Survey Research and Design in Psychology

Lecture 6 - Psychometric Instrument Development

Dr Ananthan Ambikairajah

University of Canberra

Focus/Overview for today

- Exploratory factor analysis (revision)
- What concepts are and how to measure them
- Measurement error
- Psychometrics
- Reliability and validity
- Creating composite scores

Recap

UNIVERSITY OF CANBERRA

Factor analysis

- Correlational methods used to identify clusters of covariance (called factors)

Two main purposes

- Theoretical (Principal axis factoring)
- Data reduction (Principal components analysis)

Two main types of extraction methods

- Exploratory factor analysis (EFA)
- Confirmatory factor analysis (CFA)

Steps to conduct an EFA

8 key steps

1) Test the assumptions of EFA. Can you do the analysis on your data?
2) Select your method of extraction. Are factors likely to be related?
3) Determine the number of factors. What is your data telling you?
4) Select the items that fit each factor.
5) Name and define each factor. Do these make sense?
6) Examine correlations amongst factors. To what extent are they related to each other?
7) Analyse internal reliability. Do the items in each factor contribute to a reliable subscale?
8) Compute composite scores

Test assumptions

Sample size

- At a minimum: more than 5 cases per variable (1:5 variables:cases ratio)
- Ideal: more than 20 cases per variable (1:20 variables:cases ratio)
- The total number of cases should be at least 200, ideally more by Comrey & Lee (2013) guidelines:
  - 50 cases = very poor, 100 = poor, 200 = fair, 300 = good, 500 = very good, 1000+ = excellent

Test assumptions

UNIVERSITY OF CANBERRA

Level of measurement

- In order to be included in an exploratory factor analysis, your variables must be suitable for Pearson product-moment correlational analyses
  - Interval or ratio level of measurement

Normality

- Factor analysis can cope with minor violations of normality (i.e. it is relatively robust)
- However, it works best if the variables are normally distributed

Test assumptions

UNIVERSITY OF CANBERRA

Outliers

- Factor analysis is sensitive to outliers including:
  - Bivariate outliers
  - Multivariate outliers
- Identify outliers (e.g. scatterplots, Mahalanobis distance), remove them or recode them if they are influential

Linearity

- Factor analysis is based on correlations between variables
- Remember that correlations are a measure of the linear association between variables
- It is important to check that the variables are linearly related
  - Check the scatterplots

Test assumptions

Factorability

- Factorability assesses whether there are sufficient intercorrelations amongst the items to warrant factor analysis.
- Assess factorability via one or more of:
  - Correlation matrix correlations $> .3$?
  - Anti-image matrix diagonals $> .5$?
  - Measures of sampling adequacy (MSAs)?
    - Bartlett's significance?
    - KMO $> .5$ or $.6$?

Choose method of extraction

UNIVERSITY OF CANBERRA

Which extraction method do you use?

- There are two main approaches to extracting factors:
  - Principal axis factoring (PAF) analyses shared variance
    - Principal axis factoring is used to discover the underlying structure of a set of variables
    - It analyses only the common (shared) variance, so leaves out variance that is unique to each measurement item
  - Principal components analysis (PCA) analyses all variance
    - Principal components analysis is more common - it is used to reduce many variables down to a smaller number of factor scores, which can then be used in other analyses
    - It analyses all of the variance in each variable (common and unique)

Which should you use?

- Often there is little difference between principal components and principal axis factoring
- If they come to different solutions, try to work out why, and decide which is more appropriate for your data

Remember! The goal = explain the variance

- A good factor solution is one that explains the lion's share of the variance with the fewest factors
  - In general, we tend to be happy if we can explain 50 to 75% of the variance (but don't be too strict on this rule)

Rotating the solution

- A rotated solution is generally easier to interpret

Orthoganal

- Varimax in jamovi
- Minimises the factor covariation
- Produces uncorrelated factors

Oblique

- Oblimin in jamovi
- Allows factors to covary
- Allows correlations between factors

Orthogonal vs oblique

UNIVERSITY OF CANBERRA

- Are you expecting factors to be related or unrelated?
  - If related, choose oblique
  - If unrelated, choose orthogonal
  - If unsure, choose oblique, then switch to orthogonal if the correlations between factors are smaller than .3
  - You can try both, then see which solution is most interpretable

How many factors do you have?

- Jamovi will give you the same number of factors as variables that you have
- There is no definitive statistic that tells you. Make your decision based on:
  - Which solution explains the largest amount of variance with the smallest number of factors
  - What does the theory say? (predictions/expectations)
  - Kaiser's criterion? Which Eigenvalues are greater than 1?
  - Where is the elbow in the scree plot?
  - How interpretable is your last factor?
  - Which solution fits best? (type of EFA, rotation, number of factors)
- Your factors must be interpretable and make theoretical sense.

Selecting items depends on. . .

- Use factor loadings to help identify which item belongs in which factor.
- An exploratory factor analysis of a good measurement instrument ideally has:
  - A simple factor structure (each variable loads strongly ($>$ .50) on only one factor)
  - Each factor has multiple loading variables (more loadings -> greater reliability)
  - Target factor loadings are high ($>$ .5) and cross-loadings are low ($<$.3), with few intermediate values (.3 to .5)
- Eliminate one variable at a time, then re-run the exploratory factor analysis

Steps

UNIVERSITY OF CANBERRA

1) Examine assumptions
2) Choose extraction method and rotation
3) Determine number of factors
4) Select items
5) Name and describe factors
6) Examine correlations amongst factors
7) Analyse internal reliability (today's lecture!)
8) Compute composite scores (today's lecture!)
9) Check factor structure across sub-groups

Concepts and their measurement

- "Concepts express common elements in the world (to which we give a name)" Bryman & Cramer (1997)
- Hypotheses specify relations between concepts

Operationalisation

UNIVERSITY OF CANBERRA

- A concept needs to be operationally defined in order to be systematically researched
- "An operational definition specifies the procedures (operations) that will permit differences between individuals in respect of the concept(s) concerned to be precisely specified..." Bryman & Cramer (1997)
- "...What we are in reality talking about here is measurement, that is, the assignment of numbers to the units of analysis - be they people, organisations or nations - to which a concept refers" Bryman & Cramer (1997)

Menopause nomenclature example

Ambikairajah *et al. Reproductive Health* (2022) 19:29
https://doi.org/10.1186/s12978-022-01336-7

Reproductive Health

REVIEW Open Access

A review of menopause nomenclature

Ananthan Ambikairajah[1,2*], Erin Walsh[1] and Nicolas Cherbuin[1]

Abstract

Menopause nomenclature varies in the scholarly literature making synthesis and interpretation of research findings difficult. Therefore, the present study aimed to review and discuss critical developments in menopause nomenclature; determine the level of heterogeneity amongst menopause definitions and compare them with the Stages of Reproductive Aging Workshop criteria. Definitions/criteria used to characterise premenopausal and postmenopausal status were extracted from 210 studies and 128 of these studies were included in the final analyses. The main findings were that 39.84% of included studies were consistent with STRAW classification of *premenopause*, whereas 70.31% were consistent with STRAW classification of *postmenopause*. Surprisingly, major inconsistencies relating to premenopause definition were due to a total lack of reporting of any definitions/criteria for premenopause (39.84% of studies). In contrast, only 20.31% did not report definitions/criteria for postmenopause. The present findings indicate that there is a significant amount of heterogeneity associated with the definition of *premenopause*, compared with *postmenopause*. We propose three key suggestions/recommendations, which can be distilled from these findings. Firstly, premenopause should be transparently operationalised and reported. Secondly, as a minimum requirement, regular menstruation should be defined as the number of menstrual cycles in a period of at least 3 months. Finally, the utility of introducing normative age-ranges as supplementary criterion for defining stages of reproductive ageing should be considered. The use of consistent terminology in research will enhance our capacity to compare results from different studies and more effectively investigate issues related to women's health and ageing.

Plain Language Summary

The meaning of *menopause* is widely understood, but often imprecisely defined in research. The present findings revealed that there is a significant amount of heterogeneity associated with the definition of *premenopause*, compared with *postmenopause*. Three key suggestions/recommendations can be distilled from these findings. Firstly, premenopause should be transparently operationalised and reported. Secondly, as a minimum requirement, regular

Ambikairajah et al. (2022)

Low education example

UNIVERSITY OF CANBERRA

SSM - Population Health 12 (2020) 100654

Contents lists available at ScienceDirect

SSM - Population Health

journal homepage: http://www.elsevier.com/locate/ssmph

ELSEVIER

Review Article

What does (low) education mean in terms of dementia risk? A systematic review and meta-analysis highlighting inconsistency in measuring and operationalising education

Janet Maccora[a,b,d,*], Ruth Peters[a,b,d], Kaarin J. Anstey[a,b,c,d]

[a] School of Psychology, University of New South Wales, Sydney, New South Wales, Australia
[b] Neuroscience Research Australia (NeuRA), Sydney, New South Wales, Australia
[c] Australian National University, Canberra, Australian Capital Territory, Australia
[d] Australian Research Council Centre of Excellence in Populating Ageing Research (CEPAR), Australia

ARTICLE INFO

Keywords:
Education
Dementia risk
Systematic review
Modifiable risk factors for dementia
Dementia risk methodology
Social determinants of dementia risk

ABSTRACT

Low education is considered an important modifiable risk factor for dementia worldwide, despite the lack of a formal consensus definition of low education. The primary aim of this systematic review was to document and address the inconsistency in measuring and operationalising education in dementia studies. A secondary aim was to consider the dose of education required to reduce dementia risk. The protocol was registered at PROSPERO with registration ID CRD42018096168. CINAHL, Cochrane, PsycInfo, and Pubmed databases were searched using terms related to education, dementia and/or MCI, and incidence. Studies were eligible for inclusion if a risk ratio for education and any dementia, Alzheimer's Disease (AD), Vascular Dementia (VaD) or Mild Cognitive Impairment (MCI) was reported in a population cognitively healthy at baseline. Sample sizes for 65 studies meeting selection criteria ranged from 152 to 12,881, representing populations from 24 countries. Risk of bias, assessed using a tool designed specifically for dementia risk studies, was found to be medium or low for all studies. There were 23 continuous, 29 dichotomous, and 31 categorical operationalisations of education reported. Random effects meta-analyses from continuous operationalisations suggested each year of education reduced risk by eight percent for AD (95% CI:5–12%) and seven percent for any dementia (95% CI:6–9%). Dichotomous operationalisations indicated an increased risk for low education of 45% (95% CI:29–63%) for any dementia and 85% (95% CI:56–118%) for AD, however definitions of low education were heterogeneous, ranging from zero to 12 years. There were too few studies to produce summary ratios for VaD or MCI. We conclude that, while the evidence of an association between low education and dementia incidence is robust, inconsistency in the definition, measurement and operationalisation of education hinders the translation of this evidence into practical policy recommendations to reduce dementia risk.

Maccora et al. (2020)

Another definition

- The act of making a fuzzy concept measurable
- Social science often uses multi-item measures to assess related but distinct aspects of a fuzzy concept

Fuzzy vs not-fuzzy constructs

Fuzzy psychological constructs

- Measured using the average (composite score) of a number of correlated survey items. Examples include:
  - Felt emotions
  - Perceptions
  - Attitudes

Not fuzzy constructs

- Can be measured objectively and directly. Examples include:
  - Physiological responses to emotions (e.g. heart rate)
  - Physical properties (e.g. height, time etc)

Nuanced constructs

- Date of symptom onset for dementia?

Operationalisation - steps to take

1) Brainstorm indicators of a concept
2) Define the concept
3) Draft measurement items
4) Pre-test and pilot test
5) Examine psychometric properties - how precise are the measures?
6) Redraft/refine and re-test

The exception: if a measure already exists for the construct you are measuring, it's best to start there

Measurement error (a.k.a "noise")

- Measurement error is statistical deviation from the true value caused by the measurement procedure
- Observed score $=$ true score $\pm$ measurement error
- Measurement error:
  - Random error $=$ systematic error $\pm$ random error
  - Systematic error $=$ sampling error $\pm$ non-sampling error

Observed value $=$ true score $\pm$ measurement error (random/systematic)

Sources of systematic error

UNIVERSITY OF CANBERRA

Non-sampling

- Test reliability and validity
- Respondent bias - e.g. social desirability
- Researcher bias - e.g. conscious or unconscious
- Paradigm e.g., assumptions, focus on positivism, collection method

Sampling

- Non-representativeness of the sample

Systematic vs random error examples

UNIVERSITY OF CANBERRA

Systematic error

- Scales not calibrated properly, so every measurement is 50 grams too high
- This same error occurs for each measurement

Random error

- Measure the same cup of flour three times and get three slightly different readings
- The amount of error differs for each measurement

Measurement precision and noise

- The lower the measurement precision, the more participants are needed to make up for the "noise" in the measurements
- Even with a larger sample, noisy data can be hard to interpret
- Especically when testing and assessing individual clients, special care is needed when interpreting results of noisy tests.

Minimising non-sampling measurement error

- Standardise administration conditions with clear instructions and questions
- Minimise potential demand characteristics (e.g. train interviewers)
- Use multiple indicators for fuzzy constructs

Minimise sampling measurement error

UNIVERSITY OF CANBERRA

- Obtain a representative sample:
  - Use probability sampling, if possible
  - For non-probability sampling, use strategies to minimise selection bias
- Maximise response rate:
  - Pre-survey contact
  - Minimise length/time/hassle
  - Rewards/incentives
  - Coloured paper
  - Call backs/reminders

Minimising measurement error

- Ensure administrative accuracy:
  - Set up efficient coding, with well-labelled variables
  - Check data (double-check at least a portion of the data)

Psychometrics

UNIVERSITY OF CANBERRA

- The goal: To validly measure differences between individuals and groups in psychosocial qualities such as attitudes and personality

Tasks

- Develop approaches and procedures (theory and practice) for measuring psychological phenomena
- Design and test psychological measurement instrumentation (e.g. examine and improve reliability and validity of psychological tests)

Methods

UNIVERSITY OF CANBERRA

- Factor analysis
  - Exploratory
  - Confirmatory
- Classical test theory
  - Reliability
  - Validity

Reliability and validity

UNIVERSITY OF CANBERRA

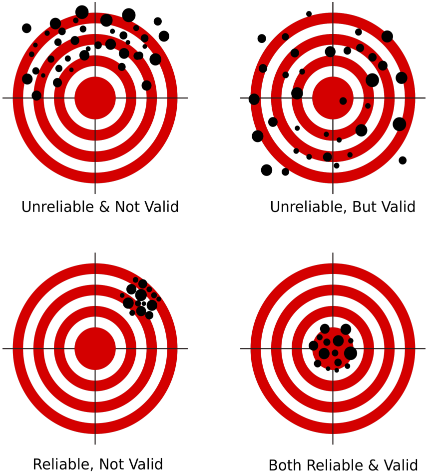
- Reliability and validity (“classical test theory”) are ways of evaluating psychological tests and measures
- Reliability is the consistency of:
  - Items within the measure - all items should be related
  - The measure over time - answer should correlate strongly over time
- Validity is about whether the measure actually measures what it is intended to measure

What is the difference?

- Reliability is generally thought to be necessary for validity, but it does not guarantee validity
- In practice, a test of a relatively changeable psychological construct such as emotions, may be valid, but not particularly reliable over time (because emotions are likely to fluctuate)

Reliability and validity

UNIVERSITY OF CANBERRA



Examples

- Reliability
  - A car which starts every time is reliable
  - A car which only starts sometimes is unreliable
- Validity
  - A car which always reaches the desired destination is valid
  - A car which misses the desired destination is not valid

Reliability and validity

- Reliability and validity are not inherent characteristics of measures. They are affected by the context and purpose of the measurement. For example, a measure that is valid for one purpose may not be valid for another purpose. Specifically, a test designed for adults and applied to children

Reliability

- Reproducibility or consistency of measurement

Types of reliability

- Internal consistency:
  - Correlation among multiple items in a factor
  - Cronbach's Alpha (α)
- Test-retest reliability:
  - Correlation between test at one time and another
  - Product-moment correlation (r)
- Inter-rater reliability:
  - Correlation between one observer and another
  - Kappa

Internal reliability/consistency

UNIVERSITY OF CANBERRA

- Internal reliability (also called internal consistency) refers to:
  - How well multiple items combine as a measure of a single concept
  - The extent to which responses to multiple items are consistent with one another
- Internal consistency is measured using:
  - Split-half reliability
  - Odd-even reliability
  - Cronbach's Alpha (α)

Recoding

UNIVERSITY OF CANBERRA

- If dealing with a mixture of positively and negatively scored items, remember to recode so that all items are measured in the same direction

Split-half reliability

- Sum the scores for the first half (e.g. 1, 2, 3) of the items
- Sum the scores for the second half (e.g. 4, 5, 6) of the items
- Compute a correlation between the sums of the two halves

Split-half reliability

1. It is okay if some groups have more of a chance in life than others
2. Inferior groups should stay in their place
3. To get ahead in life, it is sometimes okay to step on other groups
4. We should have increased social equality
5. It would be good if groups could be equal
6. We should do what we can to equalize conditions for different groups

- Correlate score between items 1-3 and 4-6
- Can also correlate scores using random items (i.e. 1,2,4 and 3,5,6)

Odd-even reliability

- Sum the scores for the odd (e.g. 1, 3, 5) items
- Sum the scores for the even (e.g. 2, 4, 6) items
- Compute a correlation between the sums of the two halves

Odd-even reliability

1. It is okay if some groups have more of a chance in life than others
2. Inferior groups should stay in their place
3. To get ahead in life, it is sometimes okay to step on other groups
4. We should have increased social equality
5. It would be good if groups could be equal
6. We should do what we can to equalize conditions for different groups

- Correlate score between items 2,4,6 and 1,3,5

Cronbach's alpha

- Averages all possible split-half reliability coefficients
- Akin to a single score which represents the degree of intercorrelation amongst the items
- Most commonly used indicator of internal reliability

Cronbach's alpha

- Rule of thumb:
  - $< .6$ = Unreliable
  - $.6$ = OK
  - $.7$ = Good
  - $.8$ = Very good, strong
  - $.9$ = Excellent
  - $> .95$ = may be overly reliable or redundant - depends on what is being measured
- Aim for between .7 and .9, but keep in mind scale length

Back to EFA: How many items per factor

- More items -> greater reliability (the more items, the more "rounded the measure")
- The minimum items to create a factor is 1
- No way to tell how reliable a single item is
- No maximum. Law of diminishing returns (i.e. each additional item will add less and less to the reliability)
- Typically $\sim$ 3 to 10 items per factor are used
- Final decision is subjective and depends on research context

Reliability analysis

- Example (simulated data): subjective memory concerns
- Note how removal of any item does not change cronbach's α

	Cronbach's alpha
scale	0.926

Item	Corrected item-total correlation	Cronbach's alpha if item dropped
Forget names of people	0.805	0.910
Misplace everyday items	0.801	0.910
Lose track of conversations	0.790	0.912
Forget why entered a room	0.817	0.907
Struggle to recall recent information	0.820	0.907

Reliability analysis

- Example (simulated data): attitudes toward healthy brain ageing
- Note how scores on item 4 do not correlate well with scores on other items and the removal of item 4 does improve cronbach's α

	Cronbach's alpha
scale	0.782

Item	Corrected item-total correlation	Cronbach's alpha if item dropped
Physical activity supports brain health	0.736	0.677
Cognitive activity supports brain health	0.697	0.695
Good sleep is important for brain health	0.720	0.681
Brain ageing is completely determined by genetics	0.039	0.883
Managing cardiometabolic health supports brain health	0.675	0.700

Question

- Which item will you delete?

Item	Corrected item-total correlation	Alpha if item deleted
V1	0.67	0.68
V2	0.77	0.55
V3	0.45	0.76
V4	0.65	0.66
V5	0.07	0.80
V6	0.63	0.72
V7	0.49	0.74

Question

- Which item will you delete?
  - Item 5

Item	Corrected item-total correlation	Alpha if item deleted
V1	0.67	0.68
V2	0.77	0.55
V3	0.45	0.76
V4	0.65	0.66
V5	0.07	0.80
V6	0.63	0.72
V7	0.49	0.74

Question

- Which items will you delete?

Item	Corrected item-total correlation	Alpha if item deleted
V1	0.67	0.68
V2	0.28	0.85
V3	0.45	0.76
V4	0.65	0.66
V5	0.55	0.69
V6	0.63	0.72
V7	0.15	0.88

Question

- Which items will you delete?
  - Items 2 and 7 - but need to do one at a time

Item	Corrected item-total correlation	Alpha if item deleted
V1	0.67	0.68
V2	0.28	0.85
V3	0.45	0.76
V4	0.65	0.66
V5	0.55	0.69
V6	0.63	0.72
V7	0.15	0.88

What if you have a short scale?

- The .7 threshold for the alpha may be hard to meet with a scale with just 3-4 items
- Sometimes short scales are accepted as reliable if they have an alpha around .6 by checking the item-rest correlations.
- Two rules of thumb:
  - Item-rest correlations should be around .5 and above – strong correlations between items belonging to the same scale
  - Item-rest correlations should be between .2 and .7

What if you have an even shorter scale?

- If your factor has only 2 items belonging to it, then it is not appropriate to calculate a Cronbach's alpha
- Instead, check the correlation between items and report this
- You would expect items to be moderately to strongly positively correlated (at around .5 or above)

Validity

- Does the measure actually measure the construct?
- Validity is multifaceted and includes:
  - Comparing wording of the items with theory and expert opinion
  - Examining correlations with similar and dissimilar measures
  - Testing how well the measure predicts things in future

Types of validity

- The following is a list of types of validity ordered from easiest to hardest to demonstrate:
  - Face validity
  - Content validity
  - Criterion validity
    - Concurrent validity
    - Predictive validity
  - Construct validity
    - Convergent validity
    - Discriminant validity

Face validity

UNIVERSITY OF CANBERRA

- Asks:
  - "At face-value, do the questions appear to measure what the tests purports to measure?"
- Important for:
  - Respondent to feel secure and buy in to the study
- How assessed:
  - Read the test items

Content validity

- Asks:
  - “Are the questions measuring the complete construct?”
- Important for:
  - Ensuring holistic assessment
- How assessed:
  - Diverse item generation (literature review, theory, interviews, expert review)

Criterion validity

UNIVERSITY OF CANBERRA

- Asks:
  - "Can a test score predict real world outcomes?"
- Important for:
  - Test relevance and usefulness
- How assessed:
  - Concurrent validity: Correlate test scores with recognised external criteria (present or past). For example:
    - Political ideology with how they voted at previous election
    - Offender risk rating with seriousness of crime
  - Predictive validity: Correlate test scores with future outcome. For example:
    - Political ideology with how they vote at the next election
    - Offender risk rating with recidivism

Construct validity

- Asks:
  - "Does the test assess the construct it purports to?" (i.e. the truth, the whole truth and nothing but the truth)
- Important for:
  - Making inferences from operationalisations to theoretical constructs
- How assessed:
  - Theoretical (is the theory about the construct valid?)
  - Statistical:
    - Convergent - correlation with similar measures i.e. other scales that measure similar constructs
    - Discriminant - not correlated with other constructs

Composite scores

- Combine item-scores into an overall factor score which represents individual differences for the target construct.
- The new composite score can then be used for:
  - Descriptive statistics and histograms
  - Correlations
  - As independent variables and/or dependent variables in inferential analyses such as multiple linear regression (MLR) models and analyses of variance (ANOVA)

How do you create composite scores?

- Unit weighting
- Regression weighting

Unit weighting

UNIVERSITY OF CANBERRA

- Average (or total) of item scores within a factor.
- Each variable is equally weighted
- $X = mean(variable_1 ... variable_n)$

Dealing with missing data

- To maximise the sample size, consider computing composite scores in a way that allows for some missing data

Compute a mean score - Jamovi

- In jamovi
- MEAN (v1, v2, v3, v4, v5, v6)
  - This will give a missing value if any items have missing data
- MEAN (v1, v2, v3, v4, v5, v6, ignore_missing=1)
  - This will calculate a mean for all participants unless they have missing scores for more than one item (not a good idea if scale is not reliable or if one item influence cronbach alpha)
- Another approach is to specify a minimum number of items. If the minimum isn't available, then jamovi will leave the composite score as missing for that participant.

How many items can be missed?

- Depends on overall reliability. A rule of thumb:
  - Allow 1 missing per 4 to 5 items
  - Allow 2 missing per 6 to 8 items
  - Allow 3+ missing per 9+ items
- Note – it is fine if you do not allow any missing data and use the mean only for full cases.
- A researcher may decide to be more or less conservative depending on the factors' reliability, sample size, and the nature of the study.

Regression weighting

- Factor score regression weighting
- The contribution of each item to the composite score is weighted to reflect responses to some items more than other items.
- $X = (.20 \times a) + (.19 \times b) + (.27 \times c) + (.34 \times d)$
- This is arguably more valid, but the advantage may be marginal, and it makes factor scores between studies more difficult to compare.
- Jamovi can calculate this for you from your EFA, with weightings based on factor loadings

Regression weighting

UNIVERSITY OF CANBERRA

- This is why the correlations between composite scores in the exploratory factor analysis output might differ from the correlations between composite scores that you calculate using mean scores

Summary

UNIVERSITY OF CANBERRA

- Psychometrics
  - Science of psychological measurement
  - Goal: Validly measure individual psychosocial difference
  - Design and test psychological measures e.g. using:
    - Factor analysis
    - Reliability and validity

Operationalisation - steps to take

UNIVERSITY OF CANBERRA

1) Brainstorm indicators of a concept
2) Define the concept
3) Draft measurement items
4) Pre-test and pilot test
5) Examine psychometric properties - how precise are the measures?
6) Redraft/refine and re-test

The exception: if a measure already exists for the construct you are measuring, it's best to start there

Measurement error (a.k.a "noise")

- Deviation of measure from true score
- Measurement error = random and systematic error
- Sources:
  - Non-sampling (e.g., paradigm, respondent bias, researcher bias)
  - Sampling (e.g., non-representativeness)
- Minimise error with:
  - Well-designed measures
  - Representative sampling
  - Reduce demand effects
  - Maximise response rate
  - Ensure administrative accuracy

Reliability

- Consistency or reproducibility
- Types:
  - Internal consistency
  - Test-retest reliability
- Internal consistency
  - Split half
  - Odd-even
  - Cronbach's α
  - Only 2 items? Correlation
- Rule of thumb for cronbach's α:
  - $< .6 =$ Unreliable
  - $.6 =$ OK
  - $.7 =$ Good
  - $.8 =$ Very good, strong
  - $.9 =$ Excellent
  - $> .95 =$ may be overly reliable or redundant - depends on what is being measured
    - Aim for between .7 and .9, but keep in mind scale length

Validity

UNIVERSITY OF CANBERRA

- Extent to which a measure measures what it is intended to measure
- Multifaceted
  - Compare with theory and expert opinion
  - Correlations with similar and dissimilar measures
  - Predicts future

Create composite scores

- Unit weighting
  - Total of items or average of items
- Regression weighting
  - Each item is weighted by its importance to measuring the underlying factor (based on regression weights)

Next week - multiple linear regression

- Correlation (review)
- Simple linear regression
- Multiple linear regression

Contributions to this course

Dr James Neill

Dr Samantha Stanley

Dr Jeroen van Boxtel

References I

Ambikairajah, A., Walsh, E., & Cherbuin, N. (2022). A review of menopause nomenclature. *Reproductive Health*, *19*(1), 29. https://doi.org/10.1186/s12978-022-01336-7

Bryman, A., & Cramer, D. (1997). Concepts and their measurement. *Quantitative Data Analysis, with SPSS for Windows*.

Comrey, A. L., & Lee, H. B. (2013). *A first course in factor analysis*. Psychology press. https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.4324/9781315827506&type=googlepdf

Maccora, J., Peters, R., & Anstey, K. J. (2020). What does (low) education mean in terms of dementia risk? A systematic review and meta-analysis highlighting inconsistency in measuring and operationalising education. *SSM - Population Health*, *12*, 100654. https://doi.org/10.1016/j.ssmph.2020.100654