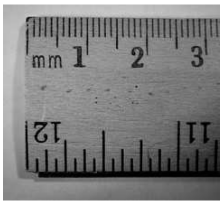
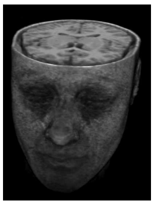
Psychometric Instrument Development





Lecture 6

Survey Research & Design in Psychology

James Neill, 2012

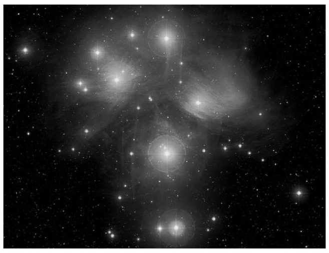
Overview

1. Recap: Exploratory factor analysis
2. Concepts & their measurement
3. Measurement error
4. Psychometrics
5. Reliability & validity
6. Composite scores
7. Writing up instrument development

Readings: Psychometrics

1. Bryman & Cramer (1997). Concepts and their measurement. [chapter - ereserve]
2. DeCoster, J. (2000). Scale construction notes. http://www.stat-help.com/scale.pdf
3. Howitt & Cramer (2005). Reliability and validity: Evaluating the value of tests and measures. [chapter – ereserve]
4. Wikiversity. Reliability and validity - http://en.wikiversity.org/wiki/Reliability_and_validity

Recap: Exploratory Factor Analysis



What is factor analysis?

- FA is:
  - –a family of multivariate correlational data analysis methods
  - –used to identify clusters of covariance (called factors)
- Two main types:
  - –Exploratory factor analysis (EFA)
  - –Confirmatory factor analysis (CFA)

EFA assumptions

- Sample size
  - – 5+ cases per variables (min.)
  - – 20+ cases per variable (ideal)
  - – Another guideline: *Or N* > 200
- Check bivariate outliers & linearity
- Factorability: check any of:
  - – Correlation matrix: Some over .3?
  - – Anti-image correlation matrix diags > .5
  - – Measures of Sampling Adequacy
    - KMO > ~ .5 to 6; Bartlett's sig?

Summary of EFA steps / process

1. Test assumptions
   - Sample size, Outliers & linearity, Factorability
2. Select type of analysis
   - PC/PAF, Orthorgonal/Oblique rotation

Summary of EFA steps / process

3. Determine no. of factors
   - Theory, Kaiser's criterion, Eigen Values, Scree plot, % variance explained, interpretability of weakest factor
4. Select items
   - Check factor loadings to identify which items belong in which factor; drop items 1-by-1 if primarily loading low and/or cross-loadings high and/or item wording doesn't belong to meaning of factor.

Summary of EFA steps / process

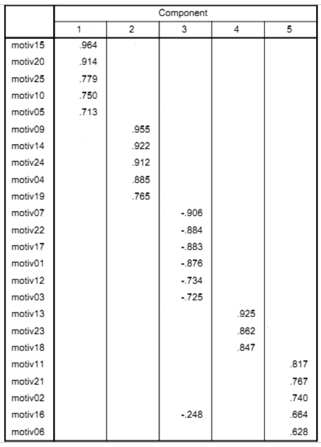
5. Name and define factors
6. Examine correlations amongst factors
7. Check factor structure for sub-groups
8. Analyse internal reliability
9. Compute composite scores

Covered in this lecture

Example EFA:
University student motivation

- 271 UC students responded to 24 university student motivation statements in 2008 using an 8-point Likert scale (False to True) e.g.,
  "I study at university … "
  - to enhance my job prospects.
  - because other people have told me I should.
- EFA PC Oblimin revealed 5 factors

Example EFA: Pattern matrix



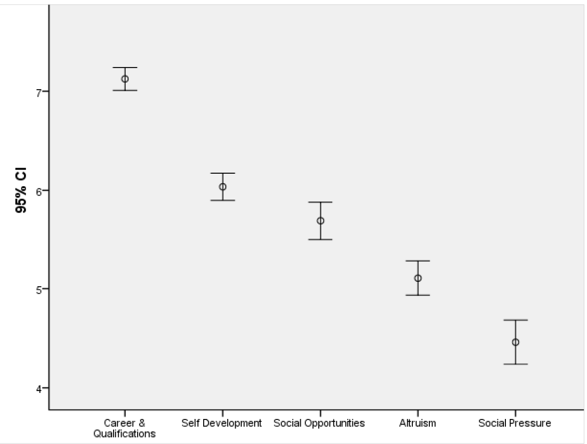
	Component				
	1	2	3	4	5
motiv15	.964				
motiv20	.914				
motiv25	.779				
motiv10	.750				
motiv05	.713				
motiv09		.955			
motiv14		.922			
motiv24		.912			
motiv04		.885			
motiv19		.765			
motiv07			-.906		
motiv22			-.884		
motiv17			-.883		
motiv01			-.876		
motiv12			-.734		
motiv03			-.725		
motiv13				.925	
motiv23				.862	
motiv18				.847	
motiv11					.817
motiv21					.767
motiv02					.740
motiv16			-.248		.664
motiv06					.628

Example EFA:
University student motivation

- Career & Qualifications
  (6 items; α = .92)
- Self Development
  (5 items; α = .81)
- Social Opportunities
  (3 items; α = .90)
- Altruism
  (5 items; α = .90)
- Social Pressure
  (5 items; α = .94)

Example EFA:
Factor correlations

Motivation	CQ	SD	SO	AL	SP
Career & Qualif.		.26	.25	.24	.06
Self Develop.			.33	.55	-.18
Social Enjoyment				.26	.33
Altruism					.11
Social Pressure					



Exploratory factor analysis: Q & A

?

Concepts & their measurement

Operationalising fuzzy concepts

Concepts & their measurement: Bryman & Cramer (1997)

Concepts

- form a linchpin in the process of social research
- express common elements in the world (to which we give a name)

Hypotheses

- express relations between **concepts**

Concepts & their measurement: Bryman & Cramer (1997)

“Once formulated, a concept … will need to be ***operationally defined***, in order for systematic research to be conducted in relation to it..."

Concepts & their measurement: Bryman & Cramer (1997)

“...An ***operational definition*** specifies the procedures (operations) that will permit differences between individuals in respect of the concept(s) concerned to be precisely specified..."

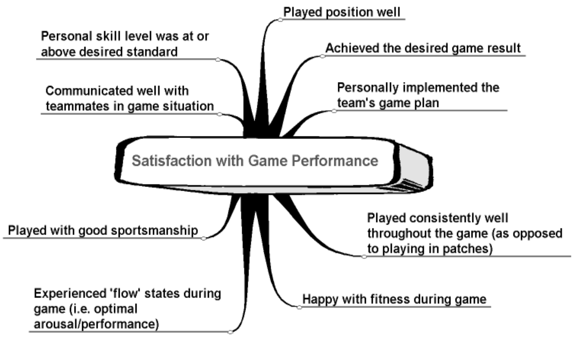
Operationalisation

- ...is the act of making a ***fuzzy concept*** measurable.
- Social sciences often use ***multi-item measures*** to assess related but distinct aspects of a fuzzy concept.

Operationalisation steps

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

Operationalisating a fuzzy concept: Example (Brainstorming indicators)



Satisfaction with Game Performance

Played position well

Achieved the desired game result

Personally implemented the team's game plan

Played consistently well throughout the game (as opposed to playing in patches)

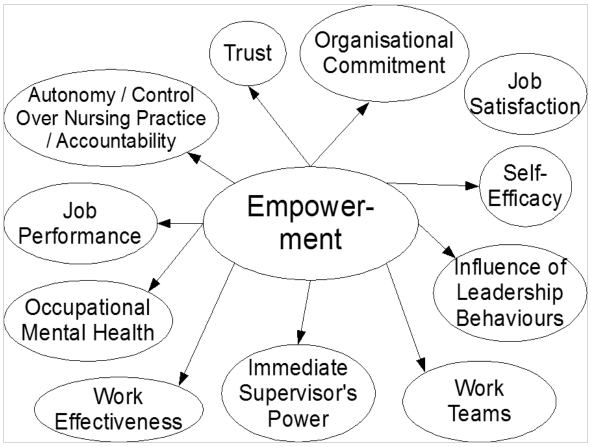
Happy with fitness during game

Experienced 'flow' states during game (i.e. optimal arousal/performance)

Played with good sportsmanship

Communicated well with teammates in game situation

Personal skill level was at or above desired standard



Empower- ment

Trust

Organisational Commitment

Job Satisfaction

Self- Efficacy

Influence of Leadership Behaviours

Work Teams

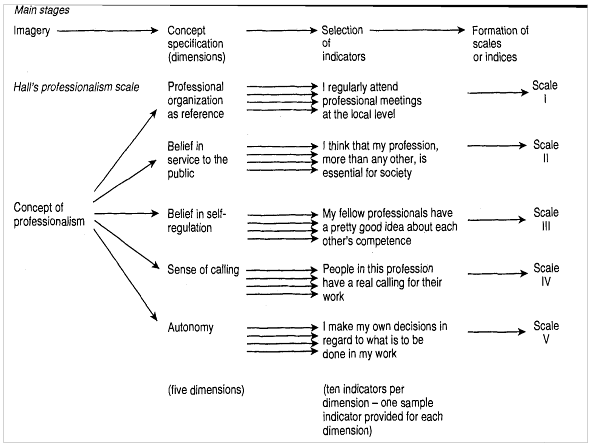
Immediate Supervisor's Power

Work Effectiveness

Occupational Mental Health

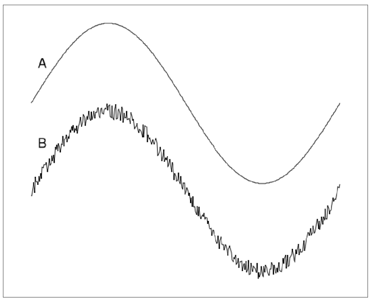
Job Performance

Autonomy / Control Over Nursing Practice / Accountability



Main stages			
Imagery	Concept specification (dimensions)	Selection of indicators	Formation of scales or indices
Hall's professionalism scale	Professional organization as reference	I regularly attend professional meetings at the local level	Scale I
	Belief in service to the public	I think that my profession, more than any other, is essential for society	Scale II
Concept of professionalism	Belief in self-regulation	My fellow professionals have a pretty good idea about each other's competence	Scale III
	Sense of calling	People in this profession have a real calling for their work	Scale IV
	Autonomy	I make my own decisions in regard to what is to be done in my work	Scale V
	(five dimensions)	(ten indicators per dimension – one sample indicator provided for each dimension)	

Measurement error



Measurement precision & noise

“The lower the precision, the more subjects you'll need in your study to make up for the "noise" in your measurements. Even with a larger sample, noisy data can be hard to interpret. And if you are an applied scientist in the business of testing and assessing clients, you need special care when interpreting results of noisy tests.”

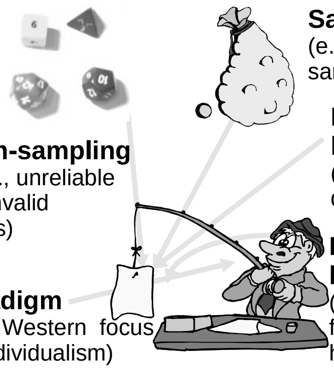
http://www.sportsci.org/resource/stats/precision.html

Measurement error

Measurement error is any deviation from the **true value** caused by the measurement procedure.

- **Observed score** =
  true score + measurement error
- Measurement error =
  systematic error + random error

Sources of measurement error



Sampling
(e.g., non-rep. sample)

Respondent bias
(e.g., social desirability)

Non-sampling
(e.g., unreliable or invalid tests)

Researcher bias
(e.g., researcher favours a hypothesis)

Paradigm
(e.g., Western focus on individualism)

To minimise measurement error

Use **well designed measures**:

- Multiple indicators for fuzzy constructs
- Sensitive to target constructs
- Clear instructions and questions

To minimise measurement error

Reduce demand effects:

- Train interviewers
- Use standard administration survey protocol

To minimise measurement error

Obtain a representative sample:

- Use probability-sampling if possible
- Minimise bias in selection for non-probability sampling

Maximise response rate:

- Pre-survey contact
- Minimise length / time / hassle
- Offer rewards / incentives
- Coloured paper
- Call backs / reminders

To minimise 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

Psychometrics: Goal

To validly measure differences between individuals and groups in psychosocial qualities such as attitudes and personality.

Psychometrics: As test-taking grows, test-makers grow rarer

> "Psychometrics, one of the most obscure, esoteric and cerebral professions in America, is now also one of the hottest."
>
> \- As test-taking grows, test-makers grow rarer, David M. Herszenhor, May 5, 2006, New York Times

e.g., due to increased testing of educational and psychological capacity and performance

Psychometric tasks

- Develop approaches and procedures (theory and practice) for measurement of psychological phenomena
- Design and test psychological measurement instrumentation

  e.g., examine and improve reliability and validity

But remember

Psychometric methods

- Factor analysis
  - Exploratory
  - Confirmatory
- Classical test theory:
  - Reliability
  - Validity
- Item response modeling

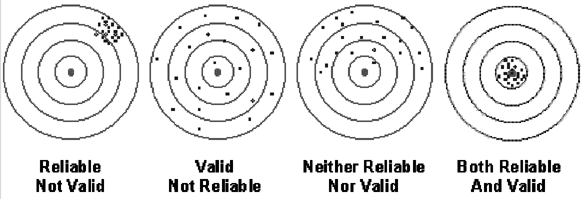
Reliability & Validity

Types of reliability

- **Internal consistency**
  - correlations amongst multiple items in a factor
    - Split-half reliability
    - Odd-even reliability
    - Cronbach's Alpha (α)
    - Alternate forms reliability
- **Test-retest reliability**
  - correlation over time
    - Product-moment correlation (*r*)

Reliability vs. validity

Reliability is generally thought to be necessary for validity, but it does not guarantee validity.



Reliability

Reproducibility of a measurement

Reliability and validity
(Howitt & Cramer, 2005)

Reliability and validity are the means by which we evaluate the value of psychological tests and measures.

- Reliability is about
  - the consistency of the items within the measure
  - the consistency of a measure over time
- Validity concerns the evidence that the measure actually measures what it is intended to measure.

Reliability and validity
(Howitt & Cramer, 2005)

- Reliability and validity are not inherent characteristics of measures. They are affected by the context and purpose of the measurement → a measure that is valid for one purpose may not be valid for another purpose.

Reliability rule of thumb

<.6 = Unreliable
.6 = OK
.7 = Good
.8 = Very good, strong
.9 = Excellent
>.95 = may be overly reliable or redundant – this is subjective and whether a scale is overly reliable depends also on the nature what is being measured

Reliability rule of thumb

Table 7 Fabrigar et al (1999).

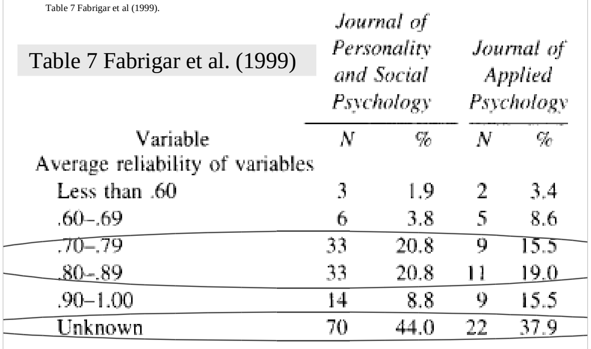


Table 7 Fabrigar et al. (1999)

Variable	*Journal of Personality and Social Psychology*		*Journal of Applied Psychology*	
	N	%	*N*	%
Average reliability of variables				
Less than .60	3	1.9	2	3.4
.60–.69	6	3.8	5	8.6
.70–.79	33	20.8	9	15.5
.80–.89	33	20.8	11	19.0
.90–1.00	14	8.8	9	15.5
Unknown	70	44.0	22	37.9

Rule of thumb - reliability coefficients should be over .70, up to approx. .90

Internal consistency
(or internal reliability)

Internal consistency is about:

- 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
(Recoding)

Remember to:

- Ensure that negatively-worded items are recoded

Types of internal consistency: Split-half reliability

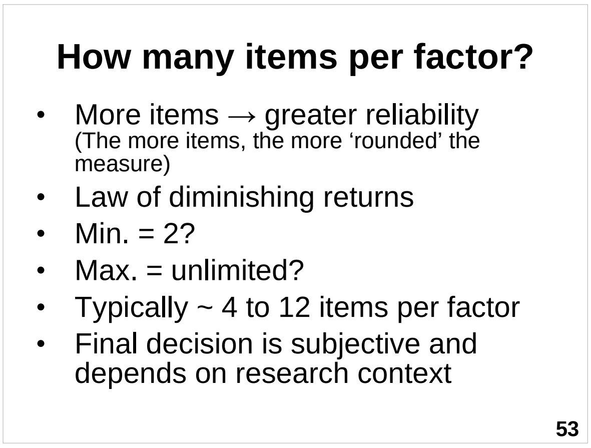
- Sum the first half of the items.
- Sum the second half of the items.
- Compute a correlation between the sums of the two halves.

Types of internal consistency - Odd-even reliability

- Sum items 1, 3, 5, etc.
- Sum items 2, 4, 6, etc.
- Compute a correlation between the sums of the two halves.

Types of internal reliability: Alpha reliability (Cronbach's α)

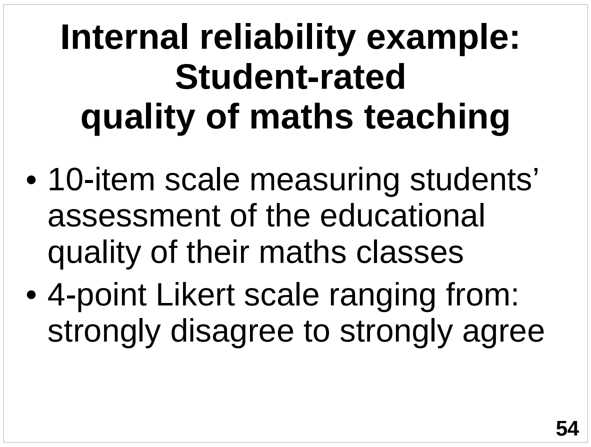
- Averages all possible split-half reliability coefficients.
- Akin to a single score which represents the degree of intercorrelation amongst the items.



How many items per factor?

- More items → greater reliability
  (The more items, the more 'rounded' the measure)
- Law of diminishing returns
- Min. = 2?
- Max. = unlimited?
- Typically ~ 4 to 12 items per factor
- Final decision is subjective and depends on research context

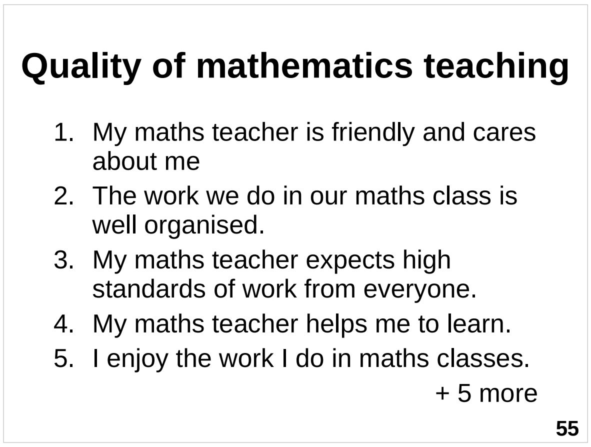
53



Internal reliability example: Student-rated quality of maths teaching

- 10-item scale measuring students' assessment of the educational quality of their maths classes
- 4-point Likert scale ranging from: strongly disagree to strongly agree

54



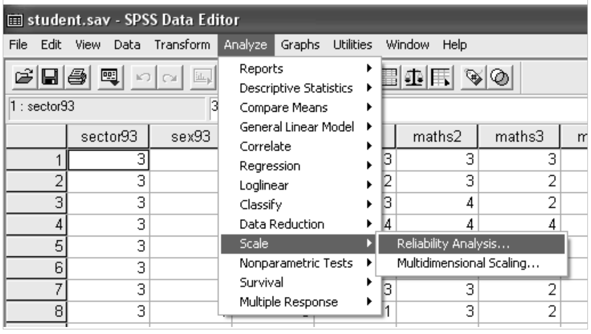
Quality of mathematics teaching

1. My maths teacher is friendly and cares about me
2. The work we do in our maths class is well organised.
3. My maths teacher expects high standards of work from everyone.
4. My maths teacher helps me to learn.
5. I enjoy the work I do in maths classes.

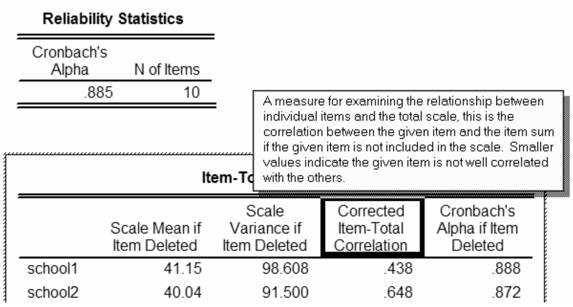
+ 5 more

55

Internal reliability example: Quality of maths teaching



SPSS: Corrected Item-total correlation



Reliability Statistics

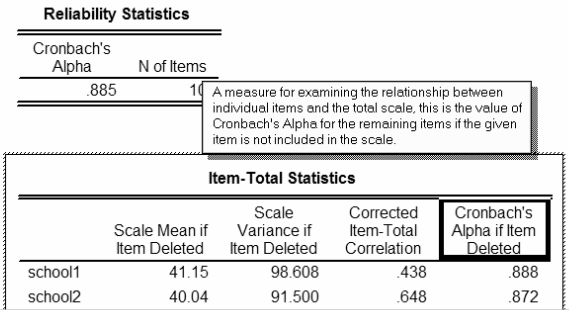
Cronbach's Alpha	N of Items
.885	10

A measure for examining the relationship between individual items and the total scale, this is the correlation between the given item and the item sum if the given item is not included in the scale. Smaller values indicate the given item is not well correlated with the others.

Item-T

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
school1	41.15	98.608	.438	.888
school2	40.04	91.500	.648	.872

SPSS: Cronbach's α



Reliability Statistics

Cronbach's Alpha	N of Items
.885	10

A measure for examining the relationship between individual items and the total scale, this is the value of Cronbach's Alpha for the remaining items if the given item is not included in the scale.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
school1	41.15	98.608	.438	.888
school2	40.04	91.500	.648	.872

SPSS: Reliability output

```
Item-total Statistics

               Scale          Scale      Corrected
               Mean         Variance       Item-            Alpha
              if Item        if Item       Total           if Item
              Deleted        Deleted    Correlation        Deleted

MATHS1        25.2749        25.5752        .6614            .8629
MATHS2        25.0333        26.5322        .6235            .8661
MATHS3        25.0192        30.5174        .0996            .9021
MATHS4        24.9786        25.8671        .7255            .8589
MATHS5        25.4664        25.6455        .6707            .8622
MATHS6        25.0813        24.9830        .7114            .8587
MATHS7        25.0909        26.4215        .6208            .8662
MATHS8        25.8699        25.7345        .6513            .8637
MATHS9        25.0340        26.1201        .6762            .8623
MATHS10       25.4642        25.7578        .6495            .8638


Reliability Coefficients

N of Cases =   1353.0                  N of Items = 10

Alpha =    .8790
```

SPSS: Reliability output

```
Item-total Statistics

               Scale          Scale      Corrected
               Mean         Variance       Item-            Alpha
              if Item        if Item       Total           if Item
              Deleted        Deleted    Correlation        Deleted

MATHS1        22.2694        24.0699        .6821            .8907
MATHS2        22.0280        25.2710        .6078            .8961
MATHS4        21.9727        24.4372        .7365            .8871
MATHS5        22.4605        24.2235        .6801            .8909
MATHS6        22.0753        23.5423        .7255            .8873
MATHS7        22.0849        25.0777        .6166            .8955
MATHS8        22.8642        24.3449        .6562            .8927
MATHS9        22.0280        24.5812        .7015            .8895
MATHS10       22.4590        24.3859        .6524            .8930


Reliability Coefficients

N of Cases =   1355.0                  N of Items =  9

Alpha =    .9024
```

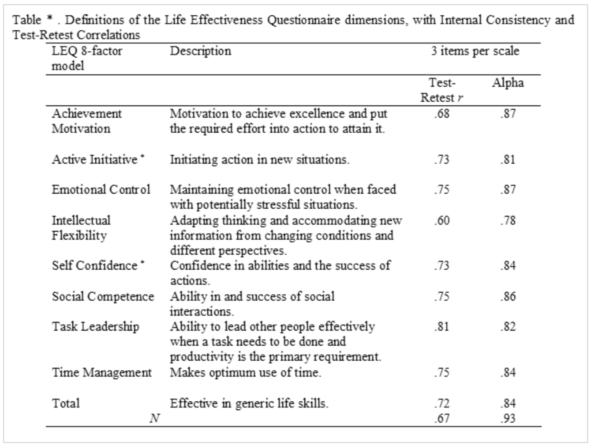


Table * . Definitions of the Life Effectiveness Questionnaire dimensions, with Internal Consistency and Test-Retest Correlations

LEQ 8-factor model	Description	3 items per scale	
		Test-Retest *r*	Alpha
Achievement Motivation	Motivation to achieve excellence and put the required effort into action to attain it.	.68	.87
Active Initiative *	Initiating action in new situations.	.73	.81
Emotional Control	Maintaining emotional control when faced with potentially stressful situations.	.75	.87
Intellectual Flexibility	Adapting thinking and accommodating new information from changing conditions and different perspectives.	.60	.78
Self Confidence *	Confidence in abilities and the success of actions.	.73	.84
Social Competence	Ability in and success of social interactions.	.75	.86
Task Leadership	Ability to lead other people effectively when a task needs to be done and productivity is the primary requirement.	.81	.82
Time Management	Makes optimum use of time.	.75	.84
Total	Effective in generic life skills.	.72	.84
N		.67	.93

Validity

Validity is the extent to which an instrument actually measures what it purports to measure.

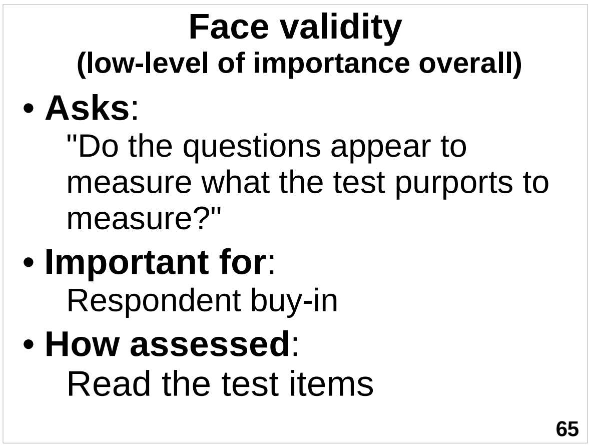
Validity = does the test measure what its meant to measure?

Validity

- Validity is multifaceted and includes:
  - Correlations with similar measures
  - How the measure performs in relation to other variables
  - How well the measure helps to predict the future

Types of validity

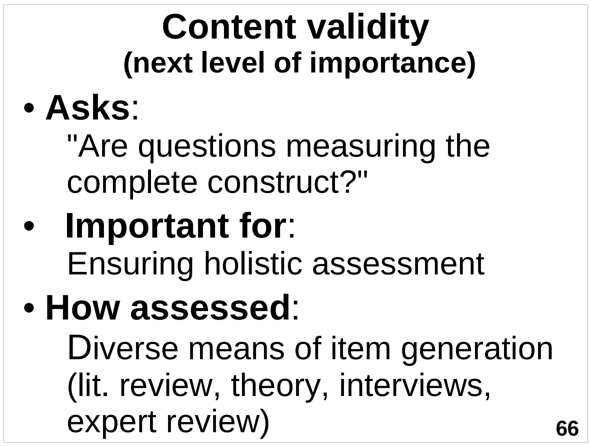
- Face validity
- Content validity
- Construct validity
- Criterion validity



Face validity

(low-level of importance overall)

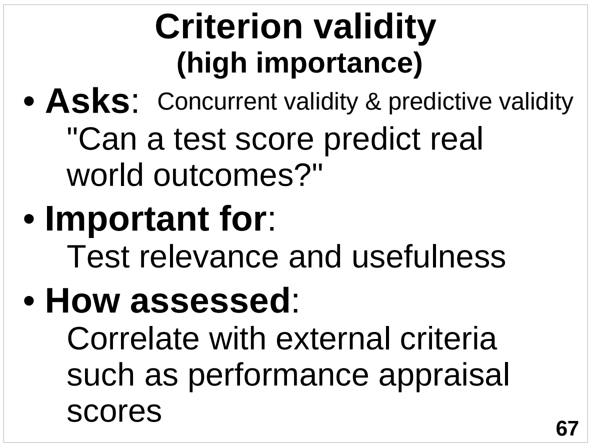
- **Asks**:
  "Do the questions appear to measure what the test purports to measure?"
- **Important for**:
  Respondent buy-in
- **How assessed**:
  Read the test items



Content validity

(next level of importance)

- **Asks**:
  "Are questions measuring the complete construct?"
- **Important for**:
  Ensuring holistic assessment
- **How assessed**:
  Diverse means of item generation (lit. review, theory, interviews, expert review)



Criterion validity

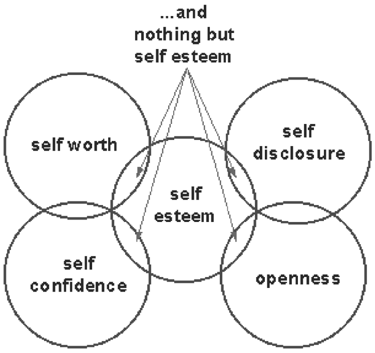
(high importance)

- **Asks**: Concurrent validity & predictive validity
  "Can a test score predict real world outcomes?"
- **Important for**:
  Test relevance and usefulness
- **How assessed**:
  Correlate with external criteria such as performance appraisal scores

Construct validity
(high importance)

- **Asks**:
  Does the test assess the construct it purports to? ("the truth, the whole truth and nothing but the truth.")
- **Important for**:
  Making inferences from operationalisations to theoretical constructs
- **How assessed**:
  Statistical (common factor underlying several measurements using different observable indicators?) and theoretical (is the theory about the construct valid?)

Construct validity
(high importance)



Composite Scores

Composite scores
(Factor scores)

Combine item-scores into overall scores which represent individual differences in the target constructs.

These new 'continuous' variables can then be used for:

- Descriptive statistics
- As IVs and/or DVs in inferential analyses such as MLR and ANOVA

Composite scores
(Factor scores)

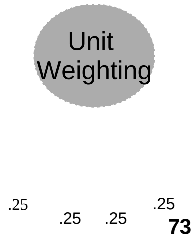
There are two ways of creating composite scores:

- Unit weighting
- Regression weighting

Unit weighting

Average (or total) of all variables in a factor.
(each variable is equally weighted)

$X = \text{mean}(y_1 \ldots y_p)$



Creating composite scores: Dealing with missing data

It can be helpful to maximise sample size by allowing for some missing data.

Reliability rule of thumb

<.6 = Unreliable
.6 = OK
.7 = Good
.8 = Very good, strong
.9 = Excellent
>.95 = may be overly reliable or redundant – this is subjective and whether a scale is overly reliable depends also on the nature what is being measured

Composite scores:
Missing data

SPSS syntax:

```
Compute X = mean (v1, v2, v3, v4, v5, v6)
```

You can specify a min. # of items. If the min. isn't available, the composite score will be missing: e.g.,

```
Compute X = mean.4 (v1, v2, v3, v4, v5, v6)
```

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

A researcher may decide to be more or less conservative depending on the factors' reliability, sample size, and the nature of the study.

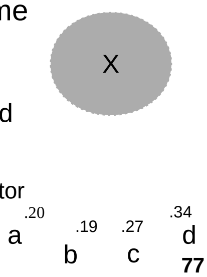
Regression weighting

Factor score regression weighting

The contribution of each item to the composite score is weighted to reflect some items more than other items.

X = 20*a + .19*b + .27*c + .34*d

This is arguably more valid, but it may be marginal, and it makes factor scores difficult to compare.

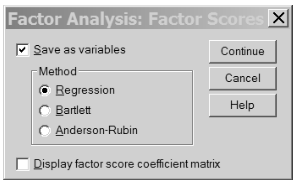


77

Regression weighting

Two calculation methods:

- Manual (use Compute)
- Automatic (use Factor Analysis – Factor Scores)



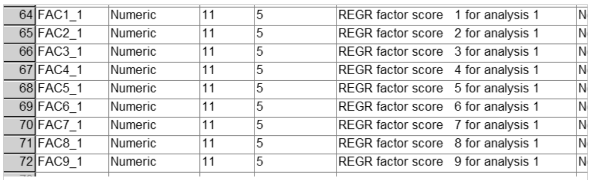
☑ Save as variables

Method

● Regression
○ Bartlett
○ Anderson-Rubin

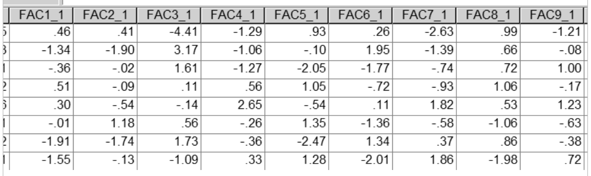
□ Display factor score coefficient matrix

78



64	FAC1_1	Numeric	11	5	REGR factor score 1 for analysis 1	N
65	FAC2_1	Numeric	11	5	REGR factor score 2 for analysis 1	N
66	FAC3_1	Numeric	11	5	REGR factor score 3 for analysis 1	N
67	FAC4_1	Numeric	11	5	REGR factor score 4 for analysis 1	N
68	FAC5_1	Numeric	11	5	REGR factor score 5 for analysis 1	N
69	FAC6_1	Numeric	11	5	REGR factor score 6 for analysis 1	N
70	FAC7_1	Numeric	11	5	REGR factor score 7 for analysis 1	N
71	FAC8_1	Numeric	11	5	REGR factor score 8 for analysis 1	N
72	FAC9_1	Numeric	11	5	REGR factor score 9 for analysis 1	N

Variable view



FAC1_1	FAC2_1	FAC3_1	FAC4_1	FAC5_1	FAC6_1	FAC7_1	FAC8_1	FAC9_1
.46	.41	-4.41	-1.29	.93	.26	-2.63	.99	-1.21
-1.34	-1.90	3.17	-1.06	-.10	1.95	-1.39	.66	-.08
-.36	-.02	1.61	-1.27	-2.05	-1.77	-.74	.72	1.00
.51	-.09	.11	.56	1.05	-.72	-.93	1.06	-.17
.30	-.54	-.14	2.65	-.54	.11	1.82	.53	1.23
-.01	1.18	.56	-.26	1.35	-1.36	-.58	-1.06	-.63
-1.91	-1.74	1.73	-.36	-2.47	1.34	.37	.86	-.38
-1.55	-.13	-1.09	.33	1.28	-2.01	1.86	-1.98	.72

Data view

Writing up instrument development

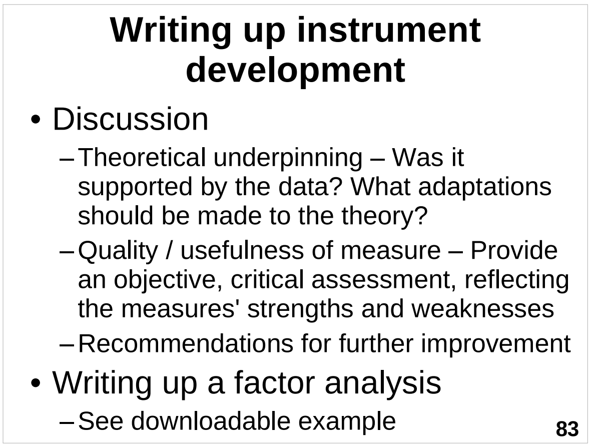
Writing up instrument development

- Introduction
  - Lit. review of underlying factors – theory and research
- Method
  - Materials/Instrumentation – summarise how the measures were developed and their expected factor structure

    e.g., present a table of the expected factors and their operational definitions.

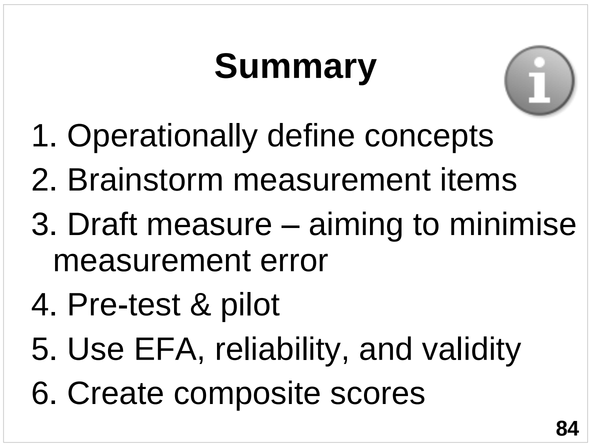
Writing up instrument development

- Results
  - Factor analysis
    - Assumption testing/ factorability
    - Extraction method & rotation
    - # of factors & items removed
    - Names & definitions of factors
    - Item factor loadings & communalities
    - Factor correlations
  - Reliability & composite scores



Writing up instrument development

- Discussion
  - Theoretical underpinning – Was it supported by the data? What adaptations should be made to the theory?
  - Quality / usefulness of measure – Provide an objective, critical assessment, reflecting the measures' strengths and weaknesses
  - Recommendations for further improvement
- Writing up a factor analysis
  - See downloadable example



Summary

1. Operationally define concepts
2. Brainstorm measurement items
3. Draft measure – aiming to minimise measurement error
4. Pre-test & pilot
5. Use EFA, reliability, and validity
6. Create composite scores

Questions

?

References

1. Allen, P. & Bennett, K. (2008). *Reliability analysis* (Ch 15) in SPSS for the health & behavioural sciences (pp. 205-218). South Melbourne, Victoria, Australia: Thomson.
2. Bryman, A. & Cramer, D. (1997). Concepts and their measurement (Ch. 4). In *Quantitative data analysis with SPSS for Windows: A guide for social scientists* (pp. 53-68). Routledge.
3. DeCoster, J. (2000). *Scale construction notes.* http://www.stat-help.com/scale.pdf (pdf)
4. Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272-299.
5. Fowler, F. (2002). Designing questions to be good measures. In *Survey research methods* (3rd ed.)(pp. 76-103). Thousand Oaks, CA: Sage. Ereserve.
6. Howitt, D. & Cramer, D. (2005). Reliability and validity: Evaluating the value of tests and measures (Ch. 13). In *Introduction to research methods in psychology* (pp. 218-231). Harlow, Essex: Pearson. eReserve.

Open Office Impress

- This presentation was made using Open Office Impress.
- Free and open source software.
- http://www.openoffice.org/product/impress.html

Version 2 Launch Programme

Welcome to the OpenOffice.org 2 Launch
Edinburgh International Conference Centre
Edinburgh, Scotland

OpenOffice.org