CLASSICAL TEST THEORY AND RELIABILITY

Cal State Northridge
Psy 320
Andrew Ainsworth, PhD

Basics of Classical Test Theory

- Theory and Assumptions
- Types of Reliability
- Example

Classical Test Theory

- Classical Test Theory (CTT) – often called the "true score model"
- Called classic relative to Item Response Theory (IRT) which is a more modern approach
- CTT describes a set of psychometric procedures used to test items and scales reliability, difficulty, discrimination, etc.

Classical Test Theory

- CTT analyses are the easiest and most widely used form of analyses. The statistics can be computed by readily available statistical packages (or even by hand)
- CTT Analyses are performed on the test as a whole rather than on the item and although item statistics can be generated, they apply only to *that* group of students on *that* collection of items

Classical Test Theory

- Assumes that every person has a true score on an item or a scale if we can only measure it directly without error
- CTT analyses assumes that a person's test score is comprised of their "true" score plus some measurement error.
- This is the common true score model

$$X = T + E$$

Classical Test Theory

- Based on the expected values of each component for each person we can see that

$$\varepsilon(X_i) = t_i$$

$$E_i = X_i - t_i$$

$$\varepsilon(X_i - t_i) = \varepsilon(X_i) - \varepsilon(t_i) = t_i - t_i = 0$$

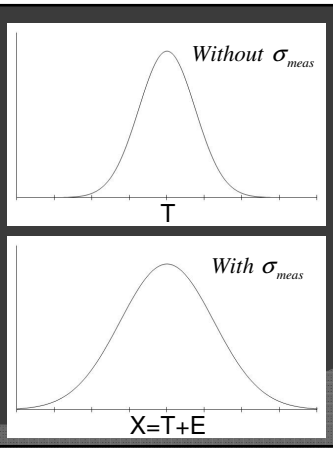
- E and *X* are random variables, *t* is constant
- However this is theoretical and not done at the individual level.

Classical Test Theory

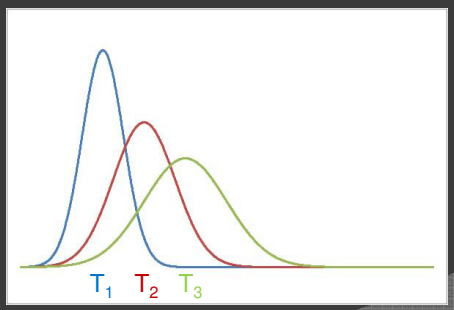
- If we assume that people are randomly selected then *t* becomes a random variable as well and we get:

$$X = T + E$$

- Therefore, in CTT we assume that the error :
  - Is normally distributed
  - Uncorrelated with true score
  - Has a mean of Zero



True Scores



- Measurement error around a T can be large or small

Domain Sampling Theory

- Another Central Component of CTT
- Another way of thinking about populations and samples
- Domain - Population or universe of all possible items measuring a single concept or trait (theoretically infinite)
- Test – a sample of items from that universe

Domain Sampling Theory

- A person's true score would be obtained by having them respond to all items in the "universe" of items
- We only see responses to the sample of items on the test
- So, reliability is the proportion of variance in the "universe" explained by the test variance

Domain Sampling Theory

- A universe is made up of a (possibly infinitely) large number of items
- So, as tests get longer they represent the domain better, therefore longer tests should have higher reliability
- Also, if we take multiple random samples from the population we can have a distribution of sample scores that represent the population

Domain Sampling Theory

- Each random sample from the universe would be "randomly parallel" to each other
- Unbiased estimate of reliability

$$r_{1t} = \sqrt{\overline{r}_{1j}}$$

- r_{1t} = correlation between test and true score
- $\overline{r}_{1j}$ = average correlation between the test and all other randomly parallel tests

Classical Test Theory Reliability

- Reliability is theoretically the correlation between a test-score and the true score, squared
- Essentially the proportion of X that is T

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

- This can't be measured directly so we use other methods to estimate

CTT: Reliability Index

- Reliability can be viewed as a measure of consistency or how well as test "holds together"
- Reliability is measured on a scale of 0-1. The greater the number the higher the reliability.

CTT: Reliability Index

- The approach to estimating reliability depends on
  - Estimation of "true" score
  - Source of measurement error
- Types of reliability
  - Test-retest
  - Parallel Forms
  - Split-half
  - Internal Consistency

CTT: Test-Retest Reliability

- Evaluates the error associated with administering a test at two different times.
- *Time Sampling Error*
- How-To:
  - Give test at Time 1
  - Give SAME TEST at Time 2
  - Calculate *r for the two scores*
- • Easy to do; one test does it all.

CTT: Test-Retest Reliability

- Assume 2 administrations X_1 and X_2

$$\varepsilon(X_{1i}) = \varepsilon(X_{2i}) \qquad \sigma^2_{E_{1i}} = \sigma^2_{E_{2i}}$$

$$\therefore \rho_{X_1X_2} = \frac{\sigma_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}} = \frac{\sigma^2_T}{\sigma^2_X} = \rho_{XT}$$

- The correlation between the 2 administrations is the reliability

CTT: Test-Retest Reliability

- Sources of error
  - random fluctuations in performance
  - uncontrolled testing conditions
    - extreme changes in weather
    - sudden noises / chronic noise
    - other distractions
  - internal factors
    - illness, fatigue, emotional strain, worry
    - recent experiences

CTT: Test-Retest Reliability

- Generally used to evaluate constant traits.
  - Intelligence, personality
- Not appropriate for qualities that change rapidly over time.
  - Mood, hunger
- Problem: Carryover Effects
  - Exposure to the test at time #1 influences scores on the test at time #2
- Only a problem when the effects are random.
- If everybody goes up 5pts, you still have the same variability

CTT: Test-Retest Reliability

- Practice effects
  - Type of carryover effect
  - Some skills improve with practice
    - Manual dexterity, ingenuity or creativity
  - Practice effects may not benefit everybody in the same way.
- Carryover & Practice effects more of a problem with short inter-test intervals (ITI).
- But, longer ITI's have other problems
  - developmental change, maturation, exposure to historical events

CTT: Parallel Forms Reliability

- Evaluates the error associated with selecting a particular set of items.
- *Item Sampling Error*
- How To:
  - Develop a large pool of items (i.e. Domain) of varying difficulty.
  - Choose equal distributions of difficult / easy items to produce multiple forms of the same test.
  - Give both forms *close in time.*
  - Calculate *r for the two administrations.*

CTT: Parallel Forms Reliability

- Also Known As:
  - Alternative Forms *or Equivalent Forms*
- Can give parallel forms at different points in time; produces error estimates of time and item sampling.
- One of the most rigorous assessments of reliability currently in use.
- Infrequently used in practice – too expensive to develop two tests.

CTT: Parallel Forms Reliability

- Assume 2 parallel tests X and X'

$$\varepsilon(X_i) = \varepsilon(X_i^{'}) \qquad \sigma^2_{E_i} = \sigma^2_{E_i^{'}}$$

$$\therefore \rho_{XX'} = \frac{\sigma_{XX'}}{\sigma_X \sigma_{X'}} = \frac{\sigma^2_T}{\sigma^2_X} = \rho_{XT}$$

- The correlation between the 2 parallel forms is the reliability

CTT: Split Half Reliability

- What if we treat halves of one test as parallel forms? (Single test as whole domain)
- That's what a split-half reliability does
- This is testing for *Internal Consistency*
  - Scores on one half of a test are correlated with scores on the second half of a test.
- Big question: "How to split?"
  - First half vs. last half
  - Odd vs Even
  - Create item groups called testlets

CTT: Split Half Reliability

- How to:
  - Compute scores for two halves of single test, calculate *r*.
- Problem:
  - Considering the domain sampling theory what's wrong with this approach?
  - A 20 item test cut in half, is 2 10-item tests, what does that do to the reliability?
  - If only we could correct for that…

Spearman Brown Formula

- Estimates the reliability for the entire test based on the split-half
- Can also be used to estimate the affect changing the number of items on a test has on the reliability

$$r^* = \frac{j(r)}{1+(j-1)r}$$

Where r* is the estimated reliability, r is the correlation between the halves, j is the new length proportional to the old length

Spearman Brown Formula

- For a split-half it would be

$$r^* = \frac{2(r)}{(1+r)}$$

- Since the full length of the test is twice the length of each half

Spearman Brown Formula

- Example 1: a 30 item test with a split half reliability of .65

$$r^* = \frac{2(.65)}{(1+.65)} = .79$$

- The .79 is a much better reliability than the .65

Spearman Brown Formula

- Example 2: a 30 item test with a test re-test reliability of .65 is lengthened to 90 items

$$r^* = \frac{3(.65)}{1+(3-1).65} = \frac{1.95}{2.3} = .85$$

- Example 3: a 30 item test with a test re-test reliability of .65 is cut to 15 items

$$r^* = \frac{.5(.65)}{1+(.5-1).65} = \frac{.325}{.675} = .48$$

Detour 1: Variance Sum Law

- Often multiple items are combined in order to create a composite score
- The variance of the composite is a combination of the variances and covariances of the items creating it
- General Variance Sum Law states that if X and Y are random variables:

$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\sigma_{XY}$$

Detour 1: Variance Sum Law

- Given multiple variables we can create a variance/covariance matrix
- For 3 items:

$$\begin{array}{cccc} & X_1 & X_2 & X_3 \\ X_1 & \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ X_2 & \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ X_3 & \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{array}$$

Detour 1: Variance Sum Law

- Example Variables X, Y and Z
- Covariance Matrix:

	X	Y	Z
X	55.83	29.52	30.33
Y	29.52	17.49	16.15
Z	30.33	16.15	29.06

- By the variance sum law the composite variance would be:

$$\sigma^2_{X+Y+Z} = \sigma^2_{Total} = \sigma^2_X + \sigma^2_Y + \sigma^2_Z + 2\sigma_{XY} + 2\sigma_{XZ} + 2\sigma_{YZ}$$

Detour 1: Variance Sum Law

	X	Y	Z
X	55.83	29.52	30.33
Y	29.52	17.49	16.15
Z	30.33	16.15	29.06

- By the variance sum law the composite variance would be:

$$s^2_{Total} = 55.83 + 17.49 + 29.06 +$$
$$+2(29.52) + 2(30.33) + 2(16.15) =$$
$$= 254.41$$

CTT: Internal Consistency Reliability

- If items are measuring the same construct they should elicit similar if not identical responses
- Coefficient OR Cronbach's Alpha is a widely used measure of internal consistency for continuous data
- Knowing the a composite is a sum of the variances and covariances of a measure we can assess consistency by how much covariance exists between the items relative to the total variance

CTT: Internal Consistency Reliability

- Coefficient Alpha is defined as:

$$\alpha = \frac{k}{k-1}\left(\frac{\sum s_{ij}}{s^2_{Total}}\right)$$

  - s^2_{Total} is the composite variance (if items were summed)
  - s_{ij} is covariance between the *i*th and *j*th items where i is not equal to j
  - *k* is the number of items

CTT: Internal Consistency Reliability

- Using the same continuous items X, Y and Z
- The covariance matrix is:

	X	Y	Z
X	55.83	29.52	30.33
Y	29.52	17.49	16.15
Z	30.33	16.15	29.06

- The total variance is 254.41
- The sum of all the covariances is 152.03

$$\alpha = \frac{k}{k-1}\left(\frac{\sum s_{ij}}{s_{Total}^2}\right) = \frac{3}{3-1}\left(\frac{152.03}{254.41}\right) = .8964$$

CTT: Internal Consistency Reliability

- Coefficient Alpha can also be defined as:

$$\alpha = \frac{k}{k-1}\left(\frac{s_{Total}^2 - \sum s_i^2}{s_{Total}^2}\right)$$

  - s_{Total}^2 is the composite variance (if items were summed)
  - s_i^2 is variance for each item
  - *k* is the number of items

CTT: Internal Consistency Reliability

- Using the same continuous items X, Y and Z
- The covariance matrix is:

	X	Y	Z
X	55.83	29.52	30.33
Y	29.52	17.49	16.15
Z	30.33	16.15	29.06

- The total variance is 254.41
- The sum of all the variances is 102.38

$$\alpha = \frac{k}{k-1}\left(\frac{s_{Total}^2 - \sum s_i^2}{s_{Total}^2}\right) = \frac{3}{3-1}\left(\frac{254.41-102.38}{254.41}\right) = .8964$$

CTT: Internal Consistency Reliability

- From SPSS

****** Method 1 (space saver) will be used for this analysis ******

R E L I A B I L I T Y A N A L Y S I S - S C A L E (A L P H A)

Reliability Coefficients

N of Cases = 100.0 N of Items = 3

Alpha = .8964

CTT: Internal Consistency Reliability

- Coefficient Alpha is considered a lower-bound estimate of the reliability of continuous items
- It was developed by Cronbach in the 50's but is based on an earlier formula by Kuder and Richardson in the 30's that tackled internal consistency for dichotomous (yes/no, right/wrong) items

Detour 2: Dichotomous Items

- If Y is a dichotomous item:
- P = proportion of successes OR items answer correctly
- Q = proportion of failures OR items answer incorrectly
- $\overline{Y}$ = P, observed proportion of successes
- s_Y^2 = PQ

CTT: Internal Consistency Reliability

- Kuder and Richardson developed the KR_{20} that is defined as

$$\alpha = \frac{k}{k-1}\left(\frac{s_{Total}^2 - \sum pq}{s_{Total}^2}\right)$$

  - Where pq is the variance for each dichotomous item
- The KR_{21} is a quick and dirty estimate of the KR_{20}

CTT: Reliability of Observations

- What if you're not using a test but instead observing individual's behaviors as a psychological assessment tool?
- How can we tell if the judge's (assessor's) are reliable?

CTT: Reliability of Observations

- Typically a set of criteria are established for judging the behavior and the judge is trained on the criteria
- Then to establish the reliability of both the set of criteria and the judge, multiple judges rate the same series of behaviors
- The correlation between the judges is the typical measure of reliability
- But, couldn't they agree by accident? Especially on dichotomous or ordinal scales?

CTT: Reliability of Observations

- Kappa is a measure of inter-rater reliability that controls for chance agreement
- Values range from -1 (less agreement than expected by chance) to +1 (perfect agreement)
- +.75 “excellent”
- .40 - .75 “fair to good”
- Below .40 “poor”

Standard Error of Measurement

- So far we’ve talked about the standard error of measurement as the error associated with trying to estimate a true score from a specific test
- This error can come from many sources
- We can calculate it’s size by:

$$s_{meas} = s\sqrt{1-r}$$

- *s* is the standard deviation; *r* is reliability

Standard Error of Measurement

- Using the same continuous items X, Y and Z
- The total variance is 254.41
- s = SQRT(254.41) = 15.95
- α = .8964

$$s_{meas} = 15.95 * \sqrt{1-.8964} = 5.13$$

CTT: The Prophecy Formula

- How much reliability do we want?
- Typically we want values above .80
- What if we don't have them?
- The Spearman-Brown can be algebraically manipulated to achieve

$$j = \frac{r_d(1-r_o)}{r_o(1-r_d)}$$

  - j = # of tests at the current length, r_d = desired reliability, r_o = observed reliability

CTT: The Prophecy Formula

- Using the same continuous items X, Y and Z
- α = .8964
- What if we want a .95 reliability?

$$j = \frac{r_d(1-r_o)}{r_o(1-r_d)} = \frac{.95(1-.8964)}{.8964(1-.95)} = \frac{.0984}{.0448} = 2.2$$

- We need a test that is 2.2 times longer than the original
- Nearly 7 items to achieve .95 reliability

CTT: Attenuation

- Correlations are typically sought at the true score level but the presence of measurement error can cloud (attenuate) the size the relationship
- We can correct the size of a correlation for the low reliability of the items.
- Called the Correction for Attenuation

CTT: Attenuation

- Correction for attenuation is calculated as:

$$\hat{r}_{12} = \frac{r_{12}}{\sqrt{r_{11} r_{22}}}$$

- $\hat{r}_{12}$ is the corrected correlation
- r_{12} is the uncorrected correlation
- r_{11} and r_{22} the reliabilities of the tests

CTT: Attenuation

- For example X and Y are correlated at .45, X has a reliability of .8 and Y has a reliability of .6, the corrected correlation is

$$\hat{r}_{12} = \frac{r_{12}}{\sqrt{r_{11} r_{22}}} = \frac{.45}{\sqrt{.8 * .6}} = \frac{.45}{\sqrt{.48}} = .65$$