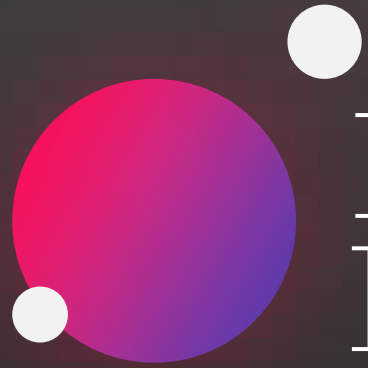
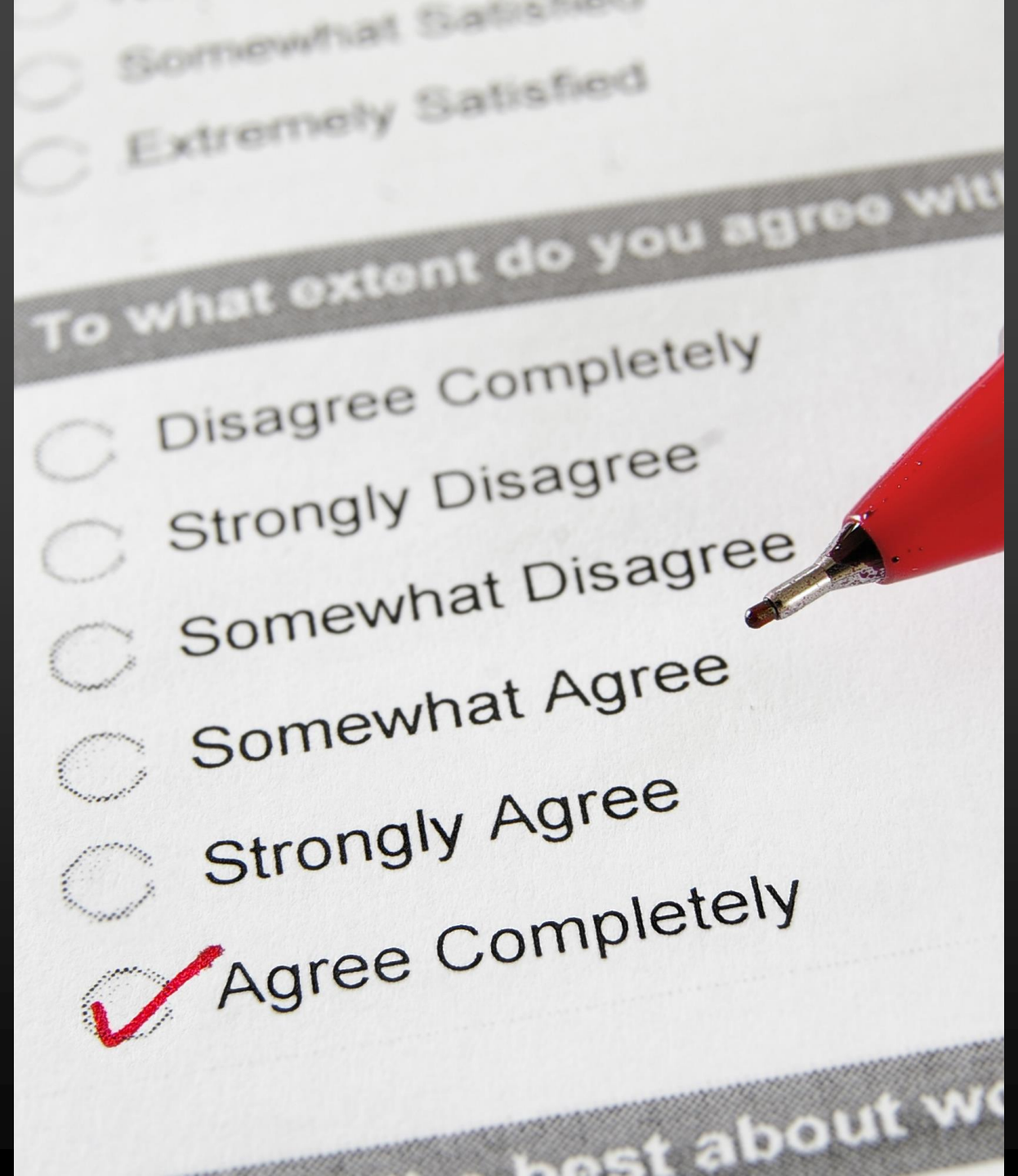




Research Methods I

Scales and Psychometrics

Stacey Coleman



Lecture outline

- Use of questionnaires in psychology research
- Choosing appropriate measures
- Validity and reliability
- Scales, sub-scales and negatively worded items, and calculating scale totals



Questionnaires

- Questionnaires are often used in psychological research as a research instrument to assess human characteristics
- Participants read questions/statements and provided self-report responses
- Purpose to gather information from participants
 - Responses are examined or scored and can then be used in analyses



Questionnaires

- Choosing an appropriate questionnaire (measure) for your study is another vital aspect of study design
- Need to think carefully about what it is you actually want to measure (think about your operational definition!) and what the best way of measuring this is...

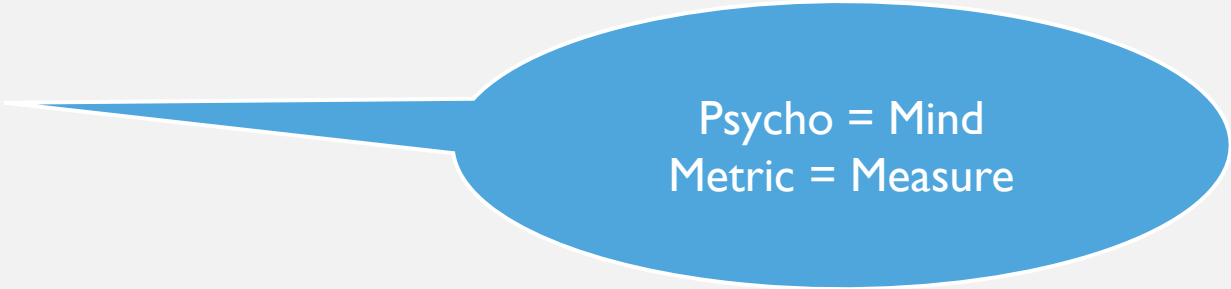
E.g.,

- Stress
- Memory
- Intelligence



Examples of Psychological Measures

- Method for measuring individual differences related to some psychological concept, based on a sample of relevant behaviour in a scientifically designed and controlled situation.
- **Psychometric** tests
 - IQ test
 - Personality tests
 - Five Factor Personality Test
 - Openness to experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism
- Questionnaire measuring psychological distress
- Questionnaire measuring self-esteem



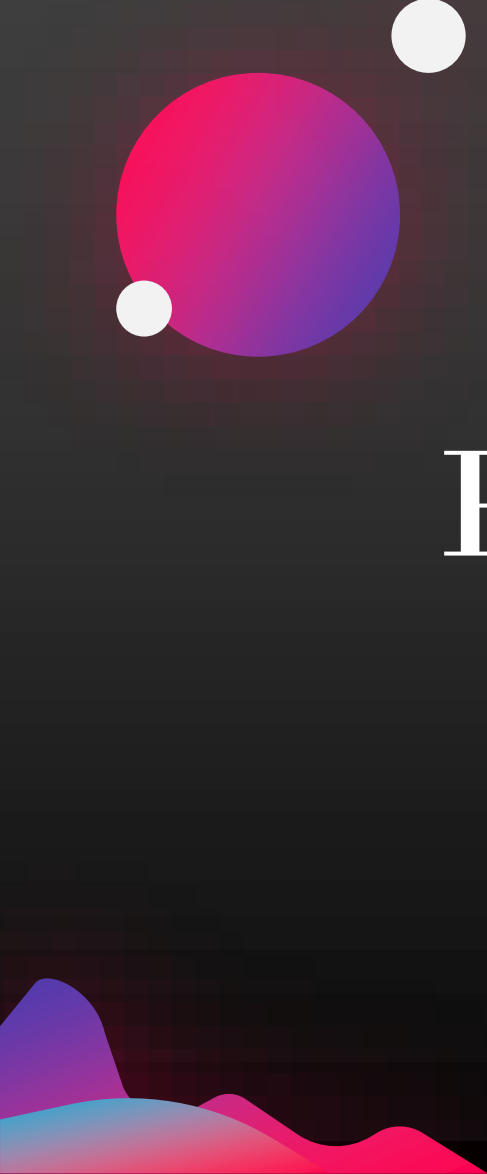
Psycho = Mind
Metric = Measure



Psychological Measures

- Scales and tests that aim to measure characteristics, such as:
 - Attitudes
 - Ability
 - Intelligence
 - Personality
- They take a **long time** to develop and require expert knowledge to devise





Psychometric Properties

Psychometric Properties

- The intrinsic components of a test are termed its psychometric properties
- These are typical characteristics of tests that identify and define critical aspects of a test
 - E.g., its suitability or reliability for use in a certain context
- Put simply, psychometric properties provide information about a test's adequacy, relevance, and validity



Psychometric Properties

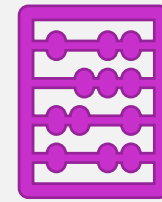
- A good psychometric test must have three fundamental properties:



**Psychometric
Reliability**



**Psychometric
Validity**



**The Practice
of Norming**

Psychological Measures

Measures capturing psychological constructs are expected to be **reliable** and **valid**

RELIABILITY

the extent to which a test yields consistent results.

VALIDITY

the extent to which a test measures or predicts what it is supposed to.



Choosing Measures

- Choosing an **existing** reliable and valid scale is always the best option:
 - You know that you are measuring something in a robust way (you haven't just made it up!)
 - You have existing data from other studies to compare your findings with
 - Sensitive to change within studies (if appropriate)
 - Allows for stronger statistical analyses



Scale Items

The items within a psychological scale often consist of a set of statements:

e.g., “I take a positive attitude towards myself”

Responses are given on a response scale:

Strongly disagree	Disagree	Nether agree or disagree	Agree	Strongly disagree
------------------------------	-----------------	-------------------------------------	--------------	------------------------------

This is known as a Likert Scale (Likert, 1932)



Scale Items

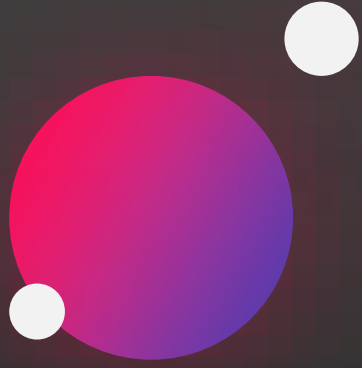
- Each individual item within a scale measures a small aspects of the given characteristics
- An overall score can be calculated by summing all responses to all items
 - Overall score can then be used to determine how much they tend to lean towards one end of the scale or the other
 - e.g., Introversion - Extraversion



Scales and Tests

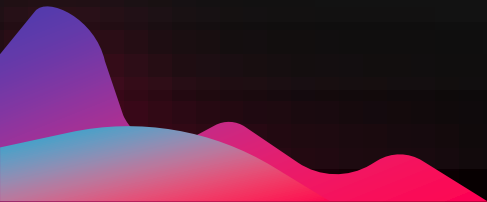
- A variety of psychological scales and tests are used for:
 - Selecting employees
 - Measuring levels of depression
 - Measuring cognitive ability
- Responses may be use to examine differences between groups of people, or to examine differences within the same group of people at different time points





E.g., measuring levels
of self-esteem

Defined as ‘an individual’s subjective evaluation of their own worth’



The How Much I Like Myself Scale

1. I really like myself all of the time

Strongly disagree Disagree Agree Strongly agree

2. My friends think I am awesome

Always Sometimes Not sure Not often Never

3. I know I am the best person I can be at this time

Yes No



OR



Rosenberg's Self-Esteem Scale

On the whole, I am satisfied with myself.

Strongly Agree

Agree

Disagree

Strongly Disagree

I feel that I have a number of good qualities.

Strongly Agree

Agree

Disagree

Strongly Disagree

I am able to do things as well as most other people.

Strongly Agree

Agree

Disagree

Strongly Disagree

I feel that I'm a person of worth, at least on an equal plane with others.

Strongly Agree

Agree

Disagree

Strongly Disagree

I take a positive attitude toward myself.

Strongly Agree

Agree

Disagree

Strongly Disagree





Reliability

Consistency
or stability
of a
measure

Reliability

- Refers to the consistency or stability of a measure
 - Across time / different testing (external)
 - Within itself (internal)
- Any measure should yield a similar estimate on repeated use if the measured trait has not changed



Test-retest reliability

CONSISTENCY OF A MEASUREMENT OVER TIME

- Assessed by administering the measure to the same group of participants on two or more separate occasions and correlating the sets of scores
- If the construct being measured is stable, scores should be consistent over time
- E.g., you would not expect the results of a test of IQ or personality traits to be different if measured more than once



Examples of The Big Five Inventory Personality measure

In General, I am someone who...	Disagree Strongly 1	Disagree a little 2	Neither agree nor disagree 3	Agree a little 4	Agree strongly 5
Is talkative					
Tends to find fault with others					
Does a thorough job					
Is depressed, blue					
Is original, comes up with new ideas					
Is reserved					
Is helpful and unselfish with others					
Can be somewhat careless					
Is relaxed, handles stress well.					
Is curious about many different things					
Is full of energy					
Starts quarrels with others					
Is a reliable worker					



Internal Consistency

CONSISTENCY OF MEASUREMENT WITHIN THE TEST ITSELF

- All items on a test should measure the same thing
- Items should correlate highly with each other

Measure of the homogeneity of the test - i.e., the degree to which the various parts of an instrument measure the same variable

- Cronbach's alpha – response to each item is correlated with the response to every other item



Example: Femininity as measured using the SRI

Item	Almost never true	Rarely true	Less than half the times true	Neutral	More than half the times true	Often true	Almost always true
Cheerful							
Shy							
Affectionate							
Flatterable							
Loyal							
Feminine							
Sympathetic							

The response to each item should be correlated with the response to every other item – otherwise not tapping into the same thing (concept)

Reliability

Interpretation of Cronbach's Alpha

Certain benchmark values have been proposed in order to interpret Cronbach's alpha:

$\alpha < .60 = \text{poor!}$

$.60 < \alpha < .70 = \text{adequate}$

$.70 < \alpha < .80 = \text{good}$

$.80 < \alpha < .90 = \text{very good!}$

$\alpha > .90 = \text{perhaps too good...}$



Reliability – Item Analysis

- We can assess internal consistency using SPSS (reliability analysis)
- To increase the reliability of a scale, we can remove items that are causing the low reliability
- Running reliability analysis – Table called **Item-Total Statistics**
 - ‘Cronbach’s alpha if item deleted’
 - Select item that would most increase the alpha and simply delete it from the scale
 - May need to be done more than once



Reliability – Item Analysis

- Remove items until reliability achieved (alpha) is satisfactory
- Issue? Ideally, we should have at least 10 items left in a scale after removing those causing low reliability
 - What if we need to delete more?
- Also – we cannot just delete items without paying attention to their content.
 - We may end up altering the underlying construct



Other forms of reliability...

- **SPLIT-HALF RELIABILITY**

- Items divided randomly
- Internal consistency between two sets of scores

- **INTER-RATER RELIABILITY**

- Extent to which two different raters agree in their evaluations
- Consistency of measurement when different people observe the same event or score the same test





Validity

How well a test
actually
measures what
it is designed to
measure

Face Validity

The test looks plausible at face value

- i.e., does it **look like** it is measuring what it claims to be measuring?
- E.g., Life Orientation Test (Scheier & Carver, 1985); a measure of dispositional optimism
 - In uncertain times I usually expect the best
 - If something can go wrong for me it will
 - I'm always optimistic about my future
 - I hardly ever expect things to go my way



Construct Validity

Does the test successfully measure the psychological construct it is designed to measure?

- E.g., intelligence
- Assessed by looking at the relationship between test scores and other behaviours that it **should be related to**
 - Do scores on a new IQ test relate to academic performance?
 - Does a measure of psychological distress report similar scores to a psychiatric interview?



Construct Validity

- Should be related to measures of the same construct – *convergent validity*
- Should be unrelated to measures of unrelated constructs – *discriminant / divergent validity*



Content Validity

- Whether the items on a test measure all the knowledge or skills that are assumed to underlie the construct of interest
- E.g., does a measure of sex roles capture every psychological trait related to males and females, or just certain types



Criterion-related Validity

The ability of test scores to correlate with meaningful criterion measures

- Are they able to predict future outcomes?
- The test should predict present or future outcomes that are influenced by the construct being measured!
- E.g., does a measure of IQ successfully predict future performance/success in that job?
- E.g., do scores on a measure of conscientiousness predict future academic performance?



Criterion-related Validity

- Degree to which a test forecasts or is statistically related to a criterion

Concurrent

- How well a predictor can predict a criterion
- When a new measure is validated by comparing with an existing measure of the construct

Predictive

- Collect predictor information and use it to forecast future criterion performance





The Importance of Norming

(Standardisation)

The Importance of Norming

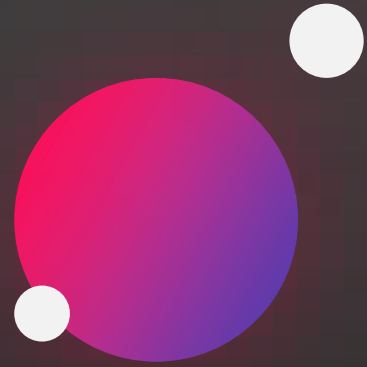
- Psychometric tests often normed against groups for comparison
- Avoids evaluation of individual items and observes the total score of an individual against a representative sample
 - Samples generally made representative based on demographic factors; age, gender, education, religion, etc
- Considered important because score of 50 out of 75 on a psychometric test means nothing unless it is compared to performance of others at a similar level on the same test
 - Practice of using relative scores gains importance further when interpreting ability test results



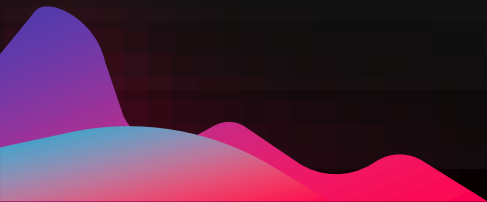
The Importance of Norming

- Psychological constructs such as personality have no right/wrong answers, so cannot be ‘marked’ using percentages
 - This is where norming becomes important to make sense of some assessment scores
- Standardisation:
 - Important scientifically and ethically
 - Standardisation to populations – cannot necessarily be applied *across* populations
 - What about cross-cultural research?
 - A test standardised for one population may not be applicable to another





Measures and data.....



Recap: Measures

- Choosing an **existing** reliable and valid scale is always the best option:
 - You know that you are measuring something in a robust way (you haven't just made it up!)
 - You have existing data from other studies to compare your findings with
 - Sensitive to change within studies (if appropriate)
 - Allows for stronger statistical analyses



BUT BEWARE! Scales and sub-scales

- Some psychological measures may include sub-scales
- Sometimes we can either add up all the items for a total score, or create separate sub-scales for each
- Sometimes we have to keep each sub-scale **separate** from the others



e.g., Bem's Sex Role Inventory

- FEMININITY SUB-SCALE: Twenty items measure feminine traits
 - “*Understanding*”; “*Soft spoken*”; “*Loves children*”
- MASCULINITY SUB-SCALE: Twenty items masculine traits
 - “*Independent*”; “*Self-reliant*”; “*Willing to take risks*”

These sub-scales HAVE to be kept separate



But beware (again...)

- Some scales contain items which are *negatively worded* and need to be **reverse-scored** prior to analysis (to prevent demand characteristics/lazy responses)
- E.g., Life Orientation Test (Scheier & Carver, 1985): a measure of dispositional optimism
- In uncertain times I usually expect the best
- If something can go wrong for me it will ←
- I'm always optimistic about my future
- I hardly ever expect things to go my way ←

Negatively
worded
items



But beware (again...)

- Negatively worded DOES NOT mean that the item contains (what we would generally consider to be) a negative word within it...
- Negatively worded refers to an item being worded in a way that is not typical of measuring what the scale overall is supposed to measure...
- E.g., Arnetts Inventory of Sensation Seeking (AISS; Arnett 1994)



AISS

- Twenty item measure of sensation seeking
 - Contains two sub-scales; novelty seeking and intensity seeking
- Scores can be calculated for novelty seeking, intensity seeking, or overall levels of sensation seeking (total of all items)
- Novelty seeking items:
 - I would like to travel to places that are strange and far away.
 - I think it's best to order something familiar when eating in a restaurant
 - Item simply gets reverse scored

Negatively
worded item



Why does it matter?

- It will influence
 - The quality of your results,
 - The type of conclusions you are able to draw from your results
 - The type of statistical analysis you are able to conduct



Final points to consider

Issues with psychological measurements / tests



Social Desirability / Demand Characteristics

- Participants aiming to guess what is regarded as the socially acceptable response and give an answer that matches this in order to be seen as ‘good’/a ‘good person’, or in an attempt to please the researcher
- Participants may try to interpret the aim of the questions/scale and attempt to work out what is expected of them and try to answer accordingly



Social Desirability

- Participants may feel pressure to respond in a certain way
 - Even if it does not truly represent how they feel
- In an attempt to gain genuine responses, measures can be cleverly disguised; if the responder is not aware of the true intentions of the measure, their answers are more likely to be honest
- Disguising a measure is likely to be effective when responses involved disclosing highly sensitive, embarrassing, or potentially damaging information
 - But – Deception!



Social Desirability

- How?
- Use a different type of scale
- E.g., Semantic Differential Scale (Osgood, Suci, & Tannenbaum, 1957)
 - Works by asking respondents to select a point between two opposing adjectives

For me, donating my organs would be....

Beneficial - - - - - Harmful

Enjoyable - - - - - Boring



Response Bias

- It is more difficult to disagree with statements compared to agreeing
- To deal with this confound, items in a scale need to combine a mix of positive and negative statements
 - Respondents need to think carefully before answering
- Perception/reactions to some scales can be unfavourable;
 - Respondents may think you believe in the statements on the scale (tricky if the statements are sensitive/extreme)
 - Give a clear explanation at the start



Scale Construction: What to avoid

- Phrasing of items need to be clear and unambiguous, so the person answering will have no doubt what answer to give.
- Ideally we want all responders to interpret the scale items in the same way.
- In reality we know this does not always happen.
- What we can do to alleviate this is ensure certain things are avoided in order to maximise clarity...



Scale Construction: What to avoid

- 1. Complex Items
- 2. Technical Terms
- 3. Ambiguity
- 4. Double-barrelled questions
- 5. Double Negatives
- 6. Emotive Language
- 7. Leading Questions
- 8. Invasion of Privacy
- 9. Unbalancing
- 10. Sensitivity



Personality Measures / Intelligence Tests

- These measures undergo a high degree of standardisation and scrutiny to ensure they are valid and reliable
 - They tend to be used in professional practice and could have a profound impact on a person's life
 - Job selection, educational assessment, clinical diagnosis
- Intelligence tests need continual revision
 - Sensitive to bias from culture, class, language, etc



Summary

- Choosing appropriate measures for your study is another vital aspect of study design
- Choosing an **existing** reliable and valid scale is always the best option
- Scales should aim to be reliable and valid
- Beware scales and sub-scales
- And beware negatively worded items!
- Choose your scale and, subsequently, type of data carefully...



Questions?

