Chapter 3 Psychometrics: Reliability & Validity

The purpose of classroom assessment in a physical, virtual, or blended classroom is to measure (i.e., scale and classify) examinees' knowledge, skills, and/or attitudes. For example in achievement testing, one measures, using points, how much knowledge a learner possesses (called scaling) and then his or her total raw point score equates to a grade, e.g., "A," "B," "C," etc. (called classifying).

In this chapter, we will consider essential attributes of any measuring device: reliability and validity. Classical reliability indices (i.e., test-retest [stability], parallel forms [equivalence], internal consistency, and inter-rater) are most routinely used in classroom assessment and are hence, discussed. Item Response Theory (IRT) and other advanced techniques for determining reliability are more frequently used with high-stakes and standardized testing; we don't examine those. Four types of validity are explored (i.e., content, criterion-related [predictive or concurrent], and construct). Content validity is most important in classroom assessment. The test or quiz should be appropriately reliable and valid. The test or quiz should be appropriately reliable and valid.

I. Classical Reliability Indices

A. Introduction

1. Reliability is an indicator of consistency, i.e., an indicator of how stable a test score is across applications or time. A measure should produce similar or the same results consistently if it measures the same "thing."
   a. A measure can be reliable without being valid. A measure cannot be valid with out being reliable. The book, *Standards for Educational and Psychological Testing* (2014), provides guidance for all phases of test development.
   b. Data collection tools (i.e., tests, examinations, indexes, scales, measures, etc.) must be constructed so that a "score" (i.e., item, total score, and subtest/subscale scores can be computed). These scores are necessary to compute reliability coefficients. Essay and brief extended responses, scored using a rubric, are included in computing a reliability coefficient.
   c. Open-ended questions an items on an interview guide are excluded in computing a reliability coefficient.

2. <u>The Four Types of Reliability</u>
   a. <u>Test-Retest reliability</u> (also called Stability) answers the question, "Will the scores be stable over time." A test or measure is administered. Some time later the same test or measure is re-administered to the same or highly similar group. One would expect that the reliability coefficient will be highly correlated. For example, a classroom achievement test is administered. The test is given two weeks later with a reliability coefficient of $r = 0.70$, giving evidence of consistency (i.e., stability).
   b. <u>Parallel forms reliability</u> (also called Equivalence) answers the question, "Are the two forms of the test or measure equivalent?" If different forms of the same test or measure are administered to the same group; one would

expect that the reliability coefficient will be high. For example, Form A and Form B of a test of customer service knowledge or reading achievement are administered. The scores correlate at *r* = 0.77, giving evidence of equivalency.

c. Internal consistency reliability answers the question, "How well does each item measure the content or construct under consideration?" It is an indicator of reliability for a test or measure which is administered once. One expects the correlation between responses to each test item to be highly correlated with the total test score. For example, individual items (questions) on an employee job satisfaction (attitude) scale or a classroom achievement test which is administered once, should measure the same attitude or knowledge.

d. Different trained raters, using a standard rating form, should measure the object of interest consistently; this is called inter-rater reliability. Inter-rater agreement answers the question, "Are the raters consistent in their ratings?" The reliability coefficient will be high, if the observers rated similarly. For example, three senior sales trainers rating the closing skills of a novice sales representative or master teachers rating the teaching effectiveness of a first or second year teacher should agree in their ratings.

B. The Theoretical Basis for Classical Reliability Indices

1. The Classical True Score Model is the theoretical basis for classical reliability.
   a. The Classical True Score Model is O = T + E, where "O" = observed score; "T" = true score (what an examinee really knows) + "E" = error.
   b. An individual's observed score is composed of a true score and error. Or, add the true score and measurement error to get the observed score (i.e., the earned test score, 88% or 88/100 points).
   c. The error term is due to systematic and/or random error.
      (1) Error prevents a measure (e.g., test or scale) from perfect reliability.
      (2) We try to keep measurement error very small so a true score almost equals an observed score. A high reliability coefficient indicates lower measurement error: the true and observed scores are more similar.

2. True Scores
   a. A true score is or reflects what the examinee actually knows or more formally, "the examinee's true score can be interpreted as the average of the observed scores obtained over an infinite number of repeated testing with the same test" (Crocker & Algina, 1986, p. 109).
   b. An examinee's true score is unrelated to measurement errors which affect the examinee's observed score.

3. Error Scores
   a. An error score is that part of the observed test score due to factors other than what the examinee knows or can do. There are two types of error: random and systematic.
   b. Random error exerts a differential effect on the same examinee across different testing sessions; because of this inconsistent effect, reliability is affected. Random errors vary from examinee to examinee; there is no consistency in the source of error. Sources of random error include
      (1) Individual examinee variations, e.g., mood changes, fatigue, stress, perceptions of importance;
      (2) Administration condition variation such as noise, temperature, lighting, seat comfort;
      (3) Measurement device bias which favors some and places others at a disadvantage due to gender, culture, religion, language or other factors such as ambiguous wording of test items;
      (4) Participant bias e.g., guessing, motivation, cheating, and sabotage; and
      (5) Test administrator bias such as nonstandard directions, inconsistent proctoring, scoring errors, inconsistent score or results interpretation.
   c. Systematic error is that error which is consistent across uses of the measurement tool (i.e., test or scale) and is likely to affect validity, but not reliability. Examples include an incorrectly worded item, poorly written directions, inclusion of items unrelated to the content, theory, etc. upon which the measurement tool is based.

4. Measuring Error: Standard Error of Measurement (S_E or SEM)
   a. We can think of a standard error as the standard deviation of the error term from the Classical True Score Model.
      (1) The closer to zero the standard error is, the better. Zero reflects an absence of measurement error, thus O (Observed Score) = T (True Score). A standard error is never larger than its standard deviation.
      (2) The standard error is only computed for a group, not an individual.
      (3) Once computed the SEM can be used to construct an interval, wherein we expect an examinee's true score to lie.
      (4) The smaller the SEM is, the narrower the interval. Narrow intervals are more precise at estimating an individual's true score (T).

$$S_E = \delta_x \sqrt{1 - \rho_{xx}}$$

   b. Formula 3.1
      Where: S_E = standard error of measurement
      δ = standard deviation
      ρ_{xx} = test reliability coefficient

   c. Interpreting the Standard Error of Measurement
      (1) The magnitude of the standard error of measurement is inversely related to the reliability coefficient. As "r" increases, S_E decreases.
      (2) Measurement tools with a large S_E tend to be unreliable.

(3) The S_E tends to remain stable across populations and reminds the researcher that any test score (or other score) is nothing more than an estimate which can vary from a subject's "True Score."

(4) Constructing Intervals:

±1s = We are 68% sure or confident that an examinee's true scores falls within one standard error of measurement, plus or minus.

±2s = We are 95% confident that an examinee's true scores falls within two standard errors of measurement.

±3s = We are 99.7% sure that an examinee's true scores falls within three standard errors of measurement.

(5) For example, if Heather's score on a history test is 80 and $S_E = 1.3$ points, then, the intervals would be:

(1) We are 68% sure that Heather's true score lies between 78.7 to 81.3 points.

(2) We are 95% confident that Heather's true score falls between 77.4 to 82.6 points.

(3) We are 99.7% sure that Heather's true score is between 76.1 to 83.9 points.

5. Measuring Error Variance: Standard Error of the Estimate (SE_{yx} or $\delta_{2.1}$)

a. When using a test score to predict a criterion value (e.g., using SAT to predict a college applicant's first semester GPA), the standard error of the estimate indicates how well the test score (SAT) predicts the criterion value (GPA).

(1) When $r = \pm 1.00$ (SE_{yx} equals "0.00"), there is no difference between the predicted and obtained criterion value; they are the same. The 1.0 means prefect prediction; the ± means *r* ranges from -1.0 to +1.0.

(2) High correlations between variables are needed for the prediction to have much accuracy. The standard error of the estimate is associated with criterion-related validity.

b. Formula 3.2

$$\delta_{2.1} = \sqrt{1 - \rho_{xx}^2}$$

Where: $\delta_{2.1}$ = predicted score on test two from test one (Forms must be parallel.)

ρ^2_{xx} = squared reliability estimate

c. The SE_{yx} is used to construct prediction intervals around *r* and is interpreted in the same manner as the standard error of measurement.

C. Threats to and Techniques for Improving Reliability Indices

1. Threats to Reliability

a. Group homogeneity: When a test is given to a very similar (homogeneous) group, the resulting scores are closely clustered and the reliability coefficient will be low. The more heterogeneous the examinee group, the higher the correlation coefficient.

   b. Time limits: The rate at which an examinee works will systematically influence performance, as some will finish the test and some not.
   c. Test or measure length: If a test is too short, then the reliability coefficient will be low.
   d. Scoring errors (e.g., inconsistent scoring) will depress a reliability estimate. Keep scoring simple and consistently applied.
   e. Item Quality: Poorly constructed test items introduce ambiguity into the testing situation, thus affecting examinee performance.
   f. Other threats: These include differences in content on test or measurement forms; administration, examinee, and/or scoring errors; guessing, effects of memory, practice, boredom, etc.

2. To Improve Reliability
   a. Group homogeneity: Ensure that the group is as heterogeneous, respecting the skill or content to be tested as is natural.
   b. Time limits: Ensure there is enough time for almost all examinees to finish.
   c. Test or measure length: Ensure the test is long enough to accurately reflect the content or [assessment] domain. Test length tends to increase the magnitude of the reliability coefficient. Test length is a balance between time and content. It is wisest to select the most important content and test for that. Leave the minor points alone.
   d. Test or measure items: Use only high quality test items which conform to generally accepted editorial guidelines. Prepare examinees by teaching test taking strategies. Well written items will contribute significantly to reliability.
   e. Examinees or respondents: Poorly motivated, fatigued, uncomfortable examinees will perform below their capability. Examinees should be encouraged to get sufficient sleep and food; the testing environment should be comfortable; and everything possible should be done to motivate examinees.
   f. Item difficulty: Target p-values should be between 0.40 and 0.60, generally.

D. Reliability Estimation

1. The type of reliability estimation procedure employed is driven by the intended use of the test score.

2. Considered will be estimation procedures for Equivalence (alternate forms), Stability (test-retest), Stability and Equivalence, single instrument administration (international consistency) and selected inter-rater reliability indices. See Table 3.1a and 3.1b.

3. Reliability indices indicate a test's degree of consistency in assessing examinee performance. It is also an indication of the amount of measurement error present in scores generated by a test. The reliability coefficient is used to quantify a measure's reliability.
   a. It has the same interpretative properties as the Pearson's *r*.
   b. A test publisher (or you) computes *r* and reports a reliability coefficient of 0.90. This tells us that 81% of the observed score variance (i.e., the examinee's test score) is attributable to true score variance (i.e., what the examinee actually knows).
   c. Reliability Standards
      (1) For instruments where groups are concerned: 0.80 or higher is adequate.
      (2) For decisions about individuals: 0.90 is the bare minimum: 0.95 is the desired standard.

Table 3.1a
Reliability Estimation Procedures

Reliability Type	**Estimation Procedure**
Alternate Forms (Equivalence)	Pearson's *r*
Test–retest (Stability)	Pearson's *r*
Test-retest with Alternate Forms (Stability & Equivalence)	Pearson's *r*
Single Administration (Internal Consistency)	Spearman Brown & Prophecy Coefficient Alpha KR-20 or KR-21

Table 3.1b
Inter-Rater (Two-raters) Reliability Estimation Procedures

Type of Date	Estimation Procedure
Dichotomous Data (e.g., Yes or No)	Odds Ratio & Yule's Q Raw Agreement Indices McNemar's Test for Marginal Homogeneity
Nominal Data (e.g., Meets Expectations, Does Not Meet Expectations, or Exceeds Expectations)	Raw Agreement Indices McNemar's Test for Marginal Homogeneity
Ordinal Data (e.g., Low, Medium, High)	McNemar's Test for Marginal Homogeneity
Interval Data (e.g., Likert Scale)	Pearson's *r*

4. <u>Coefficient of Equivalence (Alternate Forms) Estimation</u>
   a. It is common practice to administer two forms (hopefully parallel) to examinees to reduce the chances of cheating. Differences in content within the test items on the two forms are usually the main source of measurement error.
   b. To assess alternate forms reliability:
      (1) Two forms of the same test are constructed. It is common practice to reorder test items on Form B from Form A. Another strategy is to develop duplicate sets of items over the same content and/or skill domain.

(2) Once developed, both forms are administered to the same examinees at the same time. Randomly assign one-half of examinees, Form A and the other half, Form B. Repeat (after a short rest) so that all examinees take both forms of the test.
(3) Using test scores from both forms, compute a Pearson's *r* to serve as the coefficient of equivalence. High coefficients (e.g., 0.80's or 0.90's) indicate that scores from either form can be used interchangeably.

5. <u>Coefficient of Stability (Test–retest) Estimation</u>
   a. In those instances where an examiner needs to administer the same test at a different point in time, the chief source of measurement error is a difference in examinee performance due to examinee changes over time.
   b. To assess the impact of time and other sources of error,
      (1) The examiner administers the instrument to the examinees and waits.
      (2) At the appointed time, the same instrument is administered to the same examinees.
      (3) Using both pairs of scores (omit scores of any examinee who didn't complete both testing sessions), compute a Pearson's *r*. High coefficients (e.g., 0.80's or 0.90's) indicate that scores were relatively unaffected (stable) due to changes in examinees over time.
   c. Factors which influence coefficient of stability magnitude include the amount of time between testing sessions, age (i.e., examinee developmental level), and the "stability" of what (e.g., attribute, trait, etc.) is being measured. Memory and practice effects may adversely affect coefficient magnitude.
   d. Aptitude and achievement tests will generally generate high stability coefficients, whereas personality, interest, or attitude measures tend to produce lower coefficients. However, typical performance tests, which are satisfactorily constructed with well written items and measure a clearly defined, stable trait, should yield high coefficients of stability.

6. <u>Coefficient of Stability and Equivalence Estimation</u>
   a. There are instances when an examiner wants to administer Form A, wait, and then give Form B. If possible, randomly assign one-half of examinees, Form A and the other half, Form B. Wait the required time and then reverse the order of administration. This way all examinees complete both Form A and Form B.
   b. Next, correlate, using Pearson's *r*, testing session 1 scores with those of session 2. The Coefficient of Stability and Equivalence is produced and is interpreted as above.
   c. The Coefficient of Stability and Equivalence is likely to be lower than either the coefficient of stability or coefficient of equivalence since its magnitude will be affected by content discrepancies, the effects of time, and other effects of random and systematic error.

E. Internal Consistency (IC) Reliability Estimation

1. Single Instrument Administration (Internal Consistency)
   a. Most educational and psychological tests are administered once. Here, it is essential to have an estimate as to how an examinee's observed score variance reflects his or her true score variance. In other words, the extent to which the examinees' observed scores are similar to their true scores.
   b. Drawing Conclusions from Internal Consistency Indices
      (1) While an examiner is interested in an examinee's observed score, he or she is at least equally interested in how an observed score will generalize to the domain or cluster of all possible items that could have been included on the examination or measure.
      (2) We base our conclusion on how well examinees performed or scored across items on the whole test or its subtests.
   c. Factors affecting Internal Consistency Reliability Estimation
      (1) The lead source of error is in content sampling, i.e., items drawn from different knowledge or skill domains (e.g., English, history, science, math, etc.) are included on the same test, or significant content and skills are overlooked in favor of minor content and/or skill sets.
      (2) Do not use IC estimation procedures on a test which combines different content (e.g., spelling, math, reading comprehension, etc.). Differing content affects reliability measurement. If examinees perform consistently across test items, then the test has item homogeneity, i.e., the items measure the same knowledge or skills. Item homogeneity is critical to internal consistency reliability estimation.
      (3) Item quality will materially impact error. Items should be constructed based on appropriate guidelines. IC indices should not be used with speeded tests as the reliability estimate is inflated. Speeded tests are those tests where the time in which the test is completed affects the final score (e.g., a timed, typing test).
   d. Internal consistency (IC) is an index of both test item homogeneity and item quality.
      (1) Internal consistency for dichotomously scored items, i.e., correct or incorrect, can be estimated:
         (a) By first computing a Pearson's *r* and then applying the Spearman-Brown Prophecy Formula to estimate reliability for the whole test.
         (b) An alternative is the KR-20 or KR-21 (Kuder & Richardson, 1937). A test (if unidimensional) is divided into halves, and then these halves are correlated. If an examination is composed of subtests, as in a test battery, then each subtest is halved.
         (c) Cronbach's alpha can also be used for dichotomously scored items.
      (2) For polytomously scored items, i.e., items where differing points can be earned as with short answer or essay items, or item response formats on attitude, interest, or personality instruments Cronbach's alpha is used.

2. Dichotomously Scored Items: The Spearman Brown Prophecy Formula
   a. When a test is halved, the resulting IC coefficient is going to be underestimated in magnitude. The Spearman correction is applied to yield an estimated IC coefficient for the entire test.
   b. The Spearman Brown Prophecy coefficient is computed by Formula 3.3

$$\rho_{xx} = \frac{2\rho_{AB}}{1 + \rho_{AB}}$$

where
ρ_{AB} = Pearson correlation between both halves of the test
ρ_{xx} = Projected reliability coefficient for the entire test

   c. For convenience's sake, let's assume that ρ_{xx} = .71. Therefore

$$\rho_{xx} = \frac{2\,(.71)}{1 + .71} = \frac{1.42}{1.71} = .83$$

The estimated total test IC coefficient is .83. Remember, we used two equal halves of a 20 item test. This estimate is based on the assumption that each half of the test is parallel, i.e., no error variance is present.

3. Dichotomously Scored Items: KR-20 & KR-21
   a. One of the problems with split-half methods of which Spearman-Brown is but one is their failure to produce a unique reliability coefficient estimate. The problem is similar to following the "floating IC coefficient" which changes with each manner in which a test is halved. The KR-20 and KR-21 solve this problem.
   b. The KR-20 is very labor intensive to compute as the p-value for each item is required. The use of computer programs renders this a less significant issue. Of the two formulae, KR-21 yields a slightly higher, IC coefficient than KR-20, as the KR-21 formula assumes that all test item p-values are identical. The KR-21 is much easier to compute and hence more widely used.
   c. KR-21 coefficient is computed by Formula 3.4

$$KR_{21} = \frac{k}{k-1} \bullet 1 - \frac{\bar{x}\,(k-\bar{x})}{(k)\,(s^2)}$$

where: k = number of test items
s^2 = total test variance
$\bar{x}$ = total test score mean

   d. A recent 20 item test has yielded the following summary statistics $\bar{x}$ = 16 and s^2 = 4.1.

$$KR_{21} = \frac{k}{k-1} \bullet 1 - \frac{\bar{x}\,(k-\bar{x})}{(k)\,(s^2)} = \frac{20}{19} \bullet \frac{16(20-16)}{(20)\,(4.1)}$$

$$(1.0526) \bullet (1-- .7804) = (1.0526)\,(.2196) = .23$$

We can say that at least 23% of total score variance is due to true score variance. We can say the same thing about Cronbach's alpha as both methods are based on item covariance. This is not good.

e. The KR-21 is a lower bound estimate of the "true" IC coefficient also known as the coefficient of precision. The KR's will usually yield a slightly lower estimate of the coefficient of precision than the split-half approaches, but will produce a slightly higher estimate than Pearson's *r*.

4. <u>Internal Consistency Estimation for Polytomously Scored Items</u>
   a. In applying Cronbach's alpha, we are assuming that each item is actually a subtest with its own variance producing a composite estimate. If all items were strictly parallel (no error variance), then alpha is a perfect estimate of reliability. But this is not the case. At best, alpha is a lower band estimate of the "true" reliability coefficient.
   b Coefficient alpha (α) is computed by Formula 3.5

$$\alpha = \frac{k}{k-1} \bullet 1 - \frac{\sum \sigma_i^2}{\sigma_x^2}$$

where: k = number of test items
σ_i^2 = variance of item *i*
σ_x^2 = total test variance

   c. Since Cronbach's alpha is very labor intensive to compute by hand we will assume the following values: $\sum\sigma_i^2 = 80$; $\sigma_x^2 = 120$; k = 9. Therefore

$$\alpha = \frac{9}{9-1} \bullet 1 - \frac{80}{120} = \frac{9}{8} \bullet 1 - .667 = (1.125)(.333)$$

$$\alpha = 0.375 \text{ (a low IC coefficient)}$$

Since alpha = .375, we can say that at least 38% of total score variance is due to true score variance.

   d. Computing Item Variance
      (1) The formula for computing item variances for dichotomously (right or wrong) scored items is $\sigma^2 = p(1-p)$. *P* is the proportion of examinees or respondents correctly endorsing the item (i.e., p-value).
      (2) To compute σ_i^2 (variance of item *i*), for polytomously scored items, such as essays or some attitude inventories, first compute the standard

deviation for each item and then square the standard deviation to produce the item's variance.

(3) To compute $\sum\sigma_i^2$, simply sum the individual item variances.

II. Validity Indices

A. Introduction

1. Essential properties of any measure are that the measure (e.g., instrument, test, scale, index, etc.) be both valid and reliable. An instrument can be reliable, but not valid; but, validity is really a matter of degree and not "all or nothing." The process of validation is a combination of logical argument and empirical validation. There is content, criterion-related, and construct validity.

2. Strictly speaking one validates the use to which a measure (e.g., test, scored direct performance, scale, index, etc.) is used, rather than the instrument itself.
   a. Each unique use must be specifically documented, i.e., validated.
   b. Tests or measures valid for one use aren't necessarily valid for another.
   c. Measures are often validated after use. However, one should conduct at least one appropriate validation study during a measure's developmental phase.

3. Face validity is not a measurement concept. It is the degree to which an instrument <u>appears</u> to measure what it says it does. It is a "political" concept.
   a. Content validity is ensured by the process through which the measure is constructed. Face validity is a conclusion drawn after the measure is constructed (Nunnally & Bernstein, 1994, pp. 108-110).
   b. An instrument designed to measure word-processing performance may be challenged if it contained computer programming items.
   c. Courts typically confuse face validity with content validity.
   d. For most testing applications, an instrument should have face validity.

B. Content Validity

1. Rationale and Application
   a. A content valid test should have at least moderate to high levels of internal consistency. This suggests that the items measure a common element.
   b. Content validity primarily rests upon logical argument and expert judgment, and frequently empirical research. The degree of content validity is largely a function of the extent to which test items are a complete and representative sample of the content and skills to be learned.
   c. Representativeness is critical as the examiner will generalize from the sample of items on the test to the degree of content mastery possessed by an individual examinee. This inference is called "generalizability."
      (1) To permit maximum generalizability of results, a representative sampling of content and incorporation of the most appropriate testing methods are necessary.

(2) The key to establishing generalizability is an adequate representation of items over content, and a testing method which produces results similar to possible alternative approaches. The testing method should be as close to "real world" as possible.

d. Its most common management and education applications are:
   (1) Examinations, standardized tests, and/or locally developed measures or tests that require content validity. Examples are:
      (a) End of course examinations as in finals or qualifying examinations to be admitted into a course of study.
      (b) Standardized tests such as the GRE, TOFEL, ACT, SAT.
      (c) Locally developed tests such as a midterm or unit test.
   (2) Training, organizational development, and HR functions.
      (a) Applicant screening, proficiency, certification, and/or licensure tests.
      (b) Organization climate, employee attitude, management aptitude or style, communication style, and/or leadership style inventories (also called surveys, scales, indexes, etc.)

e. Other names for content validity are: intrinsic validity, curricular validity, representativeness, and relevance.

2. Empirical Evidence for Content Validity
   a. Learning should cause posttest scores to rise, if examinees were pretested.
   b. A correlation between scores of two different measures which measure the same content or skill, e.g., reading readiness, can be used as evidence of the content validity of one of the measures provided the other is widely seen as credible.
   c. One would expect a positive correlation between final examination scores of students, each of whom took sections of the same course, all other things being equal:
      (1) If the correlation between test scores was zero then we would assume no association existed. We would further expect that there might be something amiss about the students, course, instruction, etc.
      (2) If there was a positive correlation between the test scores, then we would have evidence of content validity, coupled with the prior logical argument that the test items adequately sampled the domain of all possible test items related to the content, and the professional judgment of content experts that the test was suitable.

3. A Strategy for Ensuring Content Validity
   a. A well-formulated plan and procedure for test construction must be laid out before actually constructing the test. Arguments for content validity are strengthened when most experts would agree that the plan was well devised, items were representative of the domain in content and number, and the plan was well executed.

b. Potential users should agree that the procedure ensures a reasonably representative collection of items from the content and/or skill domain.
c. Agree, a priori, upon which test item formats are most appropriate.
   (1) Item format should be consistent with the test purposes and knowledge or skill to be measured.
   (2) Items should conform to established editorial guidelines.
   (3) There should be at least a modest level of internal consistency among the items.
d. Remember that item selection involves values as to what is important, etc. Accordingly, such values should be made explicit.
e. The number of items per content domain sub-strata and item format should be summarized in a Test Blue Print or Table of Specifications.
f. An item analysis (Chapter 5) is essential to improving the content validity of a test. However, an item analysis (IA) can only be done after the administration of a test or other similar measure. An IA provides evidence as to how examinees responded to each test item. Common IA indices used in content validity studies include:
   (1) Internal Consistency reliability
   (2) p-values (proportion of correct endorsements)
   (3) Index of Discrimination (identifies most discriminating items)

4. The same procedures that are required to ensure content validity are intimately related to ensuring criterion-related and construct validity. It is also vital that a predictor instrument, such as a placement or university admissions test be content valid over the ability or skills upon which a prediction in based.

C. Criterion-related Validity

1. Examinee or employee performance on one measure is correlated with performance on another measure, known to be both valid and reliable, i.e., a criterion. Performance on the first measure should be highly correlated with performance on the second (i.e., known to be valid and reliable) measure. Criterion-related validity is established empirically by a correlation coefficient which is called a validity coefficient.
2. Types of Criterion Related Validity
   a. Predictive: Performance is predicted based on one or more known measured variables. A common example is the MAT, GRE or GMAT. Student test scores are "plugged" into a formula and the first semester grade point average (GPA) is predicted. If the predicted GPA is at or higher than the minimum admission GPA, the applicant has a better chance of admission.
   b. Concurrent: Performance on one measure is measured against performance on another measure, known to be valid, reliable, and considered a "standard." An example is comparing a short aptitude test to a much longer version. If scores on the short version correlate highly with

scores on the longer version, then the shorter version could be used. Time and money are saved.

c. Concurrent validity is often confused with a construct validation strategy.

d. Measures used for predictive purposes require either content and/or construct validity; such are essential if the measure is used for personnel selection or entry into special programs. Content and/or construct validity provide preliminary evidence of criterion-related validity especially if the predictor test must be selected before its criterion-related validity has been established. They also provide criterion-related validity evidence, aside from the validity coefficient.

2. The Criterion Related Validity Validation Process
   a. Before using a score to make a decision about a person (e.g., admission, hiring, etc.), the user needs to have evidence of a relationship between the test score and criterion performance. Steps in a criterion-related validation study are:
      (a) Select an appropriate criterion (e.g., behavior, indicator, attitude, etc.) and measurement method, such as a well established instrument or commonly accepted procedure as in performance measurement.
      (b) Obtain a representative sample from the same examinee population for whom the test is intended.
      (c) Administer the test and record each examinee's score'
      (d) Once criterion data are available, collect a measurement from each examinee.
      (e) Compute the validity coefficient to determine the correlation's strength between the test scores and criterion.
   b. The empirical evidence is usually a validity coefficient which will vary in type (point-biseral or Pearson) given the nature of the data (e.g., ordinal or interval).

D. Construct Validity

1. One can view a theory as a statement of the hypothesized relationship between and/or among constructs which attempt to explain or describe a behavior (e.g., leadership), disposition (conservative vs. liberal), phenomena (e.g., intelligence or personality), etc.
   a. A construct is not directly observable or measurable; it is the variables which compose the construct that are actually measured.
   b. A construct is a statement about the relationship between and among the variables which comprise it. A construct is more general than a specific variable.
   c. Every construct has a "domain of observables", measurable variables. A complex construct will require a larger domain of observables which will in turn require more variables to define the construct. Less complex constructs have smaller domains of observables and are easier to define.

d. Constructs, while abstract, are helpful in understanding behavior, motivation, communication styles, etc. They provide a frame of reference for predicting what might happen; for why something has happened; for intervention; and for personal and professional growth; and in medicine, for treatment.
e. Nunnally and Bernstein (1994, p. 85) write that "science's two major concerns are (1) developing measures of individual constructs and (2) finding functional relations between measures of different constructs."

2. A theory may be composed of one construct (also called unidimensional) or several constructs (called multi-dimensional).
   a. A Unidimensional theory (single construct) will typically have one full, usually brief measure devoted to it.
   b. A multidimensional theory (two or more constructs) will most likely have one sub-test devoted to each construct. Thus, the measure of a theory, based on five constructs will likely have five sub-tests.
   c. Typical Applications of Theory to Management and Education
      (1) Management, leadership, and learning theories
      (2) Organizational development (e.g., communication, change, group behavior, job satisfaction, and culture) theories.
      (3) Individual behavior (e.g., motivation, attitudes, etc.) theories.
      (4) Marketing (e.g., consumer behavior, product attractiveness, consumer satisfaction, etc.) theories.
      (5) Intelligence testing (IQ).

3. Establishing Construct Validity
   a. Introduction
      (1) A construct is "validated" by logical argument and empirical methods, usually a validity coefficient and/or factor analysis. Nunnally and Bernstein (1994, pp. 86-87) described three aspects of the construct validation process:
         (a) The researcher must specify indicators or characteristics of the construct.
         (b) Through empirical research and statistical analysis, the researcher assesses the extent to which the indicators or observables measure the construct's characteristics.
         (c) The researcher or researchers repeat studies (difference or empirical) to either confirm or disconfirm earlier findings so as to evaluate the degree of fit between measures of the construct and that which is predicted by the construct.
      (2) The process outlined by Nunnally and Bernstein is complex, involves significant resources, takes a substantial period of time, and "lots" of studies.

(3) To make competent assertions that construct validity is evident, a researcher must show (1) the relationship, attribute, phenomena, or behavior exits as the theory posited and (2) that the instrument measures it.

b. The Four Methods to Establish Construct Validity
   (1) Method 1: Correlation between a Measure of the Construct and Designated Variables
      (a) Within this method, two measures are administered. One is a measure of the construct and the other a specified variable (e.g., measure of job satisfaction or school performance and salary or family income, etc.)
      (b) If the resulting correlation coefficient “runs” as expected, then evidence of construct validity may be argued.
      (c) There is no generally accepted standard as to what defines an acceptable validity coefficient, but the higher the better.
   (2) Method 2: Difference Studies
      (a) One measure is administered to two different groups. If the resulting validity coefficient “runs” as expected, then an argument for construct validity can be proffered.
      (b) For example, administrative personnel are divided into two groups. One is given intensive supervisory training and the other group is not. After the training is completed, all administrative personnel are given a measure of supervisory performance. One would expect that those receiving the training would score as more effective supervisors than those not receiving the training. If these expected results (a low validity coefficient) were found, an argument for the construct validity of the theory upon which the training was based could be made. If not, then either the theory is inaccurate, the test is defective, or the training was ineffective.
   (3) Method 3: Factor Analysis
      (a) Factor analysis is an advanced correlational statistical procedure which is used to identify unobserved or latent variables called factors which are predicted by a theory.
      (b) Factor analysis is most commonly used in the development of measuring devices in which the goal of the researchers is either to confirm (confirmatory factor analysis) or identify (exploratory factor analysis) factors included within a measure which is said to operationally define a theory (Nunnally & Berstein , 1994, p. 220).
      (c) Factor analysis seeks to identify the interrelationships between a set of variables with the goal of uncovering any underlying relationship patterns or structures between or among the variables.
         [1] A structure or structures within a set of variables are factors. Factors are sets of interrelated (i.e., correlated) variables in the form of clusters. There may be either one or more clusters.

[2] The interpretation of the factors tends to be subjective. This is why the researcher needs to fully understand his or her theory and related empirical research and conceptual literature.

(d) If the measure possesses all of the factors and variables cluster as the underlying theory posits, then there is evidence of the theory's construct validity. Of course, more than one research project is needed to confirm.

(4) <u>Method 4: Convergent and Divergent Validity Strategies</u>

(a) <u>Convergent Validity</u>: A standard measure (of known validity and reliability) is administered to the same subjects as the measure being validated. The two measures are intended to measure the same construct, but are <u>not</u> administered in the same fashion.

(1) Suppose a researcher wants to validate an observational measure of math anxiety. She could first administer a written measure of math anxiety which is well established. Next, she would observe small groups of students talking about and working out math problems, while rating, observationally, math anxiety indicators.

(2) If the scores on the two measures are highly positively correlated, then the case for construct validity is strengthened. However, this strategy is often confused with concurrent validity. The key difference between the two is the purpose for which the comparison is made and how the study is conducted.

(b) <u>Divergent Validity</u>: Nunnally and Bernstein (1994, pp. 93) write, "in order to justify novel measures of attributes [constructs] a measure should have divergent validity in the sense of measuring something different from existing methods."

(1) Measures of different constructs should not highly correlate.

(2) Minimally, two attributes, each measured by at least two different methods are required for a divergent validity study.

III. Direct Performance Assessment Reliability and Validity

A. Performance Assessment Reliability

1. Reliability is vital to ensure consistency in decision-making across examinees. An observed score is the composite (i.e., combination) of an examinee's true score (what he or she actually knows and can do) and an error score. The error term is the sum of random and systematic errors, which are unrelated to the true score. We never really know the source, amount, or effect of error, but we do want to keep it small.

2. Factors which Adversely Affect Performance Assessment Reliability
   a. Vague and/or confusing performance assessment purpose, specifications, classification levels, and scoring criteria.
   b. Performance inaccuracies due to examinee unfamiliarity with the assessment's purpose, expectations, etc.

c. Inappropriate examinee performance specifications, e.g., criteria which require significant physical coordination of 4 and 5 year olds.
d. Attempts to assess an unmanageable number of performance indicators, thereby introducing random error.

3. Haladyna (1997, pp. 139-141) identified several bias errors which adversely affect reliability. He defines bias as "directional error, whereby a subset of performers may be consistently overrated or underrated." Bias sources are:
   a. Response set occurs when a rater marks at the same place on the rating scale regardless of examinee performance, e.g. always circling 5 on a 6 point scale.
   b. Leniency error is the tendency to give high range (positive) ratings despite differences among examinee performance, e.g., one judge overrates performance as compared to other raters.
   c. Severity error is the tendency to give low range (negative) ratings regardless of examinee performance differences, e.g., one judge underrates performance as compared to other raters.
   d. Central tendency error is the tendency to give middle range ratings regardless of examinee performance differences, e.g. consistently circling 3 on a five point Likert Scale.
   e. Proximity error happens when differing skills are rated similarly when sequentially ordered as in a process, e.g., rating a sequentially ordered simple and a complex skill equally because the simple skill was rated first.
   f. Halo error occurs when a performance rating is influenced by unrelated impressions, e.g., rating a well mannered examinee higher than one who is less well mannered.
   g. Logical error is interpreting a poorly worded skill specification in an unintended manner, e.g., substituting your definition of a term because you don't understand the term's definition as written in the rubric.
   h. Lack of interest error results when a rater is really not interested in the process, skill, or product being rated or believes the rating exercise is not valued, e.g., unmotivated rater who is tired, just giving whatever rating he or she wants.
   i. Idiosyncratic error results when unexpected and unpredictable ratings are given for a number of reasons, e.g., ratings which are inconsistent.

4. Guidelines for Improving Performance Assessment Reliability
   a. Use structured checklists, rating scales, or scoring rubrics to record and score examinee performance.
   b. Select raters who are subject matter experts who are also both motivated and willing. Compensate raters, if possible.
   b. Conduct very effective rater training, the objective of which is consistent student behavior or response classification.
   c. Validate the raters. Study rating patterns for consistency and compliance with performance level definition.

d. Again, train and debrief until consensus on understanding of what is to be rated is reached and rating patterns are consistent.

5. The generalizability of performance tests can be a problem. Popham (2000) recommends that when a performance test is constructed, the author clearly specifies what indicators are needed to make a judgment regarding examinee mastery. Then, the task or exercise is constructed so as to produce assessment data, suitable for decision-making, which enables inferences to be drawn. However, it is unwise to generalize to other examinees or groups of examinees based on individual or a single group performance.

B Direct Performance Assessment Validity

1. Performance assessments, like traditional tests, must be at least content valid. The establishment of other forms of validity is dependent on the intended use of the performance assessment score.
2. If an examinee's current performance is to be used to predict future performance, then criterion related predictive validity must be established.

3. If a complex disposition is to be assessed, via performance assessment, then construct validity must be established.

4. Principal threats to performance assessment validity (i.e., producing valid information for decision-making) are
   a. The failure to instruct and inform examinees regarding the specific processes, skills, outcomes, affective dispositions, or social skills to be assessed.
   b. The failure to inform examinees about the purpose of the assessment, performance expectations, scoring procedures, etc.
   c. The rater biases identified by Haladyna (1997, pp. 139-141) which are presented below.
   d. The influence of unrelated information or impressions (e.g., superior writing or language skills, attractive physical appearance, personality characteristics, etc.) on scoring which thereby inflates or deflates examinee scores, which likely leads to poor decision-making.
   e. Procedures, performance definitions, scoring criteria, or performance settings that give advantage to one examinee or examinee group over others due to examinee culture, language, economic status, age, or gender.
   f. Memory error will affect scores unless the performance is rated as it occurs.
5. Thus strategies which minimize or eliminate validity threats will lead to the production of valid data for decision-making.

Review Questions

Directions. Read each item carefully; either fill-in-the-blank or circle letter associated with the term that best answers the item.

1. A group of employees took skills Test A on Monday and again on Wednesday. Assuming no practice or other instruction had occurred, and given that the test had high reliability (r = .90), then differences between the two test scores for each employee should be, "on average."
   a. Quite small c. Negative
   b. Medium d. Less than 0.90

2. Reliability of measurement is concerned with:
   a. Systematic error c. Consistency
   b. Relevance d. Proven value

3. The longer the time interval between a prediction and the event to be predicted, the:
   a. Smaller are the chances of making a good prediction
   b. Greater are the chances of making a successful prediction
   c. Fewer are the influencing variables to be considered
   d. Smaller is the standard error

4. A grammar test is given twice within a few days to a group of word-processors. The highest possible score is 40. One examinee received a score of 35 items correct the first time, and 18, the second. If these results are similar for other students in the class, then the test is most likely:
   a. Lacking in relevance c. Too difficult
   b. Insufficiently comprehensive d. Unreliable

5. This reliability coefficient is usually greater over the short-term (time) than the long-term:
   a. Test-retest c. Split-halves
   b. Alternate Forms d. Internal Consistency

6. Which of the following is **not** a method of building reliability into a test?
   a. Adding items of good quality
   b. Administering the test to a heterogeneous group
   c. Comparing the test with existing measures
   d. Controlling the conditions of test administration

7. A human resource professional wants to compute the reliability of a test she has constructed after a single administration. What type of reliability is applicable?
   a. Test-retest c. Internal consistency
   b. Inter-rater d. Parallel forms

8. If a thermometer measured the temperature in an oven as 400º F five days in a row when the temperature was actually 397º F, this measuring instrument would be considered:
   a. Reliable but not valid
   b. Valid, but not reliable
   c. Reliable and valid
   d. Unreliable and invalid

9. When reliability is determined by correlating scores on WJ-R Tests of Cognitive Ability (an achievement test, like the GRE or GMAT) Form L and Form M, the type of reliability we are assessing is:
   a. Test-retest
   b. Parallel Forms
   c. Stability
   d. Internal Consistency

10. What is the lowest reliability coefficient that is considered acceptable for a group?
    a. 0.50
    b. 0.60
    c. 0.80
    d. 0.90

11. Which of the following is most likely a threat to reliability?
    a. Including many items in one test
    b. Measuring the wrong skill
    c. Two supervisors rating the same employee's performance differently
    d. Using a formal assessment when an informal once might have been used

12-13. Reliability and Validity Relationship
   a. Concurrent
   b. Construct
   c. Content
   d. Test-retest
   e. Internal Consistency
   f. Alternate Forms

At a minimum all tests must have (12) ______________ validity and (13) ____________ reliability.

14. Errors of measurement are caused by:
    a. Changes in test items
    b. Changes in the examinees
    c. Changes in instrument administration
    d. All of the above

15. The standard error of measurement is closely related to:
    a. Central tendency
    b. Difficulty
    c. Objectivity
    d. Reliability

16. The standard error of measurement is useful for
    a. Reporting an individual's scores within a band of the score range
    b. Converting individual raw scores to percentile ranks
    c. Reporting a groups' average score
    d. Comparing group differences

17. If we knew the exact error of measurement in an individual's observed score and we subtracted this measure from his or her observed score, we would get his or her
    a. Derived score
    b. Raw score
    c. Standard score
    d. True score

18. Suppose an employee scored 87 (also the test's mean) on a test where SEM = 2 points. If SEM were carried to ± 3 SEM, the point range would be:
    a. 85-89
    b. 83-91
    c. 81-95
    d. 81-93

19. Measurement error that is consistent across uses of a measure is said to be ____?
    a. Random error
    b. Classic error
    c. Systematic error
    d. Standard error

20. Which <u>one</u> of the following statements is <u>inaccurate</u>?
    a. The size of the standard error is positively related to the reliability coefficient.
    b. The standard error tends to remain stable across populations.
    c. The standard error is used to construct individual score bands.
    d. According to the Classical True Score Model, when O = T, the standard error is zero.

21. Which of these types of validity is based almost entirely upon rational analysis and logical argument?
    a. Content
    b. Predictive
    c. Concurrent
    d. Construct

22. A test of job performance is positively related to past supervisor ratings, job attendance, and positive attitudes for similar employees. This information pertains most directly to:
    a. Construct validity
    b. Content validity
    c. Concurrent validity
    d. Predictive validity

23. A group of employees takes the Wilson self-concept inventory and the Harry self-concept inventory which are known to measure the same "thing." If the self-concept scores from the Wilson and the Harry inventories are unrelated, which kind of validity appears to be lacking?
    a. Construct
    b. Face
    c. Content
    d. Stability

24. Suppose that employees who do well in training courses (as measured by test score) tend to become better employees (as rated by their supervisors). What kind of validity is most important here?
    a. Content validity
    b. Construct validity
    c. Criterion-related validity
    d. Face validity

25. An indispensable quality for a good measuring instrument (e.g., a test, survey, index, inventory, etc.) is that it produces evidence that it is:
    a. Standardized
    b. Valid
    c. Objective
    d. Criterion-referenced

26. If someone says, "These test scores are valid," what would be the best response?
    a. But, how valid are they?
    b. How high is the validity?
    c. What are they valid for?
    d. What is the validity coefficient?

27. Face validity is most likely to contribute to better:
    a. Interpretation
    b. Employee Motivation
    c. Reliability
    d. Sampling

28. If we want to substitute a short test for a longer version of the same test, we are primarily concerned with:
    a. Concurrent validity
    b. Predictive validity
    c. Content validity
    d. Construct validity

29. A construct refers to behavior that is
    a. Not observable and directly measurable
    b. Not observable and indirectly measurable
    c. Observable and directly measurable
    d. Observable and indirectly measurable

30. Which type of validity evidence is most likely to include most if not all of the other types of validity?
    a. Construct validity
    b. Content validity
    c. Concurrent validity
    d. Predictive validity

Items 31 to 36 are related and are application items.
Match the type of validity most directly suggested by the following situations. Note that not all of the types of validity are included.

a. Content validity
b. Concurrent validity
c. Construct validity
d. Predictive validity

31. Four experts agree that a test measures the skills needed by entry-level staff assistants.
32. An external standardized test of job competence correlates positively with a supervisor's rating.
33. A test is constructed to measure leadership style. Scores from this test correlate positively with another recognized leadership inventory (i.e., test).
34. A job counselor looks over the results of John's test and informs him of his chances of success in the company's management training program.
35. Mrs. Smith was careful to construct her test items so that it measured each learning objective she developed for staff training class.
36. Build a test covering the job competencies using agreed upon content and skills.

Answers: 1. a, 2. c, 3. a, 4. d , 5. a, 6. c, 7. c, 8. a, 9. b, 10. c, 11. c, 12. c, 13. e, 14. d, 15. d, 16. a; 17. d, 18. d, 19. c, 20. c, 21. a, 22. d, 23. a, 24. c , 25. b, 26. c, 27. b, 28. a, 29. b, 30. a, 31. a, 32. b, 33. c, 34. d, 35. a, and 36. a.

References

Standards for educational and psychological testing. (2014). Retrieved from http://www.aera.net.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, & Winston.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Needham Heights, MA: Allyn & Bacon.

Kuder, G. F. & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2*, 151-160.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Popham, W. J. (2000). *Modern educational measurement* (3rd ed.). Boston, MA: Allyn & Bacon.

Appendix 3.1
Group Contribution Index

Read each statement carefully. Next, circle the number that best represents your degree of agreement or disagreement with the statement for the individual whose group contribution you are measuring, using the following options:

1 = Strongly Disagree	3 = No Opinion	4 = Agree
2 = Disagree		5 = Strongly Agree

When you have completed rating each statement, total the individual ratings and record it the blank provided.

1.	The group member's participation was focused on the task at hand.	1	2	3	4	5
2.	The group member usually exhibited a respectful demeanor.	1	2	3	4	5
3.	The group member contributed an acceptable quantity of data, e.g., research articles, URLs, books, etc., given the team's task.	1	2	3	4	5
4.	The quality of the group member's data (e.g., research articles, URLs, books, etc.) contribution was high, given the task.	1	2	3	4	5
5.	The group member's contribution of data (e.g., research articles, URLs, books, etc.) was relevant to the team's task.	1	2	3	4	5
6.	The group member acceptably met the team's deadlines.	1	2	3	4	5
7.	When required, the member exhibited appropriate mediating skills.	1	2	3	4	5
8.	The member followed team directions in an acceptable manner.	1	2	3	4	5
9.	The group member exhibited appropriate listening skills which assisted the team in accomplishing its task.	1	2	3	4	5
10.	The team member was sufficiently flexible so as to enable the work group to complete the task at hand.	1	2	3	4	5
11.	The team member demonstrated writing skills, which helped the work group meet its objective.	1	2	3	4	5
12.	By providing constructive feedback to team mates, the member contributed towards accomplishing the team's task.	1	2	3	4	5

Scoring Total:________

54-60: Very Significant Contribution
48-53: Significant Contribution
42-47: Average Contribution
36-41: Poor Contribution
$\leq$ 35: No Meaningful or Useful Contribution

Performance Category Definitions

Very Significant Contribution: The group member contributed towards the common purpose or project in such an outstanding or extraordinary manner so that the contribution was exceptional in terms of its meaning, usefulness, and timeliness.

Significant Contribution: The group member contributed towards the common purpose or project in an above average (i.e., more than would normally or reasonably be expected) manner so that the contribution was very meaningful, useful, and timely.

Average Contribution: The group member contributed towards the common purpose or project in a typical (i.e., as one would normally or reasonably be expected) manner so that the contribution was meaningful, useful, and timely but not atypical.

Poor Contribution: The group member made a modestly meaningful, useful, and timely contribution towards the common purpose or project but the contribution was less than what would normally or reasonably have been expected.

No Contribution: The group member made no meaningful, useful, or timely contribution towards the common purpose or project.