To θ or not to θ:

A simulation study on the validity of IRT

Jonathan J. Park, M.A.,

Netasha K. Pizano, M.A.,

Kathleen S. J. Preston, PhD.

Table of Contents

General Research Question

- How do IRT ability estimates perform against classically derived estimates when poorly functioning items are present
- How well do these different testing paradigms identify and eliminate poorly functioning items

Impetus for Research – Applied Research

- Typically assesses test frameworks on established scales
  - (ex., Dumenci & Achenbach 2008; Ferrando & Chico 2007)
- Scale modification not discussed at length beyond factor analytic approaches to ensure unidimensionality
- Common finding that IRT performs similar to classically derived methods

Impetus for Research – Simulated Research

- Utilize ‘ideal’ scales or forego scalar modification
  - Ex., Macdonald & Paunonen, 2002; Xu & Stone, 2012
- Item difficulty and discrimination parameters are optimal for simulated populations
- Circumvent the effects that poorly functioning items may have on model fit

Similar Research

- Mead and Meade (2010) generated algorithm for identifying ‘optimal’ tests with prioritization of item information and discrimination
- Algorithms developed tests from larger test bank at fixed final length (i.e., $N_{\text{items}} = 50$)
- Selected optimal items across range of ability
  - Best items first approach

Purpose

- Do away with finite test length
- Present information regarding the efficacy of various testing methodologies using rudimentary unsupervised selection algorithms including:
  - Unweighted summed scores
  - Item selection via maximization of Cronbach's α
  - Factor analytic approach with maximization of factor loadings
  - 2PLM with prioritization of item information and item fit

Simulation Methodology

Table 1.

Simulation Conditions for Modification Study

Simulated Ability	$X \sim N(0, 1)$
Sample Sizes	$N = 100$ $N = 250$ $N = 500$
Test Length	20-items 40-items 80-items
Poor Item Proportions	10% 30% 50%
Item Discriminations (a)*	Poor [0.10; 0.50] Ideal [1.50; 2.50]
ρ_{xx}	1.00
ρ_{xy}	0.60

*See, Baker, 1985

Selection Procedures

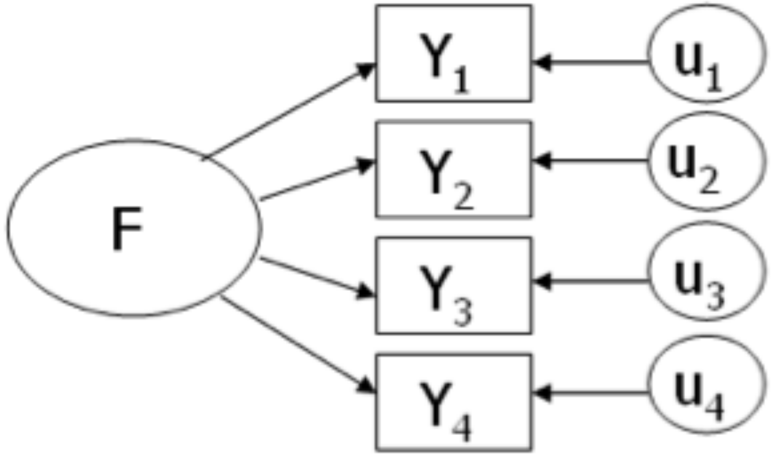
- Simulated Responses
  - Participant responses were simulated using the function *sim.raschtype*() from the *sirt* package prior to beginning selection procedures
- Unweighted Summed Scores
  - Simulated responses to all items were summed using the *rowSums*() function from the *base* package

Selection Procedures

- α-Adjustment
  - Applied *itemAnalysis*() function from the *CTT* package
  - Selected items that improved α-coefficient by greatest magnitude if removed
- Continue iterative procedure until either condition met:
  - α-coefficient >= 0.80
  - α-coefficient would decrease if any other item were to be removed from the measure

Selection Procedures

- Factor Analytic Approach
  - Items selected for removal via iterative process via the *factanal*() function in the *stats* packages
  - Remove items with lowest factor loadings until all items were acceptable
    - i.e., (λ >= 0.30; see, Brown, 2014)



Selection Procedures

- 2PLM Approach (see, Preston, 2018)
  - Items tested for unidimensionality/Local Dependence
  - Scalar optimization via prioritization of item information and fit
    - *estimate.mml.2pl*() function within the *TAM* package
    - *IRT.informationCurves*() function for item/test information
    - *IRT.itemfit*() function within the CDM package
- Default settings for all other options
- Iterative Log-Likelihood tests to confirm no significant change to model fit per each item removal

Selected Output ($N = 500$)

- Final Test Length
- Estimated-to-True Score Correlations
- Estimated-to-Outcome Correlations
- False Positive Rate
- False Negative Rate

Final Test Length (N = 500)

- α-adjustment performed best at small test sizes
- At larger test sizes, CFA achieved best approximations of optimal test length
- IRT procedure consistently generated shortest tests

Table 2.

Final Test Length for Large Sample Condition (N = 500)

Test Length

20-Items

	Proportion of Poorly Functioning Items		
Condition	10%	30%	50%
SS	20.00 (0.00)	20.00 (0.00)	20.00 (0.00)
Alpha	16.80 (0.02)	12.73 (0.02)	10.01 (0.01)
CFA	15.71 (0.04)	12.21 (0.05)	9.25 (0.02)
IRT	14.78 (0.05)	11.34 (0.06)	7.92 (0.08)
Expected	*18.00*	*14.00*	*10.00*

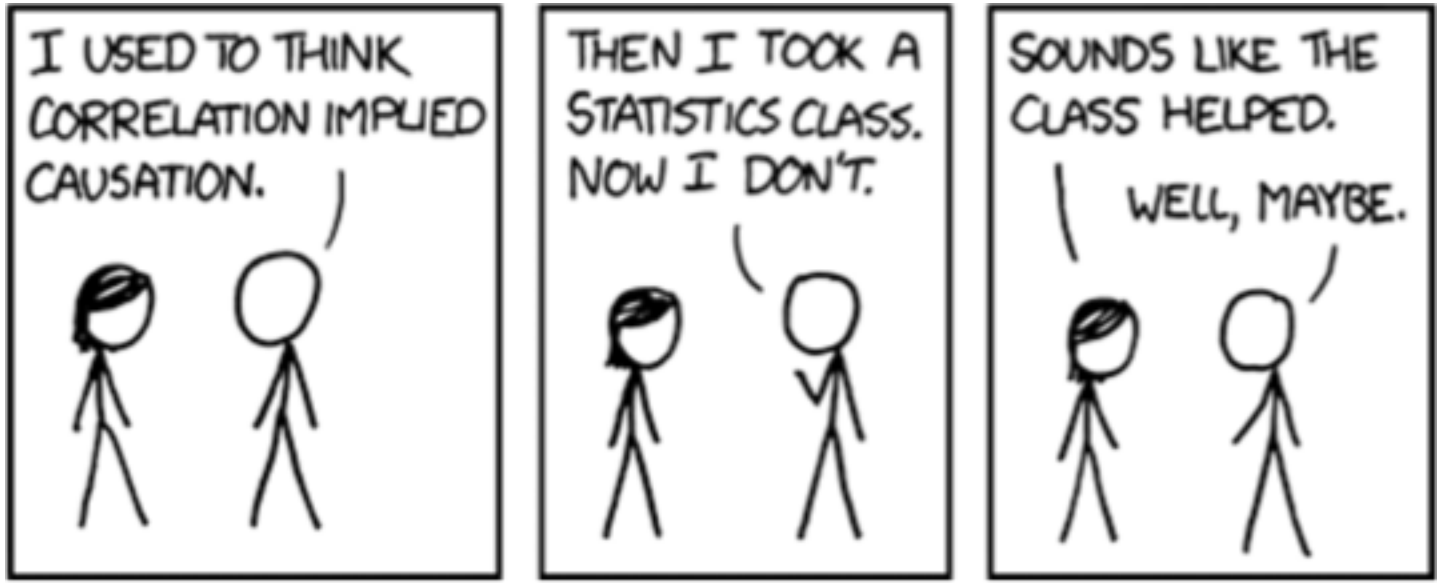
40-Items

Condition	10%	30%	50%
SS	40.00 (0.00)	40.00 (0.00)	40.00 (0.00)
Alpha	40.00 (0.00)	37.58 (0.09)	19.56 (0.14)
CFA	31.58 (0.10)	24.61 (0.06)	17.41 (0.11)
IRT	28.93 (0.13)	22.83 (0.08)	16.12 (0.16)
Expected	*36.00*	*28.00*	*20.00*

80-Items

Condition	10%	30%	50%
SS	80.00 (0.00)	80.00 (0.00)	80.00 (0.00)
Alpha	80.00 (0.00)	80.00 (0.00)	79.67 (0.10)
CFA	63.53 (0.22)	48.58 (0.36)	34.91 (0.17)
IRT	50.73 (0.23)	42.60 (0.42)	31.90 (0.20)
Expected	*72.00*	*56.00*	*40.00*

Estimated-to-True Score Correlations



Estimated-to-True Score Correlations

- Accuracy improves with test length
- α-derived tests performed best with short baseline scales
- IRT and CFA emerged as more accurate methods under longer baseline scales

Table 3.

Estimate-to-True Score Correlations for Large Sample Condition (*N* = 500)

Test Length

20-Items

	Proportion of Poorly Functioning Items		
Condition	10%	30%	50%
SS	0.87 (< 0.001)	0.81 (< 0.001)	0.72 (0.001)
Alpha	0.87 (< 0.001)	0.85 (< 0.001)	0.81 (0.001)
CFA	0.86 (< 0.001)	0.84 (< 0.001)	0.80 (0.001)
IRT	0.86 (0.001)	0.83 (0.001)	0.77 (0.004)

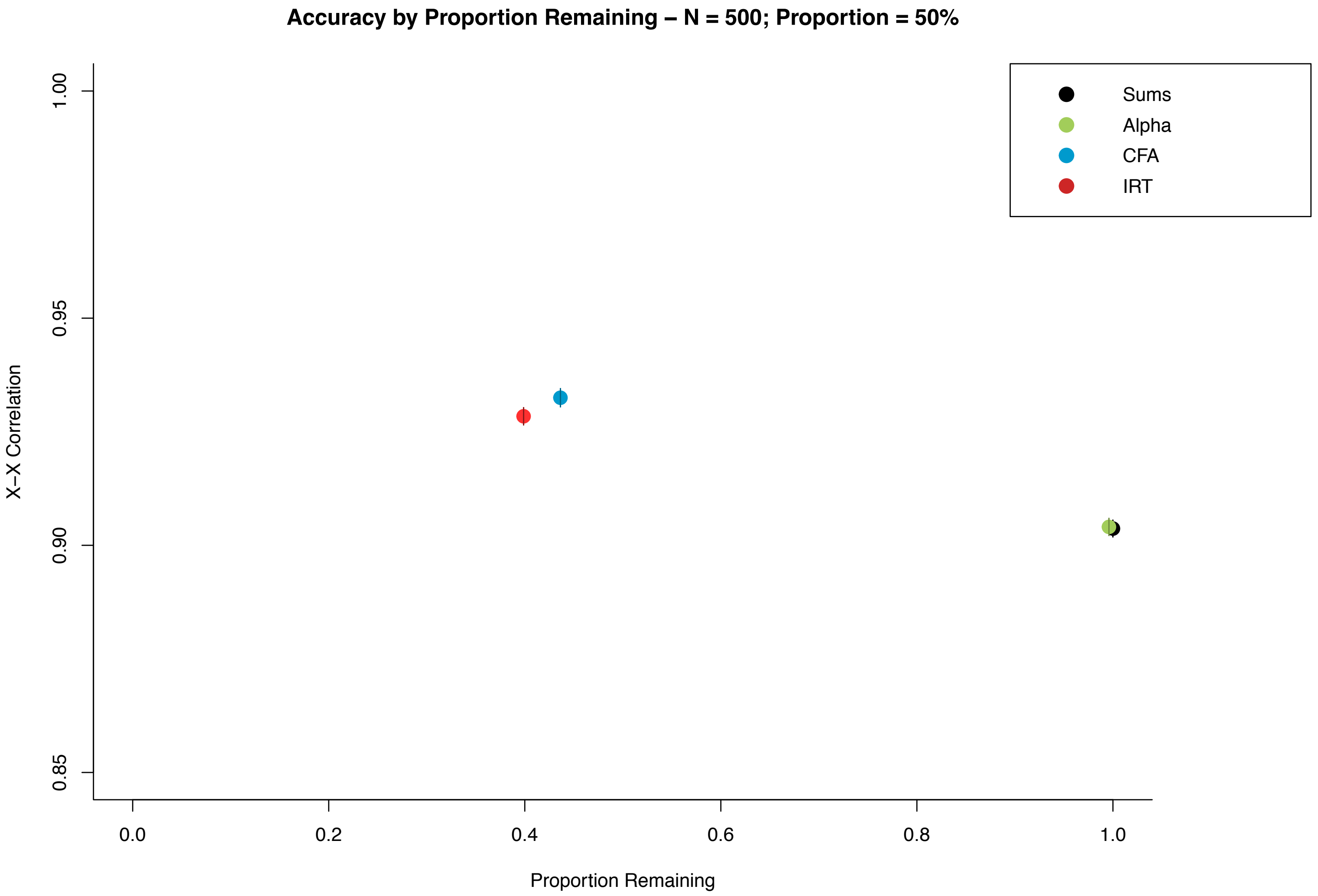
40-Items

Condition	10%	30%	50%
SS	0.93 (< 0.001)	0.89 (< 0.001)	0.83 (0.001)
Alpha	0.93 (< 0.001)	0.90 (< 0.001)	0.88 (< 0.001)
CFA	0.93 (< 0.001)	0.91 (< 0.001)	0.88 (0.001)
IRT	0.92 (< 0.001)	0.90 (0.001)	0.87 (0.002)

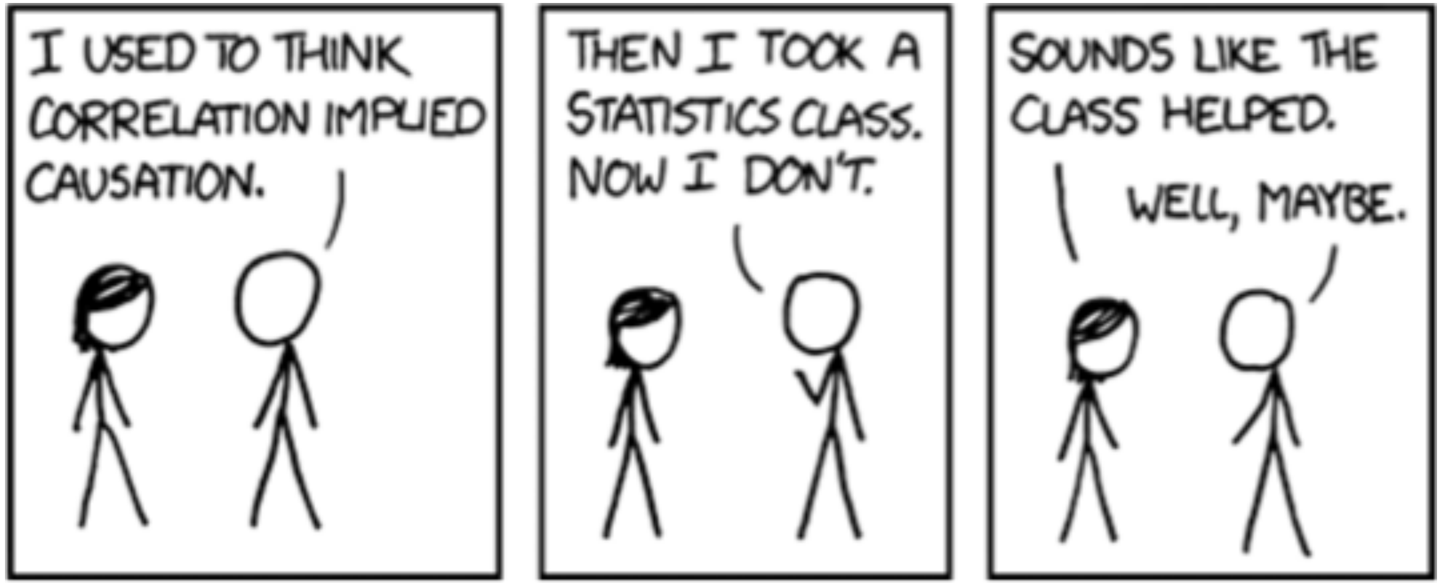
80-Items

Condition	10%	30%	50%
SS	0.96 (< 0.001)	0.94 (0.001)	0.90 (0.001)
Alpha	0.96 (< 0.001)	0.94 (0.001)	0.90 (0.001)
CFA	0.96 (< 0.001)	0.95 (0.001)	0.93 (0.001)
IRT	0.95 (< 0.001)	0.94 (0.001)	0.93 (0.001)

Estimate-to-True Score by Test Length



Estimated-to-Outcome Correlations



Estimated-to-Outcome Correlations

- Greater accuracy with test length for all methods
- α-adjustment and Factor Analysis strongest for short- to moderate-tests
- Factor Analysis and IRT at longer test lengths

Table 4.

Estimate-to-Outcome Correlations for Large Sample Condition (N = 500)

Test Length

20-Items

	Proportion of Poorly Functioning Items		
Condition	10%	30%	50%
SS	0.52 (0.001)	0.49 (0.001)	0.44 (0.002)
Alpha	0.53 (0.001)	0.51 (0.001)	0.49 (0.001)
CFA	0.52 (0.001)	0.51 (0.001)	0.48 (0.002)
IRT	0.52 (0.001)	0.5 (0.001)	0.47 (0.003)
ρ_{xy}	0.60	0.60	0.60

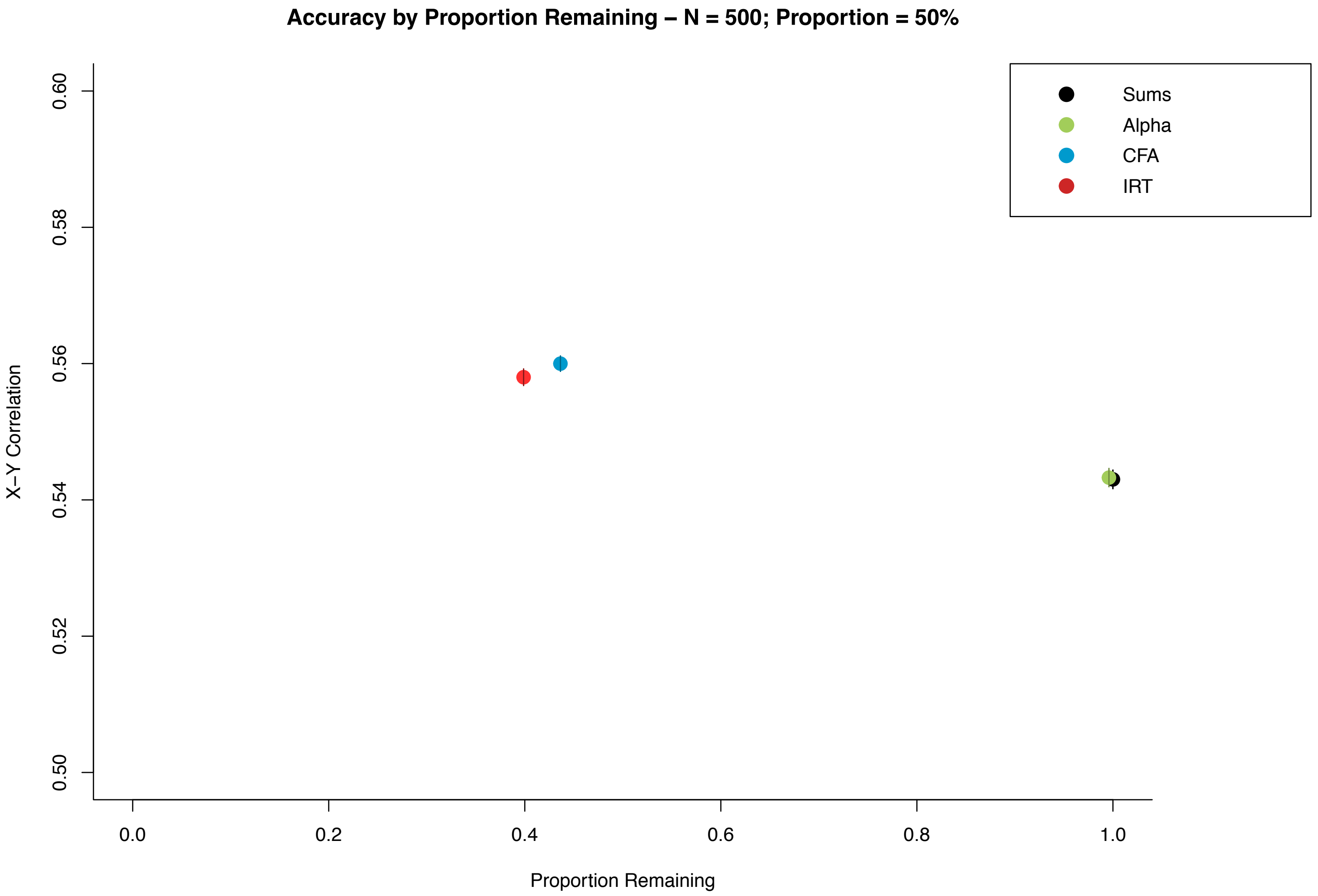
40-Items

Condition	10%	30%	50%
SS	0.55 (0.001)	0.53 (0.001)	0.50 (0.002)
Alpha	0.55 (0.001)	0.54 (0.001)	0.53 (0.002)
CFA	0.56 (0.001)	0.55 (0.001)	0.53 (0.002)
IRT	0.55 (0.001)	0.54 (0.001)	0.52 (0.002)
ρ_{xy}	0.60	0.60	0.60

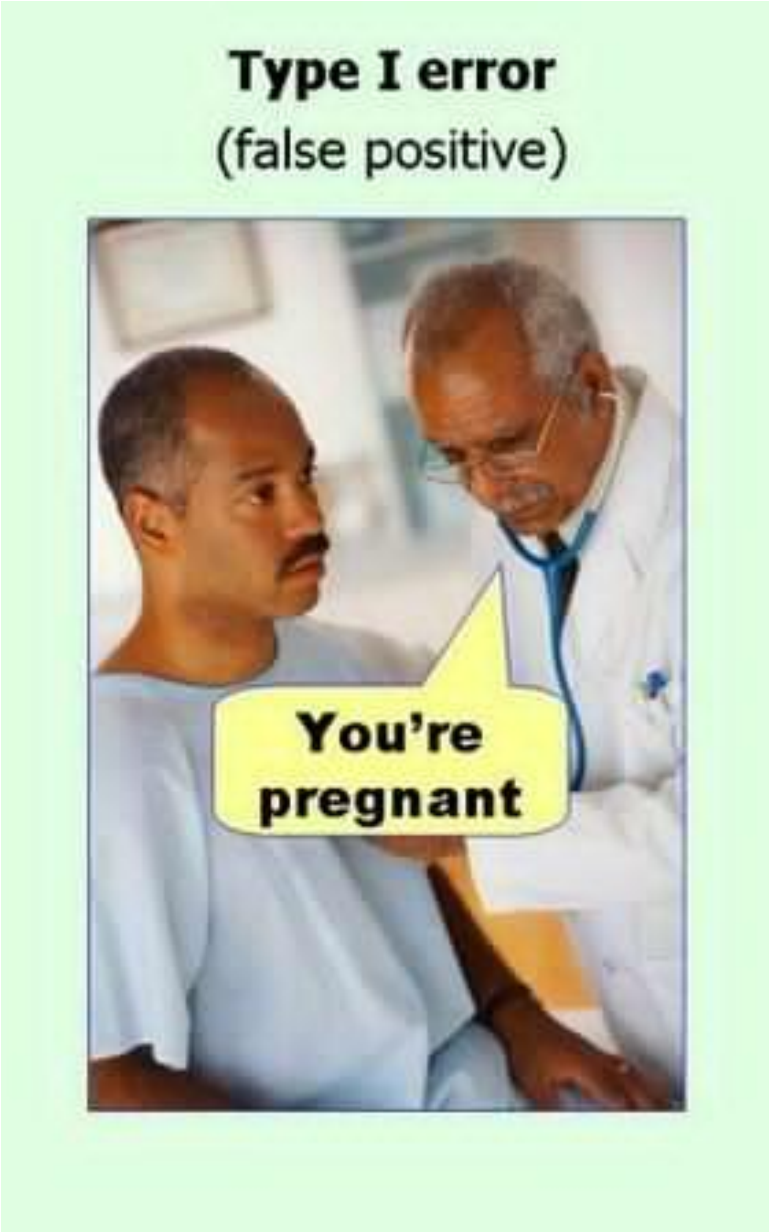
80-Items

Condition	10%	30%	50%
SS	0.58 (0.001)	0.56 (0.002)	0.54 (0.001)
Alpha	0.58 (0.001)	0.56 (0.002)	0.54 (0.001)
CFA	0.58 (0.001)	0.57 (0.002)	0.56 (0.001)
IRT	0.57 (0.001)	0.57 (0.002)	0.56 (0.001)
ρ_{xy}	0.60	0.60	0.60

Estimate-to-Outcome Accuracy by Test Length



False Positive Rate



False Positive Rates

- Unweighted sums and α-derived tests typically outperform other methods across all test length conditions
- IRT performs similarly to the aforementioned methods under 80-item condition

Table 5.

False Positive Rates for Large Sample Condition (N = 500)

Test Length

20-Items

	Proportion of Poorly Functioning Items		
Condition	10%	30%	50%
SS	4.38 (0.04)	5.27 (0.06)	6.56 (0.10)
Alpha	4.14 (0.04)	4.29 (0.06)	4.44 (0.09)
CFA	6.26 (0.02)	6.92 (0.03)	7.56 (0.02)
IRT	6.47 (0.02)	7.12 (0.04)	7.57 (0.08)

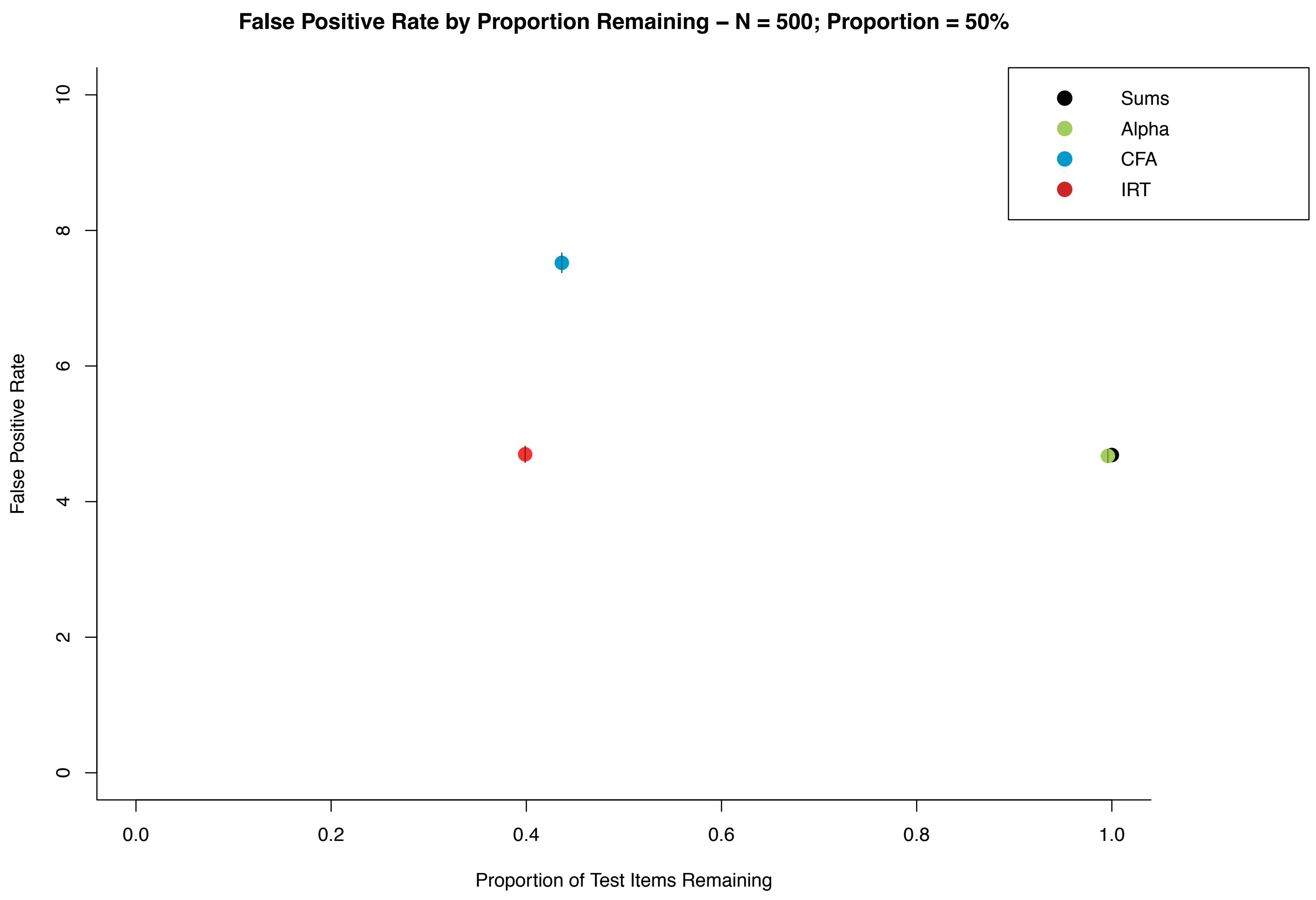
40-Items

Condition	10%	30%	50%
SS	3.64 (0.04)	4.49 (0.04)	5.75 (0.09)
Alpha	3.65 (0.04)	4.38 (0.04)	3.99 (0.09)
CFA	6.26 (0.03)	6.83 (0.04)	7.46 (0.06)
IRT	4.88 (0.03)	5.42 (0.03)	6.24 (0.07)

80-Items

Condition	10%	30%	50%
SS	2.92 (0.05)	3.39 (0.12)	4.69 (0.07)
Alpha	2.92 (0.04)	3.39 (0.12)	4.67 (0.07)
CFA	6.24 (0.05)	6.91 (0.04)	7.52 (0.02)
IRT	3.74 (0.05)	4.00 (0.08)	4.70 (0.05)

False Positive Rate by Test Length



False Negative Rate

False Negative Rate

- Unweighted sums and α-derived tests performed best at shortest test length
- IRT outperforms all methods in 40- and 80-item conditions across all conditions of poorly functioning items

Table 6.
False Negative Rates for Large Sample Condition (N = 500)

Test Length

20-Items

	Proportion of Poorly Functioning Items		
Condition	10%	30%	50%
SS	8.24 (0.05)	9.83 (0.06)	11.49 (0.10)
Alpha	8.16 (0.05)	9.48 (0.07)	10.95 (0.13)
CFA	6.28 (0.02)	7.08 (0.03)	8.15 (0.02)
IRT	6.57 (0.03)	7.52 (0.05)	9.18 (0.13)

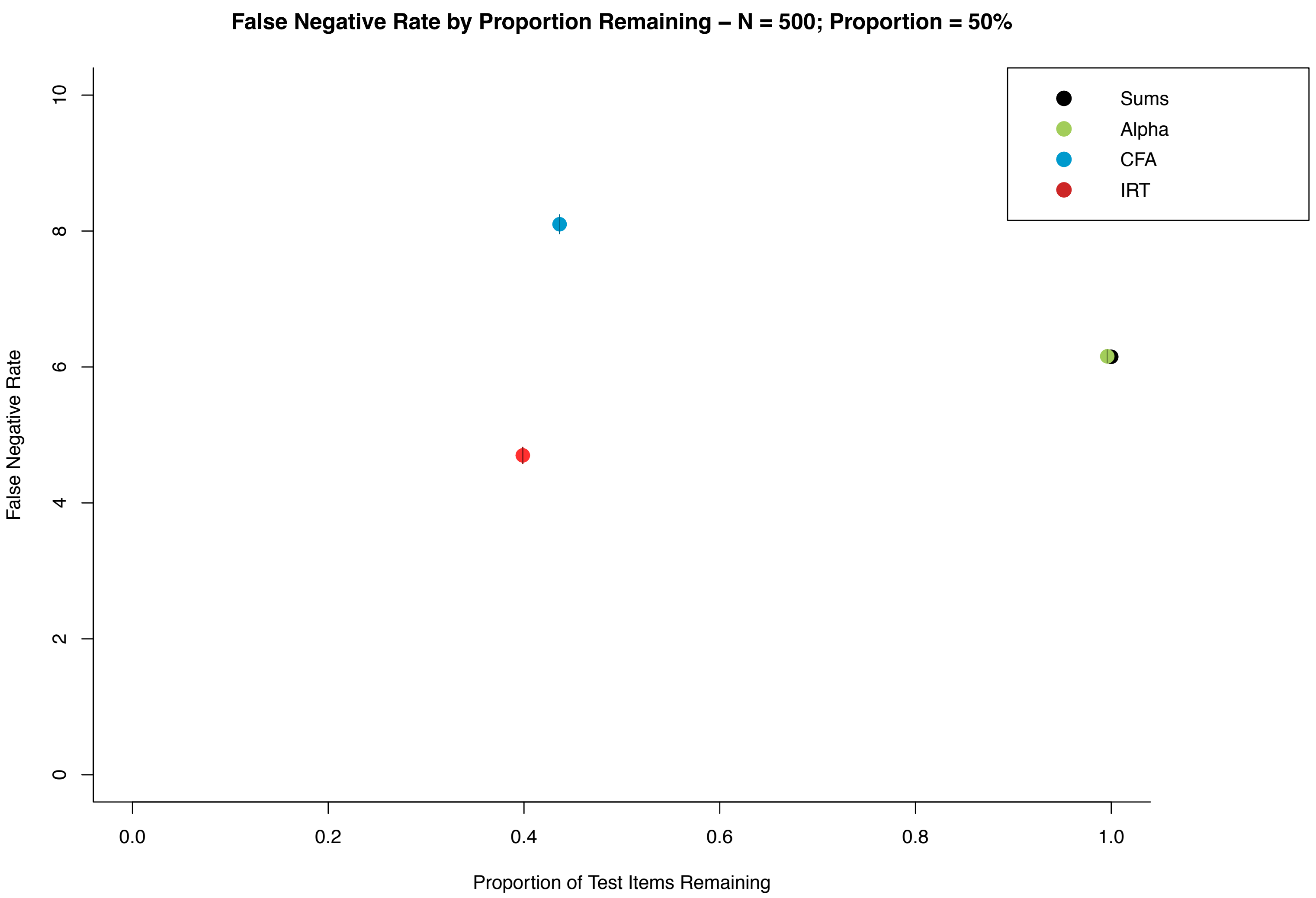
40-Items

Condition	10%	30%	50%
SS	5.76 (0.05)	6.96 (0.04)	8.46 (0.11)
Alpha	5.76 (0.05)	6.81 (0.04)	7.86 (0.13)
CFA	6.29 (0.03)	6.97 (0.05)	8.09 (0.08)
IRT	4.90 (0.04)	5.42 (0.03)	6.31 (0.08)

80-Items

Condition	10%	30%	50%
SS	3.95 (0.06)	4.84 (0.12)	6.15 (0.07)
Alpha	3.95 (0.06)	4.84 (0.11)	6.16 (0.07)
CFA	6.28 (0.05)	7.04 (0.05)	8.09 (0.03)
IRT	3.74 (0.05)	4.00 (0.08)	4.70 (0.05)

False Negative Rate by Test Length



Concluding Remarks

- All testing methodologies performed adequately when presented with varying proportions of poorly functioning items
  - Internal-, external-correlations were acceptable across all testing paradigms
- For 20-item conditions, unweighted sums and α-derived scales performed best
- In 80-item conditions, IRT selected shortest scales without the cost of accuracy (i.e., r_{xx}, r_{xy}, *False positive rate*, *False negative rate*)

Concluding Remarks

- Recommendation for application of IRT when modifying large test banks
- Possible utility in generating parsimonious scales without the cost of precision
- General recommendation for future research in small-scale test modification

Limitations

- Operational definitions for ‘poorly’ functioning items is dependent on the trait of interest
- Non-normal trait- and item-distributions were not simulated
- Simulation results assume algorithmic modification of scales rather than man-based modification
  - Likely overestimates relationships one is likely to find due to lack of human error

Scan to Learn About the Author

For References and Supplemental Information: