Reliability: Theoretical Basis

PSYC3302: Psychological Measurement and Its Applications

Mark Hurlstone
Univeristy of Western Australia

Week 3

Learning Objectives

- Introduction to the theoretical basis of reliability:
- Classical Test Theory
  - observed score, true score, and error score variability
  - four conceptions of reliability
  - standard error of measurement
  - parallel tests

Reliability

- Reliability refers to the *consistency* of a measuring tool
  - the precision with which the test measures
  - the degree to which error is present in the measurement
- For example, suppose we want to test the reliability of three measuring scales (Scale A, B, & C)
- Using each scale, we weigh, on three separate occasions, a gold bar certified to weigh exactly 1000 grams

Reliability: An Intuitive Example

Gold Bar = 1000 grams

	Scale A	Scale B	Scale C
1st Measurement	1000 grams	1300 grams	1100 grams
2nd Measurement	1000 grams	1300 grams	800 grams
3rd Measurement	1000 grams	1300 grams	1200 grams

Reliability: An Intuitive Example

Gold Bar = 1000 grams

Reliable

	Scale A	Scale B	Scale C
1st Measurement	1000 grams	1300 grams	1100 grams
2nd Measurement	1000 grams	1300 grams	800 grams
3rd Measurement	1000 grams	1300 grams	1200 grams

Reliability: An Intuitive Example

Gold Bar = 1000 grams

	Scale A	Scale B	Scale C
	Reliable	Reliable	
1st Measurement	1000 grams	1300 grams	1100 grams
2nd Measurement	1000 grams	1300 grams	800 grams
3rd Measurement	1000 grams	1300 grams	1200 grams

Reliability: An Intuitive Example

Gold Bar = 1000 grams

	Reliable	Reliable	Unreliable
	Scale A	Scale B	Scale C
1st Measurement	1000 grams	1300 grams	1100 grams
2nd Measurement	1000 grams	1300 grams	800 grams
3rd Measurement	1000 grams	1300 grams	1200 grams

Reliability

- Whether we are measuring gold bars, behaviour, or anything else, unreliable measurement is to be avoided
- We want to know that a measuring tool or test we are using is reasonably consistent
- Reliability is not an all-or-nothing thing—it exists on a continuum
- That is, a measuring tool or test will be more or less reliable

Reliability

- In every day usage, reliability connotes something positive
- In psychometrics, it only refers to something that is consistent
- Not necessarily consistently "good" or "bad"
- Just consistent

Classical Test Theory (CTT)

- Classical Test Theory (CTT; also known as *True Score Theory*) is a theory of measurement that defines the theoretical basis for reliability
- It also outlines procedures for estimating the reliability of psychological measures
- According to CTT, a test's reliability reflects the extent to which differences in respondent's test scores are a function of their true psychological differences, as opposed to measurement error

Classical Test Theory

- CTT assumes that each respondent's *observed score* on a psychological test reflects the sum of two components:
  1. their *true score* on the psychological characteristic being measured
  2. *measurement error*
- The true score is the actual amount of the psychological characteristic being measured by a test that a respondent possesses
- Measurement error refers to the component of the observed score that does not have to do with the psychological characteristic being measured

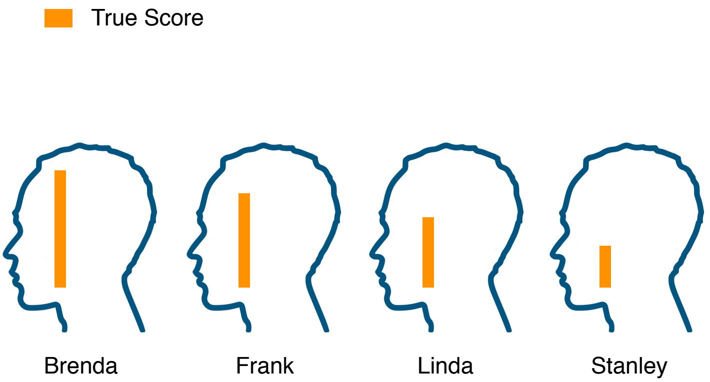
Classical Test Theory

- According to CTT, reliability is the degree to which *differences* in respondents' observed scores on a test are consistent with *differences* in their true scores
- More specifically, it is the extent to which differences in respondents' observed scores are attributable to differences in their true scores, as opposed to measurement error
- Measurement error creates inconsistency between observed scores and true scores
- When measuring psychological attributes—or anything else—the results of the measurement will always be unreliable to some degree

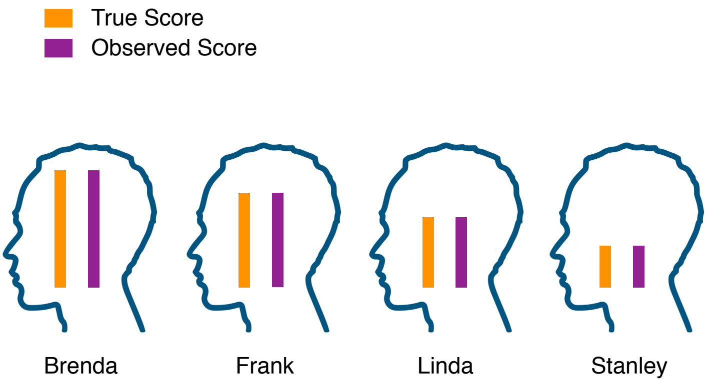
Working Memory Example

- Before discussing CTT in more depth, let's consider an example to illustrate these ideas
- Suppose I want to compare the working memory abilities of four people—Brenda, Frank, Linda, and Stanley
- To measure their working memory, I administer the Operation Span task described in our Week 1 lecture

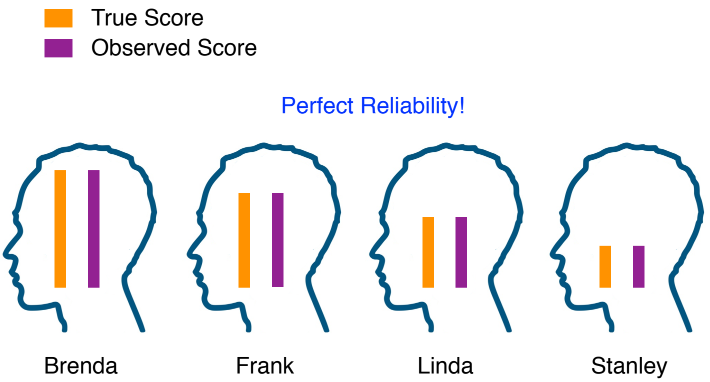
Working Memory Example



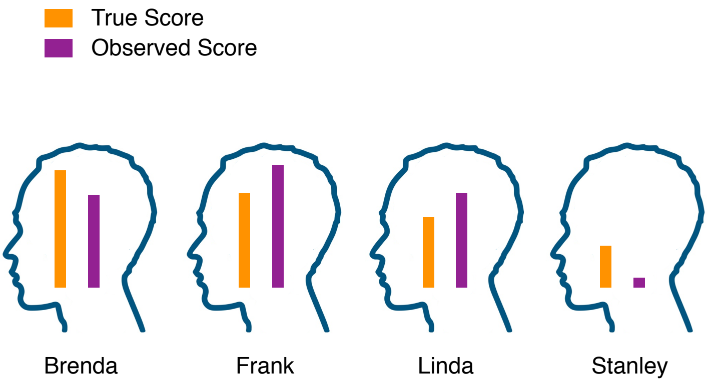
Working Memory Example



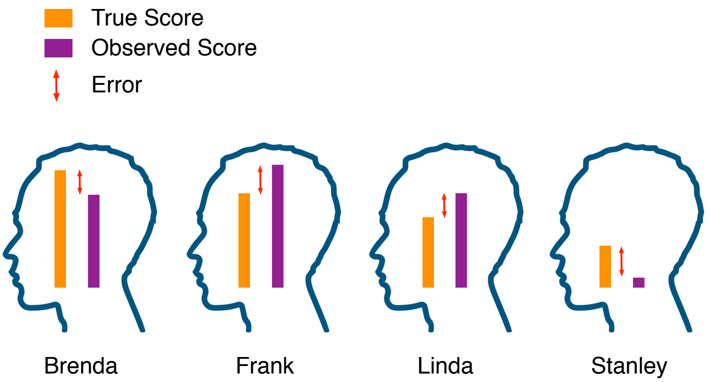
Working Memory Example



Working Memory Example



Working Memory Example



Sources of Measurement Error

There are various sources of Measurement Error:

1. Test construction
   - item sampling (variation among items in a test)
   - content sampling (variation among items between tests)
2. Test administration
   - test environment (temperature, lighting, noise)
   - events of the day (positive vs. negative events)
   - test-taker variables (physical discomfort, lack of sleep)
   - examiner-related variables (physical appearance & demeanour)
3. Test scoring and interpretation
   - subjectivity in scoring (grey area responses)
   - recording errors (technical glitches)

Classical Test Theory

- Reliability depends on two things:
  1. The extent to which differences in test scores can be attributed to real individual differences
  2. The extent to which differences in test scores are due to error

- A person's observed score on a test is that person's true score, plus error, which can be expressed as:

$$X_o = X_t + X_e, \tag{10}$$

- Where X_o represents a person's observed score, X_t represents a person's true score, and X_e represents error

Classical Test Theory

Table: Responses to a Self-Esteem Questionnaire

Respondent	Observed Score (X_o)		True Score (X_t)		Error Score (X_e)
Ashley	120	=	130	+	–10
Bob	145	=	120	+	25
Carl	95	=	110	+	–15
Denise	85	=	100	+	–15
Eric	115	=	90	+	25
Felicia	70	=	80	+	–10
Mean	105.00		105		0
Variance	608.33		291.67		316.67
Std. Dev	24.66		17.08		17.80
Reliability	$R_{xx} = .48$		$r_{ot} = .69$		$r_{oe} = .72$
	$r_{te} = .00$		$r^2_{ot} = .48$		$r^2_{oe} = .52$

Classical Test Theory

- This is an "all-knowing" example—we are pretending to know certain things that we don't actually know
- It assumes that we know a person's true score, which of course we do not—it is a hypothetical amount that cannot be directly observed
- The same is also true of measurement error
- The only amount that we do know is a respondent's observed score on a test or measurement

Classical Test Theory

- Several key assumptions of CTT are illustrated in this data set:
  1. Observed scores on a psychological measure are determined by a respondent's true scores plus measurement error
  2. Measurement error is random—it is just as likely to inflate a score as to deflate it
     - (A) error tends to cancel itself out across respondents ($\overline{X}_e = 0$)
     - (B) error scores are uncorrelated with true scores ($r_{te} = 0$)

Classical Test Theory

- Several key assumptions of CTT are illustrated in this data set:
  1. Observed scores on a psychological measure are determined by a respondent's true scores plus measurement error
  2. Measurement error is random—it is just as likely to inflate a score as to deflate it
     - (A) error tends to cancel itself out across respondents ($\overline{X}_e = 0$)
     - (B) error scores are uncorrelated with true scores ($r_{te} = 0$)

Table: Responses to a Self-Esteem Questionnaire

Respondent	Observed Score (X_o)		True Score (X_t)		Error Score (X_e)
Ashley	120	=	130	+	–10
Bob	145	=	120	+	25
Carl	95	=	110	+	–15
Denise	85	=	100	+	–15
Eric	115	=	90	+	25
Felicia	70	=	80	+	–10
Mean	105.00		105		0
Variance	608.33		291.67		316.67
Std. Dev	24.66		17.08		17.80
Reliability	$R_{xx} = .48$		$r_{ot} = .69$		$r_{oe} = .72$
	$r_{te} = .00$		$r^2_{ot} = .48$		$r^2_{oe} = .52$

Classical Test Theory

- Several key assumptions of CTT are illustrated in this data set:
  1. Observed scores on a psychological measure are determined by a respondent's true scores plus measurement error
  2. Measurement error is random—it is just as likely to inflate a score as to deflate it
     - (A) error tends to cancel itself out across respondents ($\overline{X}_e = 0$)
     - (B) error scores are uncorrelated with true scores ($r_{te} = 0$)

Classical Test Theory

- Several key assumptions of CTT are illustrated in this data set:
  1. Observed scores on a psychological measure are determined by a respondent's true scores plus measurement error
  2. Measurement error is random—it is just as likely to inflate a score as to deflate it
     - (A) error tends to cancel itself out across respondents ($\overline{X}_e = 0$)
     - (B) error scores are uncorrelated with true scores ($r_{te} = 0$)

Classical Test Theory

- Several key assumptions of CTT are illustrated in this data set:
  1. Observed scores on a psychological measure are determined by a respondent's true scores plus measurement error
  2. Measurement error is random—it is just as likely to inflate a score as to deflate it
     - A. error tends to cancel itself out across respondents ($\overline{X}_e = 0$)
     - B. error scores are uncorrelated with true scores ($r_{te} = 0$)

Classical Test Theory

Table: Responses to a Self-Esteem Questionnaire

Respondent	Observed Score (X_o)		True Score (X_t)		Error Score (X_e)
Ashley	120	=	130	+	–10
Bob	145	=	120	+	25
Carl	95	=	110	+	–15
Denise	85	=	100	+	–15
Eric	115	=	90	+	25
Felicia	70	=	80	+	–10
Mean	105.00		105		0
Variance	608.33		291.67		316.67
Std. Dev	24.66		17.08		17.80
Reliability	$R_{xx} = .48$		r_{ot} = .69		r_{oe} = .72
	r_{te} = .00		r^2_{ot} = .48		r^2_{oe} = .52

Classical Test Theory

- Several key assumptions of CTT are illustrated in this data set:
  1. Observed scores on a psychological measure are determined by a respondent's true scores plus measurement error
  2. Measurement error is random—it is just as likely to inflate a score as to deflate it
     - (A) error tends to itself cancel out across respondents ($\overline{X}_e = 0$)
     - (B) error scores are uncorrelated with true scores ($r_{te} = 0$)

Classical Test Theory

- Several key assumptions of CTT are illustrated in this data set:
  1. Observed scores on a psychological measure are determined by a respondent's true scores plus measurement error
  2. Measurement error is random—it is just as likely to inflate a score as to deflate it
     - (A) error tends to itself cancel out across respondents ($\overline{X}_e = 0$)
     - (B) error scores are uncorrelated with true scores ($r_{te} = 0$)

Classical Test Theory

Table: Responses to a Self-Esteem Questionnaire

Respondent	Observed Score (X_o)		True Score (X_t)		Error Score (X_e)
Ashley	120	=	130	+	–10
Bob	145	=	120	+	25
Carl	95	=	110	+	–15
Denise	85	=	100	+	–15
Eric	115	=	90	+	25
Felicia	70	=	80	+	–10
Mean	105.00		105		0
Variance	608.33		291.67		316.67
Std. Dev	24.66		17.08		17.80
Reliability	$R_{xx} = .48$		$r_{ot} = .69$		$r_{oe} = .72$
	$r_{te} = .00$		$r^2_{ot} = .48$		$r^2_{oe} = .52$

Classical Test Theory

- Several key assumptions of CTT are illustrated in this data set:
  1. Observed scores on a psychological measure are determined by a respondent's true scores plus measurement error
  2. Measurement error is random—it is just as likely to inflate a score as to deflate it
     - (A) error tends to itself cancel out across respondents ($\overline{X}_e = 0$)
     - (B) error scores are uncorrelated with true scores ($r_{te} = 0$)

Classical Test Theory (CTT)

- Reliability reflects the degree to which *differences* in observed scores are consistent with *differences* in true scores, as opposed to error
- Stated another way, it depends on the links between observed score, true score, and error score variability
- We can describe the variability of the three different types of scores using the variance s^2
- The relationship of the three variances can be expressed as:

$$s_o^2 = s_t^2 + s_e^2, \tag{11}$$

- Where s_o^2 is observed variance, s_t^2 is true variance, and s_e^2 is error variance

Four Conceptions of Reliability

- We can use the links between the observed score, true score, and error score variances to compute a reliability coefficient, denoted R_{xx}
- The reliability coefficient varies between 0 and 1
- Larger R_{xx} values indicate greater reliability
- If the true score variance is equal to observed score variance then $R_{xx} = 1$ and reliability is perfect
- This would indicate that there is no measurement error affecting observed scores

Four Conceptions of Reliability

- There is no clear cut-off separating "good" from "bad" reliability
- A perfect reliability of 1 will not occur—there will always be measurement error
- A reliability coefficient of .7 or .8 is acceptable for research purposes
- A reliability of .9 or greater is needed for applied purposes

Four Conceptions of Reliability

Table: A 2 × 2 Framework for Conceptualising Reliability

		Conceptual Basis of Reliability	
		True Scores	**Measurement Error**
Statistical Basis of Reliability	**Proportions of variance**	Reliability is the ratio of true score variance to observed score variance $R_{xx} = \frac{s_t^2}{s_o^2}$	Reliability is the lack of error variance $R_{xx} = 1 - \frac{s_e^2}{s_o^2}$
	Correlations	Reliability is the (squared) correlation between observed scores and true scores $R_{xx} = r_{ot}^2$	Reliability is the lack of correlation between observed scores and error scores $R_{xx} = 1 - r_{oe}^2$

Four Conceptions of Reliability

Table: A 2 × 2 Framework for Conceptualising Reliability

		Conceptual Basis of Reliability	
		True Scores	**Measurement Error**
Statistical Basis of Reliability	**Proportions of variance**	Reliability is the ratio of true score variance to observed score variance $R_{xx} = \frac{s_t^2}{s_o^2}$	Reliability is the lack of error variance $R_{xx} = 1 - \frac{s_e^2}{s_o^2}$
	Correlations	Reliability is the (squared) correlation between observed scores and true scores $R_{xx} = r_{ot}^2$	Reliability is the lack of correlation between observed scores and error scores $R_{xx} = 1 - r_{oe}^2$

Four Conceptions of Reliability

Table: A 2 × 2 Framework for Conceptualising Reliability

		Conceptual Basis of Reliability	
		True Scores	**Measurement Error**
Statistical Basis of Reliability	**Proportions of variance**	Reliability is the ratio of true score variance to observed score variance $R_{xx} = \frac{s_t^2}{s_o^2}$	Reliability is the lack of error variance $R_{xx} = 1 - \frac{s_e^2}{s_o^2}$
	Correlations	Reliability is the (squared) correlation between observed scores and true scores $R_{xx} = r_{ot}^2$	Reliability is the lack of correlation between observed scores and error scores $R_{xx} = 1 - r_{oe}^2$

Ratio of True Score Variance to Observed Score Variance

$$\text{Reminder:} \quad s_o^2 = s_t^2 + s_e^2,$$

- This is perhaps the most common way of expressing reliability
- Reliability is the proportion of observed score variance attributable to true score variance:

$$R_{xx} = \frac{s_t^2}{s_o^2}, \tag{12}$$

- For our example:

$$R_{xx} = \frac{291.67}{608.33} = .48,$$

Ratio of True Score Variance to Observed Score Variance

Reminder: $s_o^2 = s_t^2 + s_e^2,$

- This is perhaps the most common way of expressing reliability
- Reliability is the proportion of observed score variance attributable to true score variance:

$$R_{xx} = \frac{s_t^2}{s_o^2}, \tag{12}$$

- For our example:

$$R_{xx} = \frac{291.67}{608.33} = .48,$$

Ratio of True Score Variance to Observed Score Variance

$$\text{Reminder:} \quad s_o^2 = s_t^2 + s_e^2,$$

- This is perhaps the most common way of expressing reliability
- Reliability is the proportion of observed score variance attributable to true score variance:

$$R_{xx} = \frac{s_t^2}{s_o^2}, \tag{12}$$

- For our example:

$$R_{xx} = \frac{291.67}{608.33} = .48,$$

Table: Responses to a Self-Esteem Questionnaire

Respondent	Observed Score (X_o)		True Score (X_t)		Error Score (X_e)
Ashley	120	=	130	+	–10
Bob	145	=	120	+	25
Carl	95	=	110	+	–15
Denise	85	=	100	+	–15
Eric	115	=	90	+	25
Felicia	70	=	80	+	–10
Mean	105.00		105		0
Variance	608.33		291.67		316.67
Std. Dev	24.66		17.08		17.80
Reliability	$R_{xx} = .48$		$r_{ot} = .69$		$r_{oe} = .72$
	$r_{te} = .00$		$r^2_{ot} = .48$		$r^2_{oe} = .52$

Ratio of True Score Variance to Observed Score Variance

$$\text{Reminder:} \quad s_o^2 = s_t^2 + s_e^2,$$

- This is perhaps the most common way of expressing reliability
- Reliability is the proportion of observed score variance attributable to true score variance:

$$R_{xx} = \frac{s_t^2}{s_o^2}, \tag{12}$$

- For our example:

$$R_{xx} = \frac{291.67}{608.33} = .48,$$

Four Conceptions of Reliability

Table: A 2 × 2 Framework for Conceptualising Reliability

		Conceptual Basis of Reliability	
		True Scores	**Measurement Error**
Statistical Basis of Reliability	**Proportions of variance**	Reliability is the ratio of true score variance to observed score variance $R_{xx} = \frac{s_t^2}{s_o^2}$	Reliability is the lack of error variance $R_{xx} = 1 - \frac{s_e^2}{s_o^2}$
	Correlations	Reliability is the (squared) correlation between observed scores and true scores $R_{xx} = r_{ot}^2$	Reliability is the lack of correlation between observed scores and error scores $R_{xx} = 1 - r_{oe}^2$

Four Conceptions of Reliability

Table: A 2 × 2 Framework for Conceptualising Reliability

		Conceptual Basis of Reliability	
		True Scores	**Measurement Error**
Statistical Basis of Reliability	**Proportions of variance**	Reliability is the ratio of true score variance to observed score variance $R_{xx} = \frac{s_t^2}{s_o^2}$	Reliability is the lack of error variance $R_{xx} = 1 - \frac{s_e^2}{s_o^2}$
	Correlations	Reliability is the (squared) correlation between observed scores and true scores $R_{xx} = r_{ot}^2$	Reliability is the lack of correlation between observed scores and error scores $R_{xx} = 1 - r_{oe}^2$

Lack of Error Variance

$$\text{Reminder:} \quad s_o^2 = s_t^2 + s_e^2,$$

- Reliability is conceived as the degree to which measurement error is minimised in comparison with the variance of observed scores:

$$R_{xx} = 1 - \frac{s_e^2}{s_o^2}, \tag{13}$$

- For our example:

$$R_{xx} = 1 - \frac{316.67}{608.33} = 1 - .52 = .48,$$

Lack of Error Variance

Reminder: $s_o^2 = s_t^2 + s_e^2$,

- Reliability is conceived as the degree to which measurement error is minimised in comparison with the variance of observed scores:

$$R_{xx} = 1 - \frac{s_e^2}{s_o^2}, \tag{13}$$

- For our example:

$$R_{xx} = 1 - \frac{316.67}{608.33} = 1 - .52 = .48,$$

Lack of Error Variance

Reminder: $s_o^2 = s_t^2 + s_e^2,$

- Reliability is conceived as the degree to which measurement error is minimised in comparison with the variance of observed scores:

$$R_{xx} = 1 - \frac{s_e^2}{s_o^2}, \quad (13)$$

- For our example:

$$R_{xx} = 1 - \frac{316.67}{608.33} = 1 - .52 = .48,$$

Classical Test Theory

Table: Responses to a Self-Esteem Questionnaire

Respondent	Observed Score (X_o)		True Score (X_t)		Error Score (X_e)
Ashley	120	=	130	+	–10
Bob	145	=	120	+	25
Carl	95	=	110	+	–15
Denise	85	=	100	+	–15
Eric	115	=	90	+	25
Felicia	70	=	80	+	–10
Mean	105.00		105		0
Variance	608.33		291.67		316.67
Std. Dev	24.66		17.08		17.80
Reliability	$R_{xx} = .48$		$r_{ot} = .69$		$r_{oe} = .72$
	$r_{te} = .00$		$r^2_{ot} = .48$		$r^2_{oe} = .52$

Lack of Error Variance

$$\text{Reminder:} \quad s_o^2 = s_t^2 + s_e^2,$$

- Reliability is conceived as the degree to which measurement error is minimised in comparison with the variance of observed scores:

$$R_{xx} = 1 - \frac{s_e^2}{s_o^2}, \tag{13}$$

- For our example:

$$R_{xx} = 1 - \frac{316.67}{608.33} = 1 - .52 = .48,$$

Four Conceptions of Reliability

Table: A 2 $\times$ 2 Framework for Conceptualising Reliability

		Conceptual Basis of Reliability	
		True Scores	**Measurement Error**
Statistical Basis of Reliability	**Proportions of variance**	Reliability is the ratio of true score variance to observed score variance $R_{xx} = \frac{s_t^2}{s_o^2}$	Reliability is the lack of error variance $R_{xx} = 1 - \frac{s_e^2}{s_o^2}$
	Correlations	Reliability is the (squared) correlation between observed scores and true scores $R_{xx} = r_{ot}^2$	Reliability is the lack of correlation between observed scores and error scores $R_{xx} = 1 - r_{oe}^2$

Four Conceptions of Reliability

Table: A 2 × 2 Framework for Conceptualising Reliability

		Conceptual Basis of Reliability	
		True Scores	**Measurement Error**
Statistical Basis of Reliability	**Proportions of variance**	Reliability is the ratio of true score variance to observed score variance $R_{xx} = \frac{s_t^2}{s_o^2}$	Reliability is the lack of error variance $R_{xx} = 1 - \frac{s_e^2}{s_o^2}$
	Correlations	Reliability is the (squared) correlation between observed scores and true scores $R_{xx} = r_{ot}^2$	Reliability is the lack of correlation between observed scores and error scores $R_{xx} = 1 - r_{oe}^2$

The (Squared) Correlation Between Observed Scores and True Scores

$$\text{Reminder:} \quad s_o^2 = s_t^2 + s_e^2,$$

- Reliability is conceived as the (squared) correlation between observed scores and true scores:

$$R_{xx} = r_{ot}^2, \tag{14}$$

- For our example:

$$R_{xx} = .69^2 = .48$$

The (Squared) Correlation Between Observed Scores and True Scores

$$\text{Reminder:} \quad s_o^2 = s_t^2 + s_e^2,$$

- Reliability is conceived as the (squared) correlation between observed scores and true scores:

$$R_{xx} = r_{ot}^2, \tag{14}$$

- For our example:

$$R_{xx} = .69^2 = .48$$

The (Squared) Correlation Between Observed Scores and True Scores

$$\text{Reminder:} \quad s_o^2 = s_t^2 + s_e^2,$$

- Reliability is conceived as the (squared) correlation between observed scores and true scores:

$$R_{xx} = r_{ot}^2, \tag{14}$$

- For our example:

$$R_{xx} = .69^2 = .48$$

Table: Responses to a Self-Esteem Questionnaire

Respondent	Observed Score (X_o)		True Score (X_t)		Error Score (X_e)
Ashley	120	=	130	+	–10
Bob	145	=	120	+	25
Carl	95	=	110	+	–15
Denise	85	=	100	+	–15
Eric	115	=	90	+	25
Felicia	70	=	80	+	–10
Mean	105.00		105		0
Variance	608.33		291.67		316.67
Std. Dev	24.66		17.08		17.80
Reliability	$R_{xx} = .48$		$r_{ot} = .69$		$r_{oe} = .72$
	$r_{te} = .00$		$r^2_{ot} = .48$		$r^2_{oe} = .52$

The (Squared) Correlation Between Observed Scores and True Scores

Reminder: $s_o^2 = s_t^2 + s_e^2,$

- Reliability is conceived as the (squared) correlation between observed scores and true scores:

$$R_{xx} = r_{ot}^2, \tag{14}$$

- For our example:

$$R_{xx} = .69^2 = .48$$

Four Conceptions of Reliability

Table: A 2 × 2 Framework for Conceptualising Reliability

		Conceptual Basis of Reliability	
		True Scores	**Measurement Error**
Statistical Basis of Reliability	**Proportions of variance**	Reliability is the ratio of true score variance to observed score variance $R_{xx} = \frac{s_t^2}{s_o^2}$	Reliability is the lack of error variance $R_{xx} = 1 - \frac{s_e^2}{s_o^2}$
	Correlations	Reliability is the (squared) correlation between observed scores and true scores $R_{xx} = r_{ot}^2$	Reliability is the lack of correlation between observed scores and error scores $R_{xx} = 1 - r_{oe}^2$

Four Conceptions of Reliability

Table: A 2 × 2 Framework for Conceptualising Reliability

		Conceptual Basis of Reliability	
		True Scores	**Measurement Error**
Statistical Basis of Reliability	**Proportions of variance**	Reliability is the ratio of true score variance to observed score variance $R_{xx} = \frac{s_t^2}{s_o^2}$	Reliability is the lack of error variance $R_{xx} = 1 - \frac{s_e^2}{s_o^2}$
	Correlations	Reliability is the (squared) correlation between observed scores and true scores $R_{xx} = r_{ot}^2$	Reliability is the lack of correlation between observed scores and error scores $R_{xx} = 1 - r_{oe}^2$

Lack of (Squared) Correlation Between Observed Scores and Error Scores

$$\text{Reminder:} \quad s_o^2 = s_t^2 + s_e^2,$$

- Reliability is conceptualised as the absence of (squared) correlation between observed scores and error scores:

$$R_{xx} = 1 - r_{oe}^2, \tag{15}$$

- For our example:

$$R_{xx} = 1 - (.72)^2 = 1 - .52 = .48$$

Lack of (Squared) Correlation Between Observed Scores and Error Scores

$$\text{Reminder:} \quad s_o^2 = s_t^2 + s_e^2,$$

- Reliability is conceptualised as the absence of (squared) correlation between observed scores and error scores:

$$R_{xx} = 1 - r_{oe}^2, \tag{15}$$

- For our example:

$$R_{xx} = 1 - (.72)^2 = 1 - .52 = .48$$

Lack of (Squared) Correlation Between Observed Scores and Error Scores

$$\text{Reminder:} \quad s_o^2 = s_t^2 + s_e^2,$$

- Reliability is conceptualised as the absence of (squared) correlation between observed scores and error scores:

$$R_{xx} = 1 - r_{oe}^2, \tag{15}$$

- For our example:

$$R_{xx} = 1 - (.72)^2 = 1 - .52 = .48$$

Classical Test Theory

Table: Responses to a Self-Esteem Questionnaire

Respondent	Observed Score (X_o)		True Score (X_t)		Error Score (X_e)
Ashley	120	=	130	+	–10
Bob	145	=	120	+	25
Carl	95	=	110	+	–15
Denise	85	=	100	+	–15
Eric	115	=	90	+	25
Felicia	70	=	80	+	–10
Mean	105.00		105		0
Variance	608.33		291.67		316.67
Std. Dev	24.66		17.08		17.80
Reliability	$R_{xx} = .48$		$r_{ot} = .69$		$r_{oe} = .72$
	$r_{te} = .00$		$r^2_{ot} = .48$		$r^2_{oe} = .52$

Lack of (Squared) Correlation Between Observed Scores and Error Scores

$$\text{Reminder:} \quad s_o^2 = s_t^2 + s_e^2,$$

- Reliability is conceptualised as the absence of (squared) correlation between observed scores and error scores:

$$R_{xx} = 1 - r_{oe}^2, \tag{15}$$

- For our example:

$$R_{xx} = 1 - (.72)^2 = 1 - .52 = .48$$

Four Conceptions of Reliability

- The book provides a detailed discussion of the links between the four conceptions of reliability
- It also includes mathematical "proofs" of their relationships to one another
- You don’t need to know the proofs—you just need to remember the four formulas and be able to perform the reliability calculations

Standard Error of Measurement

- The reliability coefficient R_{xx} does not directly reflect the size of measurement error associated with a test
- By contrast, the standard deviation of error scores tell us in "test score units" the average size of error scores we can expect to find when a test is administered to a group of people
- For our example, the standard deviation of error scores is 17.80
- The standard deviation of error is also known as the *standard error of measurement* and is a crucially important concept in psychometrics

Standard Error of Measurement

- The reliability coefficient R_{xx} does not directly reflect the size of measurement error associated with a test
- By contrast, the standard deviation of error scores tell us in "test score units" the average size of error scores we can expect to find when a test is administered to a group of people
- For our example, the standard deviation of error scores is 17.80
- The standard deviation of error is also known as the *standard error of measurement* and is a crucially important concept in psychometrics

Classical Test Theory

Table: Responses to a Self-Esteem Questionnaire

Respondent	Observed Score (X_o)		True Score (X_t)		Error Score (X_e)
Ashley	120	=	130	+	–10
Bob	145	=	120	+	25
Carl	95	=	110	+	–15
Denise	85	=	100	+	–15
Eric	115	=	90	+	25
Felicia	70	=	80	+	–10
Mean	105.00		105		0
Variance	608.33		291.67		316.67
Std. Dev	24.66		17.08		17.80
Reliability	$R_{xx} = .48$		$r_{ot} = .69$		$r_{oe} = .72$
	$r_{te} = .00$		$r_{ot}^2 = .48$		$r_{oe}^2 = .52$

Standard Error of Measurement

- The reliability coefficient R_{xx} does not directly reflect the size of measurement error associated with a test
- By contrast, the standard deviation of error scores tell us in "test score units" the average size of error scores we can expect to find when a test is administered to a group of people
- For our example, the standard deviation of error scores is 17.80
- The standard deviation of error is also known as the *standard error of measurement* and is a crucially important concept in psychometrics

Standard Error of Measurement

- The standard error of measurement se_m is related to the reliability coefficient R_{xx}
- If we know the value of R_{xx} and the standard deviation of observed scores σ_o then we can calculate se_m as follows:

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

- For our example:

$$se_m = 24.66\sqrt{1 - .48} = 24.66(.72) = 17.80$$

Standard Error of Measurement

- The standard error of measurement se_m is related to the reliability coefficient R_{xx}
- If we know the value of R_{xx} and the standard deviation of observed scores σ_o then we can calculate se_m as follows:

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

- For our example:

$$se_m = 24.66\sqrt{1 - .48} = 24.66(.72) = 17.80$$

Table: Responses to a Self-Esteem Questionnaire

Respondent	Observed Score (X_o)		True Score (X_t)		Error Score (X_e)
Ashley	120	=	130	+	–10
Bob	145	=	120	+	25
Carl	95	=	110	+	–15
Denise	85	=	100	+	–15
Eric	115	=	90	+	25
Felicia	70	=	80	+	–10
Mean	105.00		105		0
Variance	608.33		291.67		316.67
Std. Dev	24.66		17.08		17.80
Reliability	$R_{xx} = .48$		$r_{ot} = .69$		$r_{oe} = .72$
	$r_{te} = .00$		$r^2_{ot} = .48$		$r^2_{oe} = .52$

Standard Error of Measurement

- The standard error of measurement se_m is related to the reliability coefficient R_{xx}
- If we know the value of R_{xx} and the standard deviation of observed scores σ_o then we can calculate se_m as follows:

$$se_m = \sigma_o \sqrt{1 - R_{xx}} \qquad (16)$$

- For our example:

$$se_m = 24.66\sqrt{1 - .48} = 24.66(.72) = 17.80$$

Standard Error of Measurement

- The standard error of measurement se_m is related to the reliability coefficient R_{xx}
- If we know the value of R_{xx} and the standard deviation of observed scores σ_o then we can calculate se_m as follows:

$$se_m = \sigma_o \sqrt{1 - R_{xx}} \qquad (16)$$

- For our example:

$$se_m = 24.66\sqrt{1 - .48} = 24.66(.72) = 17.80$$

Classical Test Theory

Table: Responses to a Self-Esteem Questionnaire

Respondent	Observed Score (X_o)		True Score (X_t)		Error Score (X_e)
Ashley	120	=	130	+	–10
Bob	145	=	120	+	25
Carl	95	=	110	+	–15
Denise	85	=	100	+	–15
Eric	115	=	90	+	25
Felicia	70	=	80	+	–10
Mean	105.00		105		0
Variance	608.33		291.67		316.67
Std. Dev	24.66		17.08		17.80
Reliability	$R_{xx} = .48$		$r_{ot} = .69$		$r_{oe} = .72$
	$r_{te} = .00$		$r^2_{ot} = .48$		$r^2_{oe} = .52$

Parallel Tests

- Given that we do not know people's true scores or the degree of measurement error, reliability cannot be estimated directly
- However estimates can be obtained by other means
- One way is by constructing a so called *parallel test*
- For example, we could construct two different measures of self esteem

Parallel Tests

- To qualify as a parallel test, the previous assumptions of CTT must be satisfied for each test, and the following must be true:
  1. participants true scores for one test must be exactly equal to their true scores on the other test—known as "tau equivalence"
  2. the tests must have the same level of error variance
- In other words, the observed scores on both tests should have the same mean and standard deviation

Parallel Tests

- If two tests are parallel, we can compute a correlation between the scores on the two tests
- According to CTT, *the correlation between parallel test scores is equal to the reliability* (see the textbook for a proof)
- The parallel test assumption therefore provides an important bridge to the real world of testing
- It allows us to use procedures to estimate reliability in real-life testing situations

Parallel Tests

- In practice, parallel tests are difficult to come by
- Two measures of the same construct rarely have the same mean and standard deviation—a necessary pre-requisite for parallel tests
- Accordingly, researchers often use a measure of internal consistency reliability known as Cronbach's α
- We will discuss this—and other methods of generating empirical estimates of reliability—in the second half of this lecture

References

Furr, M. R., & Bacharach, V. R. (2014; Chapter 5). Psychometrics: An Introduction (second edition). Sage.