

Ch. 4: Reliability

590

Psychology 402 - Fall 2023 - Dr. Michael Diehr

Copyright © 2023 Michael Diehr
All Rights Reserved
For use only by students enrolled
in my sections of Psyc 402
through the end of the semester.
May not be posted, shared or uploaded
online without permission.

591

Psychology 402 - Fall 2023 - Dr. Michael Diehr

Reliability

- Constructs & Measurement
- History
- Classical Test Score Theory
- Four Kinds of Reliability
- Standard Error of Measurement
- Increasing Reliability

610

Psychology 402 - Fall 2023 - Dr. Michael Diehr

Constructs & Measurement

- Psychology as “soft science”
- Construct
 - exists but can’t be directly measured
 - examples
- Measurement
 - “true value” - intelligence
 - measured or *observed* value (e.g. IQ test score)
 - discrepancy - “error”
- How to conceptualize *error*?

611

Psychology 402 - Fall 2023 - Dr. Michael Diehr

History 1

- 1896 - Karl Pearson - product-moment correlation (for continuous variables)
- 1904 - Charles Spearman - “*The proof and measurement of association between two things*” - *Rho* - correlation for Ordinal variables

612

Psychology 402 - Fall 2023 - Dr. Michael Diehr

History

- Pearson, Spearman, Thorndike (1900-1907)
 - Basic reliability theory
- Kuder, Richardson (1937), Cronbach (1989)
 - Reliability coefficients
- Bartholomew & Knott (1990s)
 - Latent variable theory
- Drasgow et al (late 1990s)
 - Item Response Theory (IRT)

613

Psychology 402 - Fall 2023 - Dr. Michael Diehr

Classical Test-Score Theory

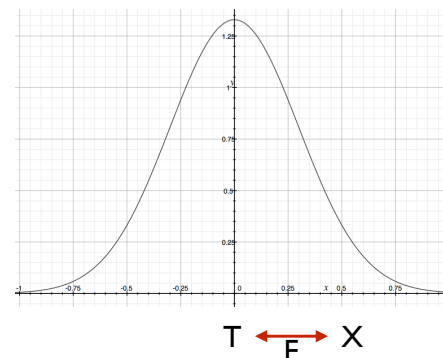
615

Psychology 402 - Fall 2023 - Dr. Michael Dohr

616 Classical Test-Score Theory

- T = True Score
- X = Observed
- E = Error

- $X = T + E$
- $E = X - T$



Psychology 402 - Fall 2023 - Dr. Michael Dohr

Classical Test-Score Theory

- True score (T) : the “actual” score that exists
- Observed score (X) : score as measured by a test
- Error (E) : difference between Observed and True score
- $X = T + E$
- $E = X - T$
- Assumptions: True scores have no variability. Errors are random (e.g. a normal distribution with mean of zero)

617

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Classical Test-Score Theory: Reliability

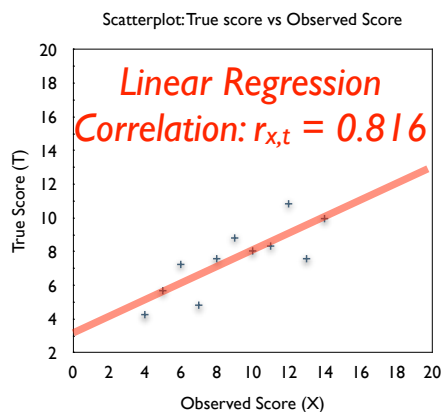
- Reliability = correlation between Observed score and True score
- $R_{X,T}$

618

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Classical Test Score Theory

X	T
10	8.04
8	7.58
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68



619

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Models of Reliability

- Most reliability measures are Correlation coefficients
- Alternate definition: Reliability is the ratio of the variance of True scores to the variance of the Observed scores
 - $\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X}$
- Or, it's the “Signal to Noise” ratio
 - $\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_E}$
- A test with reliability of $r^2=0.40$ means that 40% of variation in test scores is due to variation in the “true” score, and 60% of variation is random or chance factors.

620

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Sources of Error

- “Error” is considered the difference between True score and Observed score
- Where does Error arise?
 - Measurement errors
 - Change in True score
 - Sampling issues
 - Observer effects
 - etc...

621

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Measuring Reliability in Practice

- Since True score is hidden, can't use the direct formula: $R_{X,T}$
- Instead
 - think about sources of error
 - practical methods
 - *estimate* reliability

622

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Test-Retest Reliability

- Test-Retest
 - administer test, delay for interval, administer test again
 - compute correlation between two administrations
 - same subjects, exact same test, two administrations

623

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Test-Retest Reliability

- Pros
 - easy for experimenter to do
- Cons
 - what causes error?
 - short testing interval → practice effects
 - long testing interval → change in true score over time
 - subjects have to take test 2x

624

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Parallel Forms Reliability

- Also called “alternate” or “equivalent” forms
 - Item Sampling
 - administer two versions of the test to same subjects (can have zero delay)
 - compute correlation between two scores
 - same subjects, different test forms, two administrations
 - Pros: more rigorous method of determining reliability
 - Cons: more work: experimenter must make a second test, subjects have to take 2 tests

625

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Internal Consistency Reliability

626

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Internal Consistency Reliability

- Give single test, calculate internal consistency of various subsets of items
- Only one test, one administration, same group of subjects
- Older methods:
 - Split half method
 - Spearman-Brown formula
 - KR20 formula
 - average of all possible Split Halves
- can only handle right/wrong scoring
- newer methods are better...

627

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Internal Consistency Reliability

- New: Cronbach's Alpha (α)
 - estimates a lower bound for reliability
 - does not require right/wrong scoring
 - can be used with Likert scales
 - $\alpha \geq .90$ is good
 - $\alpha \geq .80$ is ok
 - α between .70 - .80 is borderline
 - $\alpha < .70$ is bad

628

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Inter-Rater Reliability

- Observational data differs from self-report data.
- Even though most behavioral rating systems attempt to be precise, errors occur (e.g. did the child fall down? or were they pushed?)
- We must consider the reliability of different observers (also called "raters")
- **Cohen's Kappa**
 - ranges from -1 to +1
 - "poor" $< .40$
 - "good" $.40$ to $.75$
 - "excellent" $> .75$

629

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Reliability: errors & methods

	Description	Name	Statistic
Time Sampling	1 test given two times	test-retest reliability	correlation between scores at two times
Item Sampling	2 different tests given once	Alternate or Parallel forms	correlation between scores on 2 versions
Internal Consistency	One test, multiple items	Split Half or internal reliability	Cronbach's Alpha
Observer Differences	One test w/ 2+ observers	inter-observer reliability	Kappa

630

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Quiz: What kind of Reliability?

Procedure	Source of error?	Reliability?
Olympic judges giving consistent scores for a gymnastics performance		
Correlation between your IQ test score taken at age 12 and again at age 13		
Correlation between scores on 2 versions of the midterm (assuming each student takes both versions)		
Correlation between student scores on questions 1-25 vs 26-50 of the midterm.		

632

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Summary

- Reliability
 - how consistent measured scores are
- Error
 - $E = X - T$
- What kind of Error?
 - test-retest, item sampling, internal consistency, observer-differences

634

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Standard Error of Measurement

- Desire to answer question “how close is this test result to the true result”
- If we know the Reliability (r) of the test, we can estimate the likely range of true values
- Given
 - S = std dev of measured scores
 - r = reliability coefficient of test

$$SEM = S\sqrt{1-r} \quad \sigma_{meas}$$

635

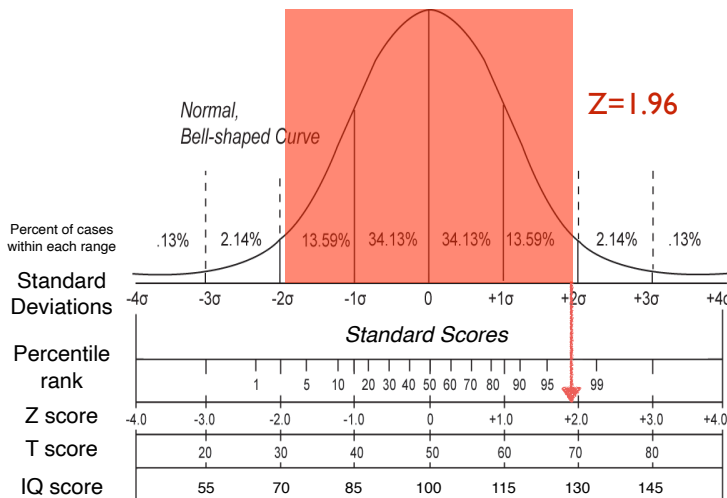
Psychology 402 - Fall 2023 - Dr. Michael Dohr

SEM example: IQ

- Example: a person scored 106 on an IQ test, that has a reliability of 0.89. What is the 95% confidence interval of the their true score
- S = 14
r = 0.89
 $SEM = S\sqrt{1-r}$
 $SEM = 14\sqrt{1-0.89}$
 $SEM = 4.64$
- Next, compute a confidence interval

636

Psychology 402 - Fall 2023 - Dr. Michael Dohr



637

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Confidence Interval

- “How likely is a true score to fall within a range”
- Z = z-score associated with % range
- Confidence interval = Z * SEM
- Example:
 - 95% confidence interval : Z = 1.96
 - SEM = 4.64
 - 1.96 * 4.64 = 9.1
 - 95% CI = ± 9.1 points
 - Range = X ± CI
 - 106 ± 9.1 = range from 96.9 ... 115.1

638

Psychology 402 - Fall 2023 - Dr. Michael Dohr

SEM Exercise

- This is for practice, not scored for points

639

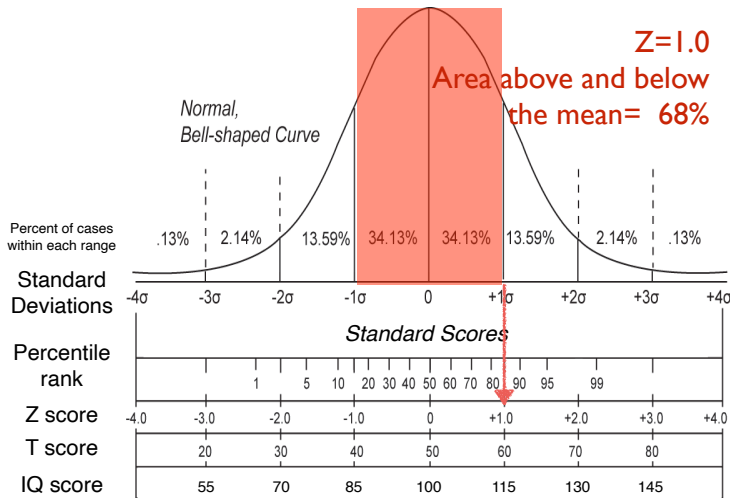
Psychology 402 - Fall 2023 - Dr. Michael Dohr

Common Z scores & Confidence Levels

Z Score	Area above mean	Area above + below Mean	Proportion as %
0.00	0.000		0%
0.13	0.051		
0.67	0.249		
1.00	0.341	0.682	68%
1.64	0.449		
1.96	0.475		95%
2.57	0.495		

640

Psychology 402 - Fall 2023 - Dr. Michael Dohr



641

Psychology 402 - Fall 2023 - Dr. Michael Dohr

How reliable?

- $r = .70$ or $.80$ or higher is often considered “good enough” for much research
- $r > .90$ is very good
- may not be worth effort to go higher
- Some real-world tests have $r > 0.9$
- example: modern IQ tests

642

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Increasing Reliability

643

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Increasing Reliability

- **Increase N** (number of questions, items or tests)...
- **Focus** on common characteristic...
- Other methods (covered later)
 - Use **Item Analysis** (“discriminability analysis”) to find items that best measure a single characteristic
 - Use **Factor Analysis** to determine sub-characteristics of a single test

644

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Increase N

- N = number of questions or items or tests
- Formula: increase N to increase reliability
- $N_d = r_d (1 - r_o) / r_o (1 - r_d)$
 N_d = new N (times old N)
 r_d = desired level of reliability
 r_o = observed level of reliability

645

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Increase N - Examples

- $N_d = r_d (1 - r_o) / r_o (1 - r_d)$
- Example:
 - 20-item CES-D has reliability of $.87$.
 - We need $r = 0.95$
 - $N_d = 2.82$
 - new N is $2.82 \times 20 = 56$ items

646

Psychology 402 - Fall 2023 - Dr. Michael Dohr

Increase N - Examples

- $N_d = r_d (1 - r_o) / r_o (1 - r_d)$
- Your 40-item test has reliability of .50.
- You want .90.
- $N_d = 9.0$
- new N is $9 \times 40 = 360!$

647

Psychology 402 - Fall 2023 - Dr. Michael Dieftr

Focus Test

- Reliability increases as a test focuses on a single concept or characteristic (“construct”)
- Trying to capture multiple concepts in a single test reduces reliability
- Methods:
 - Informal – remove items with poor face validity (chapter 5)
 - Statistical:
 - Discriminability Analysis (chapter 6)
 - Factor Analysis (chapter 13)

648

Psychology 402 - Fall 2023 - Dr. Michael Dieftr

Reliability Summary

- Measurement Error occurs in all fields -- Psychology focuses on it
- Kind of Reliability : *where* the error came from
- Improving Reliability: more items, focusing test, discriminability, factor analysis
- Reliability is useful: calculate SEM to get Confidence Intervals
- Reliability is not Validity: Reliable tests aren't automatically valid
- A reliable test *might* be valid

649

Psychology 402 - Fall 2023 - Dr. Michael Dieftr