PSYC30013
Research Methods for Human Inquiry

Week 11 (Day 2): Psychological Assessment:
Principles and applications of reliability in Psychological Assessment

Week 11: Day 2

Aim of the Psychological Assessment lectures

- To introduce you to important issues in relation to psychological assessment and testing.
  - Most importantly, issues of reliability and validity in the measurement of psychological constructs.

Agenda for today's lecture

11.1. Definitions: psychological assessment, validity, reliability

11.2. Classical Test Theory and the theory of reliability

11.3. Estimating reliability in practice

11.4. The Standard Error of Measurement, predicted true scores, and building true score confidence intervals

(11.5.) Making tests more reliable: the Spearman-Brown Prophecy formula ← Held over for next week

Psychological Assessment Defined

From the Australian Psychological Society (2018):
Ethical guidelines for psychological assessment and the use of psychological tests

From Section 1, “Introduction”:

1.1. Psychological assessment encompasses multiple sources of data generated from one or more assessment methods, such as **psychological tests**; behavioural observation; structured, semi-structured and clinical interviews; questionnaires; rating scales; checklists; behavioural simulations or games; and other structured and evidence-based approaches to gathering information about clients

Psychological Research vs Assessment

Psychological ***research*** seeks to make generalisations about representative samples of people.

Psychological ***assessment*** is a special case of psychological research which seeks to make generalisations about specific individuals, that is, with samples of $n = 1$.

Psychological Assessment Standards

From the Australian Psychological Society (2018):
Ethical guidelines for psychological assessment and the use of psychological tests

From Section 2, “Competence”:

2.2. Psychologists appreciate that the following broad areas of knowledge and understanding underpin competent use of psychological tests:

a) the psychological theory underpinning the test;

b) **the nature of the construct(s) underlying a test score**, which is essential to the way in which inferences are to be drawn from test results;

c) **basic psychometric principles and procedures**, and the technical requirements of tests;

d) **the technical properties and limitations** of the particular instrument or instruments used; and

e) the context in which the test is being used (e.g., for clinical diagnosis, forensic/legal purposes, personal/ relationship counselling, school achievement, personnel selection, diagnosis of brain functioning) in order to integrate the test results with other pertinent information.

Psychological Assessment Standards

From the Australian Psychological Society (2018):
Ethical guidelines for psychological assessment and the use of psychological tests

From Section 2, “Competence”:

2.3. Psychologists ensure that any test used as part of a formal psychological assessment:

a) has clear directions for administration and scoring, and adequate information about the properties of scores derived from the test – including the purpose of the test, **the relevant standard errors, and validity and reliability data**;

b) is valid for the purpose for which the test is being used, and is also valid for any sub-population of the total population to be included in the particular testing program (e.g., sub-populations defined according to age, gender, ethnicity, language background or socioeconomic status); and

c) has appropriate normative or reference group data to allow for the interpretation of scores in relation to a clearly defined population.

Validity and reliability

- A test is valid if it accurately measures what it purports to measure (more discussion of validity next week)

- A test is reliable if has the property of consistency in measurement.
  - Often judged by seeing if test-takers obtain consistent scores when they're re-examined
    - With the same test under repeated conditions
    - With different sets of equivalent items (Anastasi & Urbina, 1997)

- Common claim: A test ***must*** be reliable in order for it to be valid. However, not all reliable tests are valid.

11.2. Classical Test Theory and the theory of reliability

Fundamental equations of Classical Test Theory (1)

$$x_i = \tau + \epsilon_i$$

Where

- x_i is the observed score x obtained on test occasion *i*
- τ is the person's true score
- ϵ_i is error, unsystematic variance (everything besides the true score) on test occasion *i*
- Error could be
  - Endogenous to test-taker (e.g. client had a bad dream the night before taking the test)
  - Exogenous to test-taker (e.g. psychologist made mistake recording response to a test item)

Assumptions of Classical Test Theory

- Expected value of the error is zero
- Errors do not correlate with each other
- Errors do not correlate with true scores
- Expected value of the test is equal to the true score

An example test

Respondent	Observed Score (x)		True Score (τ)		Error (ε)
Aaron	120	=	130	+	-10
Blake	145	=	120	+	25
Carl	95	=	110	+	-15
Denise	85	=	100	+	-15
Eric	115	=	90	+	25
Felicia	70	=	80	+	-10
MEAN	**105**		**105**		**0**
VARIANCE	$\sigma_x^2 =$ **608.33**		$\sigma_\tau^2 =$ **291.67**		$\sigma_\epsilon^2 =$ **316.67**
STD DEV	$\sigma_x =$ **24.66**		$\sigma_\tau =$ **17.08**		$\sigma_\epsilon =$ **17.80**

Correlations		
$r_{\tau\varepsilon}$	=	.00
$r_{x\tau}$	=	.69
$r_{x\tau}^2$	=	.48
$r_{x\varepsilon}$	=	.72
$r_{x\varepsilon}^2$	=	.52

Table adapted from Furr (2021)

Fundamental equations of Classical Test Theory (2)

$$\sigma_x^2 = \sigma_\tau^2 + \sigma_\epsilon^2 + 2Cov(\tau, \epsilon)$$

Where

- σ_x^2 is the variance of observed test results
- σ_τ^2 is true score variance
- σ_ϵ^2 is error variance
- $2Cov(\tau, \epsilon)$ cancels out due to the assumption that errors do not correlate with true scores

Fundamental equations of Classical Test Theory (3)

$$Reliability = r_{x\tau}^2 = \frac{\sigma_\tau^2}{\sigma_x^2} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\epsilon^2} = \frac{Signal}{Signal + Noise}$$

Where

- $r_{x\tau}^2$ is the theoretical reliability coefficient
- σ_x^2 is the variance of observed test results
- σ_τ^2 is true score variance
- σ_ϵ^2 is error variance
- In practice estimate $r_{x\tau}^2$ using the sample reliability coefficient r_{xx}

Four perspectives on reliability

		Conceptual basis of reliability: Observed scores in relation to…	
		True scores	Measurement error
Statistical basis of reliability in terms of…	Proportions of variance	Reliability is the ratio of true score variance to observed score variance	Reliability is the lack of error variance
		$\frac{\sigma_\tau^2}{\sigma_X^2}$	$1 - \frac{\sigma_\epsilon^2}{\sigma_X^2}$
	Correlations	Reliability is the squared correlation between observed scores and true scores	Reliability is the lack of squared correlation between observed scores and error scores
		$r_{X\tau}^2$	$1 - r_{X\epsilon}^2$

Table adapted from Furr (2021)

11.3. Estimating reliability in practice

Ways of estimating reliability

- Test-retest reliability
- Alternate-forms reliability
- Split-half reliability
- Cronbach's α (or alternatives)

Test-retest reliability

Estimate reliability by obtaining scores from the original test and a retest, and work out the correlation between them.

Problems:

- Can work well for stable traits, but will work poorly for transitory states
- Carryover effects (e.g. participants googled test answers between the test and retest, participants get bored doing the test a second time and slack off)
- True score may change between test and retest
- Participants might fail to return for the retest

Alternate-forms reliability

Estimate reliability by obtaining scores from two alternate forms of the test, and work out the correlation between them.

Problems:

- Hard to know if two forms are really parallel (i.e. equivalent)
- Might not really fix the problem of carryover effects

Split-half reliability

Estimate reliability by splitting your test into two parallel subtests, and work out the correlation between the two subtests.

Problems:

- Assumption that the two subtests are parallel is often questionable
- The fact that each subtests is half the size of the main test will tend to deflate the reliability estimate.

Cronbach's α

Famous approach to reliability, popularized by Cronbach (1951)

- Loosely can be thought of as the mean of all possible split-half reliabilities, scaled up to the full test instead of a half-test.
- Historically popular, easy to implement in SPSS or R.
- Provides a *lower-bound* estimate for reliability.
- Sijtsma (2009) provides details on why it's of limited use.

More modern alternatives may be preferable, e.g. McDonald's ω (McDonald, 2013).

11.4. The Standard Error of Measurement, predicted true scores, and building true score confidence intervals

Standard Error of Measurement

The reliability coefficient r_{xx} is useful but doesn't tell us in test score units how much measurement error is 'typical'.

Standard error of measurement (SE_m) is used instead to represent the average error score.

Formula: $SE_m = \sigma_x\sqrt{1 - r_{xx}}$

(where σ_x is the standard deviation of observed scores)

Reliability vs Standard Error of Measurement

A close relationship exists between reliability (r_{xx}) and the Standard Error of Measurement (SE_m). Since $SE_m = \sigma_x\sqrt{1 - r_{xx}}$, when:

$r_{xx} = 0$, then $SE_m = \sigma_x \sqrt{(1 - 0)} = \sigma_x$

$r_{xx} = .2$, then $SE_m = \sigma_x \times .89$

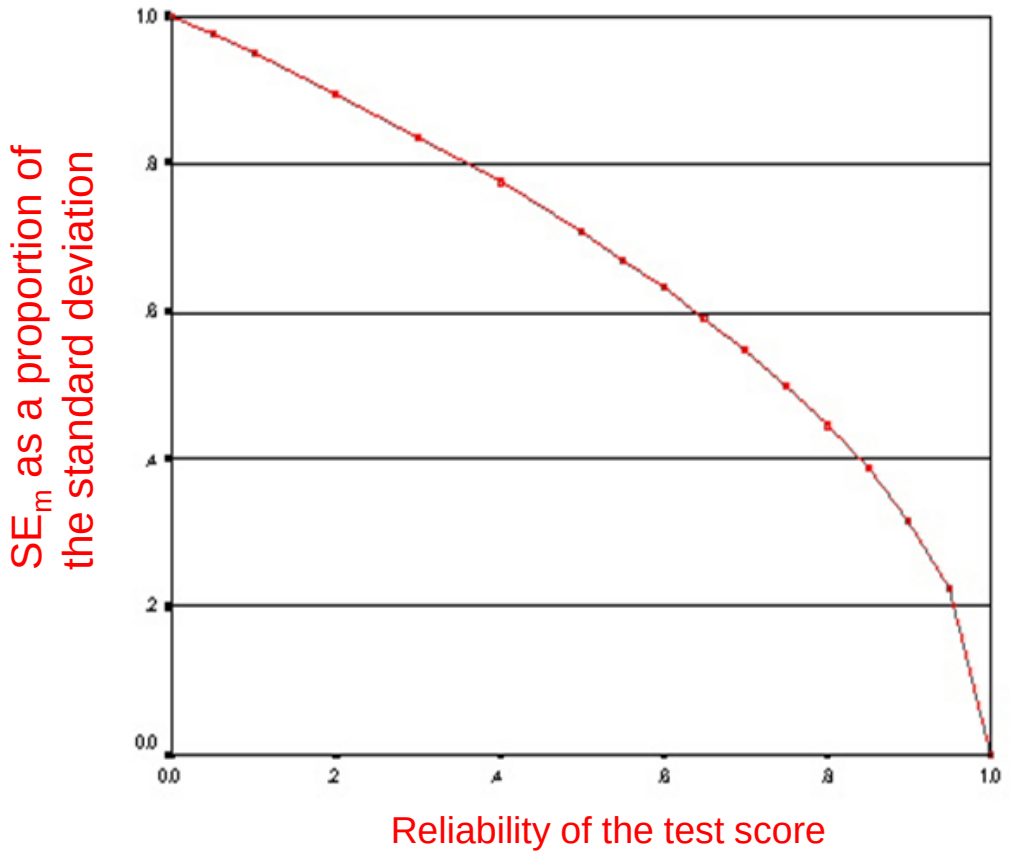
$r_{xx} = .4$, then $SE_m = \sigma_x \times .78$

$r_{xx} = .6$, then $SE_m = \sigma_x \times .63$

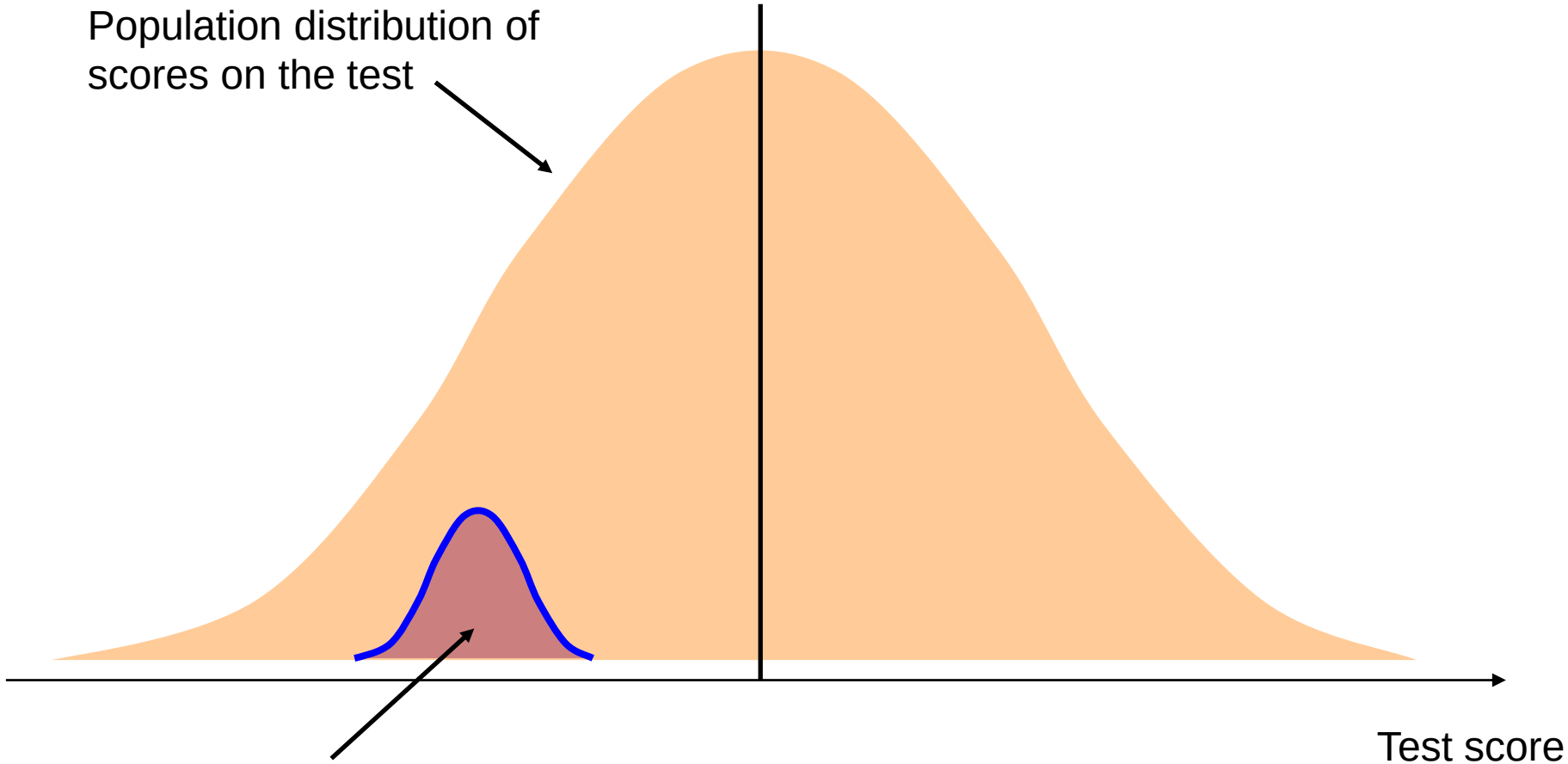
$r_{xx} = .8$, then $SE_m = \sigma_x \times .45$

$r_{xx} = 1$, then $SE_m = \sigma_x \sqrt{(1 - 1)} = 0$

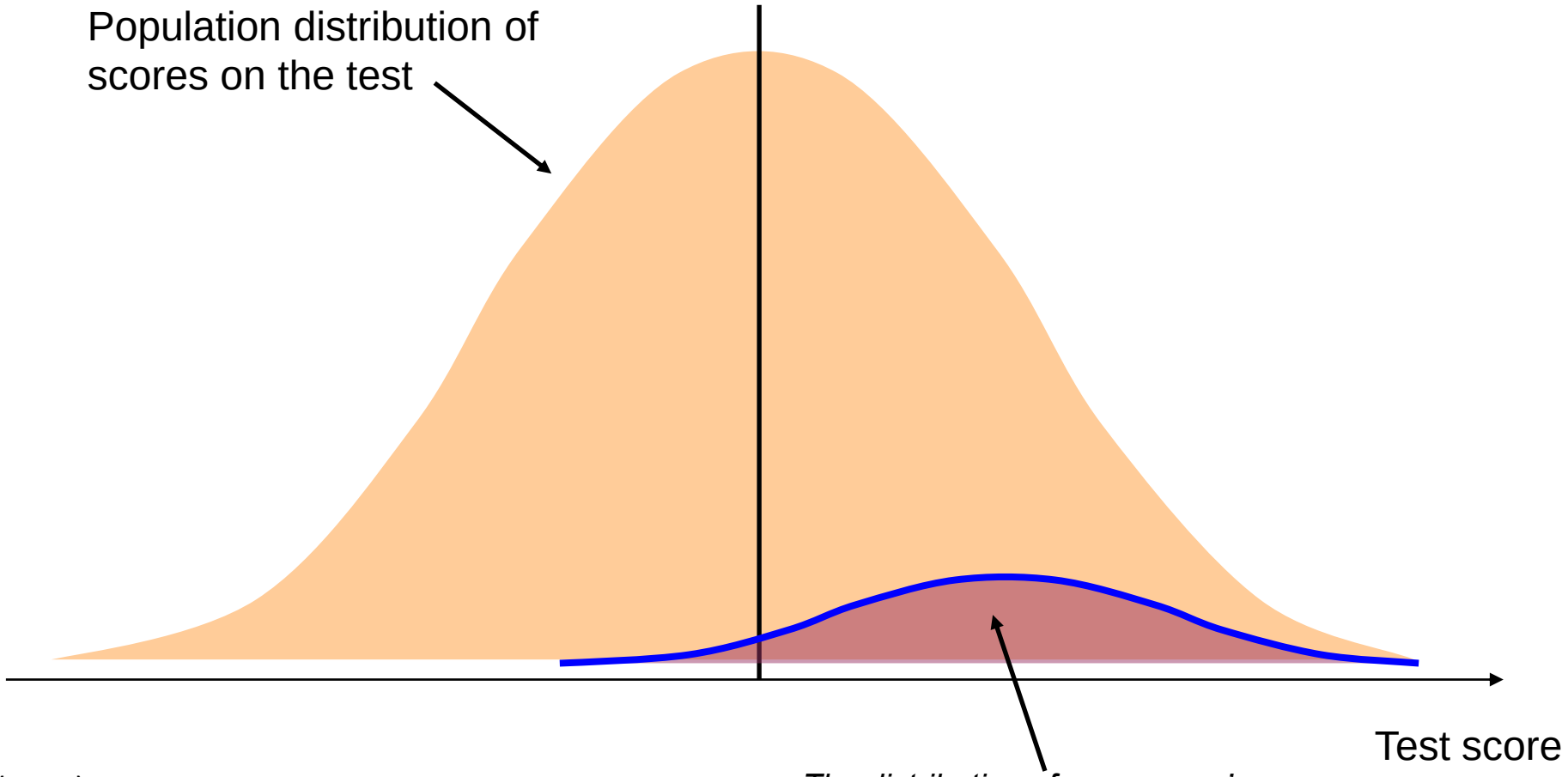
Graph of the relationship between r_{xx} and SE_m



Example of high reliability



Example of low reliability



Nunnally & Bernstein (1994):
"If important decisions are made with respect to specific test scores, a reliability of .90 is a bare minimum, and a reliability of .95 should be considered the desirable standard" (p. 265)

Predicted True Score

Given your client's observed test score x, the Kelley (1927) regression formula is used to predict their true score

$$\hat{\tau} = r_{xx} \times x + (1 - r_{xx}) \times \mu_{\tau}$$

Where

- $\hat{\tau}$ is the predicted true score
- r_{xx} is the reliability coefficient
- x is your client's observed score
- μ_{τ} is the is the population mean score for the test (e.g. 100 for an IQ test)

Predicted True Score: Worked Example

Your client scored 74 on an IQ test which has a r_{xx} of .96.

$$\hat{\tau} = r_{xx} \times x + (1 - r_{xx}) \times \mu_{\tau}$$
$$\hat{\tau} = 0.96 \times 74 + (1 - 0.96) \times 100$$
$$\hat{\tau} = 75.04$$

Where

- $\hat{\tau}$ is the predicted true score
- r_{xx} is the reliability coefficient
- x is your client's observed score
- μ_{τ} is the is the population mean score for the test (e.g. 100 for an IQ test)

Standard error of estimation

Confidence intervals for true scores can be built using the standard error of estimation

Similar to standard error of measurement, but slightly different

Standard error of estimation: $SE_e = \sigma_x \sqrt{r_{xx}(1 - r_{xx})}$

Standard error of measurement: $SE_m = \sigma_x \sqrt{1 - r_{xx}}$

Standard error of estimation: worked example

In your clinical practice you are using an IQ test which has a reliability (r_{xx}) of .96, and a standard deviation (σ_x) of 15

Standard error of estimation: $SE_e = \sigma_x\sqrt{r_{xx}(1 - r_{xx})}$

Standard error of estimation: $SE_e = 15\sqrt{.96(1 - .96)}$

Standard error of estimation: $SE_e = 15\sqrt{0.0384}$

Standard error of estimation: $SE_e = 15 \times 0.196$

Standard error of estimation: $SE_e = 2.94$

True score confidence interval

95% CI is defined as follows

Lower Bound: $\hat{\tau} - (1.96 \times SE_e)$

Upper Bound: $\hat{\tau} + (1.96 \times SE_e)$

Where

- $\hat{\tau}$ is the predicted true score
- 1.96 is chosen because 95% of observations lie with 1.96 standard deviations of the mean, assuming a normal distribution
- SE_e is the standard error of estimation

True score confidence interval: Worked example

Your client scored 74 on an IQ test which has a r_{xx} of .96.
You worked out that your client's predicted true score ($\hat{\tau}$) is 75.04.
You worked out the standard error of estimation (SE_e) is 2.94.

Lower Bound: $\hat{\tau} - (1.96 \times SE_e)$
Upper Bound: $\hat{\tau} + (1.96 \times SE_e)$

Lower Bound: $75.04 - (1.96 \times 2.94) = 75.04 - 5.76 = 69.27$
Upper Bound: $75.04 + (1.96 \times 2.94) = 75.04 + 5.76 = 80.80$

95% CI [69.27, 80.80]

References

Optional reading, not required for this course but potentially useful for your future career:

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Prentice Hall/Pearson Education (Chapter 4).

Australian Psychological Society. (2018). Ethical guidelines for psychological assessment and the use of psychological tests. Retrieved 15 May 2020 from https://www.psychology.org.au/for-members/resource-finder/resources/ethics/Ethical-guidelines-psychological-assessment-tests

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Furr, R. M. (2021). *Psychometrics: An introduction*. Los Angeles, CA: Sage.

Kelley, T. L. (1927). The interpretation of educational measurements. New York, NY: World Book.

McDonald, R. P. (2013). Test theory: A unified treatment. Psychology press.

Nunnally, J. C. , & Bernstein, I. H. (1994). *Psychometric theory* (Chapter 7). New York, NY: McGraw-Hill.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*, 199-223.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107-120.