Reliability: Importance for Psychological Measurement

PSYC3302: Psychological Measurement and Its Applications

Mark Hurlstone
Univeristy of Western Australia

Week 4

Reliability and Individual Scores

- Reliability is one cornerstone of a test's psychometric quality
- The reliability coefficient helps:
  - the test developer build an adequate measuring instrument
  - the test user to select a suitable test
- But the utility of the reliability coefficient does not end here
- It is also important for the interpretation of individual test scores

Test Score Interpretation

- Everyday, across the world, psychological tests are used to make important decisions:
  - *what is the patient's diagnosis?*
  - *is this person competent to stand trial?*
  - *who should be hired, promoted, or fired?*
  - *which student should be awarded a scholarship?*
  - *which parent should gain custody?*
- Sometimes, these psychological tests are used to make life or death decisions (see week 1 lecture)
- A test's reliability has important implications for the quality of decisions made on the basis of a person's test score

Test Score Interpretation

- We can never know a person's "true score" on a psychological construct
  - e.g., we can never know a person's true level of working memory, intelligence, or executive functioning
- A person's observed score on a test is an *estimate* of their true score
- That estimate may be used to make important decisions that affect the individual
- We therefore want to know about the *precision* of an individual's test score as an estimate of his or her true score

Test Score Interpretation

- Two sources of information can help us evaluate an individual's test score:
  1. a *point estimate*: a "best estimate" of a person's true score
  2. a *confidence interval*: the range in which the true score is likely to fall
- The confidence interval around a given observed score gives us an idea of its accuracy or precision as an estimate of a true score

Key point:

- Point estimates and confidence intervals are directly affected by the test score reliability coefficient

Test Score Interpretation

- Two sources of information can help us evaluate an individual's test score:
  1. a *point estimate*: a "best estimate" of a person's true score
  2. a *confidence interval*: the range in which the true score is likely to fall
- The confidence interval around a given observed score gives us an idea of its accuracy or precision as an estimate of a true score

Key point:

- Point estimates and confidence intervals are directly affected by the test score reliability coefficient

Test Score Interpretation

- Two sources of information can help us evaluate an individual's test score:
  1. a *point estimate*: a "best estimate" of a person's true score
  2. a *confidence interval*: the range in which the true score is likely to fall
- The confidence interval around a given observed score gives us an idea of its accuracy or precision as an estimate of a true score

Key point:

- Point estimates and confidence intervals are directly affected by the test score reliability coefficient

True Score Estimates

- There are two kinds of point estimates of a person's true score that can be computed from a person's observed score:
  1. An individual's observed test score
  2. An adjusted true score estimate

True Score Estimates

- There are two kinds of point estimates of a person's true score that can be computed from a person's observed score:

  1. An individual's observed test score
  2. An adjusted true score estimate

True Score Estimates:

- One point estimate is based solely on a person's observed score on a test
- It is the single best estimate of a person's true score at the moment she or he took the test
- For example, if you measured a person's working memory using the Operation Span task, his or her score is a point estimate of his or her true working memory
- This point estimate does not take measurement error into account

True Score Estimates

- There are two kinds of point estimates of a person's true score that can be computed from a person's observed score:
  1. An individual's observed test score
  2. An adjusted true score estimate

True Score Estimates

- There are two kinds of point estimates of a person's true score that can be computed from a person's observed score:
    1. An individual's observed test score
    2. An adjusted true score estimate

True Score Estimates:

- Because of measurement error—e.g., fatigue, distraction—sometimes a person's observed score will be lower or higher than her or his true score
- Thus, if a person took the same test on two different occasions, then she or he would likely obtain different scores
- The second point estimate—known as an *adjusted true score estimate*—takes such measurement error into account
- Using a person's observed score on a test (e.g., of working memory) and the reliability coefficient for that test, it allows us to predict that person's future score on the very same test

True Score Estimates:

- The adjusted true score estimate reflects an effect called *regression to the mean*:
- If a person's observed score on a test is *above* the mean on the first testing occasion, her or his score is likely to be somewhat *lower* (i.e., closer to the mean) on the second
- If a person's observed score on a test is *below* the mean on the first testing occasion, her or his score is likely to be somewhat *higher* (i.e., closer to the mean) on the second
- Regression to the mean follows logically from CTTs assumptions about the "randomness" of measurement error

True Score Estimates:

- The adjusted true score estimate can be calculated using the following formula:

$$X_{est} = \overline{X} + R_{xx}(X_o - \overline{X}) \tag{19}$$

- Where:
  - X_{est} = the adjusted true score estimate (i.e., an estimate of a person's score on a second testing occasion)
  - X_o = a person's observed score on the first testing occasion
  - $\overline{X}$ = the test score mean
  - R_{xx} = the reliability coefficient

True Score Estimates:

- Example: a person obtains a score X of 65 on a spelling test, with a mean $\overline{X}$ of 50, and a reliability coefficient R_{xx} of .9

$$X_{est} = 50 + .9(65 - 50)$$

$$X_{est} = 63.5$$

True Score Estimates:

- The adjusted score estimate reflects the discrepancy in an individual's observed score that is likely to arise between two testing occasions
- The size and direction of this discrepancy is a function of three factors:
    1. the size of the reliability coefficient
    2. the size of the difference between an individual's observed test score and the mean
    3. the direction of the difference between an individual's observed test score and the mean (whether the score was above or below the mean)

True Score Estimates:

- The adjusted score estimate reflects the discrepancy in an individual's observed score that is likely to arise between two testing occasions
- The size and direction of this discrepancy is a function of three factors:
  1. the size of the reliability coefficient
  2. the size of the difference between an individual's observed test score and the mean
  3. the direction of the difference between an individual's observed test score and the mean (whether the score was above or below the mean)

True Score Estimates:

- Poor reliability produces bigger discrepancies between the estimated true score and the observed score

True Score Estimates:

Example

- As per the original example, a person obtains a score X of 65 on a spelling test, with a mean $\overline{X}$ of 50
- Now the reliability coefficient R_{xx} is .45

$$X_{est} = 50 + .45(65 - 50)$$

$$X_{est} = 56.75$$

- The lower reliability coefficient—reflecting greater measurement error—brings the adjusted true score estimate closer to the mean (8.25 points vs. 1.5 points when R_{xx} = .9)

True Score Estimates:

- The adjusted score estimate reflects the discrepancy in an individual's observed score that is likely to arise between two testing occasions
- The size and direction of this discrepancy is a function of three factors:
  1. the size of the reliability coefficient
  2. the size of the difference between an individual's observed test score and the mean
  3. the direction of the difference between an individual's observed test score and the mean (whether the score was above or below the mean)

True Score Estimates:

- The adjusted score estimate reflects the discrepancy in an individual's observed score that is likely to arise between two testing occasions
- The size and direction of this discrepancy is a function of three factors:
  1. the size of the reliability coefficient
  2. the size of the difference between an individual's observed test score and the mean
  3. the direction of the difference between an individual's observed test score and the mean (whether the score was above or below the mean)

True Score Estimates:

- The difference between the estimated true score and the observed score will be larger for relatively extreme observed scores (high or low) than for relatively moderate scores

True Score Estimates:

Example

- As per the original example, the mean score on a spelling test $\overline{X}$ is 50, with a reliability coefficient R_{xx} of .9
- Now the person obtains a score X of 55

$$X_{est} = 50 + .9(55 - 50)$$

$$X_{est} = 54.5$$

- The smaller discrepancy between the person's score and the test score mean yields a smaller adjustment toward the mean (0.5 points vs. 1.5 points when X = 65)

True Score Estimates:

- The adjusted score estimate reflects the discrepancy in an individual's observed score that is likely to arise between two testing occasions
- The size and direction of this discrepancy is a function of three factors:
  1. the size of the reliability coefficient
  2. the size of the difference between an individual's observed test score and the mean
  3. the direction of the difference between an individual's observed test score and the mean (whether the score was above or below the mean)

True Score Estimates:

- The adjusted score estimate reflects the discrepancy in an individual's observed score that is likely to arise between two testing occasions
- The size and direction of this discrepancy is a function of three factors:
  1. the size of the reliability coefficient
  2. the size of the difference between an individual's observed test score and the mean
  3. the direction of the difference between an individual's observed test score and the mean (whether the score was above or below the mean)

True Score Estimates:

- If the difference between the original score and the mean is *positive* (i.e., the original is higher than the mean) the adjusted true score estimate will be *smaller* than the original score—see previous examples
- Conversely, if the difference between the original score and the mean is *negative* (i.e., the original is lower than the mean) the adjusted true score estimate will be *larger* than the original score

True Score Estimates:

Example

- As per the original example, the mean score on a spelling test $\overline{X}$ is 50, with a reliability coefficient R_{xx} of .9
- Now the person obtains a score X of 35

$$X_{est} = 50 + .9(35 - 50)$$

$$X_{est} = 36.5$$

- Unlike the previous examples, the adjusted true score estimate is higher (as opposed to lower) than the original test score

True Score Estimates: Final Remarks

- The observed score is an "unbiased" estimate of the true score—it is the best estimate of a person's true score
- The adjusted score estimate is the best estimate of a predicted true score
- If, based on a person's initial performance on a test, you wanted to predict that person's performance on a subsequent administration of the same test, then the adjusted score estimate will likely be a better predictor than the observed score estimate

Test Score Interpretation

- Two sources of information can help us evaluate an individual's test score:
  1. a *point estimate*: a "best estimate" of a person's true score
  2. a *confidence interval*: the range in which the true score is likely to fall
- The confidence interval around a given observed score gives us an idea of its accuracy or precision as an estimate of a true score

Key point:

- Point estimates and confidence intervals are directly affected by the test score reliability coefficient

Test Score Interpretation

- Two sources of information can help us evaluate an individual's test score:
  1. a *point estimate*: a "best estimate" of a person's true score
  2. a *confidence interval*: the range in which the true score is likely to fall
- The confidence interval around a given observed score gives us an idea of its accuracy or precision as an estimate of a true score

Key point:

- Point estimates and confidence intervals are directly affected by the test score reliability coefficient

True Score Confidence Intervals

- Point estimates of an individual's true score are usually reported with *true score confidence intervals*
- Confidence intervals reflect the *precision* of the point estimate of an individual's true score
- Confidence intervals are constructed using the the standard error of measurement se_m introduced in our week 3 lecture:

$$se_m = \sigma_o \sqrt{1 - R_{xx}} \tag{16}$$

- Where σ_o is the standard deviation of observed test scores and R_{xx} is the reliability coefficient of the test

True Score Confidence Intervals

Example

- For our spelling test example, suppose the standard deviation of observed scores is σ_o = 10 and the reliability coefficient R_{xx} = .84

$$se_m = 10\sqrt{1 - .84}$$

$$se_m = 4$$

True Score Confidence Intervals

- The se_m is the standard deviation of a theoretically normal distribution of test scores obtained by one person on equivalent tests
- It is an index of the extent to which one individual's scores vary over tests presumed to be parallel
- In accordance with CTT, an observed test score is one point in the theoretical distribution of scores the test-taker could have obtained
- The se_m allows us to estimate, with a specific level of confidence (typically 95%), the range in which the true score is likely to exist

True Score Confidence Intervals

- To use the se_m to estimate the confidence interval of the true score, we make an assumption
- If the individual were to take a large number of equivalent tests, scores on those tests would tend to be normally distributed, with the individual's true score as the mean
- Since the se_m functions like a standard deviation in this context, we can use it to predict what would happen if an individual took additional equivalent tests ...

True Score Confidence Intervals

- Approximately 68% (actually, 68.26%) of the scores would be expected to occur within $\pm 1 se_m$ of the true score
- Approximately 95% (actually, 95.44%) of the scores would be expected to occur within $\pm 2 se_m$ of the true score
- Approximately 99% (actually, 99.74%) of the scores would be expected to occur within $\pm 3 se_m$ of the true score

True Score Confidence Intervals

- Suppose an individual obtained a score of 50 on one spelling test and that test had a se_m of 4, then using 50 as the point estimate we can be:
  - 68% (actually, 68.26%) confident that the true score falls within 50 $\pm 1se_m$ (or between 46 and 54)
  - 95% (actually, 95.44%) confident that the true score falls within 50 $\pm 2se_m$ (or between 42 and 58)
  - 99% (actually, 99.74%) confident that the true score falls within 50 $\pm 3se_m$ (or between 38 and 62)

True Score Confidence Intervals

- We can calculate a confidence interval around an individual's estimated true score using the formula:

$$95\% \text{ confidence interval} = X_o \pm (z_{95\%})(se_m) \tag{20}$$

- Where X_o is an individual's observed test score (i.e., a point estimate of his or her true score) and se_m is the standard error of measurement
- $z_{95\%}$ is the z score from a normal distribution table corresponding to a score below which 95% of the area of the normal distribution falls...
  - in this case, $z_{95\%}$ = 1.96

True Score Confidence Intervals

Example

- Let's calculate the 95% confidence interval for our individual who obtained a score of 50 on a spelling test that had a se_m of 4

$$95\% \text{ confidence interval} = 50 \pm (1.96)(4)$$

$$= 50 \pm 7.84$$

$$= 42.16 \text{ to } 57.84$$

True Score Confidence Intervals

- When calculating confidence intervals, we typically use the 95% confidence interval (as in the previous example)
- But you can use other confidence intervals too; here are some other common intervals and their associated z scores:
  - 68% confidence interval: $z_{68\%} = 1$
  - 75% confidence interval: $z_{75\%} = 1.15$
  - 85% confidence interval: $z_{85\%} = 1.44$

True Score Confidence Intervals

- The precision of a true score estimate is closely related to reliability
- Highly reliable tests will produce narrower confidence intervals (greater precision) than less reliable tests
- We have seen that for a reliable test (R_{xx} = .85) with σ_o = 10, the se_m was 4, and the confidence interval had a range of 7.84 points
- Suppose the reliability coefficient was R_{xx} = .44. The se_m would now be:

$$se_m = 4\sqrt{1 - .44}$$

$$= 7.48$$

True Score Confidence Intervals

- The 95% confidence interval would now be:

$$95\% \text{ confidence interval} = 50 \pm (1.96)(7.48)$$

$$= 50 \pm 14.66$$

$$= 35.34 \text{ to } 64.66$$

- *The key point is that reliability affects the the confidence, accuracy, or precision with which an individual's true score is estimated*

True Score Confidence Intervals

- The 95% confidence interval would now be:

$$95\% \text{ confidence interval} = 50 \pm (1.96)(7.48)$$

$$= 50 \pm 14.66$$

$$= 35.34 \text{ to } 64.66$$

- *The key point is that reliability affects the the confidence, accuracy, or precision with which an individual's true score is estimated*

True Score Confidence Intervals

- Estimating true scores and true score intervals can have important consequences in applied contexts
- For example, recall from our week 1 lecture that in some US states a mentally retarded prisoner with an IQ of 70 or less cannot be subjected to the death penalty
- If a prisoner obtained an IQ score of 65, with a 95% confidence interval of ± 10 (55 to 75), we would not be confident enough to classify this individual as mentally retarded
- But if the same prisoner obtained an IQ score of 65 with a 95% confidence interval of ± 4 (61 to 69), we would be much more confident about this classification

Behavioural Research

- Reliability has important implications for the interpretation of behavioural research
- The interpretations we draw in psychology depend on the psychometric soundness of the measurement procedures we use
- To interpret behavioural research accurately, we need to be aware of the ways in which reliability and measurement error affect our results

Reliability, True Associations, and Observed Associations

- In psychology, we are often interested in the association between variables (e.g., IQ and SAT scores)
- In psychometrics, the most common way of measuring these associations is through a correlation coefficient

Reliability, True Associations, and Observed Associations

- According to CTT, the correlation between the observed scores on two measures ($r_{x_o y_o}$) is determined by two things:
  1. the correlation between the true scores on the two psychological constructs being assessed by the measures ($r_{x_t y_t}$) and
  2. the reliabilities of the two measures (R_{xx}, R_{yy})

$$r_{x_o y_o} = r_{x_t y_t} \sqrt{R_{xx} R_{yy}} \tag{21}$$

Example

- Suppose we measure the association between self-esteem and academic achievement
- Imagine the true correlation between the two constructs is $r_{x_t y_t}$ = .40 (we would not actually know this)
- Assume further that both scales have good reliability—.80 (self-esteem) and .86 (academic achievement). The correlation between measure is:

$$r_{x_o y_o} = .40\sqrt{(.80)(.86)}$$

$$= .40(.829)$$

$$= .33$$

Example

- Suppose we measure the association between self-esteem and academic achievement
- Imagine the true correlation between the two constructs is $r_{x_t y_t}$ = .40 (we would not actually know this)
- Assume further that both scales have good reliability—.80 (self-esteem) and .86 (academic achievement). The correlation between measure is:

$$r_{x_o y_o} = .40\sqrt{(.80)(.86)}$$

$$= .40(.829)$$

$$= .33$$

Measurement Error and Observed Associations

- The key point to understand is that **observed associations** (i.e., between measures) will always be weaker than **true associations** (i.e., between psychological constructs)
- This is because of measurement error—our measurement will *always* be prone to some degree of error
- This imperfect measurement "weakens" or "attenuates" observed associations
- However, it is possible to estimate the true association between a pair of constructs by employing a formula known as the *correction for attenuation*

Measurement Error and Observed Associations

- The formula for correction of attenuation is given by:

$$r_{x_t y_t} = \frac{r_{x_o y_o}}{\sqrt{R_{xx} R_{yy}}} \tag{22}$$

- Let's apply this formula to our previous example to estimate the true association from the observed association and the reliability of the two measures:

$$\begin{aligned} r_{x_t y_t} &= \frac{.33}{\sqrt{(.80)(.86)}} \\ &= \frac{.33}{.83} \\ &= .40 \end{aligned}$$

Measurement Error and Observed Associations

- The formula for correction of attenuation is given by:

$$r_{x_t y_t} = \frac{r_{x_o y_o}}{\sqrt{R_{xx} R_{yy}}} \tag{22}$$

- Let's apply this formula to our previous example to estimate the true association from the observed association and the reliability of the two measures:

$$r_{x_t y_t} = \frac{.33}{\sqrt{(.80)(.86)}} = \frac{.33}{.83} = .40$$

Effect Sizes and Statistical Significance

- We have seen how measurement error (i.e., low reliability) "attenuates" observed associations
- You should also know that effects sizes (e.g., Cohen's d, Eta Squared) and statistical significance tests (e.g., t test, F test) are also affected by reliability and measurement error
- Thus, whether you are interpreting correlations, or the size or significance of differences among groups or conditions, the results of a study should always be interpreted in the context of reliability
- If poor reliability biases these results, we can easily be misled into making inaccurate conclusions about our research

Implications For Behavioural Research

1. Researchers—and readers of research—should always consider the effects of reliability on their results when interpreting effect sizes and/or statistical significance
2. Researchers should strive to use highly reliable measures in their work
3. Researchers should always report reliability estimates of their measures

Test Construction and Refinement

- The textbook includes a final section on "Test Construction and Refinement" (p.186-192)
- This material is relevant to your lab class next week
- Accordingly, I recommend reading this section before you attend your lab

In Next Week's Lab ...

- ... You will:
  1. Learn how to perform internal consistency reliability analyses in SPSS using Coefficient α
  2. Learn procedures for evaluating item information

In Next Week's Lecture ...

- Introduction to the theoretical basis of a second cornerstone of psychometrics—"validity"

References

Furr, M. R., & Bacharach, V. R. (2014; Chapter 7). Psychometrics: An Introduction (second edition). Sage.