Psychometric Barriers II: Test Bias

PSYC3302: Psychological Measurement and Its Applications

Mark Hurlstone
Univeristy of Western Australia

Week 8

Don't Panic! ...

... But we need to talk about final exams

Final Exam

- **125** multiple choice questions (five alternatives each)
- Time allowed: 2:00 hours
- Questions derived from the lecture slide materials
- All lectures will be represented in the exam
- Contains a mixture of "easy", "moderate", and "hard items"
- You won't need to perform any calculations
- However, you need to be able to recognise and distinguish formulas from the lecture slides

Final Exam

INSTRUCTIONS:
There is 1 section to this exam.

It includes **125** multiple choice questions that must be answered in pencil using the multiple choice sheet provided. For each question, choose the best alternative amongst those provided. In some cases, you may believe that none of the alternatives are correct. Alternatively, you may believe that there is an error in the question.

Nonetheless, in all cases, you should do your best to answer the question based on the information provided and move onto the next question.

Final Exam

- There will be a revision lecture on the content we have covered in Week 12
- There are example questions on LMS to illustrate the item formats you will encounter; click on the following link:
  - Example Question Formats

Data Set For Laboratory Report

- The data set for the laboratory report is accessible on LMS; click on the following link:
  - PSYC3302 Lab Report Data
- Important: Negatively keyed items have already been reverse coded
- To facilitate interpretation of your PCA, you will need to add the actual VCS items in the variable labels
- The previously uploaded Excel (.xls) file has been replaced with an SPSS (.sav) file

Learning Objectives

- Introduction to the problem of test score bias:
  - Construct Bias
  - Predictive Bias
- Methods for detecting construct bias:
  - Item Discrimination Index
  - Factor Analysis
  - Differential Item Functioning
  - Rank-Order Consistency
- Methods for detecting predictive bias:
  - Regression: Intercept Bias
  - Regression: Slope Bias
  - Regression: Intercept + Slope Bias

Test Bias

- It is important that test scores do not discriminate unjustifiably against any particular group
- In this context, "group" typically refers to gender and/or race
- In psychometrics, **test bias** arises when there is a factor inherent in a test that systematically prevents accurate, impartial measurement
- Systematic is a key word in this definition—(test) bias implies systematic (not random) variation

Example: Gender Differences in Mathematics

- Suppose you administer a math test to a group of males and females
- You find, on average, that males score higher than females
- This could indicate that males tend to have greater math ability than females
- However, it is also possible the test might be biased in some way:
  - the test may overestimate the true math ability of males
  - the test may underestimate the true math ability of females

Two Forms of Test Bias

- In this lecture, two types of test bias will be discussed:
  1. Construct bias:
     - biases in the meaning of a test
  2. Predictive bias:
     - biases in the use of a test

Two Forms of Test Bias

- In this lecture, two types of test bias will be discussed:
  1. Construct bias:
     - biases in the meaning of a test
  2. Predictive bias:
     - biases in the use of a test

Construct Bias

- Construct bias occurs when a test has different *meaning* for two groups in terms of the construct it measures
  - it concerns the relationship of observed scores to true scores on a psychological test
- If this relationship is systematically different for different groups, then we might conclude that the test is biased
- Construct bias can lead to situations in which two groups have the same average "true score" but different averages on "observed scores" on a test of the construct
- For example, males and females no doubt have the same math ability, but a bias in our math test might yield larger observed scores for males

Two Forms of Test Bias

- In this lecture, two types of test bias will be discussed:
  1. Construct bias:
     - biases in the meaning of a test
  2. Predictive bias:
     - biases in the use of a test

Two Forms of Test Bias

- In this lecture, two types of test bias will be discussed:
  1. Construct bias:
     - biases in the meaning of a test
  2. Predictive bias:
     - biases in the use of a test

Predictive Bias

- Predictive bias occurs when a test's *use* has different implications for two (or more) groups
  - it concerns the relationship between scores on two different tests
- An example might be the The Australian Tertiary Admission Rank (ATAR) used to predict first year university performance
- Suppose it were discovered that the ATAR was more predictive of academic performance for males than for females
- This would suggest that the ATAR suffers from predictive bias

Bear in Mind ...

- The two types of bias are independent
- A test might have no construct bias but might suffer from predictive bias
- Conversely, a test might have no predictive bias but suffer from construct bias

Identifying Test Score Bias

- There are least two categories of procedures that can be used to identify test score bias:
  1. **Internal** methods to identify construct bias
  2. **External** methods to identify predictive bias
- There is no single method that can be used to establish bias
- This is analogous to establishing the reliability and validity of test scores
- There are, however, various accepted ways of estimating the degree to which test bias exists

Test Score Bias vs. True Group Differences

- Simply because two groups differ in their mean scores on a test does not imply that the test suffers from test score bias
- If females score higher on a test of optimism than males then it is fully possible that they are, as a group, more optimistic than men
- Similarly, a weighing scale would be expected to demonstrate that, on average, males are heavier than females
- This does not mean the scale is biased

Detecting Test Score Bias

1. Construct Bias (internal methods)
2. Predictive Bias (external methods)

Detecting Test Score Bias

1. Construct Bias (internal methods)
2. Predictive Bias (external methods)

Detecting Construct Bias: Internal Evaluation of a Test

- **Construct bias** is related to the meaning of test scores
- If a test suffers from construct bias, then the scores might have different meanings for different groups
- Under such conditions, it does not make psychological sense to compare test scores across those two groups
- For example, suppose a researcher administers a "mechanical aptitude" test to a sample of test takers and finds evidence of construct bias related to biological sex

Detecting Construct Bias: Internal Evaluation of a Test

- This would suggest that the responses of males and females on the test reflect different underlying psychological constructs
  - Male responses might reflect a single construct:
    - mechanical aptitude
  - Female responses might reflect two constructs:
    - (1) mechanical aptitude + (2) stereotype threat
- Since the mechanical aptitude test measures different constructs in males and females, it simply doesn't make sense to compare their test scores

Detecting Construct Bias: Internal Evaluation of a Test

- Evaluating construct bias typically occurs by conducting analyses at the item level
- An item on a test is biased if:
  1. people belonging to different groups respond in different ways to the item, and
  2. if it could be shown that these differing responses were not related to group differences associated with the psychological construct measured by the test
- We will consider four different procedures for detecting construct bias

Four Procedures For Detecting Construct Bias

1. Item Discrimination Index
2. Factor Analysis
3. Item Functioning Analysis
4. Rank Order Consistency

Four Procedures For Detecting Construct Bias

1. Item Discrimination Index
2. Factor Analysis
3. Item Functioning Analysis
4. Rank Order Consistency

Detecting Construct Bias: Item Discrimination Index

- An item's discrimination index—symbolised by d—reflects the degree to which the item is related to the total test score
- A high d indicates that people who answer an item correctly tend to do better on the test as a whole than people who answer the item incorrectly
- A strong d suggests an item is highly similar conceptually to most other test items
- Item discrimination indexes reflect the structure of associations among items

Detecting Construct Bias: Item Discrimination Index

- Suppose we divide the takers of the mechanical aptitude test into two groups:
  - one with high scores on the test
  - one with low scores on the test
- If the test item reflects mechanical aptitude, we would expect:
  - a relatively high proportion of "high scorers" to answer it correctly
  - a relatively low proportion of "low scorers" to answer it correctly
- If the item is "bad"—not related to mechanical aptitude—we would expect no difference between groups

Detecting Construct Bias: Item Discrimination Index

- A test item with a high d (e.g., .90) indicates that people with a:
  - high mechanical aptitude have a high probability of answering it correctly
  - low mechanical aptitude have a low probability of answering it correctly
- This indicates that the item strongly discriminates amongst people with varying levels of aptitude
- This, in turn, suggests that it is a good reflection of the construct being assessed by the test
- A test item with a low d (e.g., .10) is not a good reflection of the construct being assessed by a test

Detecting Construct Bias: Item Discrimination Index

- To test for construct bias we can compare item discrimination indexes separately for two groups
- For example, we might compute the indexes for males and females on the mechanical aptitude test
- If the indexes are approximately equal, this suggests the item reflects the construct in the same way for both genders
- If the indexes are quite different, this suggests the presence of construct bias related to gender
- In the latter case, we might remove the "bad" item or modify it to neutralise the bias

Four Procedures For Detecting Construct Bias

1. Item Discrimination Index
2. Factor Analysis
3. Item Functioning Analysis
4. Rank Order Consistency

Four Procedures For Detecting Construct Bias

1. Item Discrimination Index
2. Factor Analysis
3. Item Functioning Analysis
4. Rank Order Consistency

Detecting Construct Bias: Factor Analysis

- A second method for detecting construct bias is by performing a factor analysis (or PCA)
- Factor analysis can be used to evaluate the internal structure of a test—how the items on a test relate to each other—separately for different groups
- For example, we might find that among males the mechanical aptitude test has a clear unidimensional structure
- This would suggest that test scores reflect one, and only one psychological construct

Detecting Construct Bias: Factor Analysis

- To test for the presence of construct bias, we could then perform a factor analysis on the females responses to the test items
- If we also found a unidimensional structure for female responses, we would conclude that the test does not suffer from construct bias
- However, if we found the test exhibited a multidimensional structure, then this would suggest the test has a different internal structure for males and females
- We would then conclude that the test does suffer from construct bias—the test scores reflect different psychological constructs for males and females

Four Procedures For Detecting Construct Bias

1. Item Discrimination Index
2. Factor Analysis
3. Item Functioning Analysis
4. Rank Order Consistency

Four Procedures For Detecting Construct Bias

1. Item Discrimination Index
2. Factor Analysis
3. Item Functioning Analysis
4. Rank Order Consistency

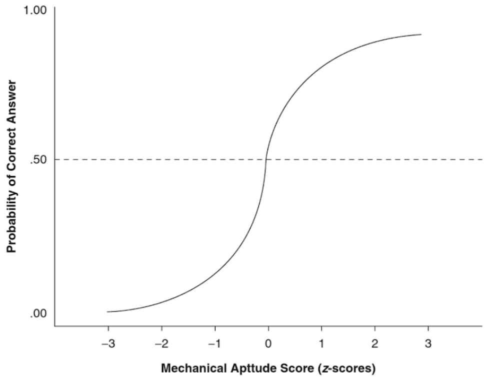
Detecting Construct Bias: Differential Item Functioning (DIF) Analysis

- DIF is a sophisticated approach to the assessment of construct bias developed within the context of Item Response Theory (IRT; Lecture 11)
- Theoretically, DIF assumes that we can estimate peoples' trait levels directly from the test data
  - the trait levels are, in essence, participants' true scores for the psychological construct being measured
- We can also estimate the trait levels for the people in different groups (e.g., males and females)
- We can then determine whether the trait levels and the item responses match up in the same way for both groups
  - if they don't, we have evidence for construct bias

Detecting Construct Bias: Differential Item Functioning (DIF) Analysis

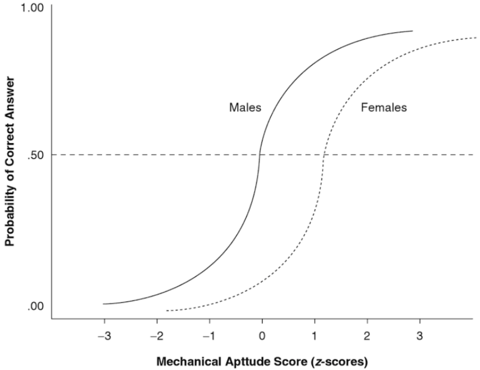
- IRT is based on the idea that there is a mathematical function relating a participants trait level to the probability he or she will answer a question correctly
- If you have a group of people take a test and you know their respective trait levels (estimated from their test scores), then you can generate an item characteristic curve (ICC) showing this function for each item
- If we have two groups (e.g., males and females) we can plot the ICCs separately for each group
- Inspection of these curves can help diagnose possible construct bias

ICC Example



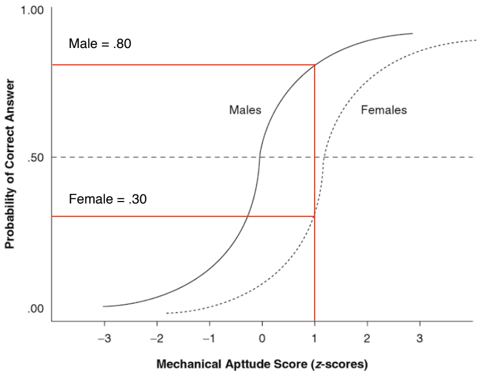
- The x-axis shows the trait levels of respondents expressed as z scores
- The y-axis shows the probability of answering a particular test item correctly

Uniform Bias Example



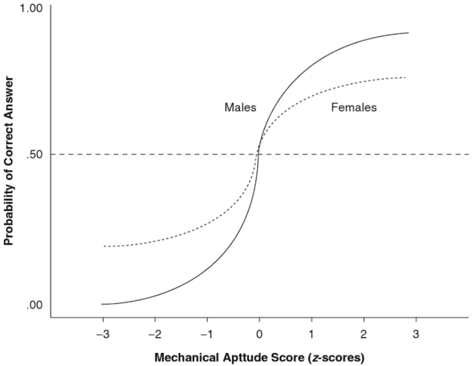
- Here the ICCs differ in location, but not shape
- A male whose mechanical aptitude is +1 standard deviation above the mean has an approximately 80% chance of answering the item correct
- The corresponding percentage for females is approximately 30%

Uniform Bias Example



- Here the ICCs differ in location, but not shape
- A male whose mechanical aptitude is +1 standard deviation above the mean has an approximately 80% chance of answering the item correct
- The corresponding percentage for females is approximately 30%

Nonuniform Bias Example



- Here the ICCs differ in shape and location
- At some levels of aptitude, females find the item easier than males
- At other levels of aptitude females find the item more difficult than males

Four Procedures For Detecting Construct Bias

1. Item Discrimination Index
2. Factor Analysis
3. Item Functioning Analysis
4. Rank Order Consistency

Four Procedures For Detecting Construct Bias

1. Item Discrimination Index
2. Factor Analysis
3. Item Functioning Analysis
4. Rank Order Consistency

Detecting Construct Bias: Rank Order Consistency

- A really simple way to test for evidence for construct bias is to calculate the means associated with each item separately for each group
- Then, calculate the Spearman Rank Correlation between the item means (i.e., "item difficulties")
- If the correlation between the item means is low, then we would have some evidence to suggest construct bias
- The textbook suggests that a low correlation between item means would be less than .90

Detecting Construct Bias: Rank Order Consistency

Item		M_f	M_m
1	Without strong emotions, life would be uninteresting to me.	3.2	2.8
2	How I feel about things is important to me.	4.6	3.8
3	I seldom pay much attention to my feelings of the moment (R)	3.9	3.0
4	I find it easy to empathise – to feel myself what others are feeling.	4.1	3.2
5	I experience a wide range of emotions or feelings.	4.4	3.6
6	I enjoy movies about relationships.	2.9	1.5

The Spearman rank correlation is 1.0 in this case, so there is absolutely no evidence for construct bias.

This is true, even though males, on average, score lower than females.

Detecting Test Score Bias

1. Construct Bias (internal methods)
2. Predictive Bias (external methods)

Detecting Test Score Bias

1. Construct Bias (internal methods)
2. Predictive Bias (external methods)

Detecting Predictive Bias: External Evaluation of a Test

- **Predictive bias** is the degree to which a test's scores are equally predictive of an outcome for two groups
- Construct bias tends to be more relevant in pure research contexts
- Predictive bias, by contrast, is a more applied consideration
- Specifically, do the scores from a particular test predict a criterion (or outcome) with equal accuracy for two or more groups?

Example

- The government employees who created the tests behind the ATAR should make sure that the scores have equal predictive validity with respect to university grade performance across all groups
- In particular, gender
- But also race would be an important variable to consider

Predictive Bias Analyses

- It involves two steps:
  1. you need to determine whether the test scores predict the dependent variable to start with
  2. you then need to determine whether the test scores predict the dependent variable equally well across groups
- Both steps involve conducting a **regression analysis**

Bivariate Regression: Review

- Extension of Pearson correlation
- Bivariate regression consists of using one continuously measured variable (X) to predict another continuously measured variable (Y)
- The main difference is that bivariate regression includes additional terms which can be used to create a regression equation:
  - Intercept
  - Slope

Intercept (b_0)

- The expected value of Y when X is 0
- Forms the "foundation" of the regression equation
- The point at which the regression line crosses the Y-axis

Slope (b_1)

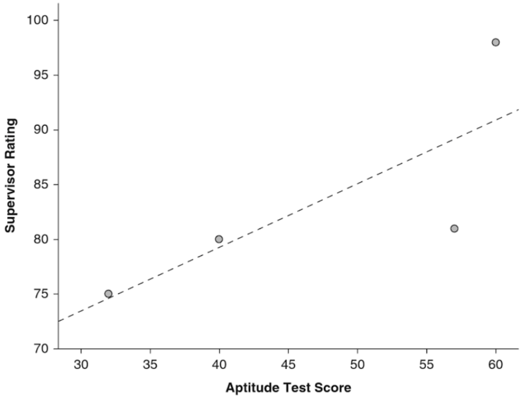
- The expected increase or decrease in the dependent variable (Y) as a function of an increase or decrease in the independent variable (X)

Example: Textbook

Trainee	Aptitude Test Score	Supervisor Rating	Predicted Supervisor Rating
1	32	75	74.59
2	40	80	79.23
3	57	81	89.09
4	60	98	90.83

- In this example, there are four trainees who completed an aptitude test (X)
- They were later rated for their work performance by their respective supervisor (Y)

Scatter Plot



- Scatter plot—and regression line—depicting the association between Aptitude Test Score and Supervisor Rating

Scatter Plot

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients		
Model	B	Std. Error	Beta	*t*	Sig.
1 (constant)	56.034	16.018		3.498	.073
AptTest	.581	.329	.781	1.766	.219

[a] Dependent Variable: Suprating

- Intercept (b_0) = 56.034
  - if you scored 0 on the aptitude test, you would be expected to receive a supervisor rating of 56.034
- Unstandardized slope (b_1) = .581
  - a one unit increase in Aptitude Test Score is associated with about half (.581) a unit increase in Supervisor Rating

Predicting Peoples' Scores

- Use the regression equation:
- $Y = b_0 + b_1(X)$
- $Y = 56.03 + .58(X)$
  - where X is equal to a trainee's aptitude score
- Example: Trainee 1 (Aptitude score of 69)
  - $Y = 56.03 + .58(69)$
  - $Y = 96.05$

One Size Fits All

- The estimation of predictive bias begins with the assumption that different groups share a common regression equation
- If a test is not biased, then one regression equation should be equally applicable to different groups of people
- Based on the idea that "one size fits all"—regardless of sex, ethnicity, or whichever group difference is being considered
- The regression equation derived from the total sample is referred to as the **common regression equation**
- Also, we use the terms **common intercept** and **common slope**

One Size Fits All?

- Next, we estimate separate regression equations for each group
- We then compare the **group-level regression equations** with the common regression equation
- If the group-level estimates do not match the common regression equation, then you might suspect that your test scores are biased

Testing Predictive Bias Statistically

- In the lab, you will learn a method based on the estimation of confidence intervals
- I will cover these techniques in this lecture only conceptually

Four Types of Predictive Bias

1. Intercept Bias
2. Slope Bias
3. Intercept + Slope Bias
4. Outcome Score Bias

Four Types of Predictive Bias

1. Intercept Bias
2. Slope Bias
3. Intercept + Slope Bias
4. Outcome Score Bias

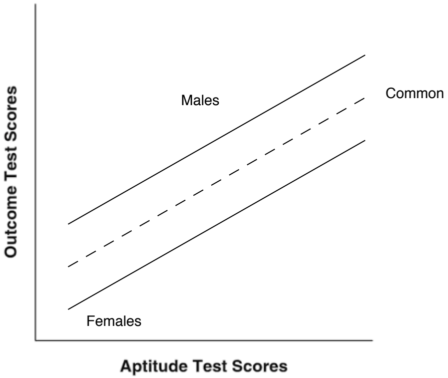
Intercept Bias

- In the aptitude test example, suppose you conducted the regression analysis separately for males and females
- For both groups, you find the slope to equal .58
  - which is equal to the common slope (so, no bias here)
- By contrast, the intercept seems to vary:
  - male intercept = 58.03
  - female intercept = 54.03
- This result would suggest **intercept bias**

Intercept Bias: Interpretation

- Because the slopes are equal, intercept bias implies that the males receive higher supervisor ratings than the females across all levels of aptitude scores
- Suppose a male and a female scored 70 on the aptitude test, what would we predict their supervisor rating to be?
  - male = 58.03 + .58(70) = 98.63
  - female = 54.03 + .58(70) = 94.63
- This result suggests that the aptitude test does not function equally for both groups (i.e., it is biased)

Graphical Illustration of Intercept Bias



- The male and female regression lines run parallel, which implies that they have the same slope
- However, they intersect the Y axis at different points, which implies that they have different intercepts

Four Types of Predictive Bias

1. Intercept Bias
2. Slope Bias
3. Intercept + Slope Bias
4. Outcome Score Bias

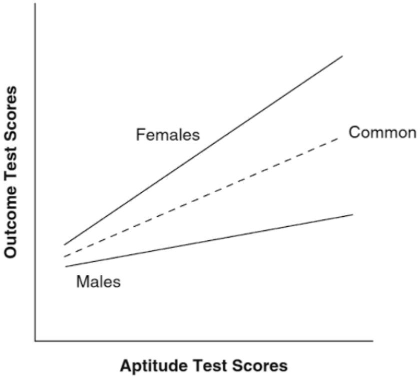
Four Types of Predictive Bias

1. Intercept Bias
2. Slope Bias
3. Intercept + Slope Bias
4. Outcome Score Bias

Slope Bias

- Suppose the two groups had similar intercepts, but their slopes differed in magnitude
- In the aptitude example, suppose the intercept was equal to 56.03 for males and females
- However, the unstandardized slopes were equal to .53 for males and .63 for females
- Graphically, the results would look like this:

Graphical Illustration of Slope Bias



- The male slope is shallower than the female slope
- This implies that we would expect different predicted scores for males and females with the same aptitude test score

Example

- Suppose a male and a female scored 70 on the aptitude test, what would we predict their supervisor rating to be?
  - male = 56.03 + .53(70) = 93.13
  - female = 56.03 + .63(70) = 100.13
    - female predicted to score 7 points higher
- Suppose a male and a female scored 60 on the aptitude test, what would we predict their supervisor rating to be?
  - male = 56.03 + .53(60) = 87.83
  - female = 56.03 + .63(60) = 93.83
    - female predicted to score 6 points higher
- This is evidence of **slope bias**

Four Types of Predictive Bias

1. Intercept Bias
2. Slope Bias
3. Intercept + Slope Bias
4. Outcome Score Bias

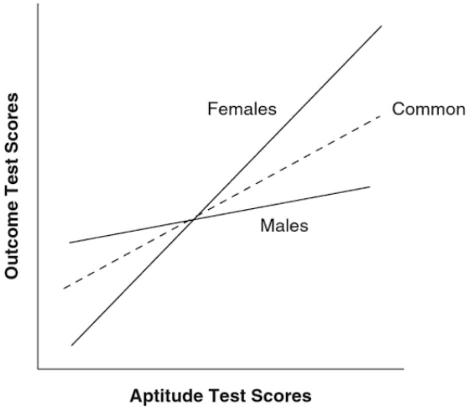
Four Types of Predictive Bias

1. Intercept Bias
2. Slope Bias
3. Intercept + Slope Bias
4. Outcome Score Bias

Intercept and Slope Bias

- In addition to "pure" intercept bias and "pure" slope bias, there are occasions where both intercept and slop bias co-occur
- In this case, there will be a complex relationship between the size of aptitude scores and the outcome scores for the different groups
- For example, we might find that for people with *low* levels of aptitude, the predicted outcome scores for males might be higher than for females
- But our analysis might reveal that for people who have *high* levels of aptitude, the predicted outcome scores for females might be higher than for males
- Graphically, the results might look like this:

Graphical Illustration of Intercept and Slope Bias



- This is evidence of **intercept + slope bias**

Four Types of Predictive Bias

1. Intercept Bias
2. Slope Bias
3. Intercept + Slope Bias
4. Outcome Score Bias

Four Types of Predictive Bias

1. Intercept Bias
2. Slope Bias
3. Intercept + Slope Bias
4. Outcome Score Bias

Outcome Score Bias

- Thus far, the discussion of predictive bias has focussed on the possibility that the predictor variable is biased
- It is also possible that the *outcome variable scores* are the ones that are biased—viz. **outcome score bias**
- For example, it is possible that the supervisor who provides the ratings is biased in favour of one group
- For a full analysis of bias, you would want to examine both possibilities:
    - predictor bias + outcome bias

Effect of Reliability

- There are occasions where one group will have less reliable scores than the other group
- This difference in reliabilities can cause the difference in slopes and intercepts
- Differences in reliability are a form of construct bias that can have an impact on predictive bias

References

Furr, M. R., & Bacharach, V. R. (2014; Chapter 11). Psychometrics: An Introduction (second edition). Sage.