Reliability for Teachers

How much can I “rely” on the scores? How stable are they?
Definition

- Reliability = The consistency or stability of assessment results
  - It is considered to be a characteristic of scores or results, not the test itself
- A test is reliable if it produces consistent scores
Factors that Influence Reliability

- Time when the test is administered
- Specific questions
- Distractions
- The person grading
- Other random/individual factors

All of these factors contribute to the score (also called an “error” score)
Reliability of Composite Scores

• When several tests or subtests contribute to an overall score, this is known as the composite score
• The reliability of the composite score is generally greater than that reliability of the individual tests (within limits; for example, a GPA – composite score – is a more reliable measure of academic achievement than a single course grade)
Methods of Estimating Reliability

• The precise reliability of an assessment cannot be known, but we can estimate it.

• Reliability coefficients can be classified in three main ways, depending on the purpose of the assessment:
  • From administering the same test on different days (test-retest)
  • From administering similar forms of the test (alternate form)
  • From the single administration of a test (internal consistency)

• Inter-rater reliability comes into play when scoring a test involves significant subjective judgment (like scoring an essay).
Test-Retest Reliability

• Usually involves the same test being administered twice (over a period of time) to the same group of students on different occasions
• Indicates the stability of the test scores over time
• Reliable tests will produce the same, or close to the same, scores both times
• The time between administration of the tests should be taken into account (not too long or too short) and depends on what is being measured by the test
Alternate-Form Reliability

• Usually involves two tests, which are the same in terms of content, difficulty, and other relevant characteristics, being administered to the same group

• Can be time consuming, and taking one test before the other can affect how the student approaches it, which affects reliability. Usually this is addressed by having two groups – one takes form A first, followed by form B; and the other takes form B followed by form A
Internal-Consistency Reliability

- Split-half reliability = Happens when the test is split into two equal halves, and each half is scored separately
  - The test should **not** be split into the first and second half, as many tests tend to get more difficult as they progress
  - Consider grouping even and odd numbered items instead

- Coefficient alpha and Kuder-Richardson reliability = Usually refers to ways of calculating the reliability of the test scores when various methods of halving / estimating the correlation between test items is used.
  - Content heterogeneity = The degree to which test items measure related characteristics
Inter-rater Reliability

- Established by having two individuals score the same set of tests
- Reflects differences in how the individuals score a test
- On some tests, such as multiple choice tests with only one correct answer for each item, inter-rater reliability is of little concern
- Essay tests and tests that require the student to “show their work” depend more on inter-rater reliability. Another area in which this is important is for rating of performances, portfolios, etc.
Sources of Measurement Error

- Content sampling error: The error that results from the differences between the sample items (what appears on the test) and the domain items (all possible items).
- Time sampling error: Errors that result from differences in student performance due to distractions, time of day, or other random outside influences.
- Inter-rater differences: Refers to whether or not a test would receive the same score if scored by two or more different graders.
Evaluating Reliability Coefficients

• A reliability coefficient of 1.0 indicates perfect correlation.
• However, due to measurement errors, perfectly reliable tests do not exist.
• The reliability coefficient depends on several factors:
  • Construct (theme or idea) to be measured (e.g., career interest is a more unstable construct than career aptitude)
  • Time available for testing
  • Test score use
  • Method of estimating reliability
How to Improve Reliability

• 0.70 is the reliability coefficient generally expected of a classroom test, though one could strive for a higher reliability

• To raise reliability, try:
  • Increasing the numbers of items on the test
  • Using multiple measurements that combine for a composite score
  • Revising poor test items (i.e., removing spelling errors, making options in a multiple choice more plausible, etc.) – this will also increase validity of scores for the purpose intended
  • Providing clear instructions regarding the administration and scoring of tests
Measurement Error

• Even if we believe a measure is exact, there is usually some degree of error

• True score= The score that would be obtained by the student if there were no errors

• Error score= The actual score is the combination of the obtained score and the error

• Measurement error reduces the usefulness of the measurement (test results)

• We consider this measurement “error” in estimating a “band” within which the true score may fall
Standard Error of Measurement

- Standard error of measurement (SEM) = A function of the reliability and standard deviation of the test, this value is often used when determining reliability because a perfect reliability is impossible to achieve.

  \[ SD = \ \text{Standard deviation of obtained scores} \]
  \[ = \ \text{The reliability of a test} \]

- Evaluating the standard error of measurement
  - The greater the reliability of a test, the smaller SEM
  - A small SEM gives us greater confidence in the precision of test scores

- Confidence intervals = A range of scores that are believed to contain the individual test taker’s “true score”
Practical Strategies for Teachers

- Most band scores are for 95% or 99% confidence. So, if an SEM is 3.5, then an obtained score of 68, for 95% confidence you multiply 1.96 with 3.5 (1.96 is the constant if you want 95% confidence). The band score would then be \((68 \pm 6.86)\), which is 61.14 to 74.86 – the student’s “true” score lies somewhere within this band. The use of band scores is usually related to placement decisions. So, if the student’s lowest possible score (within a 95% confidence band is 61.14, then placement in a special program that has a requirement of 60 or below on the test is not advisable). If the requirement is a score of 80 or below, though, the student should be placed in this program.

- For 99% confidence, the constant used to multiply with the SEM is 2.58 (which, in this case would be \(68 \pm (2.58 \times 3.5)\)).