Reliability

It is the user who must take responsibility for determining whether or not scores are sufficiently trustworthy to justify anticipated uses and interpretations.

(AERA et al., 1999)

Reliability

Refers to the consistency or stability of scores

If a test taker was administered the test at a different time, would she receive the same score?

If the test contained a different selection of items, would the test taker get the same score?

Classical Test Theory

The theory of reliability can be demonstrated with mathematical proofs

\[ X = T + E \]

\( X \) = Obtained or observed score (fallible)
\( T \) = True score (reflects stable characteristics)
\( E \) = Measurement error (reflects random error)
Random Measurement Error

All measurement is susceptible to error

Any factor that introduces error into the measurement process effects reliability

To increase our confidence in scores, we try to detect, understand, and minimize random measurement error

Sources of RME

Many theorists believe the major source of random measurement error is “Content Sampling”

Since tests represent performance on only a sample of behavior, the adequacy of that sample is very important

Sources of RME

The error that results from differences between the sample of items (the test) and the domain of items (all items) is the content sampling error

If the sample is representative and of sufficient size, then error due to content sampling will be relatively small
Sources of RME

Can also be the result “temporal instability,” or random and transient:

Situation-Centered influences (e.g., lighting & noise)
Person-Centered influences (e.g., fatigue, illness)

Other Sources of Error

Administration errors (e.g., incorrect instructions, inaccurate timing)
Scoring errors (e.g., subjective scoring, clerical errors)

All possible sources of error contribute to the lack of precise measurement

Classical Test Theory

\[ X = T + E \]

\( X \) = Obtained or observed score (fallible)
\( T \) = True score (reflects stable characteristics)
\( E \) = Error score (reflects random error)

Using group data, extended to:

\[ \sigma^2_X = \sigma^2_T + \sigma^2_E \]
\[ \sigma^2_X = \sigma^2_T + \sigma^2_E \]

\( \sigma^2_X \) = Observed score variance

\( \sigma^2_T \) = True score variance

\( \sigma^2_E \) = Error score variance

---

**Partitioning the Variance**

Percentage of observed score variance due to true score differences:

\[ \frac{\sigma^2_T}{\sigma^2_X} \]

Percentage of observed score variance due to random error:

\[ \frac{\sigma^2_E}{\sigma^2_X} \]

---

**Reliability**

\[ = \frac{\sigma^2_T}{\sigma^2_X} \]

So, how do we estimate reliability; how do we partition the variance?
The Reliability Coefficient ($r_{xx}$)

Percentage of observed score variance due to true score differences:

$$r_{xx} = \frac{\sigma^2_T}{\sigma^2_X}$$

Percentage of observed score variance due to random error:

$$1 - r_{xx} = \frac{\sigma^2_E}{\sigma^2_X}$$

Types of Reliability Coefficients

Test-Retest Reliability

Reflects the temporal stability of a measure

Most applicable with tests administered more than once and/or with constructs that are viewed as stable

It is important to consider the length of the interval between the two test administrations
Test-Retest Reliability

Subject to “carry-over effects”

Appropriate for tests that are not appreciably impacted by carry-over effects

Alternate Form Reliability

Involves the administration of two “parallel” forms; can be done in two ways:

*Delayed Administration* reflects error due to temporal stability and content sampling

*Simultaneous Administration* reflects only error due to content sampling

Alternate Form Reliability

Limitations include

- Reduces, but may not eliminate carry-over effects
- Relatively few tests have alternate forms

Some tests with alternate forms: PPVT, EVT, CVLT
Internal Consistency

Estimates of reliability that are based on the relationship between items within a test and are derived from a single administration of a test

Split-Half Reliability

Divide the test into two equivalent halves, usually an odd/even split

Reflects error due to content sampling

Since this really only provides an estimate of the reliability of a half-test, use the Spearman-Brown Formula to estimate the reliability of the complete test

<table>
<thead>
<tr>
<th>Half-Test Correlation</th>
<th>Spearman-Brown Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>.50</td>
<td>.67</td>
</tr>
<tr>
<td>.55</td>
<td>.71</td>
</tr>
<tr>
<td>.60</td>
<td>.75</td>
</tr>
<tr>
<td>.65</td>
<td>.79</td>
</tr>
<tr>
<td>.70</td>
<td>.82</td>
</tr>
<tr>
<td>.75</td>
<td>.86</td>
</tr>
<tr>
<td>.80</td>
<td>.89</td>
</tr>
</tbody>
</table>
Coefficient Alpha & KR 20
Reflects error due to content sampling
Also sensitive to the heterogeneity of the test content (or item homogeneity)
Mathematically, it’s the average of all possible split-half coefficients
Since coefficient alpha is a general formula, it is more popular

Reliability of Speed Tests
For speed tests, reliability estimates derived from a single administration of a test are inappropriate
Test-retest and alternate-form reliability are appropriate, but split-half, Coefficient Alpha and KR 20 should be avoided

Inter-Rater Reliability
Reflects differences due to the individuals scoring the test
Important when scoring requires subjective judgement by the scorer
It is possible to partition the error variance into its components:

![Error Variance Components Diagram]

- **Test-Retest Reliability**: Time Sampling
- **Alternate-Form Reliability**
  - Simultaneous Administration: Content Sampling
  - Delayed Administration: Time Sampling & Content Sampling
- **Split-Half Reliability**: Content Sampling
- **Coefficient Alpha & KR-20**: Content Sampling & Item Heterogeneity
- **Inter-Rater Reliability**: Differences Due to Raters/Scorers
Reliability of Difference Scores

Difference scores are calculated when comparing performance on two tests.

The reliability of the difference between two test scores is generally lower than the reliabilities of the two tests.

This is due to both tests having unique error variance that they contribute to the difference score.

Reliability of Composite Scores

When there are multiple scores available for an individual, one can calculate composite scores (e.g., assigning grades in class).

There are different approaches, but the important issue is that the reliability of a composite score is generally greater than the reliability of the individual scores.

Standards for Reliability

If a test score is used to make important decisions that will significantly impact individuals, the reliability should be very high - > .90 or > .95.

If a test is interpreted independently but as part of a larger assessment process (e.g., personality test), most set the standard as .80 or greater.
Standards for Reliability

If a test is used only in group research or is used as a part of a composite (e.g., classroom tests), lower reliability estimates may be acceptable, e.g., .70s

Improving Reliability

Increase the number of items (better domain sampling)

Use multiple measurements (composite scores)

Use “Item Analysis” procedures to select the best items

Increase standardization of the test

<table>
<thead>
<tr>
<th>Current Reliability</th>
<th>1.25 X</th>
<th>1.50 X</th>
<th>2.00 X</th>
<th>2.50 X</th>
</tr>
</thead>
<tbody>
<tr>
<td>.50</td>
<td>56</td>
<td>60</td>
<td>67</td>
<td>71</td>
</tr>
<tr>
<td>.60</td>
<td>60</td>
<td>65</td>
<td>71</td>
<td>75</td>
</tr>
<tr>
<td>.70</td>
<td>65</td>
<td>70</td>
<td>76</td>
<td>82</td>
</tr>
<tr>
<td>.75</td>
<td>70</td>
<td>75</td>
<td>82</td>
<td>88</td>
</tr>
<tr>
<td>.80</td>
<td>75</td>
<td>80</td>
<td>86</td>
<td>91</td>
</tr>
<tr>
<td>.85</td>
<td>80</td>
<td>85</td>
<td>90</td>
<td>94</td>
</tr>
<tr>
<td>.90</td>
<td>85</td>
<td>90</td>
<td>95</td>
<td>98</td>
</tr>
<tr>
<td>.95</td>
<td>90</td>
<td>95</td>
<td>100</td>
<td>103</td>
</tr>
<tr>
<td>1.00</td>
<td>95</td>
<td>100</td>
<td>105</td>
<td>110</td>
</tr>
</tbody>
</table>
Standard Error of Measurement

When comparing the reliability of tests, the reliability coefficient is the statistic of choice.

When interpreting individual scores, the SEM generally proves to be the most useful statistic.

It is an index of the average amount of error in test scores, or the standard deviation of error scores around the true score.

The SEM is calculated using the reliability coefficient and the standard deviation of the test.

---

Standard Error of Measurement

Since the test’s reliability coefficient is used in calculating the SEM, there is a direct relationship between $r_{xx}$ and SEM.

As the reliability of a test increases, the SEM decreases; as reliability decreases, the SEM increases.

---

Confidence Intervals

A confidence interval reflects a range of scores that will contain the test taker’s true score with a prescribed probability.

The SEM is used to calculate CIs.

Like any SD, the SEM can be interpreted in terms of frequencies represented in a normal distribution.
Confidence Intervals

\[ \text{Confidence Intervals} = \text{Obtained Score} \pm (z\text{-score}) \times \text{SEM} \]

- If z-score is 1, the result is a 68% confidence interval
- If the z-score is 1.96, the result is a 95% confidence interval

Confidence Intervals

CIs remind us that measurement error is present in all scores and we should interpret scores cautiously.

Confidence intervals are interpreted as “The range within which a person’s true score is expected to fall X% of the time”

Generalizability Theory

An extension of Classical Test Theory, but where CT says “all error is random,” GT recognizes sources of systematic error.

For example - in scoring essays, some graders are consistently more rigorous and some graders are consistently more lenient.

If you have a situation in which there is no opportunity for systematic error to enter the model, CT and GT are mathematically identical.
Generalizability Theory

CT is most useful when objective tests are administered under standardized conditions (e.g., SAT or GRE)

If these considerations are not met, consideration of the principles raised by GT may be useful (e.g., essay or projective tests)

Validity

The degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests.

(AERA et al., 1999)

Validity

Does the test measure what it is intended or designed to measure?

Refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests

Validity can be threatened in several ways
Threats to Validity

- Construct Underrepresentation, or when a test does not measure important aspects of the specified construct
- Construct-Irrelevant Variance, or when the test measures characteristics, content, or skills that are unrelated to the test construct

Reliability & Validity

- Reliability is necessary for validity, but is not sufficient
- A test can be reliable without being valid, but a test cannot be valid without being reliable

Validity analyses are often based on correlational analyses between tests and validation measures.

The reliability of the test (and validation measures) sets an upper limit on the validity coefficient.
Traditional “Types” of Validity

Content validity
Does the test provide an accurate estimate of the test taker’s mastery of the specified content domain?

Criterion-related validity
Can the test predict performance on a specified criterion?

Construct validity
Does the test accurately reflect the dimensions underlying the construct measured by the test?

Validity Evidence vs. Types

The common practice has been to refer to types of validity (e.g., content validity)

However, validity is a unitary concept
Reflects the degree to which all the accumulated evidence supports the intended interpretation of test scores for proposed purposes

Validity Evidence vs. Types

As a result, the current emphasis in the “Standards” and advanced texts is to refer to types of validity evidence rather than distinct types of validity

Instead of “Content Validity,” we have “Evidence Based on Test Content”

Instead of “Criterion-Related Validity,” we have “Evidence Based on Relations to Other Variables”
Classical Test Theory

\[ X = T + E \]

- \( X \): Obtained or observed score (fallible)
- \( T \): True score (reflects stable characteristics)
- \( E \): Error score (reflects random error)

Validity Theory

\[ T = R + I \]

- \( T \): True score (reflects stable characteristics)
- \( R \): stable relevant characteristics
- \( I \): stable irrelevant characteristics

Relevant & Irrelevant Stable Characteristics

Consider a T&M Test:

- Relevant Stable Characteristics - knowledge of T&M content (R)
- Irrelevant Stable Characteristics - reading comprehension skill (I)
Validity Theory

\[ X = R + I + E \]

- \( X \) = Obtained or observed score (fallible)
- \( R \) = relevant stable characteristics
- \( I \) = irrelevant stable characteristics
- \( E \) = Error score (reflects random error)

\[ \sigma^2_X = \sigma^2_R + \sigma^2_I + \sigma^2_E \]

- \( \sigma^2_X \) = Observed score variance
- \( \sigma^2_R \) = Variance due to relevant stable characteristics
- \( \sigma^2_I \) = Variance due to irrelevant stable characteristics (systematic error)
- \( \sigma^2_E \) = Error score variance

Reliability & Validity

We define reliability as:

\[ \frac{\sigma^2_T}{\sigma^2_X} \]

We define validity as:

\[ \frac{\sigma^2_R}{\sigma^2_X} \]
Random & Systematic Error

Reliability analysis - random measurement error is an error component of the observed score that is separate from the true score.

Validity analysis - systematic measurement error (due to stable but irrelevant characteristics) is an error component of the true score.

Validity Evidence Based on Test Content

Focuses on how well the test items sample the behaviors or subject matter the test is designed to measure.

Does the test provide an accurate estimate of the test taker’s mastery of the specified content domain?

In other words, does the test adequately measure the content, behaviors, or skills it is thought to measure?

Validity Evidence Based on Test Content

Most relevant, appropriate, and important for tests used to make inferences about the knowledge and/or skills assessed by a sample of items.

E.g., achievement tests or employment tests.

Traditionally, has involved a qualitative process in which the test is compared to a detailed description of the test domain developed by respected experts.
Face Validity

Not technically a facet of validity, but refers to a test “appearing” to measure what it is designed to measure.

Content-based evidence of validity is acquired through a systematic and technical analysis of the test content, face validity only involves the superficial appearance of a test.

Face Validity

Often desirable as it makes the test more acceptable to those taking the test and the general public.

Can be undesirable if it makes the test “transparent” to the point that it makes it easy for test takers to feign symptoms (e.g., forensic & employment settings).

Validity Evidence Based on Relations to Other Variables

Test-Criterion Evidence
- Can the test predict performance on a specified criterion?
- How accurately do test scores predict criterion performance? (traditionally considered Criterion-Related Validity)

Convergent & Discriminant Evidence
- Examines relationship with existing tests that measure similar or dissimilar constructs. (traditionally incorporated under Construct Validity)

Contrasted Group Evidence
- Do different groups perform differently on the test? (traditionally incorporated under Construct Validity)
Test-Criterion Evidence

The criterion variable is a measure of some attribute or outcome that is of primary interest, as determined by the test users

E.g., the SAT is used to predict academic performance in college, so the criterion variable is college GPA

Approach #1: Predictive Studies

Indicates how accurately test data can predict criterion scores that are obtained at a later time

When prediction is the intended application, predictive studies retain the temporal differences and other characteristics of the practical situation

Predictive Studies

Predictive studies are time-consuming and expensive

Involve the use of an experimental group, so people may be admitted to programs or given jobs without regard to their performance on the predictor, even if their scores suggest a low probability of success
Approach #2: Concurrent Studies

Obtains predictor and criterion data at the same time

Most useful when prediction over time is not important

E.g., examining the relationship between the results of a brief paper-and-pencil measure and a clinical interview

Concurrent Studies Problems

May be applied even when changes over time are important

May result in a restricted range of scores on one or both measures (e.g., SAT administered only to college freshmen)

Range restriction reduces the size of correlation coefficients

Criterion Validity Coefficients

Use the correlation between the predictor and criterion to reflect the relationship between the predictor and criterion ($r_{xy}$)

Linear regression is used to predict scores on criterion variable, but is dependent upon the strength of the correlation coefficient
Linear Regression and $S_E$

Linear regression assumes a perfect relationship between the predictor and criterion, it does not take into account prediction error.

To correct for this, we use a statistic referred to as the Standard Error of Estimate ($S_E$).

Standard Error of Estimate ($S_E$)

Represents the average amount of prediction error - the average number of points by which predicted scores differ from actual criterion scores.

A residual is the difference between the actual criterion score and its predicted value.

$S_E$ is the standard deviation of the distribution of prediction errors.

$S_E$ and Confidence Intervals

When using linear regression to predict performance, $S_E$ is used to construct confidence intervals representing a range of scores within which the actual criterion score is predicted to fall.

We use information about prediction accuracy ($S_E$) to convert a single predicted criterion score into a range within which we expect the actual criterion score to fall.
SEM and $S_E$

The SEM indicates the margin of measurement error caused by the imperfect reliability of the test.

The $S_E$ indicates the margin of prediction error caused by the imperfect validity of the test.

Predicting Criterion Scores

If you have large criterion validity coefficients, the $S_E$ will be relatively small.

As a general rule, criterion validity coefficients will not be as large as the reliability coefficients we are accustomed to.

Interpreting Validity Coefficients

How large should validity coefficients be?

- No simple answer - the relationship should be large enough so that information from the test helps predict performance on the criterion.

If the test helps predict criterion performance better than any existing predictor, the test may be useful even if the coefficients are relatively small.
Decision-Theory Models & Selection Efficiency Analysis

Does not rely on the absolute value of the criterion validity coefficients, but examines the test’s contribution in decision making situations

If the test’s use results in a high proportion of accurate decisions it is useful, otherwise get rid of it

Selection Efficiency Analysis

*Base Rate* is the overall proportion of people in the group under study who can be successful on the criterion

*Selection Rate* is the rate of people to be selected to people applying

Selection Efficiency Analysis

If you have a low selection rate (a situation where you have many applicants to fill a small number of positions) and a low base rate (success on the criterion is fairly rare), even tests with relatively small criterion validity coefficients can significantly improve decision making
Convergent & Discriminant Evidence

Convergent evidence is obtained when you have strong correlations with existing tests that measure similar constructs.

Discriminant evidence is obtained when you have low correlations with existing tests that measure dissimilar constructs.

Contrasted Group Studies

Validity evidence can also be gathered by examining groups that are expected, based on theory, to differ on the construct being measured.

- Depressed versus Control on BDI
- Examine developmental trends in performance

Internal Structure Evidence

Examining the internal structure of the test to determine if its structure is consistent with the hypothesized structure of the construct it is designed to measure.

- Factor analysis & measures of item homogeneity.

Traditionally incorporated under Construct Validity.
Response Processes Evidence

Are the responses invoked by the test consistent with the construct being assessed?

For example, does a test of math reasoning require actual analysis and reasoning, or simply rote calculations?

Traditionally incorporated under Construct Validity

Consequential Validity Evidence

Also referred to as “Validity Evidence Based on Consequences of Testing”

If the test is thought to result in some benefit, are those benefits being achieved?

A topic of controversy since some suggest that the concept should incorporate social issues and values

Sources of Validity Evidence

<table>
<thead>
<tr>
<th>Source</th>
<th>Example</th>
<th>Major Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence Based on Content</td>
<td>Analysis of item relevance and content coverage</td>
<td>used in the selection and placement of employees</td>
</tr>
<tr>
<td>Evidence Based on Relations to Other Variables</td>
<td>Test-criterion, convergent and discriminant evidence; contrasted groups studies</td>
<td>wide variety of tests</td>
</tr>
<tr>
<td>Evidence Based on Internal Structure</td>
<td>Factor analysis, analysis of test homogeneity</td>
<td>particularly useful with tests of constructs like personality or intelligence</td>
</tr>
<tr>
<td>Evidence Based on Response Processes</td>
<td>Analysis of the processes engaged in by the examinee or examiner</td>
<td>examinee to engage in a cognitive or behavioral activity</td>
</tr>
<tr>
<td>Evidence Based on Consequences of Testing Structure</td>
<td>Analysis of unintended consequences of testing.</td>
<td>other applications in tests</td>
</tr>
</tbody>
</table>