Research Methods II

SCALES AND PSYCHOMETRICS

STACEY COLEMAN

Questionnaires / Psychological Scales

In order to assess various human characteristics, psychologists often use certain measuring instruments:

- Scales
- Psychological tests
- Questionnaires

They tally up the scores or examine the responses.

Choosing appropriate measures for your study is a vital aspect of study design.

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

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

Scale Items

Psychological measures often use scale ratings

- E.g., in an attitude or a personality measure, items often consist of a statement
  - ‘I am the type of person who is outgoing, sociable’
- The respondent then answers this statement 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 aspect 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

Psychological Scales and Tests

Tend to use closed questions

- E.g., do you attend your Research Methods sessions?
  - Always Often Sometimes Never

This allows for **quantitative** analysis of data

Closed Questions

Disadvantages of using closed questions:

- Constraint of responses
- Misinterpretation
- Unrealistic
- Lacks rich information
  - **Why** did they give that response?

Need to be careful that response choices do not overlap

- If you expect only one answer to be selected then you need to ensure others cannot be selected

Types of Scales

Types of Scales

Summated Ratings - Likert Scale (Likert 1932)

- One of the most common types of scale
- Used to assess attitudes, knowledge, perceptions, values, and behavioural changes
- Respondents specify their agreement with a statement on a symmetric scale
- Anchored between two opposing (in strength) responses
  - E.g., on a continuum from ‘Strongly Disagree’ to ‘Strongly Agree’
- Normally odd-point scales are used (5-point, 7-point, etc)
  - Allows for a ‘middle’ response or a ‘Neither’ option
- But, even-point scales sometimes appropriate
  - Forced choice method

Types of Scales

Summated Ratings – Likert Scale (Likert 1932)

Statements often include both positively worded and negatively worded items

Often requirement to reverse some scores

Score for each item is summed to give overall measure score

- E.g., attitudes towards gambling

Can contain items not obviously measuring the attitude directly (known as *Diagnostic)*

Types of Scales

Equal Appearing Intervals (Thurstone, 1931)

Ps tick all they agree with from a set of statement options, then given score equivalent to the strength of the statement if they agree with that statement (known as the Scale Value).

Involves: producing large No. of statements, asking panel of judges to rate each statement, take mean value of all judges ratings, reject statements where judges' individual values have a high variance.

Types of Scales

The Semantic Differential Scale (Osgood, Suci, & Tennenbaum, 1957)

Participants select a point between opposite adjectives about where they feel they sit on that scale.

E.g., exercising for 30 minutes every day would be….

Beneficial - - - - - - - - - - Harmful

Enjoyable - - - - - - - - - - Boring

Types of Scales

The Visual Analogue Scale (VIS)

Useful when assessing characteristics that do not neatly fit into one of two categorise

- E.g., mood – may not be able to state ‘high’ or ‘low’ but can select somewhere in-between

E.g., Indicate on the line the level of your current mood

Low - High

Issues with Psychological Scales

Psychological Scales: Issues

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

Psychological Scales: Issues

Social Desirability / Demand Characteristics

Provide answers that they believe will be evaluated as indicating strength or lack of weakness/dysfunction (‘faking good’)

Provide answers that they believe will be evaluated as indicating weakness or presence of impairment/pathology (‘faking bad’),

Psychological Scales: Issues

Social Desirability

Participants may feel pressure to respond in a certain way

- Even if it does not truly represent how they feel

In at 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!

Psychological Scales: Issues

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

Constructing Psychological Scales

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.

Constructing Psychological Scales: 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

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 us in a certain context

Put simply, psychometric properties provide information about a tests adequacy, relevant, and validity

Psychometric Properties

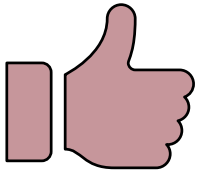
Psychologists need to be confident in the measures they are producing and using

Tests that are constructed often serve as an operational definition of the concept being examined

- E.g., our definition of reduced stress would be evident from test scores

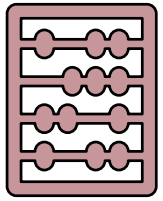
Psychometric Properties

A good psychometric test must have three fundamental properties:



Psychometric Reliability

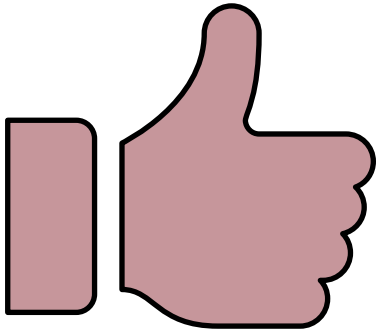
Psychometric Validity



The Practice of Norming

Psychological Measures

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



Reliability

CONSISTENCY OR STABILITY OF A MEASURE

Reliability

External Reliability

Consistently producing the same outcome on different occasions / at different time points

Test-Retest Reliability

- Testing the **same group of people** twice (or more)
- If the 2 sets of scores correlate, the test is deemed as having high external reliability

Reliability

Internal Reliability

Consistency **within** a test

Responses to similar items within a test should be similar

- E.g., if someone has a high *neuroticism* score, then they should score in the neuroticism direction of each of the items within the scale

Items should correlate highly with each other

Checks for Internal Reliability

SPLIT-HALF METHOD

Items of the scale are divided randomly into two halves

Internal consistency measures between the two halves

- Correlations used to assess similarity
- If the test is reliable, the scores for each half should be similar

CRONBACH'S ALPHA

Also called 'coefficients alpha'

Related to mean correlation between each pair of items and the numbers of items in the scale

Low alpha values indicate an unreliable test and emerge when Ps vary a lot on individual items relative to their overall variance

Checks for Internal Reliability

Cronbach's Alpha

Good reliability of a scale is represented with alpha values between .75 and 1.

As a general rule, any scale being measured should have a Cronbach's alpha of at least .70

- Drawback: even if a scale has a high alpha, individual items may still be poorly correlation with the others
- What can we do?

Internal 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

Internal 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

Validity

HOW WELL A TEST ACTUALLY MEASURES WHAT IT IS DESIGNED TO MEASURE

Validity

Face Validity

Test looks plausible at face value – it ‘makes sense’ in relation to 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

Validity

Content Validity

Is it representative of the area it intends to cover? – experts needed

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

Validity

Criterion 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?

Validity: Criterion Validity

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

Validity

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?

Validity

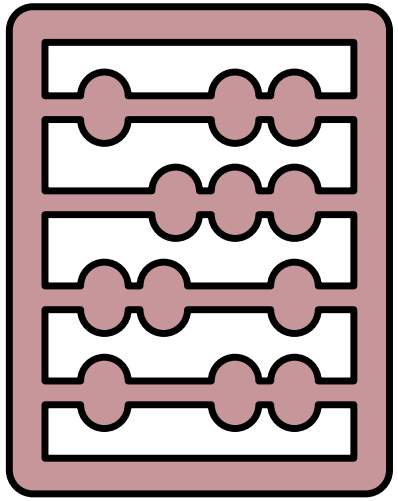
Construct Validity

Whole scientific process of establishing a psychological construct actually exists or at least theoretically sound

that it fits surrounding theory.

Not a one-off procedure that gives a numerical value – it is the development of evidence for a hypothetical psychological construct through rigorous testing.

Should be related to measures of the same construct – *convergent validity* – and it should be unrelated to measures of unrelated constructs – *discriminant validity*



The Importance of Norming

AND ALSO 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:

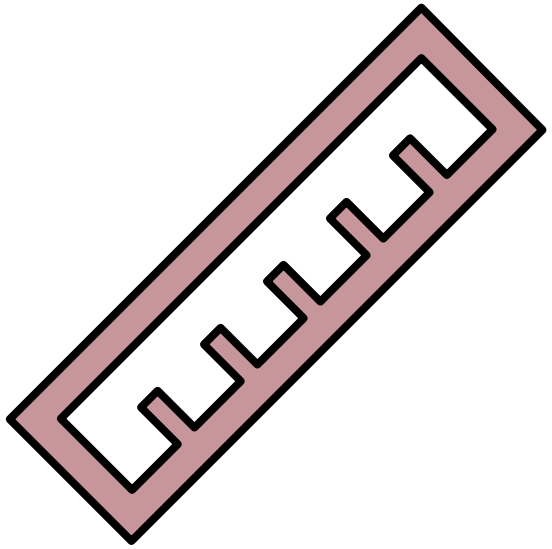
- Important scientifically and ethically
- Standardisation to populations – cannot necessarily be applied *across* populations
  - What about cross-cultural research?
  - A test standardised for one population may not be applicable to another

Normal Distribution

Psychometric tests are usually adjusted until they resemble the normal distribution.

This can be seen as a positive in that it allows us to conduct powerful statistics.

However: not all characteristics are normally distributed – attitudes towards some issues, for example, may display polarised opinions.



Scales, Sub-Scales, Reverse Scoring and Calculating Total Scores

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

Scales and Sub-Scales

Some psychological measures may include sub-scales

Sometimes we 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., Impact of Events Scale (IES; Horowitz et al., 1979)

INTRUSION SUB-SCALE:

- seven items measuring intrusive thoughts
- ‘intrusion’ of a stressful event – scoring range 0-35
  - E.g., “I thought about it when I didn’t mean to…”

AVOIDANCE SUB-SCALE:

- Eight items measuring attempts to avoid those thoughts – scoring range 0-40
  - E.g., “I stayed away from reminders of it”

Can be combined to create a ‘GLOBAL DISTRESS SCALE’ – scoring range 0-75

Reverse Scoring / Reverse Coding

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 (LOT; 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

Reverse Scoring / Reverse Coding

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.

Negatively worded item

- I think it's best to order something familiar when eating in a restaurant
  - Item simply gets reverse scored

Choosing a Psychological Measure

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....

Need to consider the type of data your study will produce

Can often measure your operational definition through data type - So, how do you decide?

Types of Data

Two broad types

Categorical

Limited number of groups to which people can belong to / can only belong to one group

E.g., gender, personality type

Continuous

Can take a vast range of values often from zero to an infinite number

E.g., ages, time taken to react to stimuli, scores on a questionnaire

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