Psychometric workup of an instrument: Late phase and CFA

Dr Cameron Hurst
cphurst@gmail.com

DAMASAC and CEU, Khon Kaen University

1st December, 2558

What we will cover....

1. Introduction
2. Assessing model fit
3. Model specification
4. Software for CFA
5. Worked example
6. Criterion-based validity
   - Case study of discriminant validity: Suicidality in Thai adolescents

Revision: What is Factor Analysis?

- A **Factor** usually refers to some latent or unobserved (or at least unmeasured) construct
- Factor Analysis: set of techniques based on correlation matrix (or modified association matrix) designed to examine the interrelationship among variables and identify or confirm existence of factors
- FA in two main flavours determined by 'purpose' or 'formality' of analysis: Exploratory and Confirmatory FA
- In psycho-social research setting, FA is closely tied in with measurement models (developing instruments for the measurement of latent constructs)

What I will discuss

- Similarities and differences between EFA and CFA
  - Similarity: EFA model (technique / model, rotation) vs CFA model
  - Purpose
- Confirmatory Factor Analysis (CFA)
  - Definition
  - Software

A review of EFA

- In some ways, CFA can be thought of as an extension of the exploratory version.
- Important to review choices made in EFA phase as they carry though to CFA.
- Each EFA model has a CFA model equivalent

Geometric interpretation of EFA and CFA

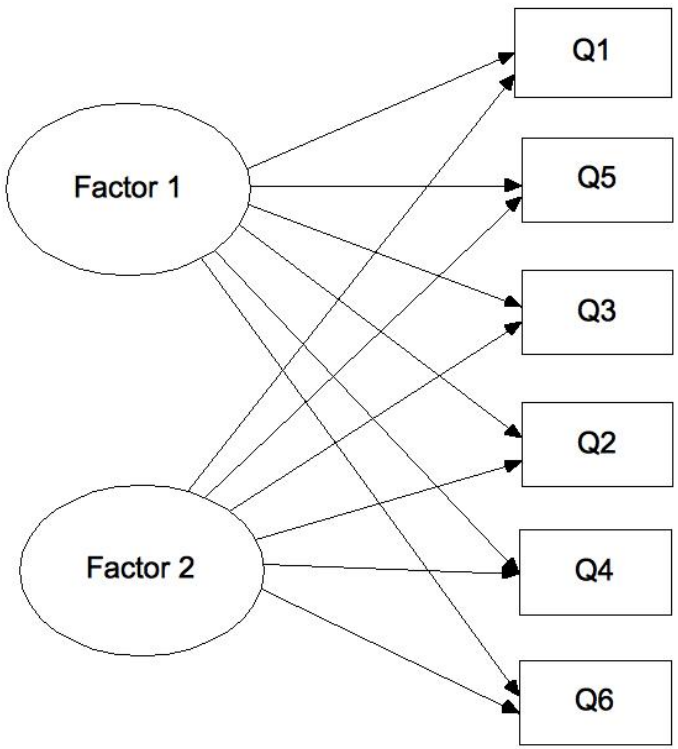


Figure: Exploratory Factor Analysis

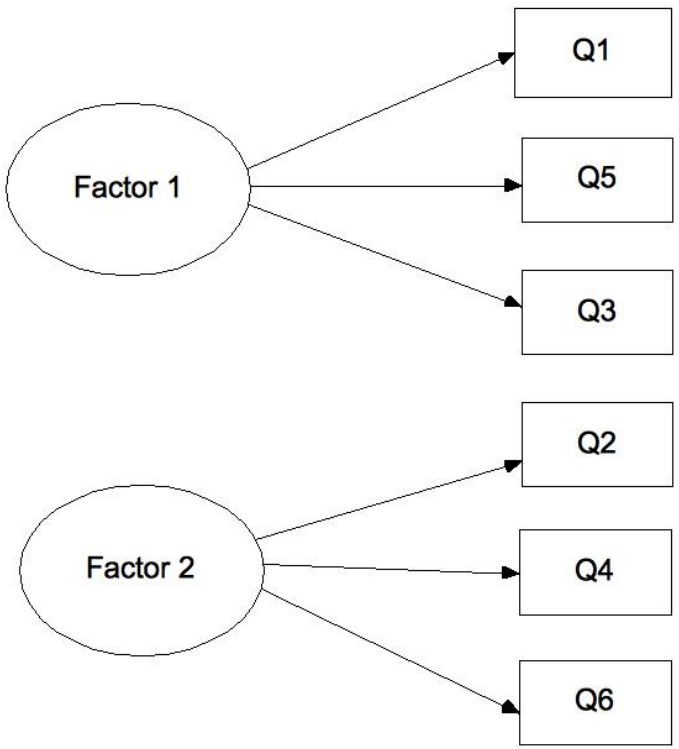


Figure: Confirmatory Factor Analysis

Confirmtory Factor Analysis

- CFA should be used when we want to confirm that we can measure something (e.g. depression) $\Rightarrow$ Construct validity
- Proposed CFA model ('Theory') may originate from the literature or from previously conducted EFA
- For example, you may want to just validate previously develop instrument on your population (e.g. Translation of an existing instrument into Thai)
- If it is from an EFA we have run, it is **important** that we **don't** use the same data for the subsequent CFA (we need to split the data)

Confrimatory Factor Analysis

- CFA is a type of Structural Equation Model (SEM)
- As CFA is a formal model (unlike EFA): Ue can use tools used to fit and test models to gauge its 'success'
- As CFA is an SEM, we can use SEM fit measures

Confirmatory Factor Analysis

The main new (beyond EFA) aspects of CFA are:

1. Gauging the 'effectiveness' of the CFA model:
2. Assessing model fitness. Does the model give a good representation of reality as represented by our data;

If model adequacy (fit and assumptions) is satisfactory we can run significance testing ofnvarious model parameters. Mainly this involves two types of hypotheses:

1. **Item-Factor Loadings**:
   Do items load significantly on the factors?
2. **Inter-factor correlations**:
   Are the factors inter-correlated?

Assessing model fit

Defn: Model fitness

For Confirmatory Factor Analysis (and all SEMs) a model fits well if there is little difference between the observed (data) correlation (or covariance) matrix and the one produced by the CFA, the implied correlation matrix

There are three different types of (Goodness of) fit statistics:

1. Absolute fit
2. Incremental fit
3. Model parsimony (technically not fit statistics)

Absolute (goodness-of-) fit statistics

These statistics measure the overall fit of the model. In this respect they are 'stand-alone' statistics. They include:

- χ^2 statitstic:
  Sensitive to N and model complexity (i.e. more complex model produces higher χ^2 (but still WIDELY quoted)
- Normed χ^2 statitstic (aka "scaled" χ^2 statitstic):
  1.0 (overfit) $<$ χ^2 (good) $<$ 2.0-3.0 (lack of fit) - also sensitive to N

Warning: χ^2 statistics in SEM

The larger the sample size, the larger the χ^2 statistic. It doesn't make sense to use χ^2 (ie good fits are from $\downarrow$N)

Although χ^2 statistics is poor, ALL CFA papers report it

Better absolute fit statistics

The **Root mean square residual** (RMSR) and **Root mean square error of approximation** (RMSEA) represent a better alternative to the χ^2 statistics.

Basic rule:
If RMSR or RSMEA are $< 0.05 \Rightarrow$ good model
If RMSR or RSMEA are $< 0.08 \Rightarrow$ adequate model

Personally, I use the RSMEA statistic

Incremental (compartitive) fit indicies

- These measures compare the current model with a baseline or previously fit model (i.e. Can be used to compare competing models....like AIC for GLMs)
- Includes a large number of statistics:
  - GOF index (GFI)
  - Adjusted GOF index (AGFI)
  - Tucker-Lewis Index (TLI)
  - Comparative Fit Index (CFI)
  - many more...
- All of these measures indicate a good fit if >0.9 (or >0.95)

Parsimony fit measures

- As with any other type of modelling (e.g. Generalized Linear Modelling), it is informative to adjust (penalize) for number of parameters (avoid overfitting)
- Parsimony measures used in CFA (and SEM) include:
  - Akaike Information Criteria (AIC)
  - Consistent Akaike Information Criteria (CAIC)
  - Bayes Information Criteria (BIC)
- ICs account for both fitness and model complexity
- best model is one where Model fit is sufficient, and model complexity low
- Problem with information criteria is values have no meaning outside context of a particular set of data

A quick note on SEM fit statistics and CFA

- As CFAs (comparative to other types of SEMs) are highly constrained. What might be though of as a good model often does not pass the ?cut-offs? for many 'SEM' fit statistics, especially when sample size is large.
- This presents problems when it comes to convincing others (e.g. Reviewers) when they tend to use *gold-standard* cut-offs (e.g. Scaled $\chi^2 < 3$, $RSMEA < 0.05$, $GFI > 0.95$ etc). It doesn't help that the psychology/psychometry reviwers are particularly **fastidious**
- A potentially interesting research area would be to develop/investigate fit statistics (specifically) useful for the CFA setting

Model (re) -specification

- If we find our initial model does not demonstrate an adequate fit, we will need to improve it (often by adding additional parameters)
- However, we need to diagnose where our model falls down (why it does not fit).
- For this purpose model respecification tools should be used.

Re-specification: How can we improve the model

I will briefly mention three types:

1. **Reduce variables** (items) in our model using Critical Ratios which are very similar to a t-test or Wald statistics (i.e. if CR $>$ 1.96 than variable does contribute and should be retained)
2. **Standardised Residuals** (difference between actual and implied covariance matrix) i.e. we can see which correlations we are not getting right
3. **Modification Indices**: $\Delta\chi^2$ (decrease in χ^2 due to new parameter or variable)

Personally, I use modification indicies (easiest)

Software

A number of specialized packages able to perform CFA as well as the usual suspects (SAS, Stata and R....but notably, not SPSS).

Using specialized packages may be preferable because of the unique way of representing and assessing SEMs (and CFAs) doesn't fit that well into the classical ?linear models? statistical framework.....at least for Stata, SAS and SPSS

Specialized packages for SEM/CFA

- **M-plus**
  - Unlike other packages, allows input of categorical data (directly into model)
  - Also good for latent growth models (i.e. Allows modelling longitudinally measured variables)
  - Closer to cutting edge
- **LISREL**
  - Mainly used by business analysis and econometrics types
- **AMOS**
  - SPSS add-in with very nice front end (easy to use)
  - Uses a graphical interface for model (re)specification
  - Generates nice figures
  - Good for quick, standard (off-the-shelf) analyses

Using R for EFA, CFA/SEM

The stistical language R is excellent for factor analysis (although it takes a little getting used to). The are four libraries that I will mention, in particular (although there are others):

1. `psych`: Good for conducting EFA (but rather limited for CFA/SEM)-**Used for worked example**
2. `sem`: Consider this the native SEM package for R. Not bad, but a little cumbersome for specifying models.
3. `lavaan`: This is a good package for CFA. Specifying models is easy, and it comes with many fit statistics-**Used for hands-on example**
4. `semPlot`: This pacakge is for generating graphs of your CFAs/SEMs. Really good-**Used for worked example**

Simple example

Dataset from study (Holzinger and Swineford, 1939) where 26 psychological tests administered to 301 year 7 and 8 students in two Chicago schools

In our example, we consider 78 girls from a single school and a subset of 6 tests (N = 78, k=6).

A simple CFA using AMOS

The six tests (items) we consider are:

1. Visperc (Visual perception score)
2. Cubes (Test of spatial visualization)
3. Lozenges (Test of spatial orientation)
4. Paragraph (paragraph comprehension score)
5. Sentence (sentence completion score)
6. Wordmean (word meaning test score)

A simple CFA using AMOS

We would expect (and let?s assume literature supports) that the first three variables represent one factor (say **'spatial appreciation'**) and the second three another (**'verbal comprehension'**)

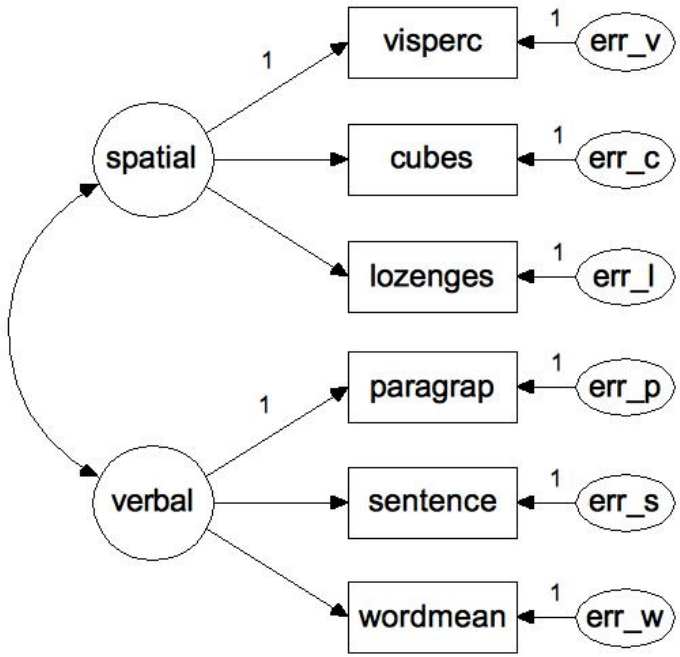
We can?t really safely assume that **spatial appreciation** and **verbal comprehension** are independent (can't assume orthogonality)

The Model

CFA model assumes:

- **SPATIAL** factor drives girls' **visprec**, **cube** and **lozenges** scores
- **VERBAL** factor drives **paragraph**, **sentence** and **word understanding**
- Also possible **SPATIAL** and **VERBAL** correlated

Q: What do you think the 'error' terms represent?

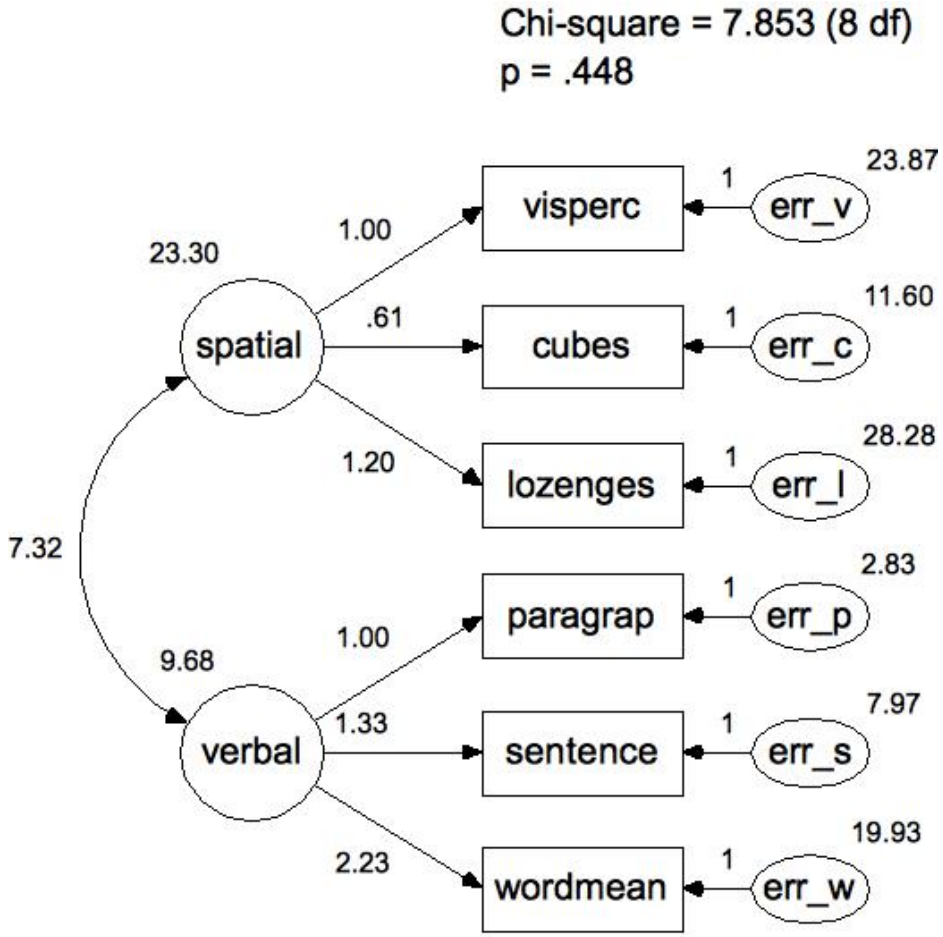


Example 8
Factor analysis: Girls' sample
Holzinger and Swineford (1939)
Model Specification

Results: Unstandardized

Results:

- $\chi^2 = 7.85 (p = 0.45) \Rightarrow$ model fits data well**?**
- Coefficients on output are in raw form
- More meaningful parameter estimates scaled (standardized βs and correlations)

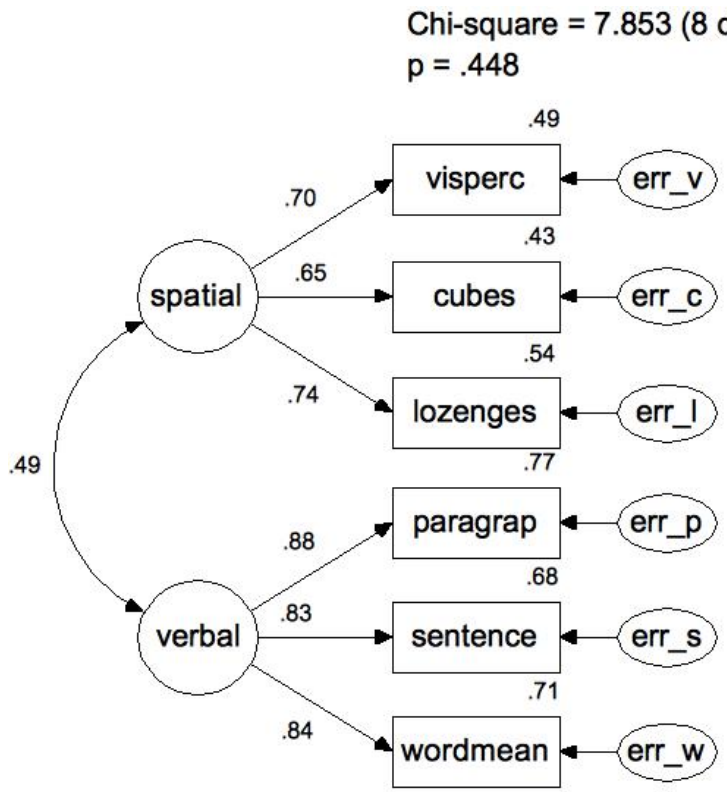


Example 8

Results: Standardized

Results:

- Verbal comprehension variables (**paragraph**, **sentence** and **wordmean**) weigh more highly on **VERBAL** and the **spatial measured item** loaded highly on **SPATIAL**
- **Spatial** and **verbal** also moderately positively correlated (r=0.49)
- Amount of variation explained in items ranges from 43% to 77%



Example 8
Factor analysis: Girls' sample
Holzinger and Swineford (1939)
Standardized estimates

CFA example: Raw output(1)

Computation of degrees of freedom (Default model

```
Number of distinct sample moments:21
Number of distinct parameters to be estimated:13
Degrees of freedom (21 - 13):8
Result (Default model)
Minimum was achieved
Chi-square = 7.9
Degrees of freedom = 8
Probability level = .4
```

			Estimate	S.E.	C.R.	P	Label
visperc	<---	spatial	1.00				
cubes	<---	spatial	.61	.14	4.25	***	par_1
lozenges	<---	spatial	1.20	.27	4.41	***	par_2
paragrap	<---	verbal	1.00				
sentence	<---	verbal	1.33	.16	8.32	***	par_3
wordmean	<---	verbal	2.23	.26	8.48	***	par_4

Raw βs

			Estimate
visperc	<---	spatial	.70
cubes	<---	spatial	.65
lozenges	<---	spatial	.74
paragrap	<---	verbal	.88
sentence	<---	verbal	.83
wordmean	<---	verbal	.84

CFA example: Raw output(2)

Squared multiple correlations (R^2)

			Estimate
visperc	<---	spatial	.70
cubes	<---	spatial	.65
lozenges	<---	spatial	.74
paragrap	<---	verbal	.88
sentence	<---	verbal	.83
wordmean	<---	verbal	.84

Eg. 71% of the variation in **word meaning** can be accounted for by factor **verbal comprehension**

CFA example: Raw output(3)

Absolute indicies

Chi-square = 7.85
Degrees of freedom = 8
Probability level = .45
Normed Chi-square = 7.85 / 8 = 0.9812

Comparitive (incremental) indicies

Model	RMR	GFI	AGFI	PGFI
Default model	1.68	.97	.91	.37
Saturated model	.00	1.00		
Independence model	13.81	.50	.29	.35

Measures of parsimony

Model	AIC	BCC	BIC	CAIC
Default model	33.85	36.65	63.63	76.63
Saturated model	42.00	46.52	90.10	111.10
Independence model	199.72	201.01	213.46	219.46

Interpretation (Fit indicies)

- Absolute fit indices
  - P(χ^2) > 0.05 implies good (enough) fit (Note: we **DON'T** want to reject here)
  - Normed $\chi^2 < 2$ (suggests good fit), but <1 suggests it might be overfit (could check using cross-validation).
- Incremental fit : Not so important since we don't have competing models, but GFI=0.97 (i.e. > 0.95) looks good
- Parsimony indices: Again only meaningful for competing models (including model respecification)

Sample size for CFA

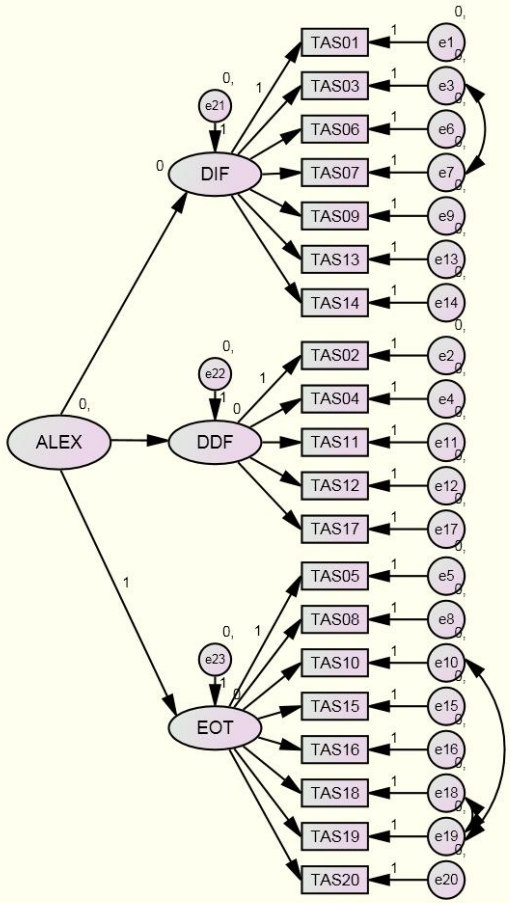
- CFAs are too complex to **formally** power
- Basic guideline: In order to give yourself the best chance of showing relationships (i.e. to ensure sufficient power), it is generally agreed that between 5-20 individuals are required **per** item (measured).

  e.g. If we have 20 questions in our instrument, we would want our instrument to be administered to, and returned by, between 100-400 people (in my experience indicates $n \rightarrow 400$).
- Another sample size approach I have seen is $N = 60 + 5k$
  where k is the number of items

  Probably a better approach for a small or large numbers of items.

Extending basic CFA: Higher order models

- So far we have only discussed first order factor analyses
- It is also possible that the first set of latent variables (factors) are driven by higher order factors
- Note the residuals on our first order factors (now they are endogenous)



Extending basic CFA: (multi-) group analysis

- Sometimes there is reason to believe there are effects modifiers that can alter the nature of relationships in FA
- For example, would boys of the same age exhibit the same patterns (of loadings and/or inter-factor correlations) in the spatial and verbal psych tests items as girls?
- This would be effectively 'examining' the effect modification of gender

Concluding remarks: Construct validty

- It is VERY important that CFA is used where it is appropriate and not EFA (and vice versa).
- If there is sufficient justification for the constructs (literature/previous analysis) and structure $=>$ CFA.
- If not we need to first uncover factor number and structures using EFA.
- If you have a large dataset, you can split the dataset run an EFA on one half (portion), and CONFIRM your findings on the other half
- Being inferential, the burden for larger samples is higher in CFA than EFA
- However, as we have seen, in the SEM setting, large sample sizes for CFA can also cause problems

Criterion-based validity

- Once we have established construct validity (our instrument measures what we believe it should measure), we need to see if it can be used for some practical purpose.
- For example, in an instrument that measures depression, can we use it to provide a **clinical diagnosis of depression**.
- To gauge whether our instrument works against some external gold standard, we need to conduct **Criterion-based validity**

Types of criterion validity

There are several different types of criterion validity. I will discuss three:

1. Discriminant validity
2. Predictive validity
3. Convergent validity

Discriminant validity

Establishing discriminant validity involves determining whether our instrument can discriminate between different types of people. For example:

- Can my new depression scale discriminate between those with a clinical diagnosis of depression (note: diagnosed using another [external] instrument or method)
- Can my new measure of 'suicidality' discriminate between those who have and have not previous attempted suicide
- Can my new instrument for breast cancer awareness discriminate between women who perform breast self examination, and those that don't

Predictive validity

Predictive validity is similar to discriminant validity, but it is concerned with 'within-subject' prediction. For example:

- Can my new instrument for Diabetes self measurement identify individuals with a longer onset-time of diabetes complications.
- Can my new measure of 'suicidality' identify individuals who will be likely to (in the future) attempt suicide

As you can see, the only real difference between discriminant validity and predictive validity is about time. Discriminant validity is established cross-sectionally, where predictive validity is more about predicting the future (for a patient)

Convergent validity

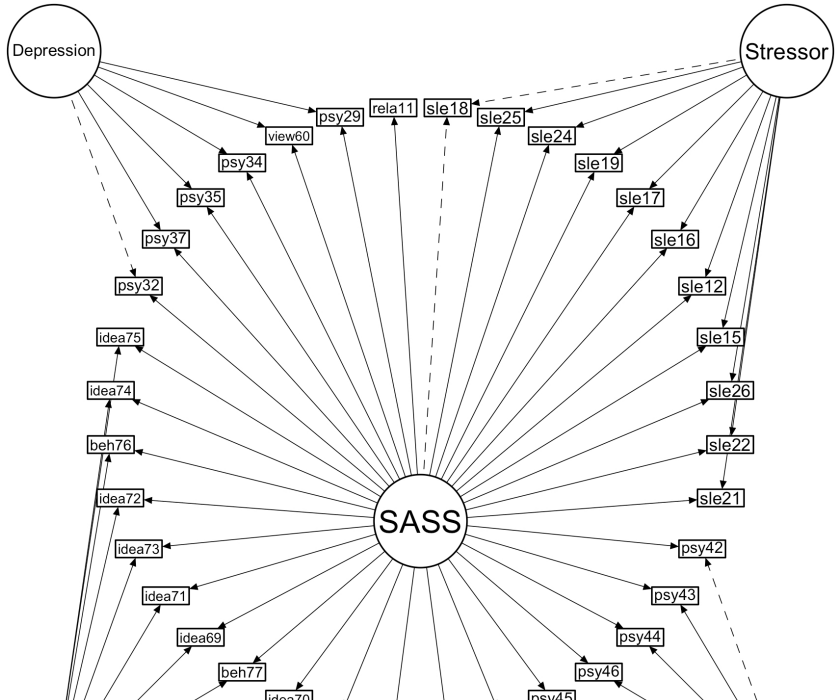
Convergent validity is about determining whether our new instrument correlates or agrees with an existing (well established and validated) instrument. For example:

How well does my new depression scale correlate with Beck Depression Inventory (BDI)

- Convergent validity is not something I do that much (I can't really work out the point).
- For example, If have have measure BDI in my dataset, why do a need a new depression instrument?

Suicidality in Thai adolescents

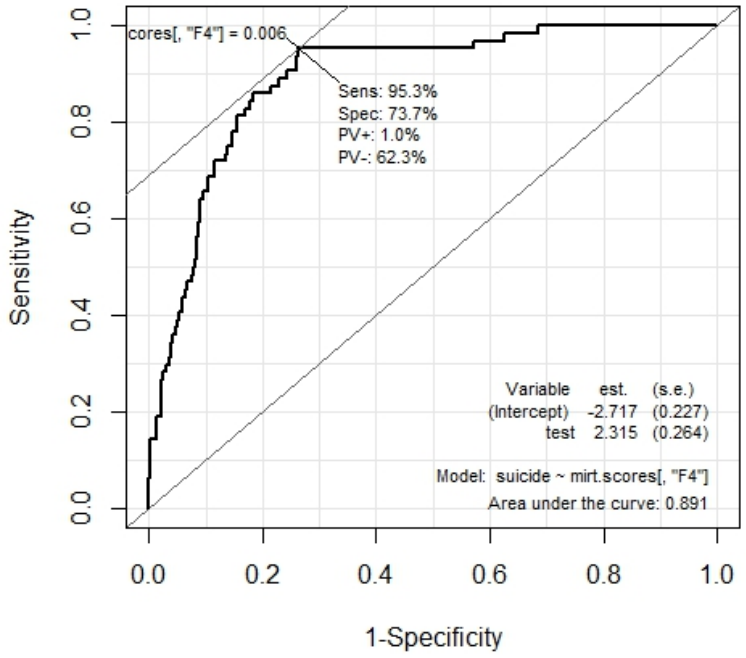
Instrument to identify adolescents at risk of attempting suicide. A Confirmatory MIRT model was fit and the resulting instrument was construct validated.

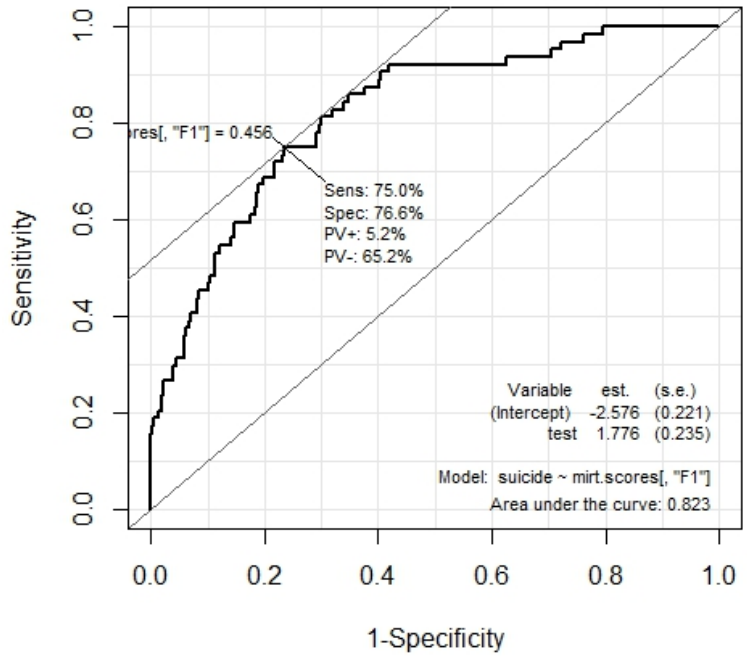


Suicidality in Thai adolescents

- After establishing construct validity, generating scores for each subscale (*Depression*, *Stressors*, *Suicidality* and *Pessimism*)
- Because this is a 'bifactor' model (special type of model) we can also get an overall measure of Suicidality (SASS)
- Once we have the score, we can see if there are associated with previous suicide attempts (yes/no)
- In this case I will use the ROC curves (but I could also just compre the scores for each suicide group using a simple t-test)
- I could generate five ROC curves, but I will provide two:
  1. The suicidality subscale
  2. The overall suicidality scale (SASS)

Extending basic CFA: Higher order models





We can see both the suicide subscale and the overall SASS can discriminate well between those who have and have not previously attempted suicide (high sensitivity and specificity).

THANK-YOU

Questions?????