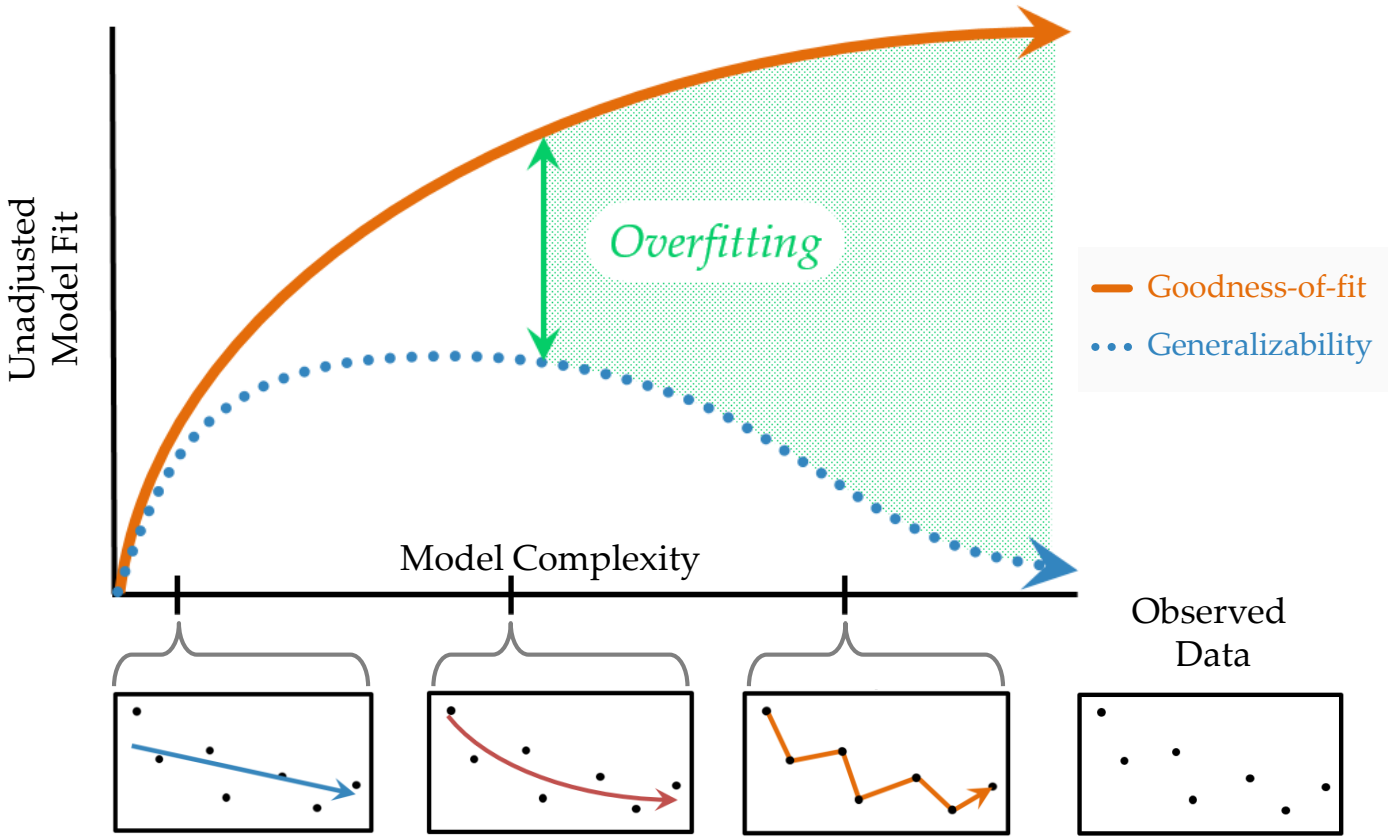
Rethinking Complexity in Item Response Theory Models

Wes Bonifay, Ph.D.

Statistics, Measurement, & Evaluation in Education
University of Missouri



Excessive flexibility makes a model "so weak that there is no way to find evidence either for or against it." (Wexler, 1978, p. 346)

Complexity

Definition: The ability of a model to fit a wide range of data patterns *(Myung, Pitt, & Kim, 2005)*

- Affected by number of parameters & functional form *(e.g., Collyer, 1985)*
- $y = x + b$ *vs.* $y = e^{xb}$

In latent variable modeling, complexity is typically gauged by counting parameters

- Model evaluation metrics like AIC, BIC, DIC penalize for complexity
- Minimum description length *(Rissanen, 1978)* also accounts for the functional form of the model

Fitting Propensity

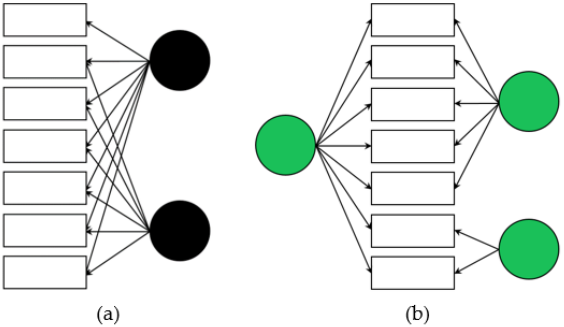
A baseline for model fit can be established by fitting models to random data *(Cutting et al., 1992)*

Fitting propensity: A model's ability to fit diverse patterns of data, all else being equal

- Models with same number of parameters but different structures may exhibit different fitting propensities *(Preacher, 2006)*

Herein, we investigate the fitting propensities of 5 popular item response theory (IRT) models

Measurement Models



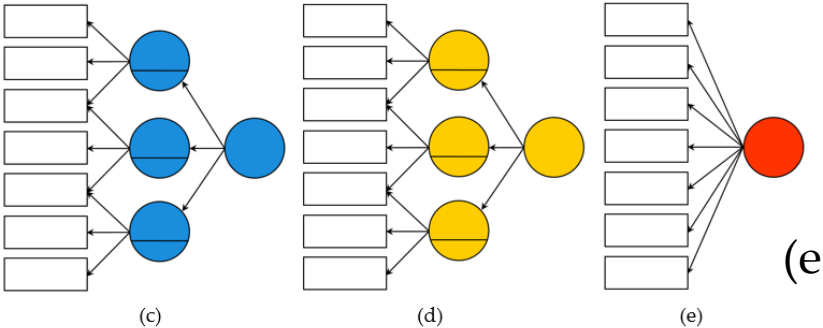
(a) EFA

(b) Bifactor

(c) DINA

(d) DINO

20 free parameters



(e) Uni 3PL → 21 free parameters

Hypotheses

***Hypothesis 1*:** The EFA model will exhibit, on average, the highest fitting propensity

***Hypothesis 2*:** The bifactor model will display higher fitting propensity than the DINA and DINO models

Unidimensional 3PL model?

Method

Data generation:

Sampling from a unit simplex *(Smith & Tromble, 2004)*

- Generate all possible item response patterns for a small number of binary response items

  ↳ $2^7 = 128$ possible patterns

- Assign to each a random weight representing the number of simulees (of $N = 10{,}000$) who supplied that particular pattern

1,000 random data sets to represent the complete data space

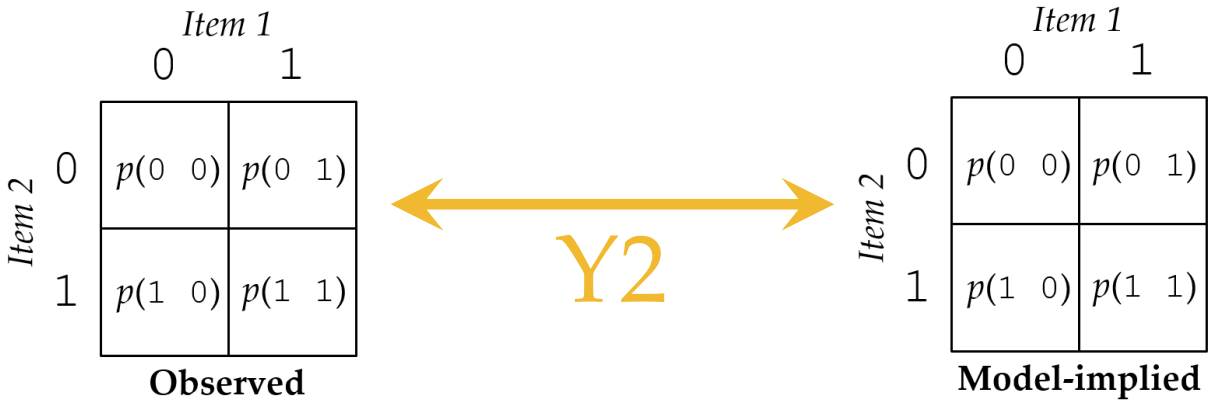
Example data set:

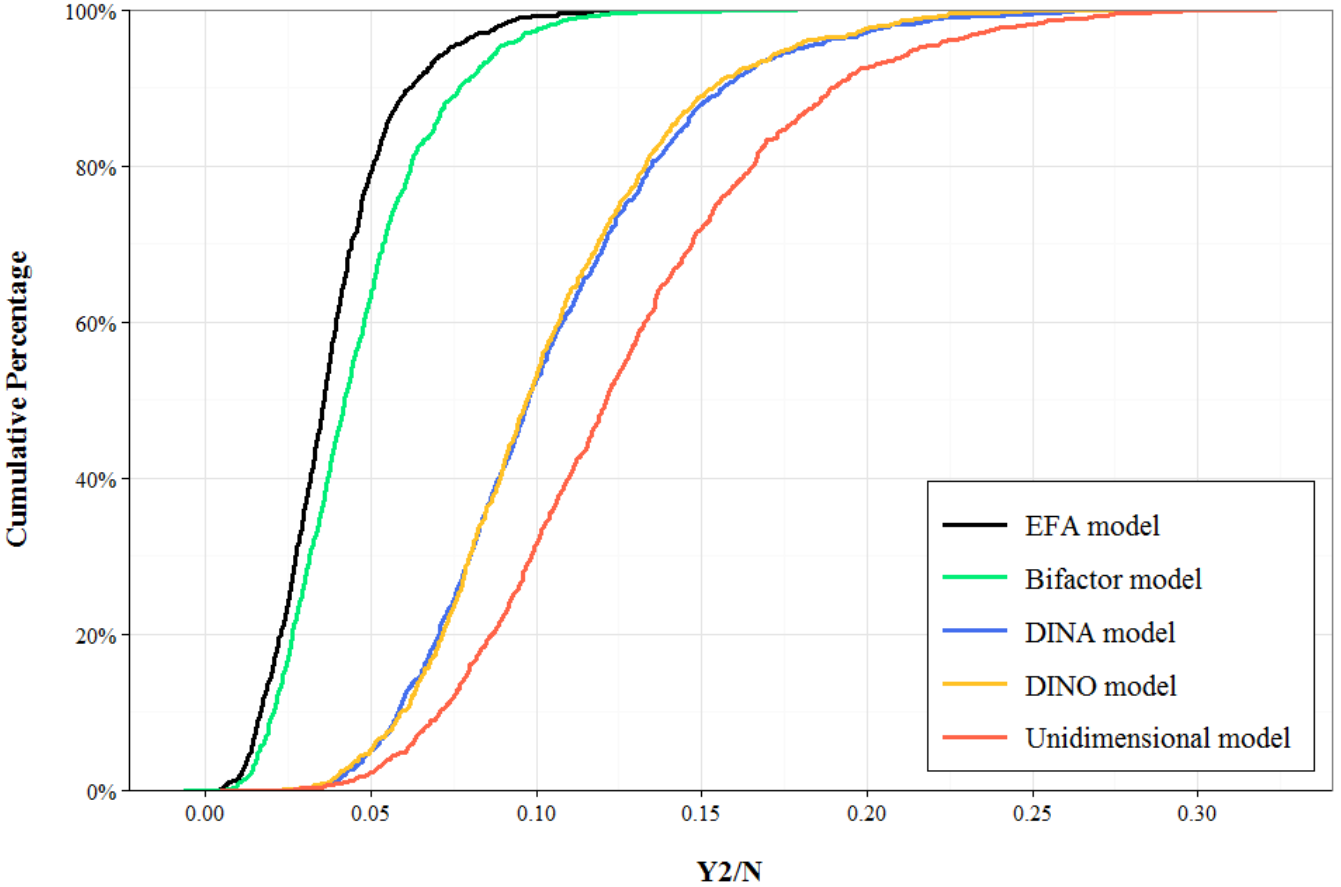
```
0 0 0 0 0 0 0 20.348
0 0 0 0 0 0 1 60.367
0 0 0 0 0 1 0 27.676
0 0 0 0 0 1 1 7.983
0 0 0 0 1 0 0 3.714
0 0 0 0 1 0 1 10.517
0 0 0 0 1 1 0 47.091
          ⋮
1 1 1 1 0 0 1 4.961
1 1 1 1 0 1 0 30.693
1 1 1 1 0 1 1 14.663
1 1 1 1 1 0 0 1.993
1 1 1 1 1 0 1 2.673
1 1 1 1 1 1 0 67.551
1 1 1 1 1 1 1 40.919
```

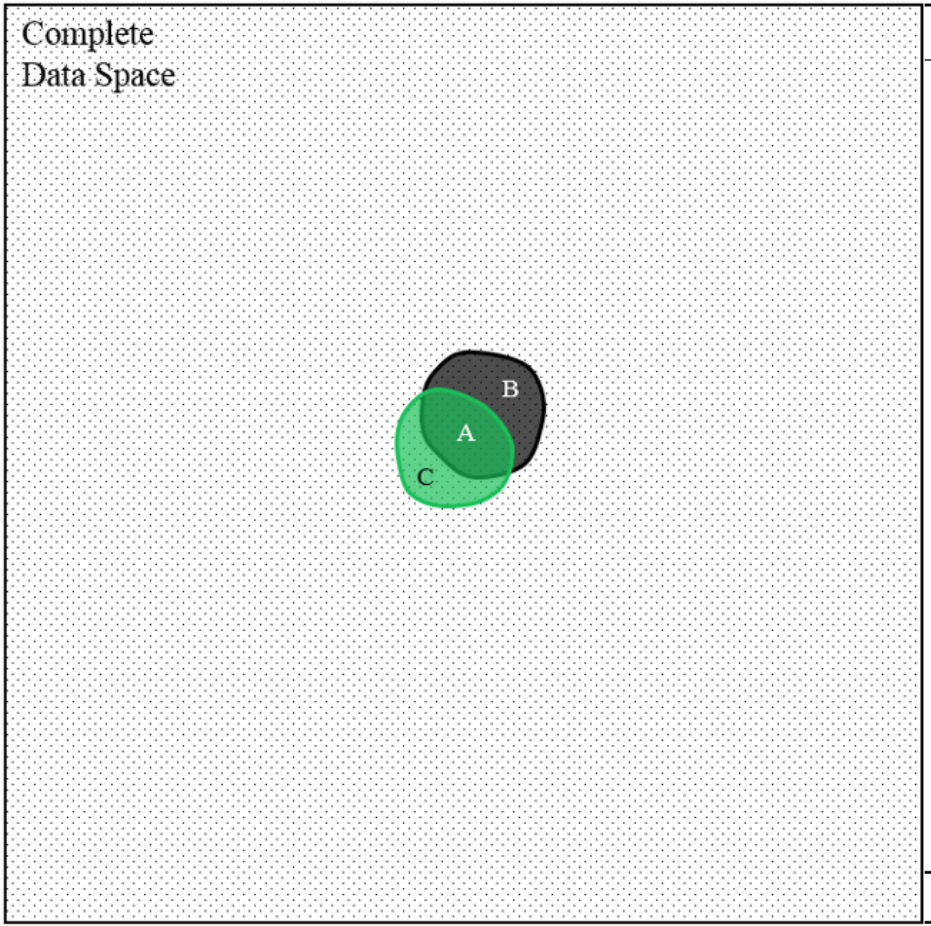
Method

Evaluation measure: Y2/*N* statistic *(Bartholomew & Leung, 2002; Cai et al., 2006)*:

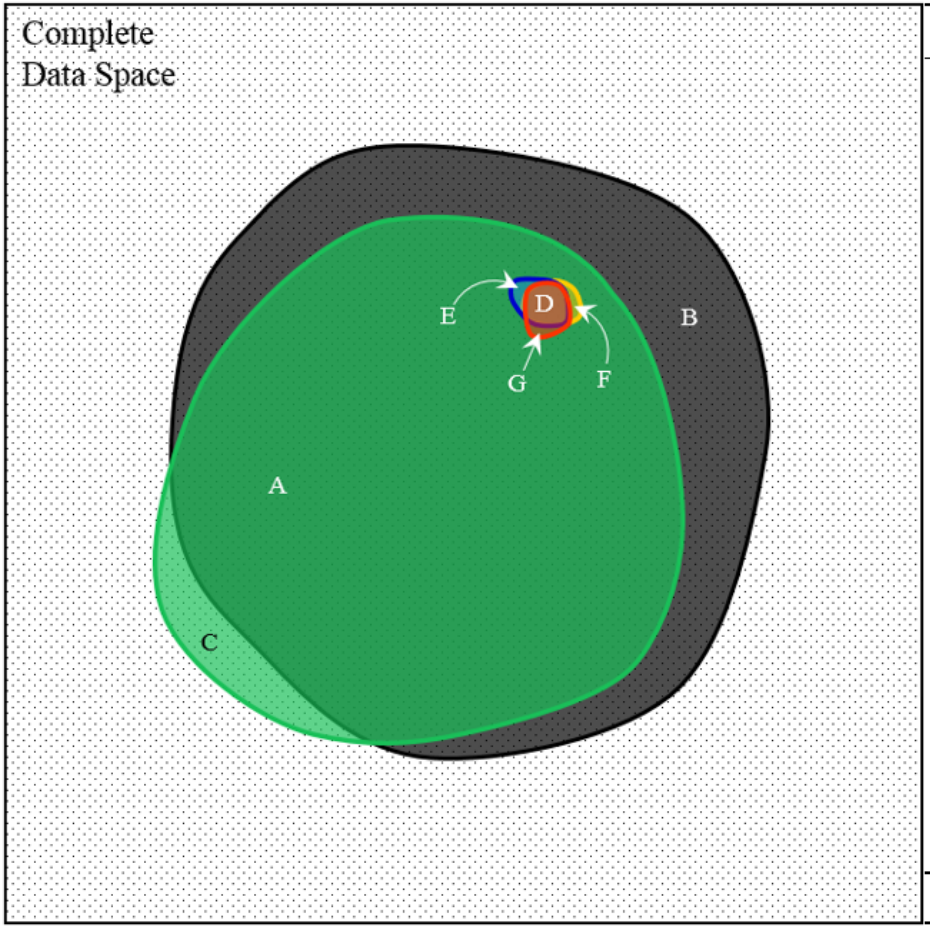
$$\text{Y2}/N = \left[\sum_{i=1}^{I} \frac{(o_i - e_i)^2}{e_i(1 - e_i)} + \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \frac{(o_{ij} - e_{ij})^2}{e_{ij}(1 - e_{ij})} \right]$$



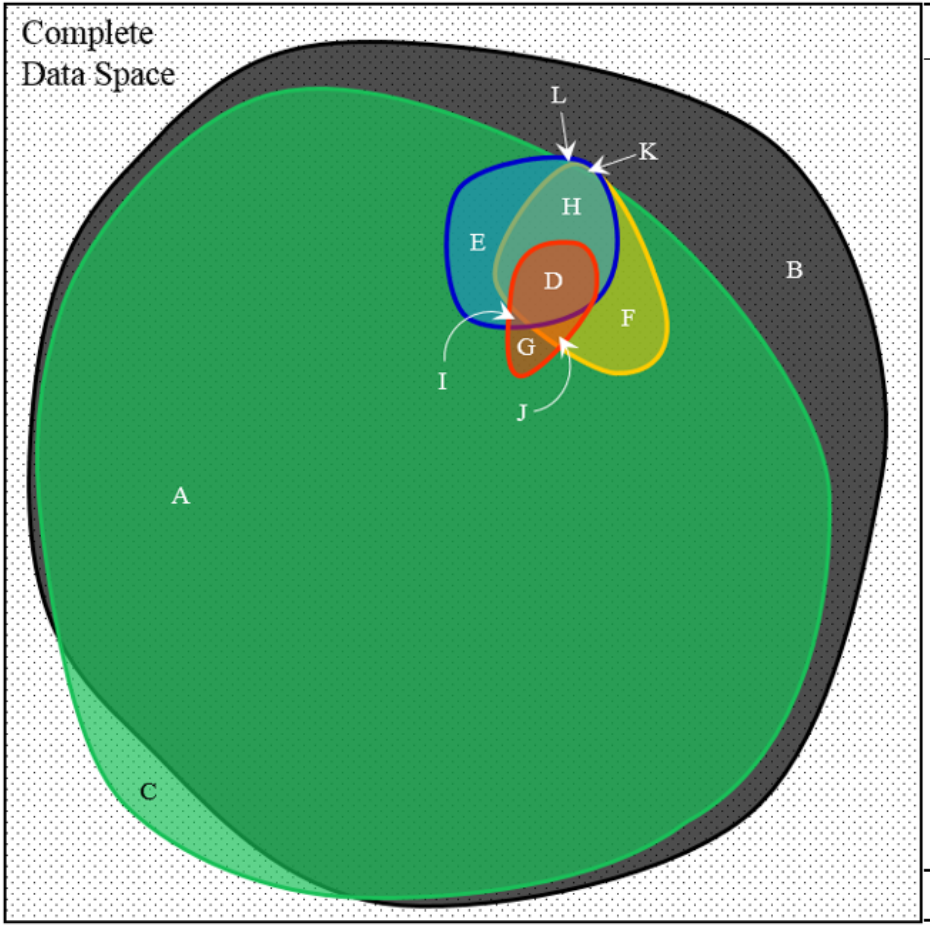

100%
80%
60%
40%
20%
0%
Cumulative Percentage
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Y2/N
EFA model
Bifactor model
DINA model
DINO model
Unidimensional model



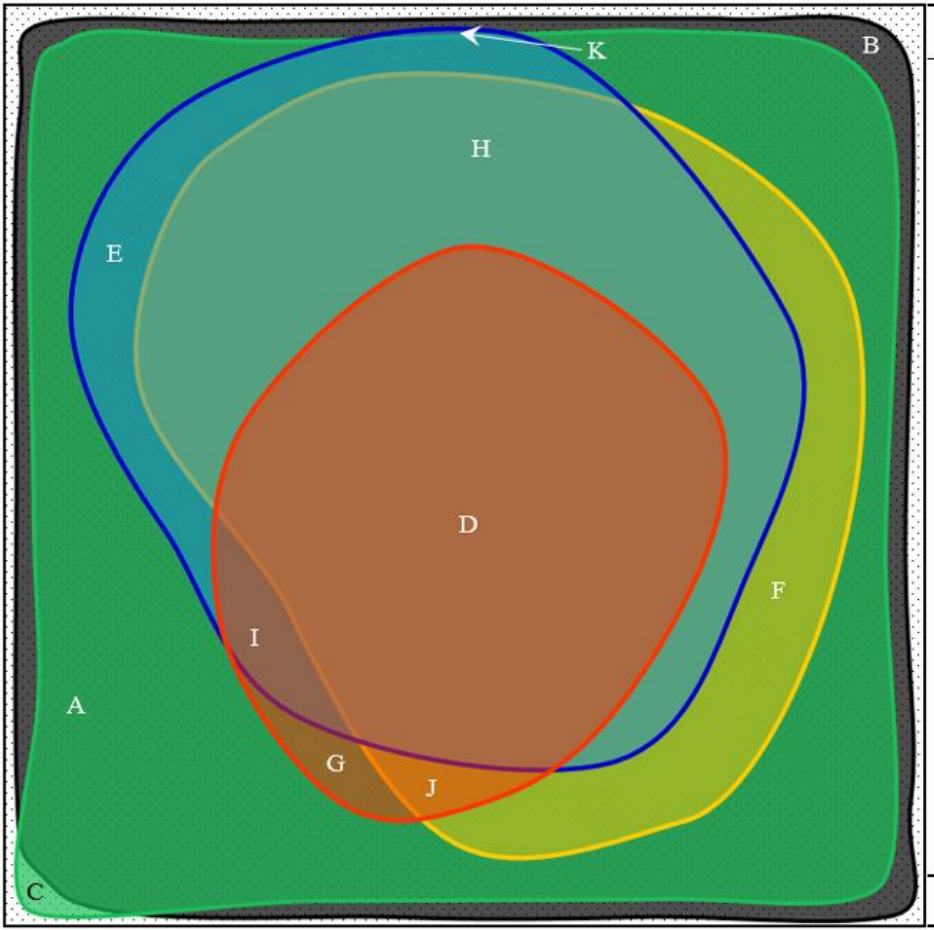
Region	Y2/N ≤ .01
EFA	1.4
Bifactor	0.9
DINA	0
DINO	0
Uni	0
A	0.4
B	1.0
C	0.5
Unoccupied	98.1



Region	Y2/N ≤ .03
EFA	36.2
Bifactor	27.0
DINA	0.3
DINO	0.3
Uni	0.3
A	22.7
B	13.5
C	4.3
D	0.2
E	0.1
F	0.1
G	0.1
Unoccupied	59.5

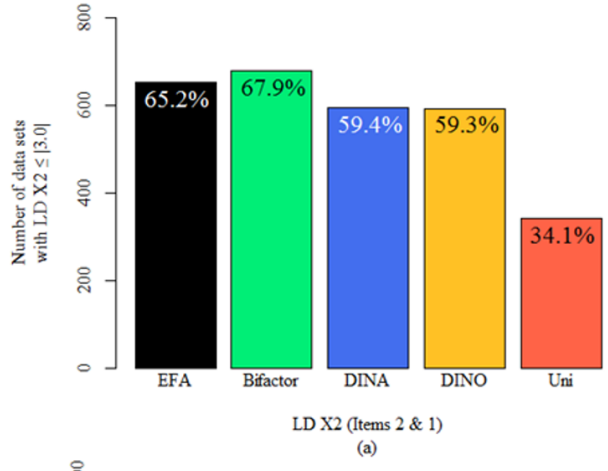


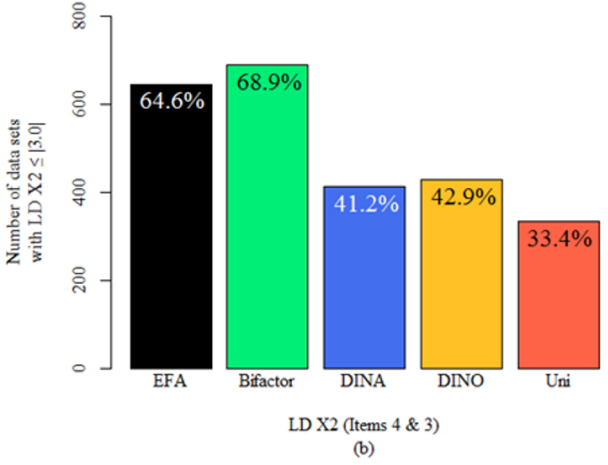
Region	Y2/N ≤ .05
EFA	79.2
Bifactor	63.5
DINA	5.1
DINO	5.2
Uni	2.3
A	51.8
B	19.2
C	3.8
D	1.1
E	1.9
F	2.0
G	0.7
H	1.7
I	0.2
J	0.3
K	0.1
L	0.1
Unoccupied	17.1

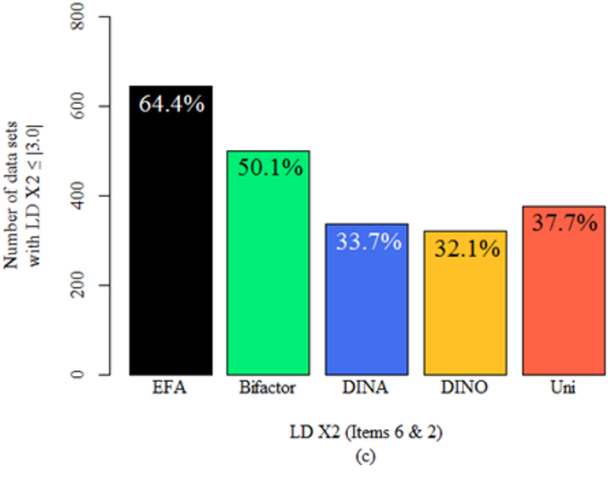


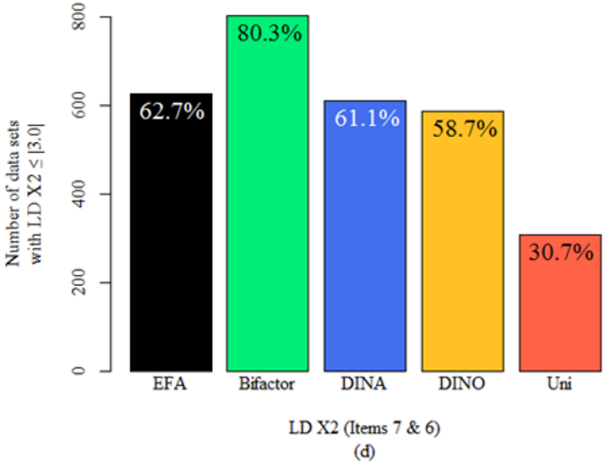
Region	Y2/N ≤ .10
EFA	99.2
Bifactor	97.3
DINA	52.6
DINO	52.9
Uni	31.5
A	96.9
B	2.1
C	0.4
D	22.8
E	6.4
F	8.1
G	2.5
H	19.4
I	3.8
J	2.4
K	0.2
Unoccupied	0.4

Local dependence LD X^2 *(Chen & Thissen, 1997)*









Discussion

Hypotheses: confirmed

The importance of ***functional form:***

- Arrangement of the variables in a model affects ability to fit well
- More complex models (EFA & bifactor) displayed high propensity to fit *any* data
- DINA & DINO models had low fitting propensity; theoretical difference → models fit different patterns
- Unidimensional model had an additional free parameter, but much lower fitting propensity!
- Strong implications re: model evaluation via goodness of fit

References

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology, 55*:1-15.

Cai, L., Maydeu-Olivares, A., Coffman, D., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology, 59*, 173-194.

Collyer, C. E. (1985). Comparing strong and weak models by fitting them to computer-generated data. *Perception & Psychophysics, 38*(5), 476-481.

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General, 121*(3), 364-381.

Myung, I. J., Pitt, M. A., & Kim, W. (2005). Model evaluation, testing and selection. In Lamberts, K. & Goldstone, R., (Eds.), *Handbook of Cognition*. London, UK: Sage Publications Ltd.

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research, 41*(3), 227-259.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*(5), 465-471.

Smith, N. A., & Tromble, R. W. (2004). Sampling uniformly from the unit simplex. *Johns Hopkins University, Tech. Rep. 29.*

Wexler, K. (1978). A review of John R. Anderson's *Language, Memory, and Thought. Cognition, 6*, 327-351.

Thanks to Dr. Li Cai for his assistance with this work.

Contact: bonifayw@missouri.edu

EFA

- 7 c (intercept) parameters
- 7 a (discrimination) parameters for Factor 1
- + 6 a (discrimination) parameters for Factor 2
- **20 free parameters**

Bifactor

- 7 c (intercept) parameters
- 7 a (discrimination) parameters for the General Factor
- 5 a (discrimination) parameters for Specific Factor 1
- + 1 a (discrimination) parameter for Specific Factor 2
- **20 free parameters**

DINA & DINO

- 7 $\lambda_{1,0}$ (intercept) parameters
- 2 $\lambda_{1,1,(1)}$ (main effect) parameters for Attribute 1
- 1 $\lambda_{1,1,(2)}$ (main effect) parameter for Attribute 2
- 2 $\lambda_{1,1,(3)}$ (main effect) parameters for Attribute 3
- 1 $\lambda_{1,2,(1,2)}$ (interaction effect) parameter for Attributes 1 & 2
- 1 $\lambda_{1,2,(2,3)}$ (interaction effect) parameter for Attributes 2 & 3
- 3 c (attribute intercept) parameters
- + 3 a (attribute discrimination) parameters
- **20 free parameters**

Uni

- 7 c (item intercept) parameters
- 7 a (item discrimination) parameters
- + 7 g (pseudo-guessing) parameters
- **21 free parameters**