of test takers*, they are related by the formula

$$\text{Reliability coefficient} = 1\text{-}\left(\frac{\text{SEM}}{\text{SD of scores}}\right)^2 .$$

(SD is an abbreviation for "standard deviation.") You can see from this formula that the smaller the SEM, the higher the reliability coefficient. If the SEM is zero, the reliability coefficient will be 1.00. You can also see that if the standard deviation of the scores is no larger than the SEM, the reliability coefficient will be zero.

Because the SEM does not differ much from one group of test takers to another, the reliability coefficient will depend heavily on the standard deviation of the scores. The larger the standard deviation of the scores, the higher the reliability coefficient. Remember, the reliability coefficient measures the consistency of the test takers' *relative positions in the group*. If the test takers differ greatly in the ability the test measures, their relative positions in the group will tend to be fairly consistent, even if there are small changes in their scores.

Another thing you can see from the formula is that if you change the units in which the scores are measured, the SEM will change, but the reliability coefficient will not. If each correct answer is worth 5 scaled-score points, the SEM of the scaled scores will be 5 times the SEM of the number-correct scores. But the SD of the scaled scores will also be 5 times the SD of the number-correct scores. The fraction on the right side of the formula will be the same for scaled scores as for number-correct scores. And so will the reliability coefficient.[9]

Test Length and Alternate-Forms Reliability

You know that if we increase the number of questions in a test, the reliability coefficient will be higher, and the SEM will be smaller. But how much higher and how much smaller? Here is a made-up but realistic example, with numbers like those we might see for a large-scale standardized test. Suppose we have a multiple-choice test consisting of 100 questions, and in the group of all test takers,

- the alternate-forms reliability coefficient is .90 and
- the alternate-forms SEM of the *percent-correct* scores is 3.0 .

How will these statistics change if we make the test longer by adding more questions or problems like those already on the test? How will they change if we make it shorter? Table 1 shows what the reliability coefficient and the SEM will be if we increase the test to 150 or 200 questions, and if we reduce it to 50, 25, or 15 questions.

Table 1 Test Length and Reliability: An Example

Number of questions	Alternate-forms reliability coefficient	SEM of percent-correct scores
15	0.57	7.7
25	0.69	6.0
50	0.82	4.2
100	0.90	3.0
150	0.93	2.4
200	0.95	2.1

Figure 3 and Figure 4 show the relationship between the length of the test and these two reliability statistics. The calculations underlying these graphs assume that if any of the questions are scored by raters, a test taker's responses to different questions will be scored by different raters.[10]

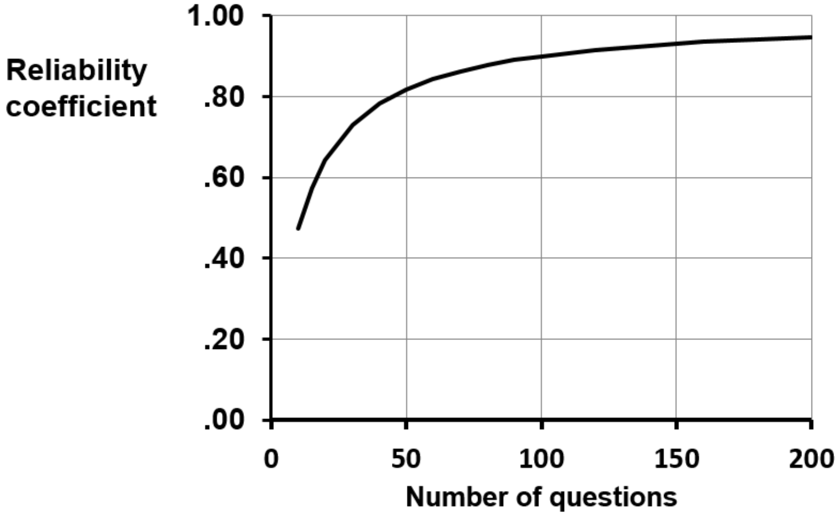

Figure 3. Test length and the reliability coefficient.

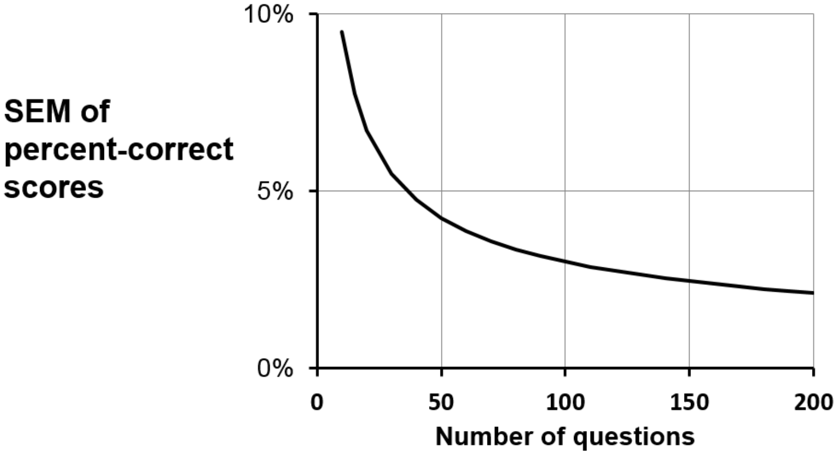

Figure 4. Test length and the standard error of measurement.

You can see that as the number of questions drops below 100, the scores become less and less reliable. Below 50 questions, the reliability gets quite weak. These graphs should help you understand why we are reluctant to report subscores when each subscore is based on a small number of items. You can also see that as the number of questions increases above 100, the improvement in reliability is gradual. The longer the test, the less we can improve its reliability by making it even longer.

Number of Raters and Interrater Reliability

What about interrater reliability? How is it related to the number of raters included in the scoring of a test-taker's responses? Here is a realistic example. Suppose we have a holistically scored test, and each response by a test taker is rated by two raters, working independently. In the group of all test takers,

- the interrater reliability coefficient is .82 and
- the standard error of scoring, as a percentage of the maximum possible score, is 3.5.

How will these statistics change if we increase the number of raters independently rating each response? How will they change if we have only a single rater rating each response? Here is a table showing what will happen (assuming that the additional raters will perform as consistently as those already involved in the scoring).

Table 2 Number of Raters and Interrater Reliability: An Example

Number of ratings of each response	Interrater reliability coefficient	SE of scoring (as percent of maximum possible score)
1	0.69	4.9
2	0.82	3.5
3	0.87	2.9
4	0.90	2.5
5	0.92	2.2
6	0.93	2.0

You can see that the difference between one rating and two ratings is large. The improvement from adding a third rating of each response is not as large, and the improvement from adding a fourth rating is quite small. The more ratings we have, the less we gain from adding another.

You may be wondering how we know what the interrater reliability of the scores is and what it would be with a different number of independent ratings of each response. Typically, we have two raters independently rate each response. For each test taker, we can compute a score that uses only the first rating of each response. We can also compute a score that uses only the second rating of each response. The correlation of these two scores is the interrater reliability coefficient of a score based on one rating of each response. We have a formula that uses this correlation to estimate the interrater reliability coefficient that would result from two, or three, or any number of independent ratings of each response.[11] For tests on which we have only a single rating of each response, we estimate the interrater reliability coefficient by selecting a sample of the test takers and having their responses rated a second time, by different raters.

Reliability of Differences Between Scores

How reliable are differences between scores? Not as reliable as the individual scores themselves. That statement is true for differences between a test taker's scores on the same test, taken at two different times. It is also true for differences between the scores of two different people taking the same test. The reason is that when you compute a difference between two scores, each of those scores is affected by measurement error. If two scores have the same SEM, the difference between those scores will have an SEM about 1.4 times as large as the SEM of the individual scores.

Demystifying the Standard Error of Measurement

The SEM does not have magical properties. This statement may seem unnecessary, but if your work involves testing, you may read or hear statements such as these:

- “A difference less than one SEM is not important.”
- “People with scores that differ by less than the SEM performed equally well on the test.”
- “The SEM tells us how much we can lower the cutscore.”

Statements like these are nonsense. The SEM is not a magic number. It is just a statistic that measures the inconsistency in the scores.

The second of the three statements above is especially misleading. If Adam has a higher test score than Bruce, then Adam performed better on the test than Bruce did. The smaller the difference between them, in relation to the SEM of the test, the greater the chance that Bruce would do better than Adam on a different edition of the test. However, that chance is still less than 50%.

Exercise: The Reliability Coefficient and the Standard Error of Measurement

For Questions 1 to 6, indicate whether the statement is true of the reliability coefficient and whether it is true of the SEM.

1. Other things being equal, the larger this statistic, the more reliable the scores.

True of the reliability coefficient? ( ) Yes ( ) No

True of the SEM? ( ) Yes ( ) No

2. It can take account of more than one source of inconsistency in the scores.

True of the reliability coefficient? ( ) Yes ( ) No

True of the SEM? ( ) Yes ( ) No

3. It is systematically related to the length of the test.

True of the reliability coefficient? ( ) Yes ( ) No

True of the SEM? ( ) Yes ( ) No

4. Its value will always be the same (or very nearly the same) for scaled scores as for raw scores (number correct, sum of ratings, etc.).

True of the reliability coefficient? ( ) Yes ( ) No

True of the SEM? ( ) Yes ( ) No

5. It measures the consistency of individual test takers' positions in a group of test takers.

True of the reliability coefficient? ( ) Yes ( ) No

True of the SEM? ( ) Yes ( ) No

6. It often differs substantially between groups of test takers.

True of the reliability coefficient? ( ) Yes ( ) No

True of the SEM? ( ) Yes ( ) No

7. Your coworker says that the reliability of a constructed-response test is .75.

What two additional pieces of information are necessary to make this statement meaningful?

Reliability of Essay Tests

How reliable are holistically scored essay tests? A few years ago, anyone in the testing profession would have answered this question by saying, "Not very reliable—a lot less reliable than a multiple-choice test requiring the same amount of testing time." However, some recent studies[12] indicate that some holistically scored essay tests used in large-scale standardized testing programs are a lot more reliable than we used to think they were. Each of the tests in these studies consisted of two essays measuring skills that were similar but not really the same. The studies used data from people who took two different editions of the test within a specified period of time. That time period was short enough that real, substantial changes in the test takers' skills would be unlikely. Reliability was estimated by the correlation between the test takers' first and second scores on the test. The test takers in each study also took a multiple-choice test requiring the same amount of testing time as the essay test.

In each of the studies, the alternate-forms reliability coefficients for the essay test and the multiple-choice test were nearly equal. And when the tests were used to predict performance in college or graduate school, the essay test predicted as well as the multiple-choice test (or, in some cases, better).

Reliability of Classifications and Decisions

Often, test scores are used to make decisions about test takers or to classify the test takers into groups. The reliability of those decisions or classifications will depend on the amount of inconsistency in the scores. But it will also depend on the score distribution and on the cut points for the decisions or classifications. If a test taker’s score is near the cut point, that person’s classification can be affected by anything that has even a small effect on the test score. If many test takers have scores near the cut points, the classifications for the group as a whole will tend to be less reliable. If fewer test takers have scores near the cut points, the classifications will be more reliable.

When we compute statistics to describe the reliability of classifications or decisions, we have to specify the sources of inconsistency we want to include as measurement error. Usually, they will be the same sources we would include in computing the reliability coefficient and the SEM:

- the selection of questions or problems on the test, and/or
- the specific day and time of testing, and/or
- the raters who score each test taker’s responses.

However, the questions we want to answer will be different from the questions answered by the reliability coefficient and the SEM. We want to know how well the classifications or decisions we are making on the basis of the scores agree with those we would make on the basis of ...

- the test takers’ scores on a different edition of the test (scored by different raters, possibly taken at a different time);
- the test takers’ “true scores” (if we could somehow know them).

These two kinds of agreement have different names. We refer to the first kind—agreement with another application of the testing process—as “classification consistency” (or “decision consistency”). We refer to the second kind—agreement with classifications based on “true scores”—as “classification accuracy” (or “decision accuracy”).

The statistics we use to describe classification consistency or accuracy are based on a classification table. In the simplest case, when the test takers are being classified into just two groups, the classification consistency table looks like Table 3. The classification accuracy table looks like Table 4. The number in each cell of the table can be either the number of test takers or the percentage of all the test takers.

Table 3 Classification Consistency Table for Two-Group Classification

		Another test edition	
		Higher group	Lower group
Test edition actually taken	Higher group	xxx	xx
	Lower group	xx	xxx

Table 4 Classification Accuracy Table for Two-Group Classification

		"True score"	
		Higher group	Lower group
Test edition actually taken	Higher group	xxx	xx
	Lower group	xx	xxx

Sometimes we want to reduce the information in the table to a single number. There is more than one way to do it. The simplest way is to add the numbers in the upper left and lower right cells of the table—the two cells that represent consistent classifications (or accurate classifications). Then divide by the total of all four cells. This statistic is often called "the percentage of agreement." Some people prefer to report a statistic called "kappa," which they call "the percentage of agreement, corrected for chance."[13]

If we compute a classification consistency table and a classification accuracy table for the same test takers taking the same test, the classification consistency table will always show *lower* agreement than the classification accuracy table. Why? In the classification accuracy table, the second classification is based on the "true score," so it is not affected by measurement error. In the classification consistency table, the second classification is based on a second testing, which *is* affected by measurement error. Trying to agree with a classification based on another testing is like shooting a basketball at a basket that moves unpredictably—not far, but enough to make the task more difficult.

The most direct way to compute classification consistency would be to have the same test takers tested twice, varying the factors to be included as measurement error—the edition of the test and/or the raters of any responses scored by raters. Of course, the two testing sessions would have to be close enough in time that the test takers did not actually change in the abilities the test measures. And the test takers' performance on the second testing would have to be unaffected by their first testing.

In the real world, we almost never get that kind of data. We have to use the data from a single testing to estimate the classification consistency and classification accuracy tables. Therefore, our classification consistency and classification accuracy tables do not include the day of testing as a source of possible inconsistency in the scores.

Summary

Here is a summary of what I think are the most important things to know about test reliability.

Test takers' scores are affected to some degree by factors that are essentially random in the way they affect the scores.

Reliability is consistency—the extent to which test takers' scores are consistent over

- different days of testing,
- different editions of the test, containing different questions or tasks, and
- different raters scoring the test takers' responses.

The reliability of test scores is the extent to which they measure something consistently. The validity of test scores is the extent to which they measure the right things for the way the scores are to be used. Test scores cannot be valid unless they are reliable.

A test taker's "true score" is the average of the scores the test taker would get over some specified set of conditions. "Error of measurement" is the difference between the score a test taker actually got and that test taker's "true score." We cannot know an individual test taker's "true score" or "error of measurement." However, we can estimate statistics for "true scores" and "errors of measurement" in a large group of test takers.

The questions, problems, or tasks on a single edition of a test are a sample of those that could possibly have been included. We can get a better sample of the test taker's performance by enlarging the test to include more questions, problems, or tasks.

Alternate-forms reliability is the consistency of scores on different editions of the test. Often we estimate it by measuring the consistency of test takers' performance on different tasks or questions in the same edition of the test. For most tests, this procedure gives us a good estimate, but for some tests, it does not.

If the test is scored by raters, the raters who score a test taker's responses are a sample of those who could possibly have scored them. We can get a better sample of the scoring by having more raters contribute independently to the test taker's score.

Interrater reliability is the consistency of scores based on ratings of the same responses by different raters. If the test is scored by raters, interrater reliability is necessary for alternate-forms reliability.

A test score can provide more than one kind of information. It can ...

- help a test user identify other people who are similar to the test taker in the skills measured by the test,
- indicate the test taker's relative position in some relevant group, and
- determine how the test taker will be classified or what decision will be made on the basis of the score.

These different kinds of information require different reliability statistics.

Every reliability statistic is based on a definition of measurement error that includes one or more of the chance factors that can affect test scores. You cannot understand a reliability statistic until you know which sources of inconsistency it includes as measurement error.

The reliability coefficient is the correlation between scores on two applications of the testing process. It measures the consistency of the relative positions of the test takers in some group. It often differs substantially between groups of test takers. You cannot understand a reliability coefficient until you know what group of test takers it refers to.

The SEM is the average (root mean square) distance of test takers' scores from their "true scores." It is expressed in the same units as the test scores. It usually does not differ much from one group of test takers to another. In a large group of test takers, about two thirds will get scores that differ from their "true scores" by less than the SEM.

Increasing the number of questions in the test will improve alternate-forms reliability. The more questions the test already includes, the less the improvement from adding an additional question.

Increasing the number of raters who contribute to a test taker's score will improve interrater reliability. The more raters who already contribute to the score, the less the improvement from adding another rater.

Differences between scores are less reliable than the individual scores being compared.

"Classification consistency" is consistency with classifications based on another application of the test. "Classification accuracy" is consistency with the classifications we would make if we knew the test takers' "true scores." Classification consistency is always less than classification accuracy.

Acknowledgments

The opinions expressed in this guide are those of the author and do not necessarily represent the position of Educational Testing Service or any of its clients. I thank Brent Bridgeman, Marna Golub-Smith, Shelby Haberman, Edward Haertel, Michael Kane, John Mazzeo, Robert Mislevy, and Michael Zieky for their comments on earlier drafts. However, they should not be considered responsible in any way for any errors or misstatements that remain.

Appendix. Answers to Exercises

Exercise: Test Scores and Chance

1. Identify three ways in which luck can influence the score of a test taker who does not guess at any of the answers.

The test taker may be unusually alert (or may be drowsy or distracted) on the day of the test.

The test taker can get an edition of the test that includes fewer (or more) questions that the test taker does not know the answers to.

On a constructed-response test, the test taker's responses may be scored by raters who especially like (or dislike) the test taker's writing style, approach to the task, and so forth.

2. Suppose you have a young friend who recently failed the driving portion of her driver's test. Her driving skills have not changed much since then, but she thinks they are good enough to pass the test. The first time, she took the test at the Greenwood testing station. That is the most convenient place for her to take the test. The Meadowbrook testing station is farther away, but she will go there if her chances of passing the test would be better there.

   Where do you think she would have a better chance of passing the test on her second try—Greenwood or Meadowbrook? Why would her chances be better there?

Reason for going to Greenwood: Familiarity with the course. (Practice effect.)

Reasons for going to Meadowbrook: To be sure of getting a different course and a different examiner. The course at Greenwood may have been a difficult course for her. The examiner who failed her at Greenwood may have unusually high standards or may dislike drivers with her particular driving style (or mannerisms or personal qualities).

Exercise: Reliability and Validity

1. Can a test produce scores that are reliable but not valid? Explain briefly why or why not.

Yes. The scores may be very reliable—consistent over different test forms, scorers, and so forth—but if the test is measuring the wrong skills, the scores will not be useful for their intended purpose.

2. Can a test produce scores that are valid but not reliable? Explain briefly why or why not.

No. If the scores are not reliable—if they depend mostly on chance factors, like which edition of the test a person takes—they will not be useful for any purpose.

Exercise: Measurement Error

For each event described below, indicate whether it is likely to be mainly a result of measurement error.

1. A student in Taiwan wants to apply to an American university that requires a score of at least 20 on each section of the TOEFL. He takes the test and finds that his speaking

score is only 19. He enrolls in a conversational English class that meets 5 hours a week. When takes the test again, a year later, his speaking score is 22.

Probably not measurement error. His improved English speaking ability would help him on any edition of the TOEFL, on any day he might take it.

2. A high school student who grew up in Vietnam has been living in the United States for only 1 year. When he takes the state's required test in mathematics, he sees that half the items are word problems; to get the right answer, he must read the question and figure out what mathematical operations to do. His score on the test is lower than the scores of the American-born students who do about as well as he does in math class.

Probably not measurement error. His limited English reading skills would give him the same disadvantage on any day and on any edition of the test. (The number of word problems in a mathematics test is not something that test makers allow to vary randomly.)

3. A high school student has a quarrel with his girlfriend on the night before he takes a college admissions test. His scores are lower than those of his classmates whose reading and math skills are similar to his.

Probably measurement error—unless he and his girlfriend quarrel nearly every Friday night! He would not have this kind of distraction on most other days that he could possibly take the test.

4. A high school student is preparing to take her final exam in U.S. history. She knows the test will include 100 multiple-choice questions and two essay questions. She practices for the test by writing essays on two of the many topics that were covered in class. When she takes the exam, one of the two essay questions is about a topic on which she practiced. Although her performance in the class has been about average, she gets the highest score in the class on the final exam.

Probably measurement error. Most pairs of essay questions that could have been on the test would not include one of the two questions on which she practiced.

Exercise: Interrater Reliability and Alternate-Forms Reliability

Three constructed-response tests all measure the **same skill**, using the **same kinds of questions**.

The scoring is holistic.

Test A consists of **one question**. Each response is scored independently by **three** raters.

Test B consists of **two questions**. Each response is scored independently by **two** raters. There are separate panels of raters for the two items.

Test C consists of **five questions**. Each response is scored by **one** rater. There are separate panels of raters for the five items.

Which test will have the highest alternate-forms reliability? Why?

Test C. It includes five separate tasks for observing the test takers' performance.

Test B includes only two tasks.

Test A includes only one task.

Which test will have the highest interrater reliability? Why?

Test C. Five separate scorers will contribute independently to each test taker's score.

On Test B, only four scorers will contribute independently to each test taker's score.

On Test A, only three.

Exercise: The Reliability Coefficient and the Standard Error of Measurement

For Questions 1 to 6, indicate whether the statement is true of the reliability coefficient and whether it is true of the SEM.

1. Other things being equal, the larger this statistic, the more reliable the scores.

True of the reliability coefficient? *Yes*

True of the SEM? *No*

2. It can take account of more than one source of inconsistency in the scores.

True of the reliability coefficient? *Yes*

True of the SEM? *Yes*

3. It is systematically related to the length of the test.

True of the reliability coefficient? *Yes*

True of the SEM? *Yes*

4. Its value will always be the same (or very nearly the same) for scaled scores as for raw scores (number correct, sum of ratings, etc.).

True of the reliability coefficient? *Yes*

True of the SEM? *No*

5. It measures the consistency of individual test takers' positions in a group of test takers.

True of the reliability coefficient? *Yes*

True of the SEM? *No*

6. It often differs substantially between groups of test takers.

True of the reliability coefficient? *Yes*

True of the SEM? *No*

7. Your coworker says that the reliability of a constructed-response test is .75.

What two additional pieces of information are necessary to make this statement meaningful?

The sources of inconsistency included as measurement error. (Questions or tasks? Raters? Anything else?)

The group of test takers that the reliability coefficient refers to. (The reliability coefficient could be very different in a different group.)

Notes

[1] I have written another booklet that describes this process in some detail. It is called *Equating Test Scores (without IRT)* and is available from Educational Testing Service at https://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf

[2] Ludwig Mies van der Rohe, the architect who famously said, “Less is more,” was not a statistician.

[3] For the statistically sophisticated reader, this assumption is that the average covariance of scores on *different questions in the same edition* of the test is equal to the average covariance of scores on different questions *in different editions* of the test.

[4] Sometimes we compute internal consistency statistics because it is useful to know that the test scores are *at least* as reliable as those statistics indicate.

[5] Sometimes only a small part of the test is scored in a way that requires judgment. On such a test, the alternate-forms reliability of the *total* scores can be greater than the interrater reliability of scores *on the judgmentally scored portion*. But it will not be greater than the interrater reliability of the *total* scores.

[6] The American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.

[7] If a group includes many test takers whose scores are close to the maximum possible score, the SEM will tend to be smaller in that group than in other groups. We don’t see that situation very often.

[8] This statement is based on the central limit theorem. Strictly speaking, it applies only to a test score computed by summing the scores on several separate test questions (e.g., a number-correct or percent-correct score), or to a scaled score based on such a score.

[9]If the conversion from number-correct scores to scaled scores is not a simple multiplication-and-addition, the reliability coefficient may be slightly different for the scaled scores and the number-correct scores. That often happens when two or more number-correct scores at the top of the score range (or at the bottom of the score range) convert to the same scaled score.

[10] The table and the graphs are based on a formula called the “Spearman-Brown formula.” If r is the reliability of the current test and k is the length of the new test divided by the length of the current test, then the reliability of the new test will be $\frac{kr}{1+(k-1)r}$.

[11] Once again, it's the "Spearman-Brown formula." If r is the interrater reliability coefficient that results from a single rating of each response, then the interrater reliability coefficient that would result from n independent ratings of each response is $\frac{nr}{1+(n-1)r}$.

[12] These studies are summarized in an article by Bridgeman, in the Winter 2016 issue of *Educational Measurement Issues and Practice* (pp. 21–24).

[13] Kappa measures the extent to which the classifications are more consistent than they would be if the test takers were classified at random, with the restriction that the percentage of the test takers classified into each group must remain unchanged. A variation often used for classifications with more than two categories is "quadratic weighted kappa."