Structural Equation Modeling

(And some other stuff)

What have we been doing?

Regression

- $Y = b_0 + b_1 X + b_2 Z + \epsilon$
- What is observed?
- Is this what you ***want*** to measure?

Latent Variable Modeling

- Structural Equation Modeling (SEM). AKA:
  - Covariance Structure Analysis
  - Covariance Structure Modeling
  - Analysis of Covariance Structures

Latent Variable Modeling

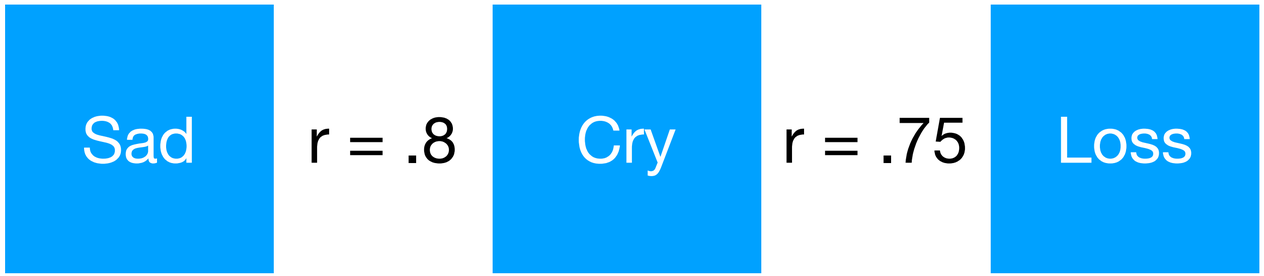
The logic

- The reason two measured variables can correlate with each other is because **they are caused by the same latent variable**
- They **share** some amount of variance
- If that shared variance is instead attributed to the **latent** variable, then the two variables will be independent of one another
- **The manifest (observed) variables are dependent upon the latent variable**

Example: The Beck Depression Inventory

Emotional Component

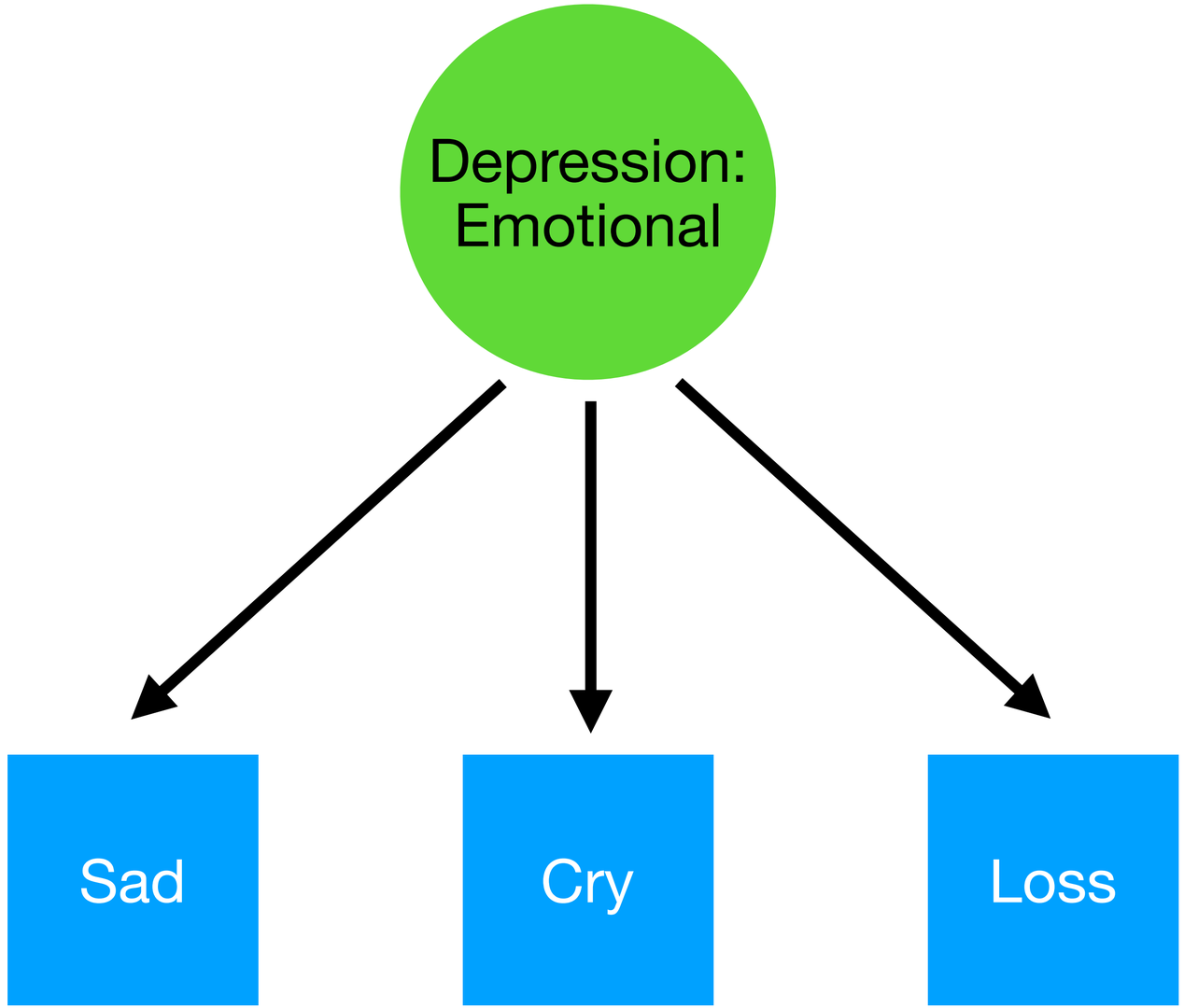
- “I feel sad”
- “I cry from sadness”
- “I no longer have interest in things I used to be interested in”



Example: The Beck Depression Inventory

Emotional Component

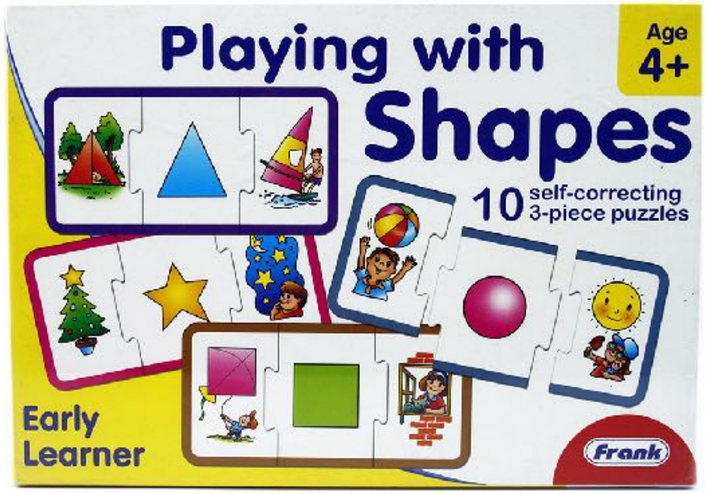
- “I feel sad”
- “I cry from sadness”
- “I no longer have interest in things I used to be interested in”
- **What’s up with these arrows?**

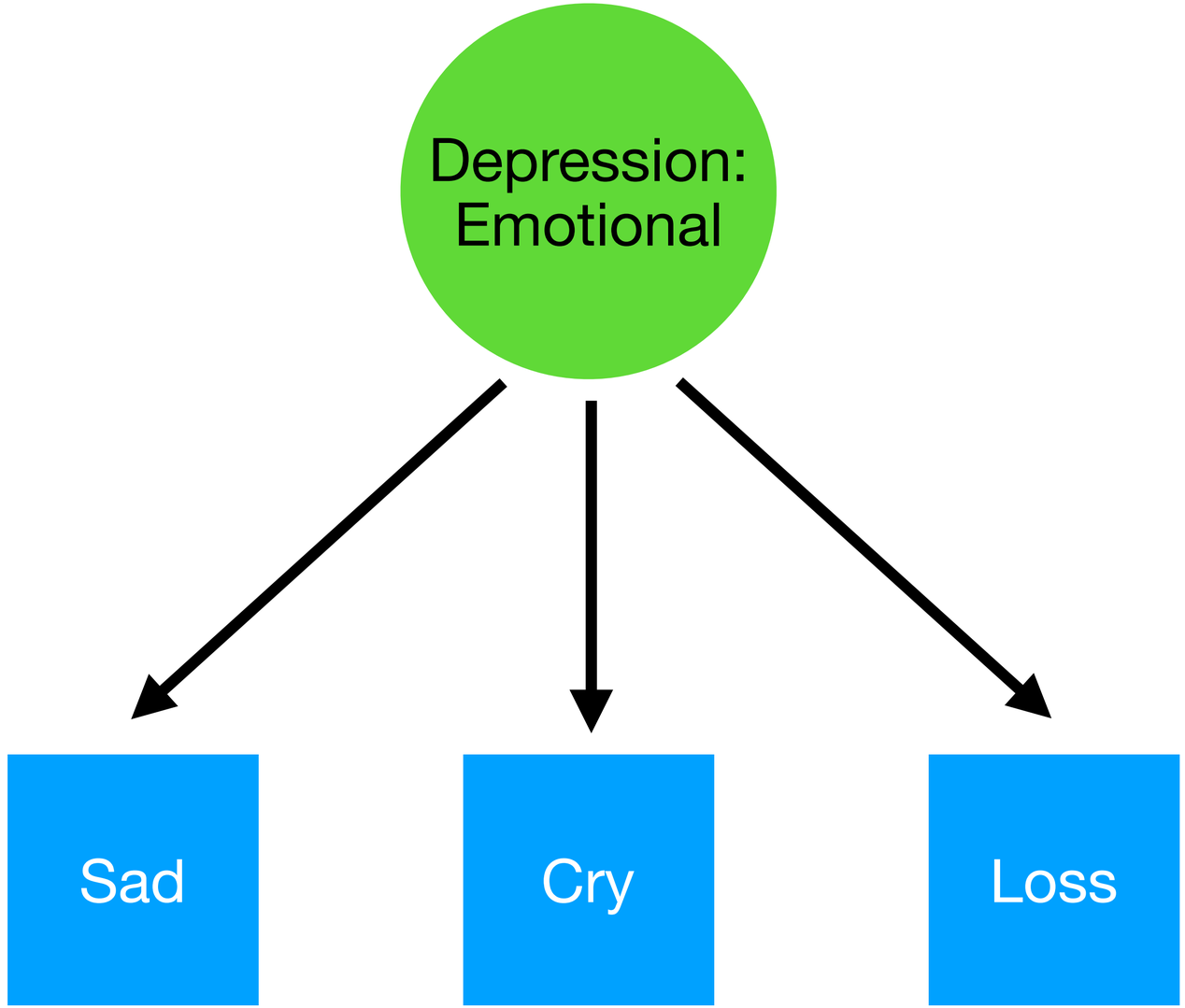


Example: The Beck Depression Inventory

Playing with shapes!

- **Squares** = **manifest** variables, **observed** variables, **indicator** variables
- **Circles** = **latent** variables

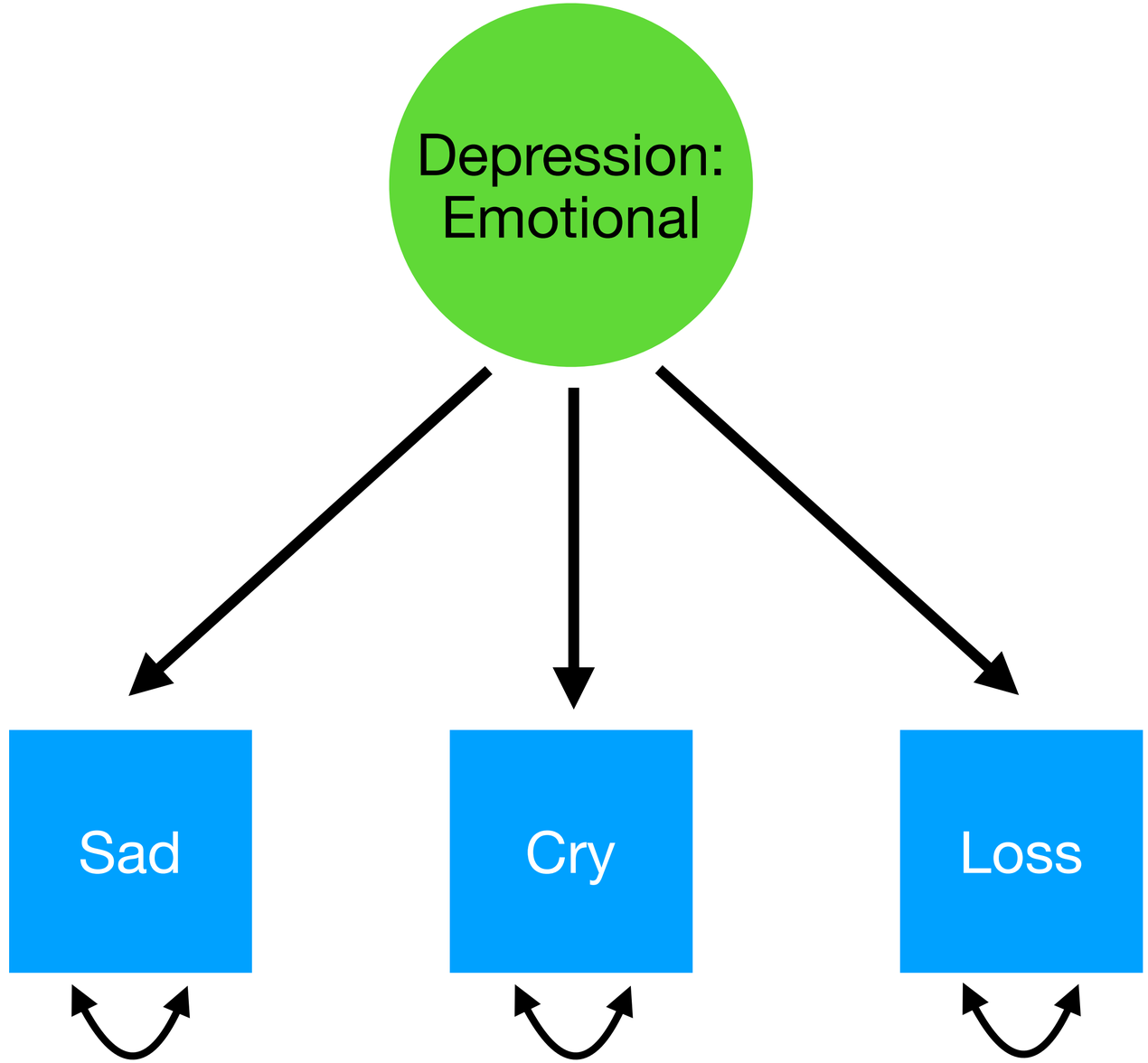




Example: The Beck Depression Inventory

Playing with shapes!

- **Squares** = **manifest** variables, **observed** variables, **indicator** variables
- **Circles** = **latent** variables
- **Curved** rams **arrows** = **residuals**/errors

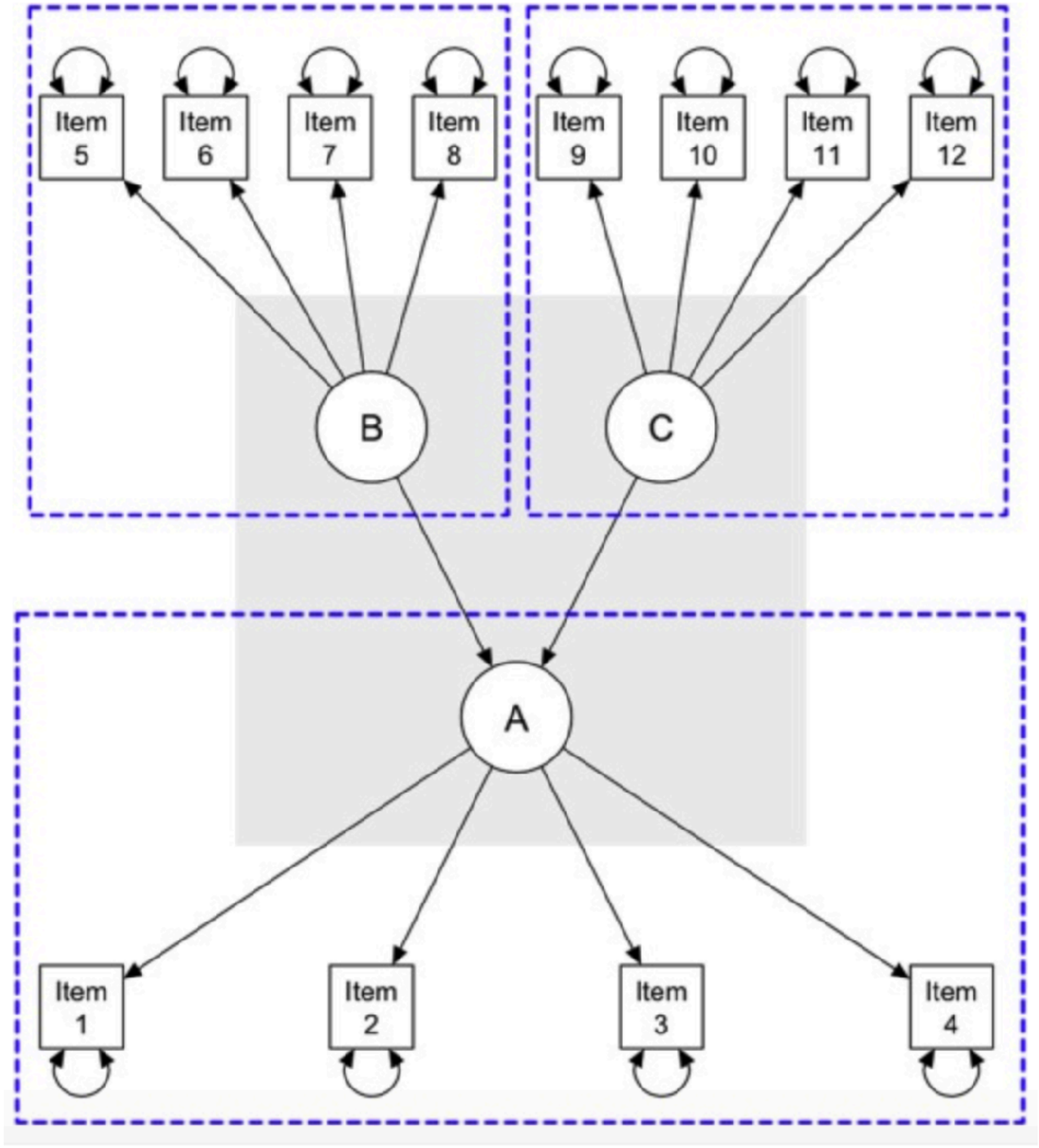


Remember Experimental Psych?

- How do we know that the green circle is actually the emotional component of depression?
- What is this called?

The Goals:

- What are the relationships between observed items and latent variables?
- What are the relationships amongst latent variables?



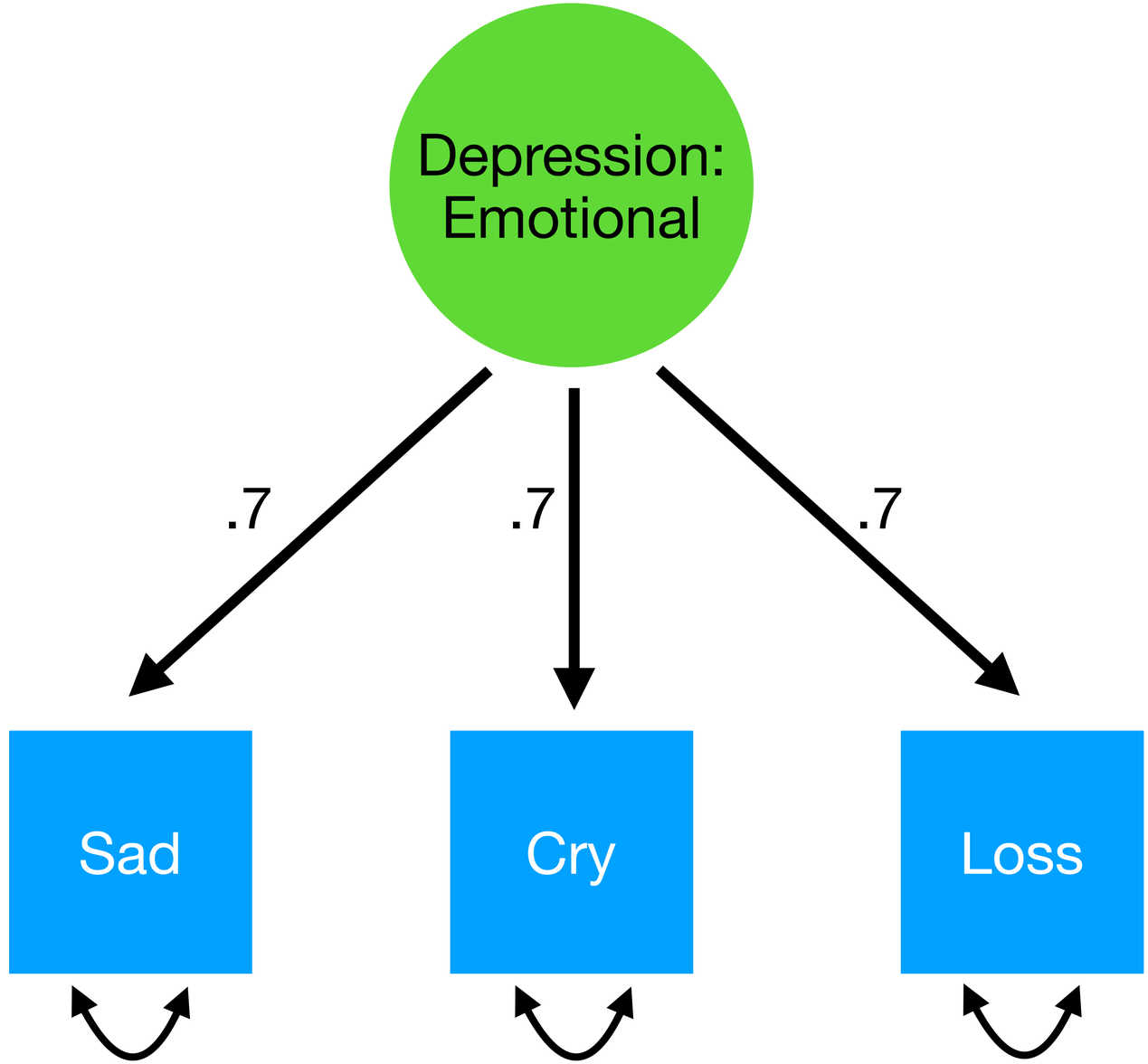
Under the hood

- Take the variance-covariance matrix of your observed data
- Take the variance-covariance matrix of the implied data
- Do they match up?

Example: The Beck Depression Inventory

Playing with shapes!

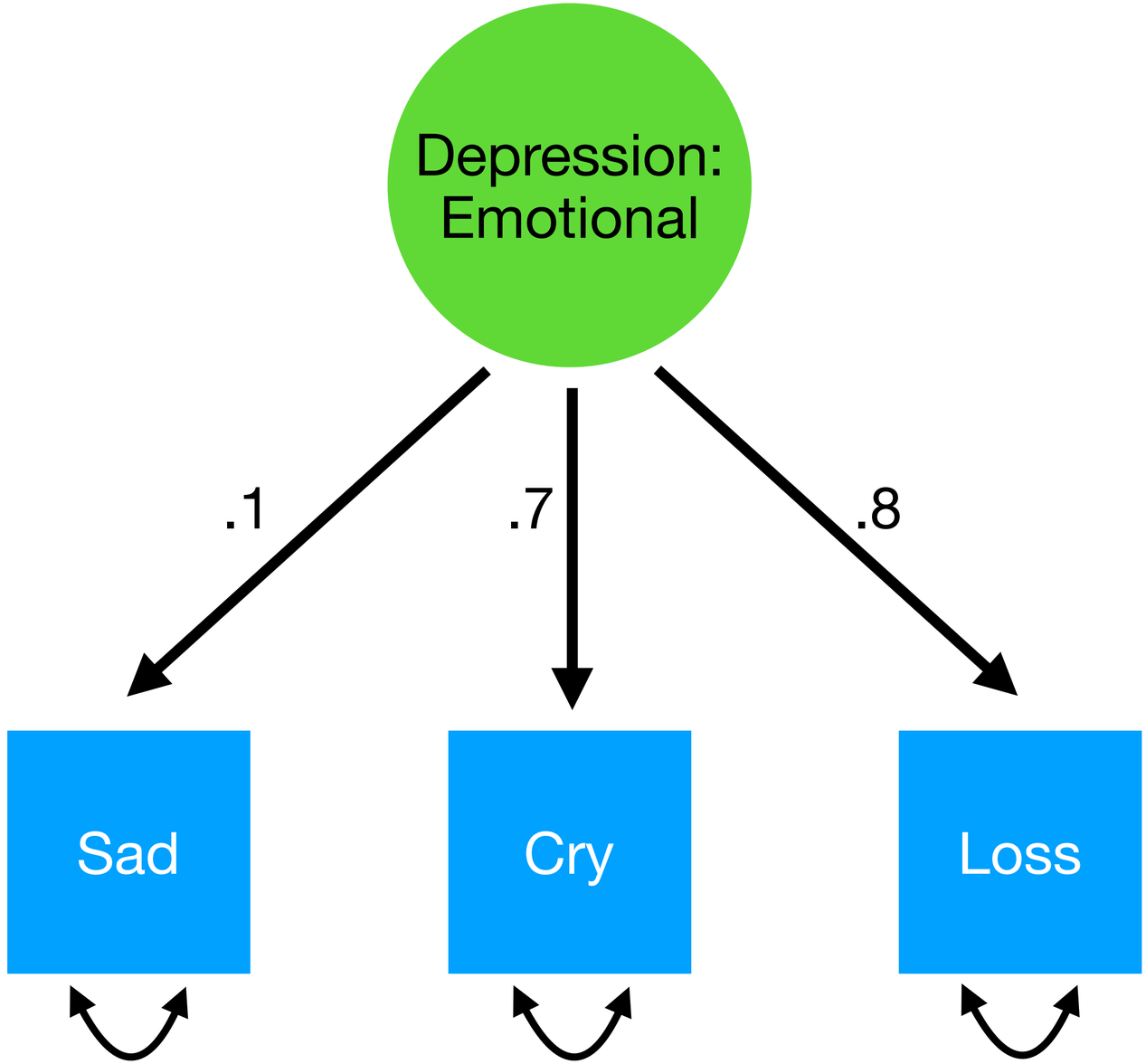
- **Factor Loadings** are extremely similar to regression coefficients



Example: The Beck Depression Inventory

Playing with shapes!

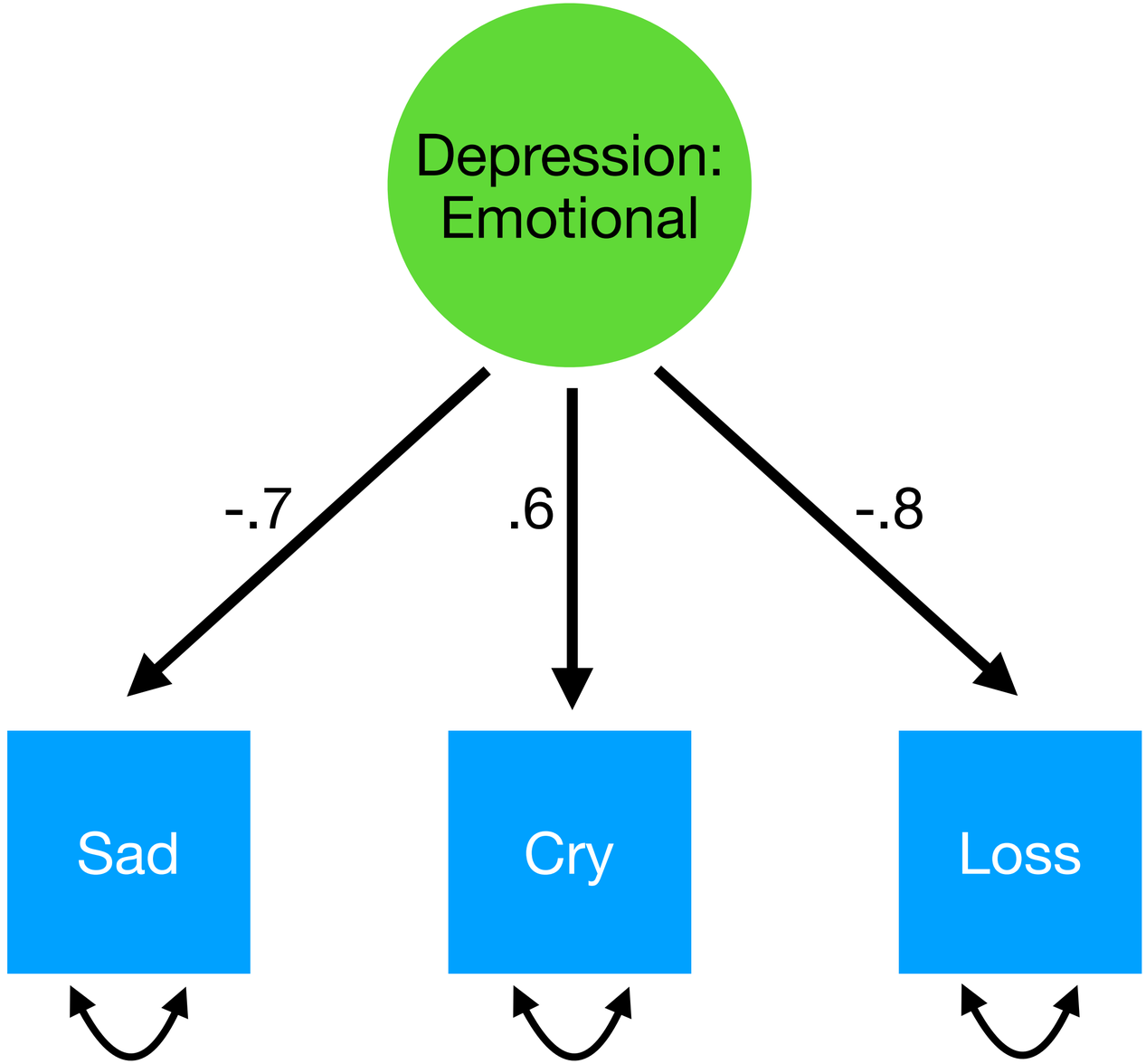
- Factor Loadings are extremely similar to regression coefficients



Example: The Beck Depression Inventory

Playing with shapes!

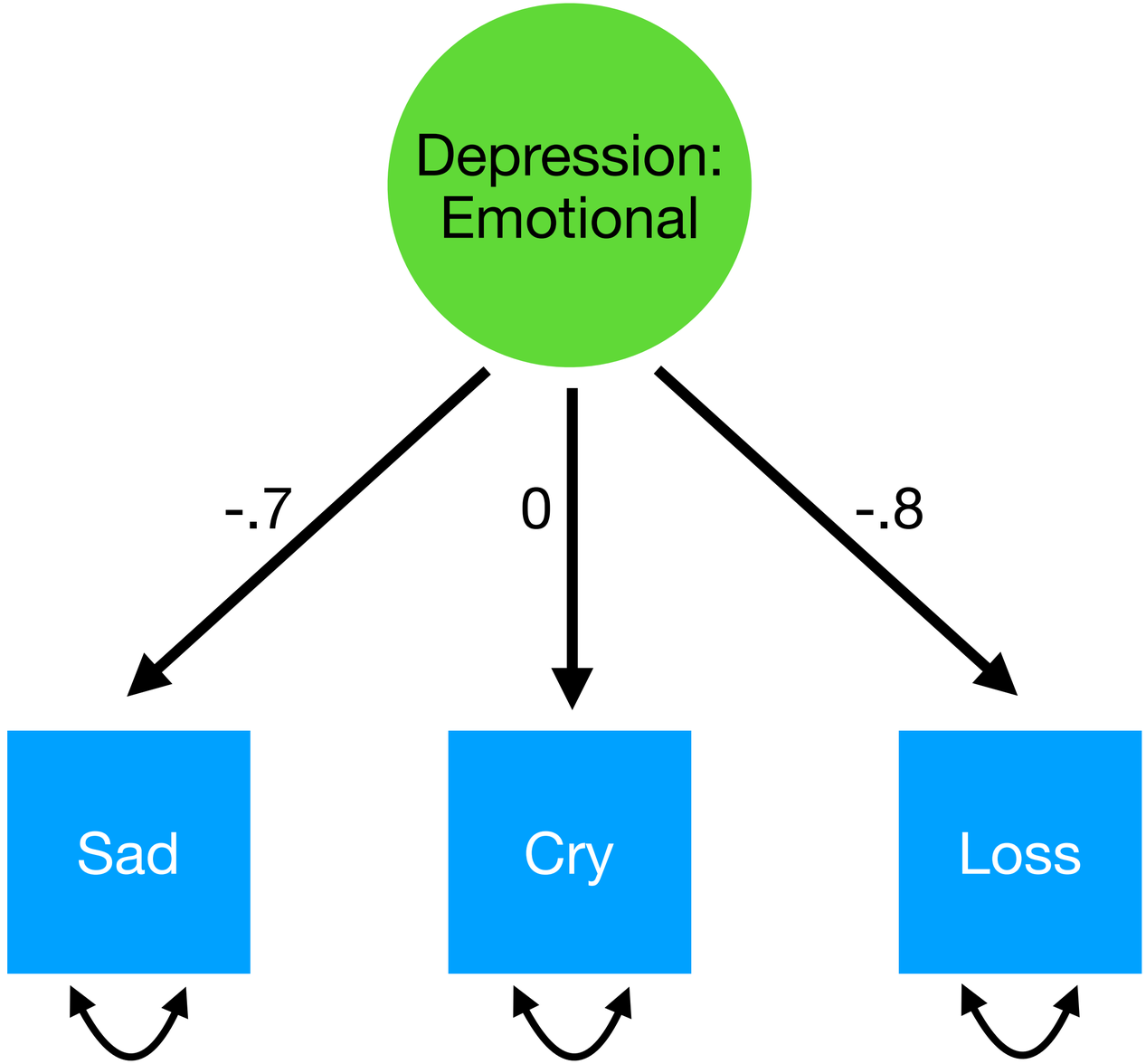
- Factor Loadings are extremely similar to regression coefficients



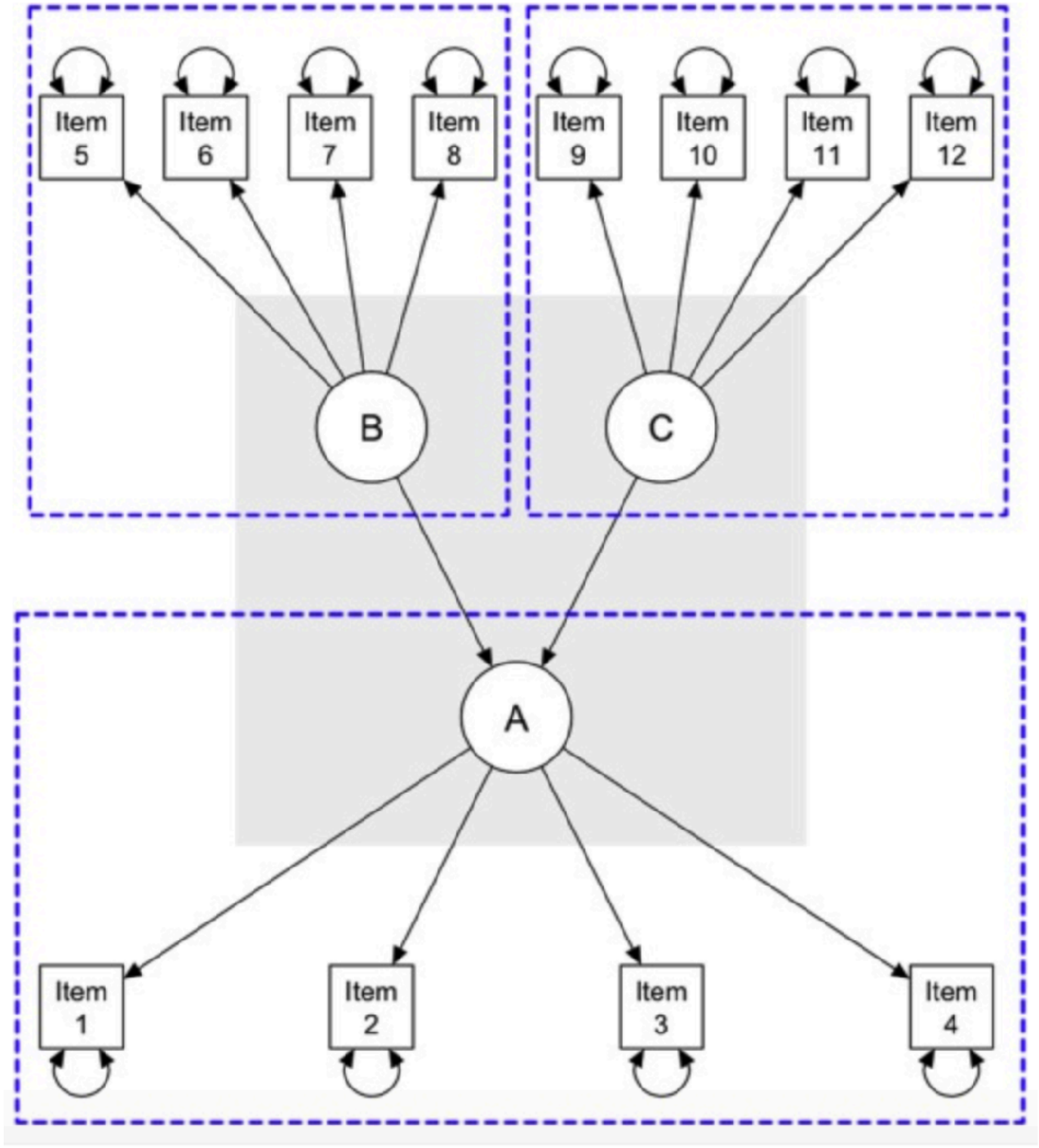
Example: The Beck Depression Inventory

Playing with shapes!

- Factor Loadings are extremely similar to regression coefficients



What is the relationship between A, B, and C?



Who decided these relationships?

You did. You are in control!

The good, the bad, and the ugly

Upsides

- Directly test hypotheses
- Use any type of data. Any. No, seriously. Any.
- Latent variables are “error-free”
- Evaluating is holistic. Hypothesized model is “good” if it fits the data.
  - Disconfirmatory procedure. Can never prove one model is “true”, but I can rule other crappier models

The good, the bad, and the ugly

Downsides

- Overfitting (wait until we get to machine learning tools!)
- Increased researcher degrees of freedom
- Need a big N!

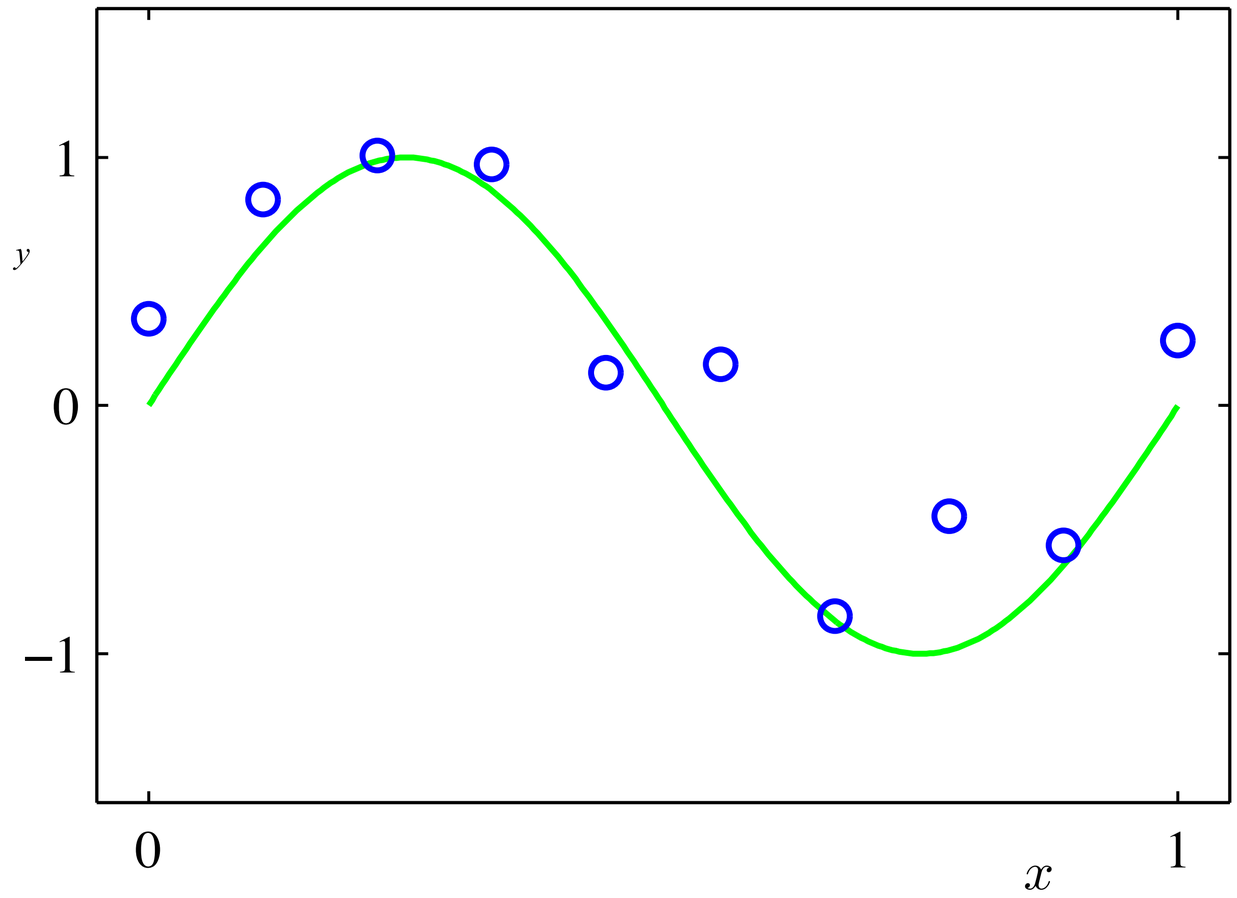
The good, the bad, and the ugly

Downsides

- **Overfitting**
- Increased researcher degrees of freedom
- Need a big N!

A brief aside about overfitting…

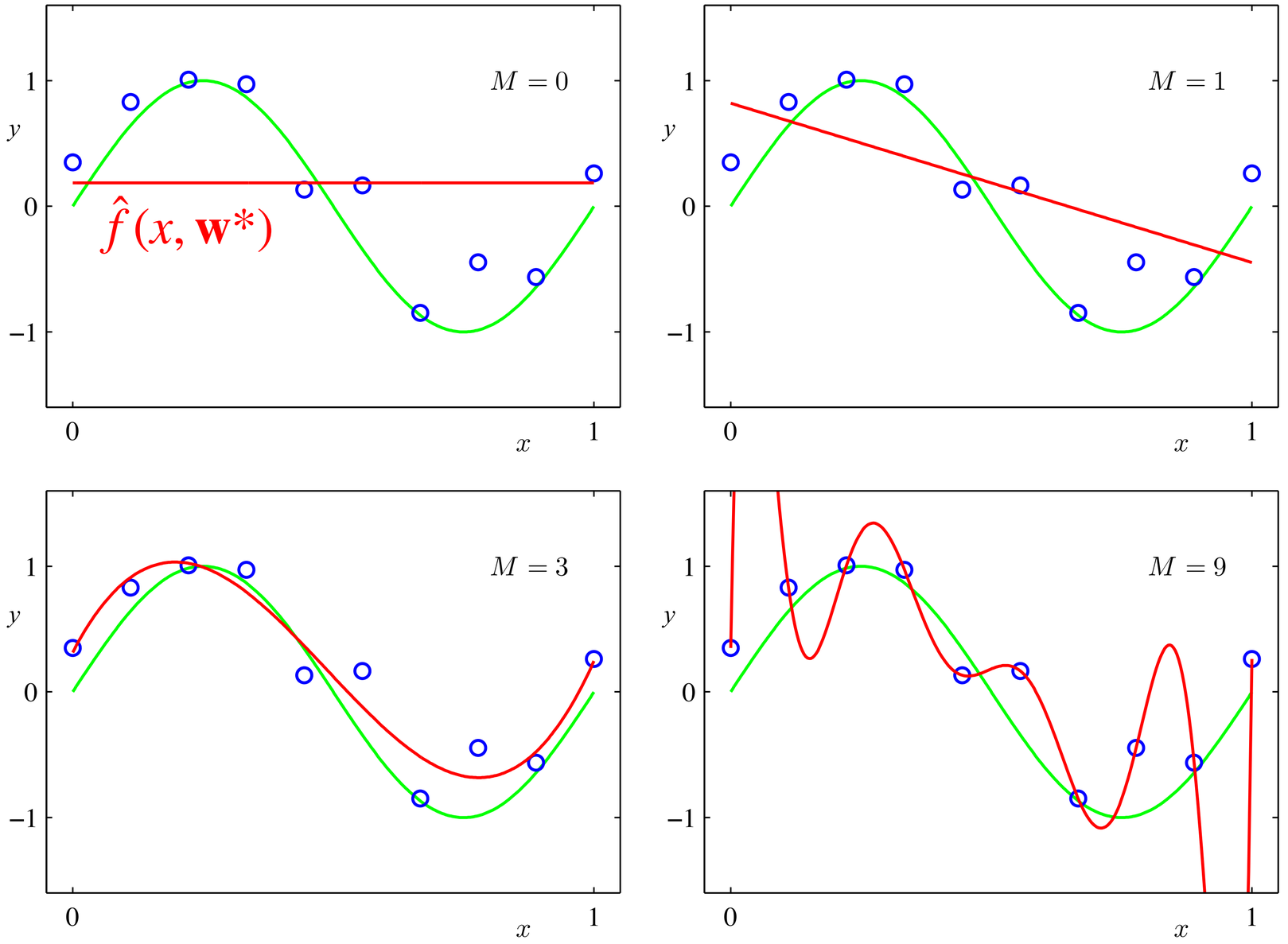
Let’s say we have these data



Slide credit: Dennis Barbour

Example

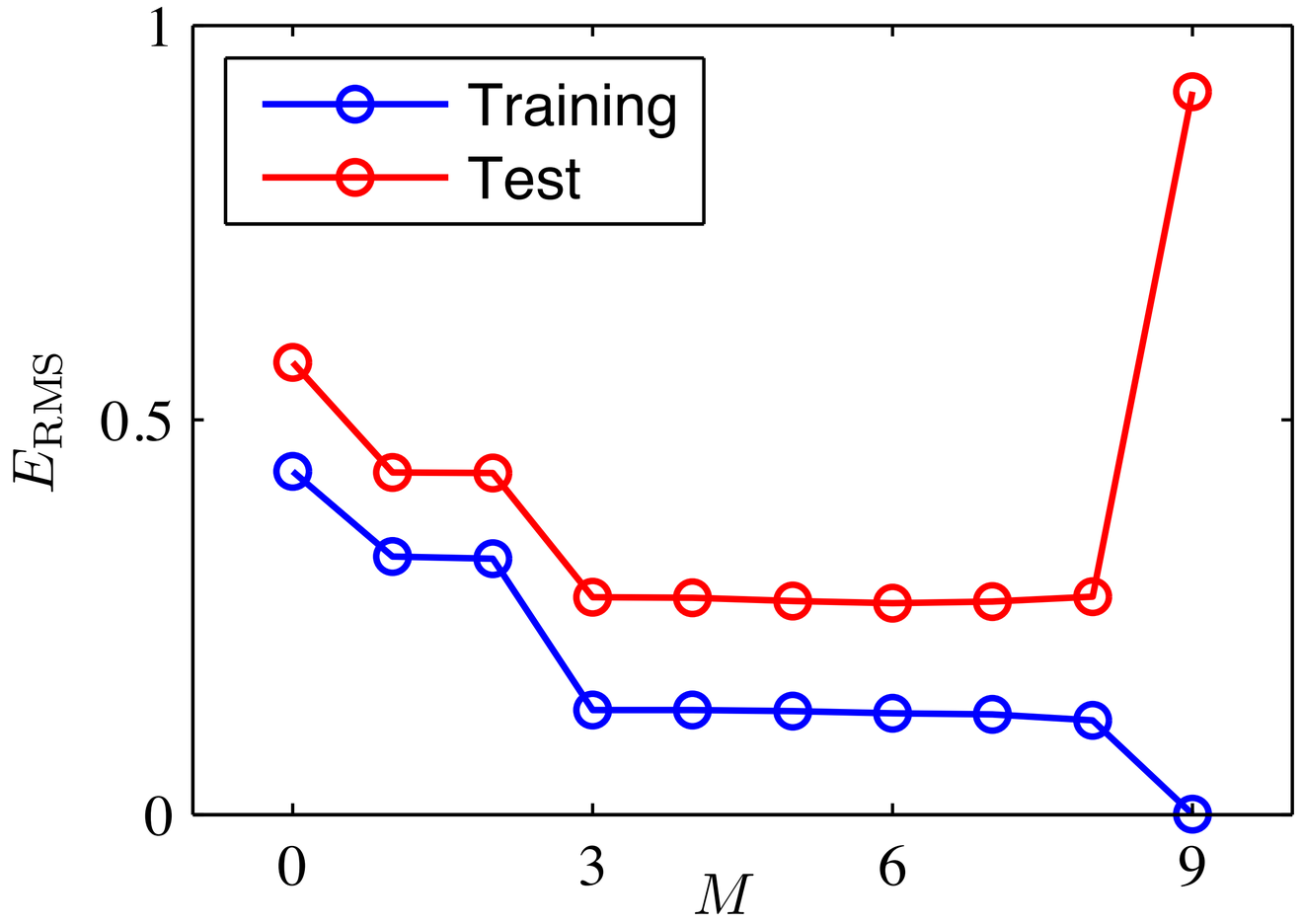
- I could fit a line using any number of parameters



Slide credit: Dennis Barbour

Example

- Why not use all 9?
- **Overfitting**
- What would happen if I used a NEW dataset instead of the training dataset?



Slide credit: Dennis Barbour

The good, the bad, and the ugly

Downsides

- Overfitting (wait until we get to machine learning tools!)
- Increased researcher degrees of freedom
- Need a big N!

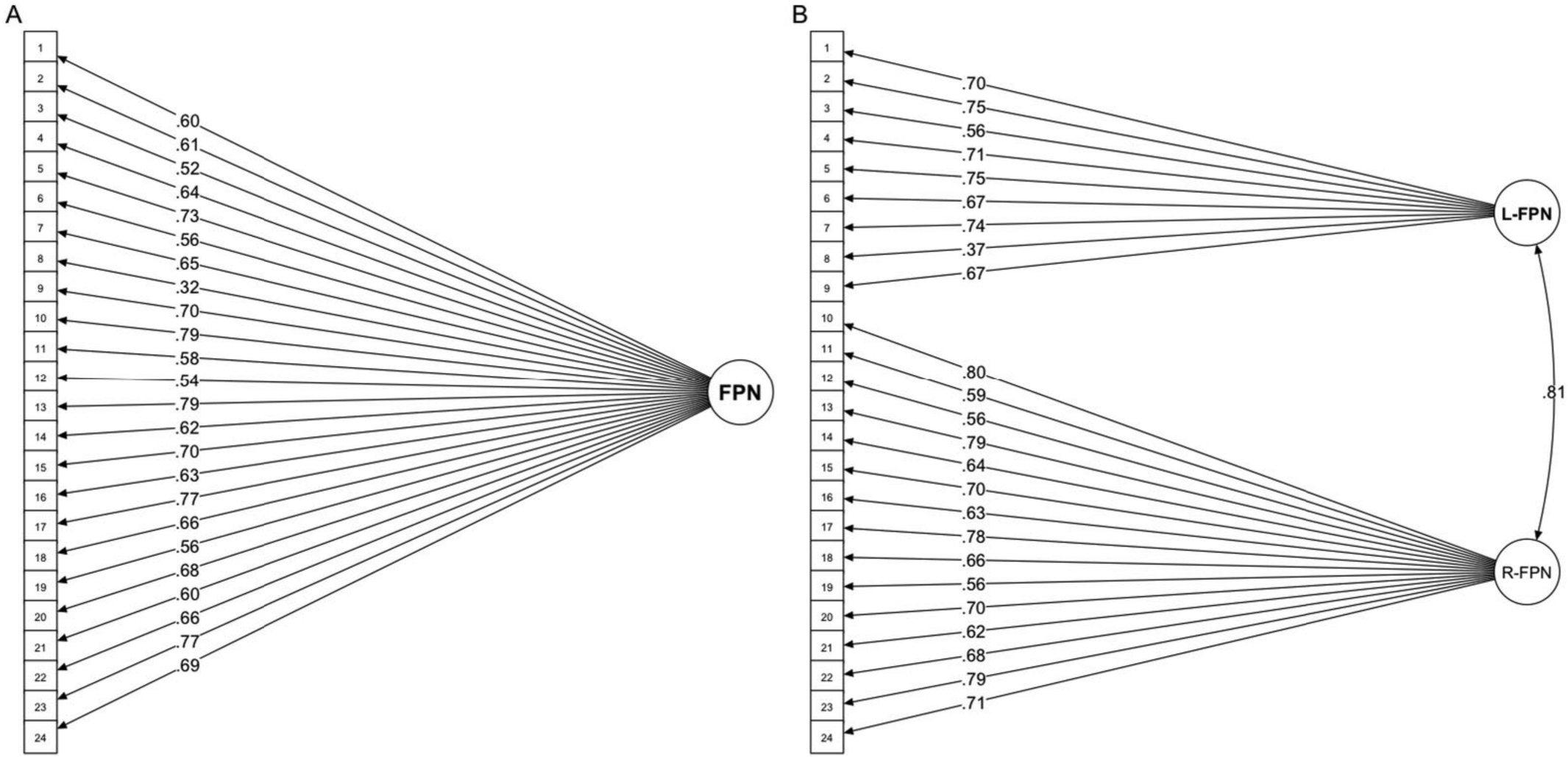
The good, the bad, and the ugly

Upsides

- **Directly test hypotheses**
- Use any type of data. Any. No, seriously. Any.
- Latent variables are “error-free”
- Evaluating is holistic. Hypothesized model is “good” if it fits the data.
  - Disconfirmatory procedure. Can never prove one model is “true”, but I can rule other crappier models

Example of testing hypotheses

Is it one brain network or 2?



Which model is better?

The good, the bad, and the ugly

Upsides

- (Mostly confirmatory). Directly test hypotheses
- Use any type of data. Any. No, seriously. Any.
- Latent variables are “error-free”
- **Evaluating is holistic**. Hypothesized model is “good” if it fits the data.
  - Disconfirmatory procedure. Can never prove one model is “true”, but I can rule other crappier models

Fit Measures

There are too many

- Chi-Squared *Sig test. It's fine, but has problems*
- Comparative Fit Index *Discrepancy between hypothesized and "baseline" model. Bigger is better. >.90*
- Tucker Lewis Index *Discrepancy between hypothesized and "baseline" model. Bigger is better. >.90*
- Root Mean Square Error of Approximation
- Standard Root Mean Square Residual

Compared to "perfect" model. Absolute fit. Smaller is better. <.08

- Akaike Information Criterion (AIC) and Bayesian Information Criteria (BIC)

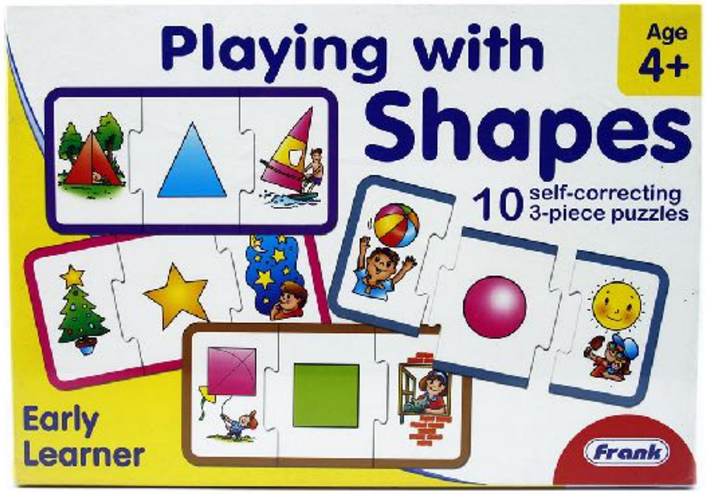
Combo of model complexity and model performance. Prefer the model with the smallest value, although note that the values themselves are meaningless

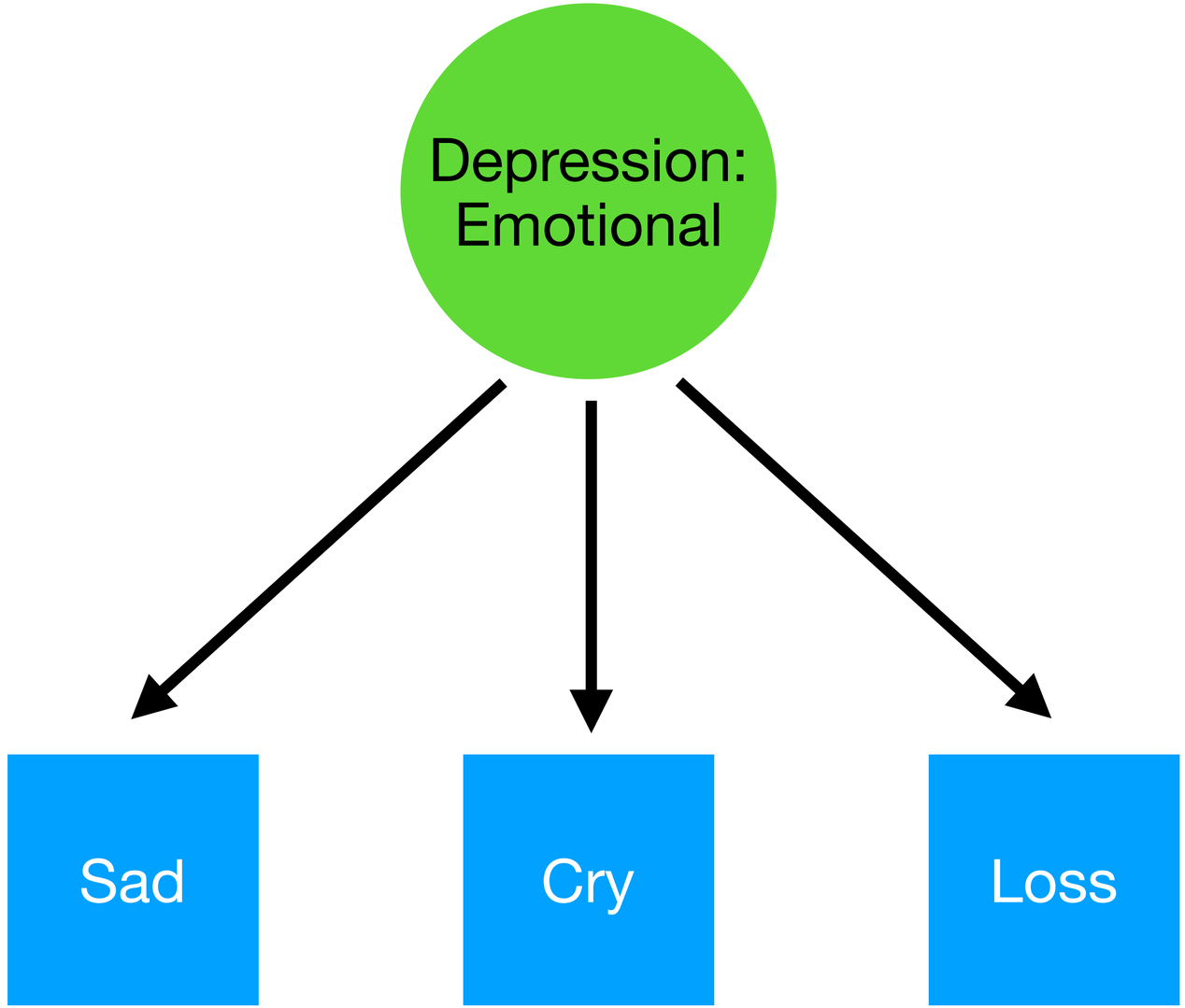
Wednesday’s Class

Example: The Beck Depression Inventory

Playing with shapes!

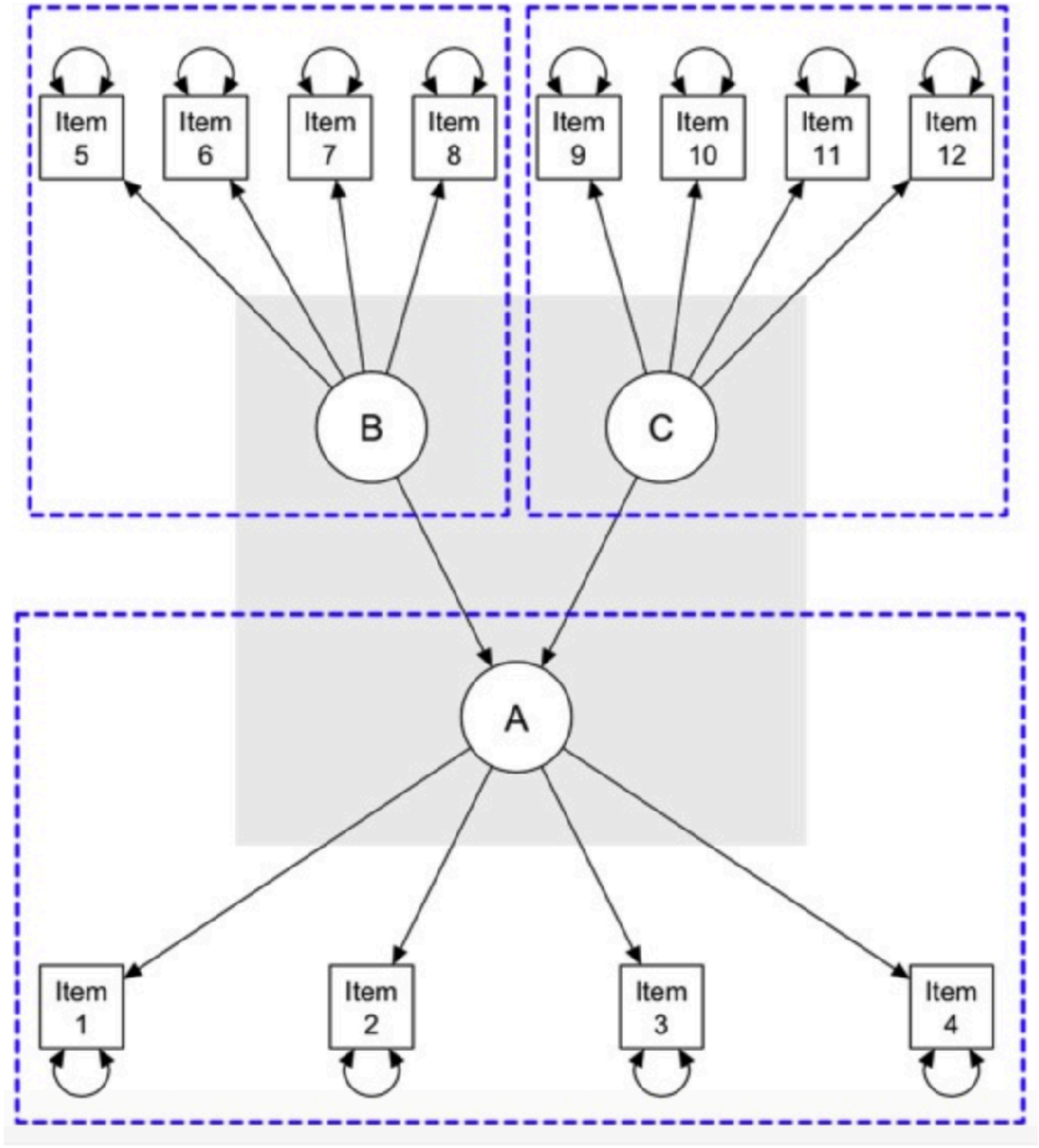
- **Squares** = **manifest** variables, **observed** variables, **indicator** variables
- **Circles** = **latent** variables





The Goals:

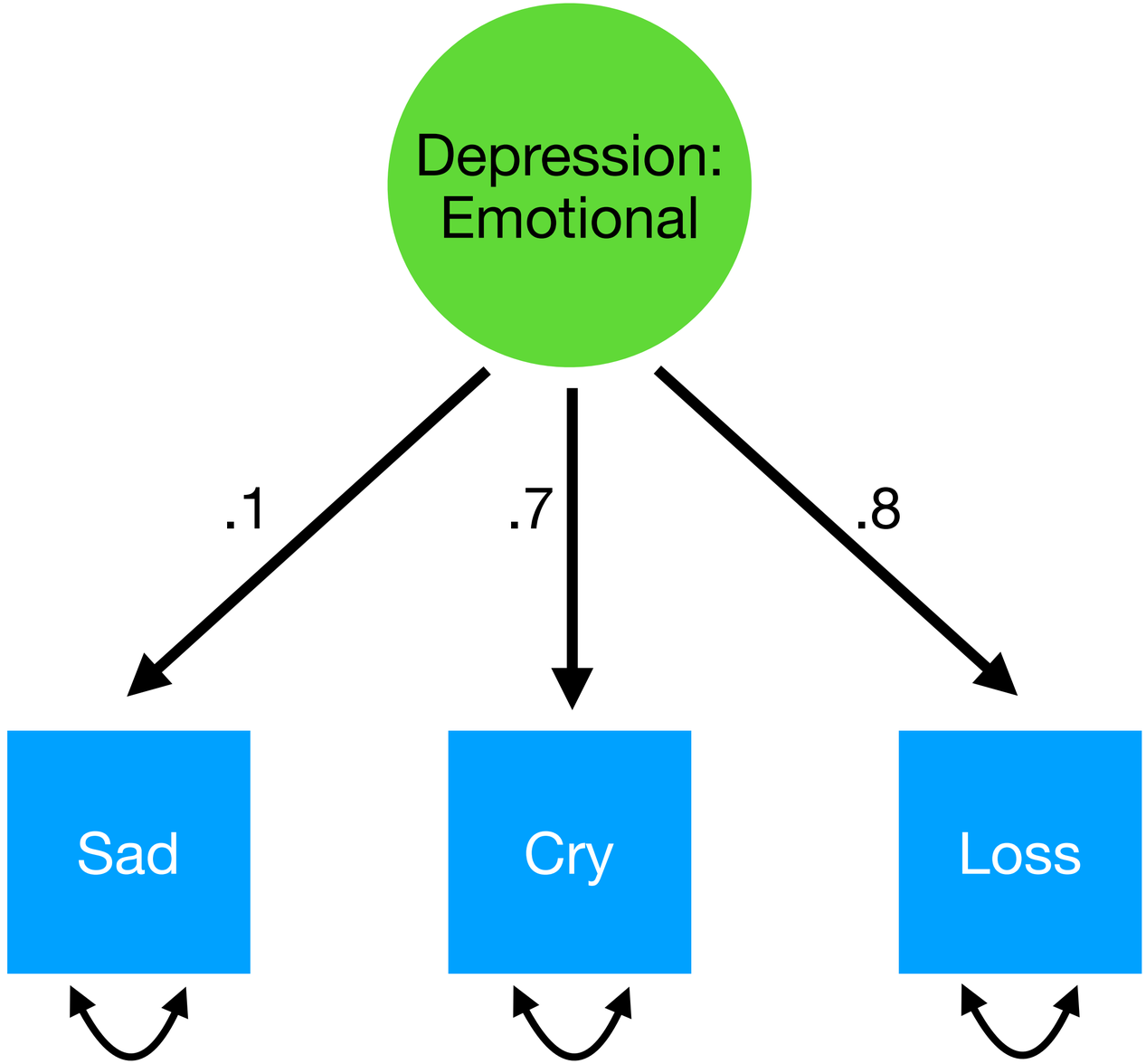
- What are the relationships between observed items and latent variables?
- What are the relationships amongst latent variables?



Example: The Beck Depression Inventory

Playing with shapes!

- Factor Loadings are extremely similar to regression coefficients



The good, the bad, and the ugly

Upsides

- Directly test hypotheses
- Use any type of data. Any. No, seriously. Any.
- Latent variables are “error-free”
- Evaluating is holistic. Hypothesized model is “good” if it fits the data.
  - Disconfirmatory procedure. Can never prove one model is “true”, but I can rule other crappier models

The good, the bad, and the ugly

Downsides

- Overfitting (wait until we get to machine learning tools!)
- Increased researcher degrees of freedom
- Need a big N!
- **WARNING: SEM cannot “prove” a model is correct or even prove causality. There are other ways of dealing with causal inference**

The good, the bad, and the ugly

Downsides

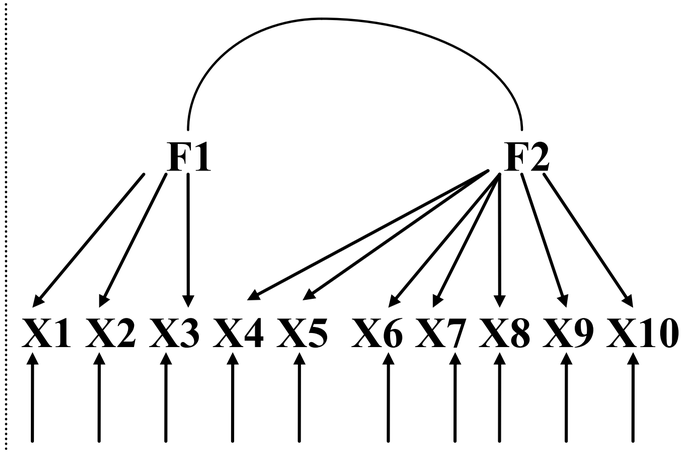
- Overfitting (wait until we get to machine learning tools!)
- Increased researcher degrees of freedom
- Need a big N!

A Summary from Shenyang Guo

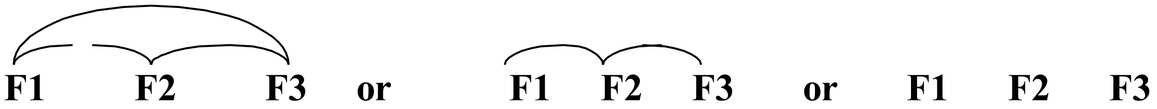
Professor of Social Work

- The goal: based on the theoretical model, you want to come up with a meaningful and parsimonious model.
- The model-fit indices for your final model should indicate that your model has a reasonably good fit.
- You should test several competing models
- You are not compelled to eliminate all nonsignificant paths.

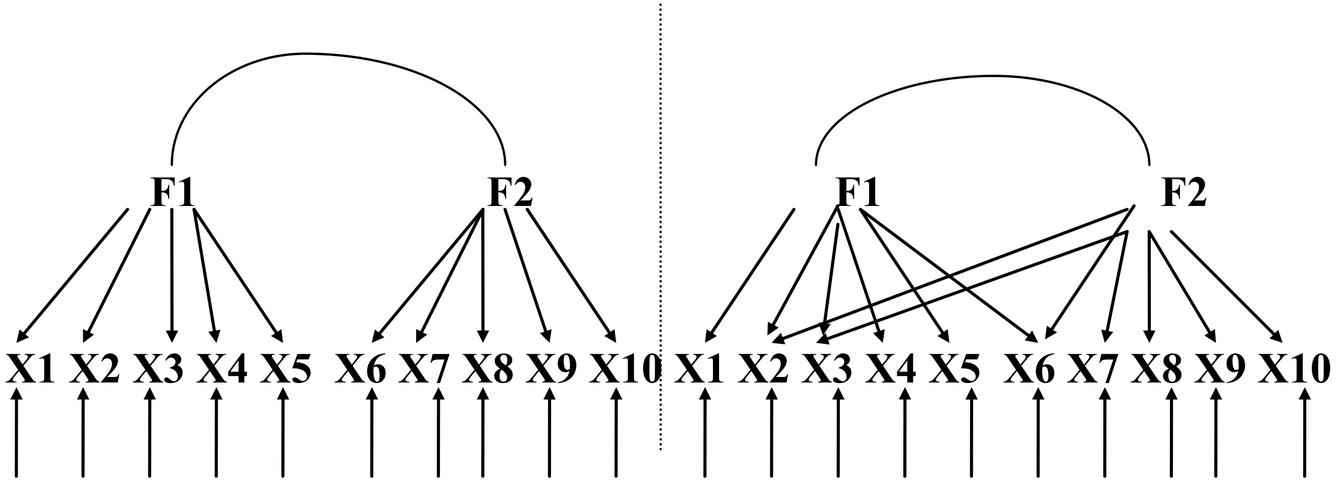
Why use this?



You can have different patterns for item groupings!



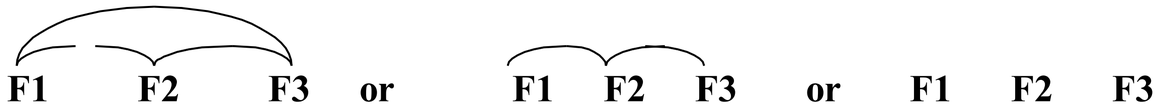
Latent factors themselves can be correlated or uncorrelated



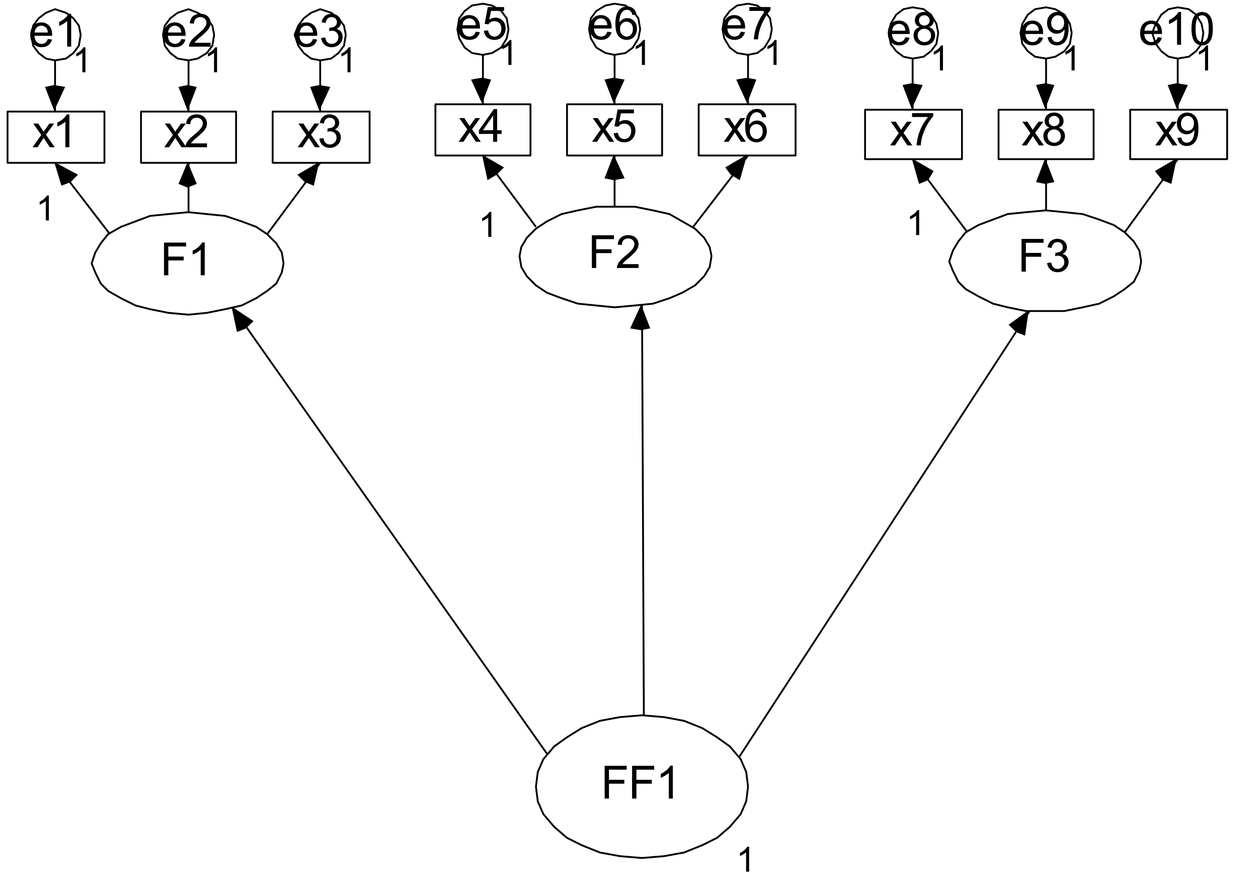
Some Conventions

- Minimum of 3 items per latent factor. 1 or 2 is possible but a pain in the butt.
- Latent factors do not inherently have a metric. You need to give it one. There are 2 ways of doing this (R will do this for you, and you can go between them):
  - Make the variance of the latent factor equal to 1
  - Make one of the factor loadings equal to 1
- Number of parameters (aka: number of factor loadings) needs to be less than the number of known data pieces (aka: N)

Why use this?



Latent factors themselves can be correlated or uncorrelated



You can have higher order factors, esp if the latent factors are correlated

Example

Singer's RVES

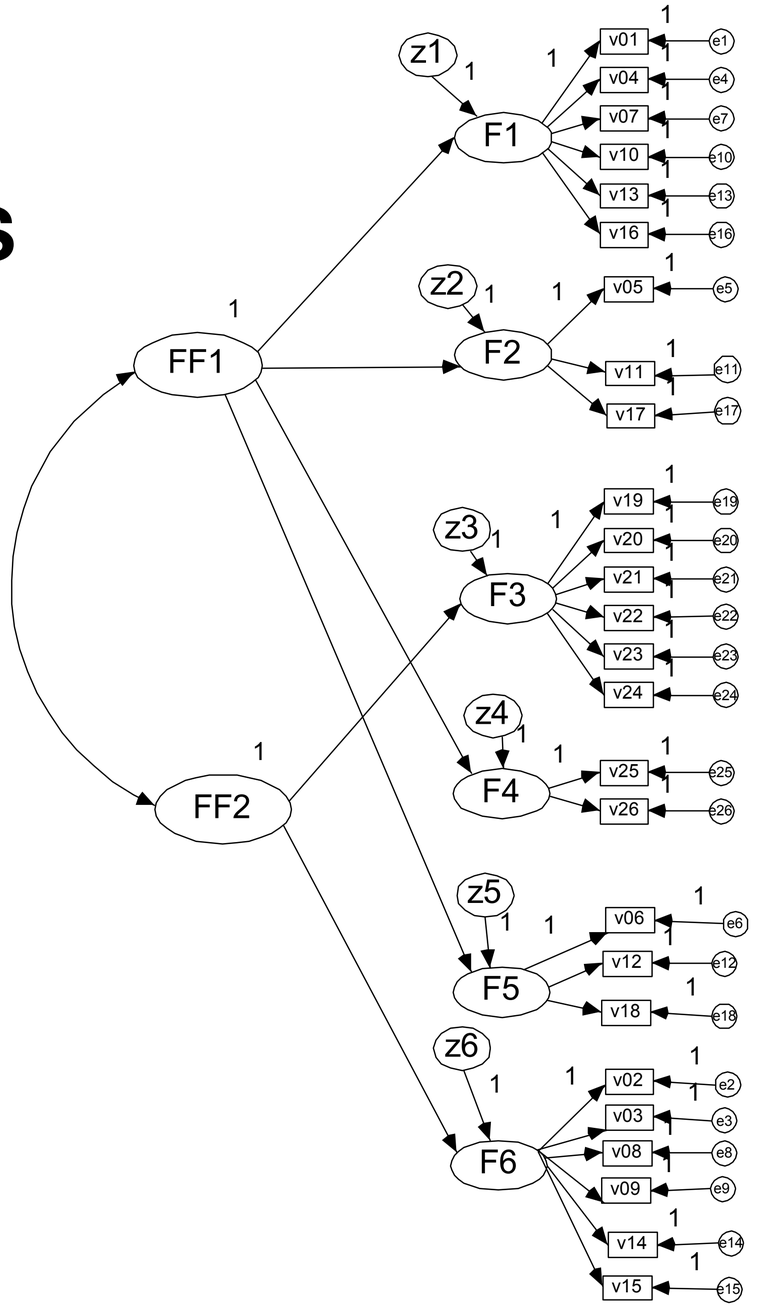
- A prior study has supported a 6-factor solution to the 26-item data, and the development of the following 6 subscales measuring recent violence exposure:
  - F1. Victimized/witnessing at home
  - F2. Witnessing at home
  - F3. Shooting/knife attack
  - F4. Sexual abuse
  - F5. Witnessing in neighborhood
  - F6. Victimized at school or in neighborhood
- Since all factors are correlated (overlapped), we want to know: whether or not a further generalizability can be arrived at?

Example

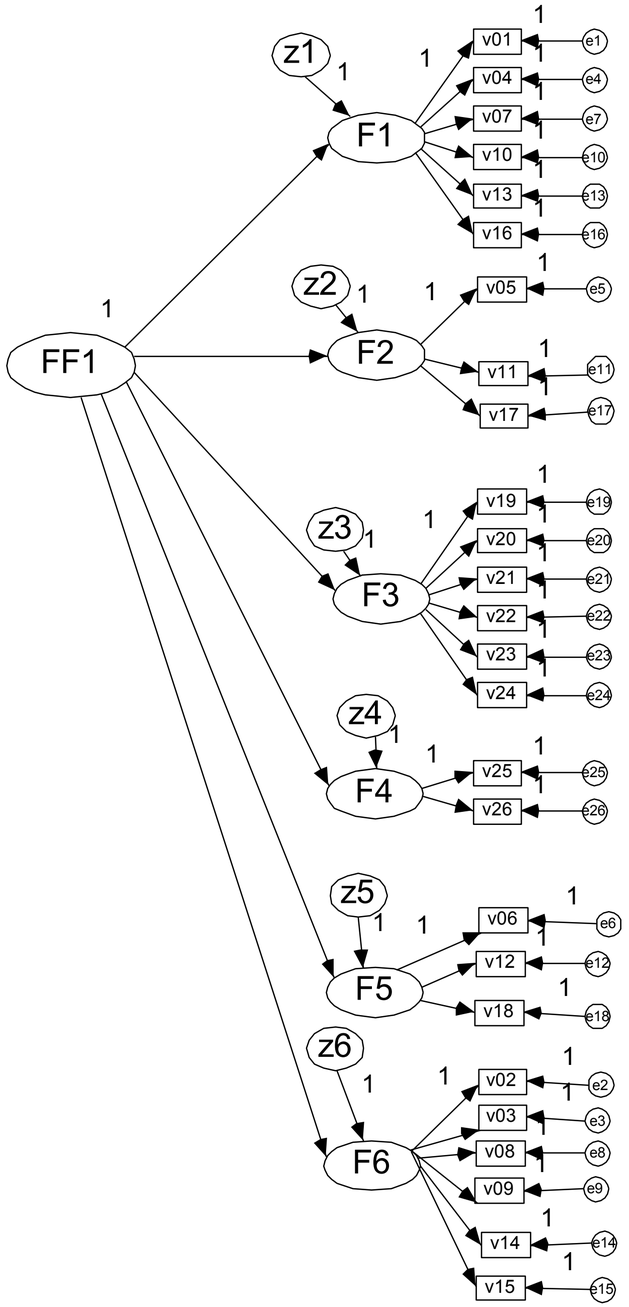
Singer's RVES

The fit indices for left figure are better than the right figure. Therefore, we prefer the left figure (aka 2 higher order factors)

If we wanted to create "total" scores, we would need 2 totals, rather than just 1



Second-order 1: Singer's RVES, 2 high-order factors



Second-order 2: Singer's RVES, 1 high-order factor

SEM in R

- `lavaan` package
- Excellent tutorial found here: **https://lavaan.ugent.be/tutorial/**
- Step 1: define the model as a text string
- Step 2: run a particular function on the defined model
- Let's walk through their SEM example together
- Graphing is a massive pain in the ass. We will not ask you to graph these. We might ask you to sketch out what you did (power point, drawing apps etc.)