Equivalence Testing of Expected Parameter Changes for Evaluating Local and Global Model Fit in SEM

Sunthud Pornprasertmanit, Chulalongkorn University, Thailand

Suppanut Sriutaisuk, Chulalongkorn University, Thailand

Moritz Heene, Ludwig Maximilian University of Munich, Germany

Wei Wu, Indiana University, Indianapolis, USA

AMPC 2026, Singapore—February 24, 2026

Outline

- The problem with current fit evaluation
- A confidence-interval-based equivalence framework
- Simulation studies
- Empirical illustration
- When does it work—and when it does not?

Introduction

- In Structural Equation Modeling (SEM), researchers evaluate whether model–data discrepancy is substantively negligible.
- **Fit indices** (e.g., RMSEA, SRMR, CFI, TLI) quantify the degree of model fit/misfit.
- Researchers typically rely on conventional cutoffs,
  - e.g., RMSEA < .06, SRMR < .08, CFI > .95, TLI > .95
  - But what do these numbers *mean substantively*?
- Fit indices are influenced by many factors: Model size, Magnitude of parameter values (e.g., factor loadings)
- Recent proposals suggest tailored cutoffs: Dynamic fit indices (McNeish & Wolf, 2022)

Introduction

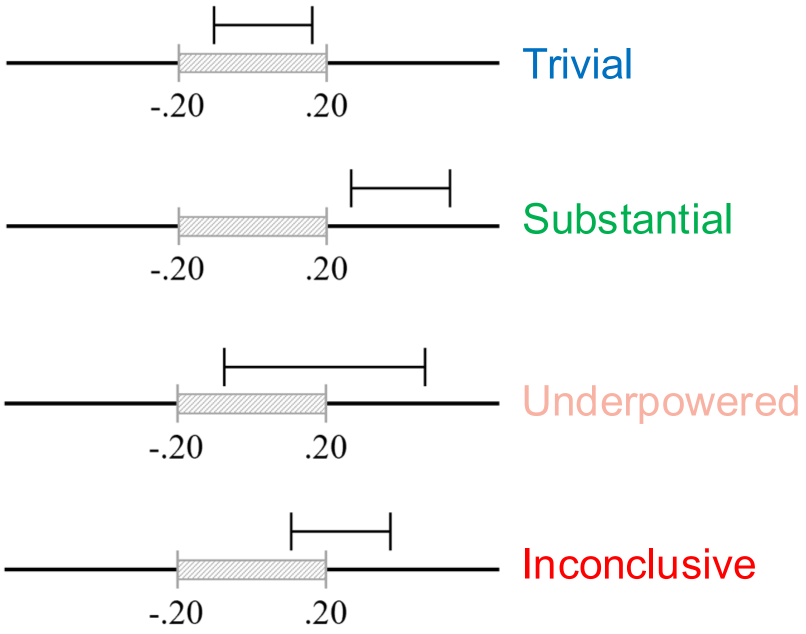
- Misspecification $\neq$ misfit
  - Good global fit does not guarantee the absence of substantively meaningful local misspecifications.
- Why not evaluate fit **directly at the local level**?
  - Examine whether each fixed parameter is appropriate.
  - If freed, its estimate should be **equivalent** to the fixed value (typically 0).

Equivalence Testing in Local Fit Evaluation

- Saris, Satorra, and Van der Veld (2009) proposed evaluating fixed parameters using:
  - Modification Indices (detectability),
  - Standardized Expected Parameter Changes (SEPCs) (magnitude),
  - A Smallest Effect Size of Interest (SESOI) (substantive importance),
    - e.g., a standardized cross-loading of .40 or a measurement error correlation of .10.
- However, when the power of the modification index is low, their decision rules may classi fy trivial misspecifications as substantial.
- Our approach reframes this using CI-based equivalence testing of EPCs.

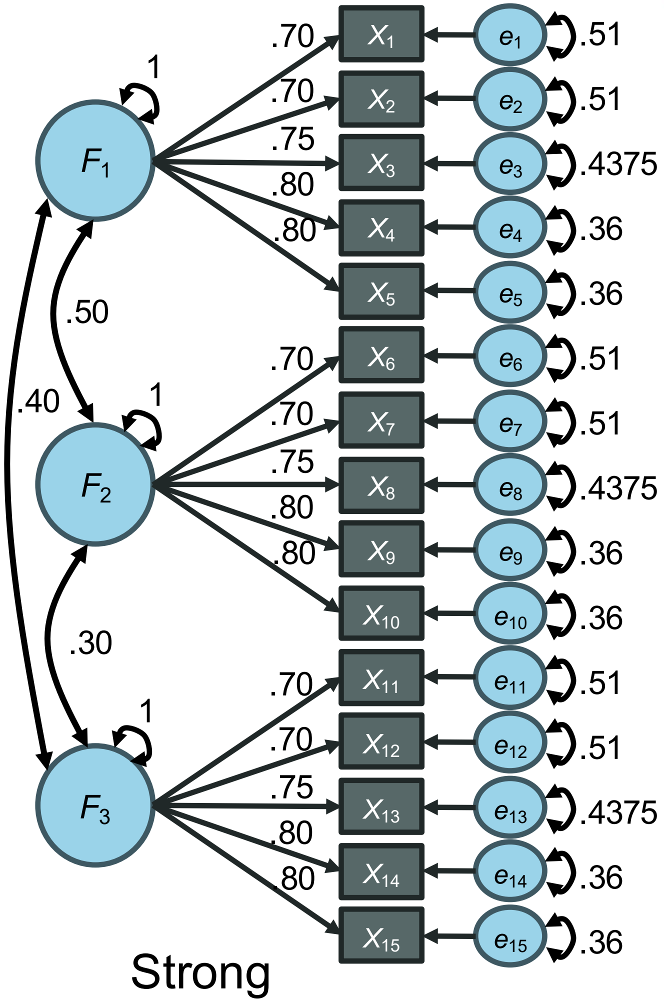
Equivalence Testing in Local Fit Evaluation

- Our proposal: Evaluate local fit using equivalence testing based on CIs of EPCs.
- Rule: use the 90% CI of each EPC (consistent with two one-tailed tests).
  - CI entirely inside SESOI ⟶ Trivial
  - CI entirely outside SESOI ⟶ Substantial
  - CI width exceeds the SESOI width ⟶ Underpowered
  - CI partially overlaps the SESOI ⟶ Inconclusive
- Proposed conservative hierarchical aggregation rules:
  - If any CI is Underpowered ⟶ Underpowered
  - Else if any CI is Substantial ⟶ Substantial
  - Else if any CI is Inconclusive ⟶ Inconclusive
  - Else -> Trivial



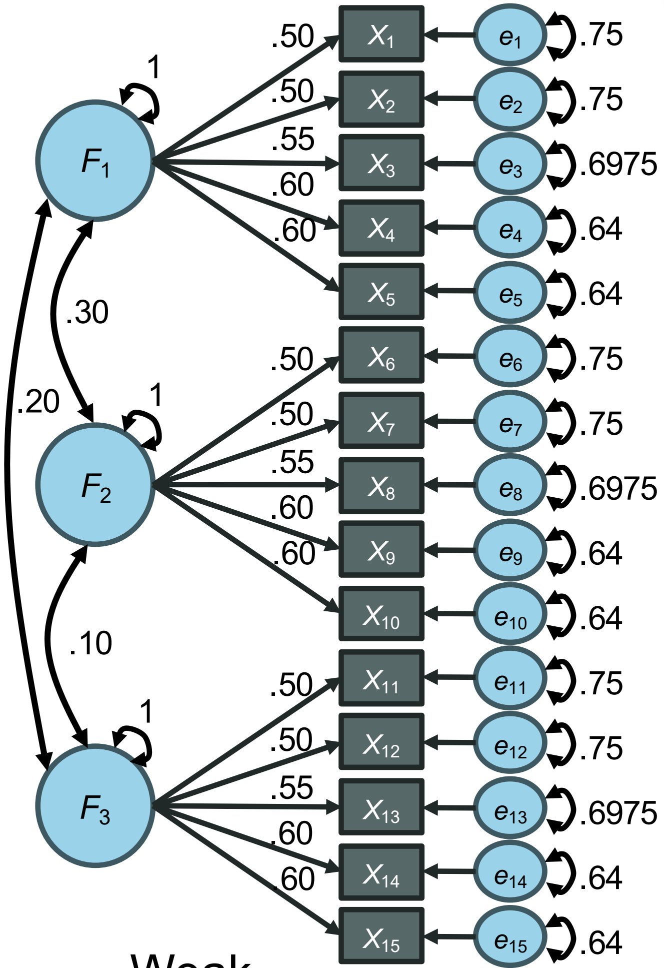
Simulation Study

- Study 1
  - Data: Three-factor CFA, 5 indicators per factor (similar to Hu & Bentler, 1999)
  - Sample size: N = 100 to 640000 (15 levels)
  - Weak vs. strong parameter values



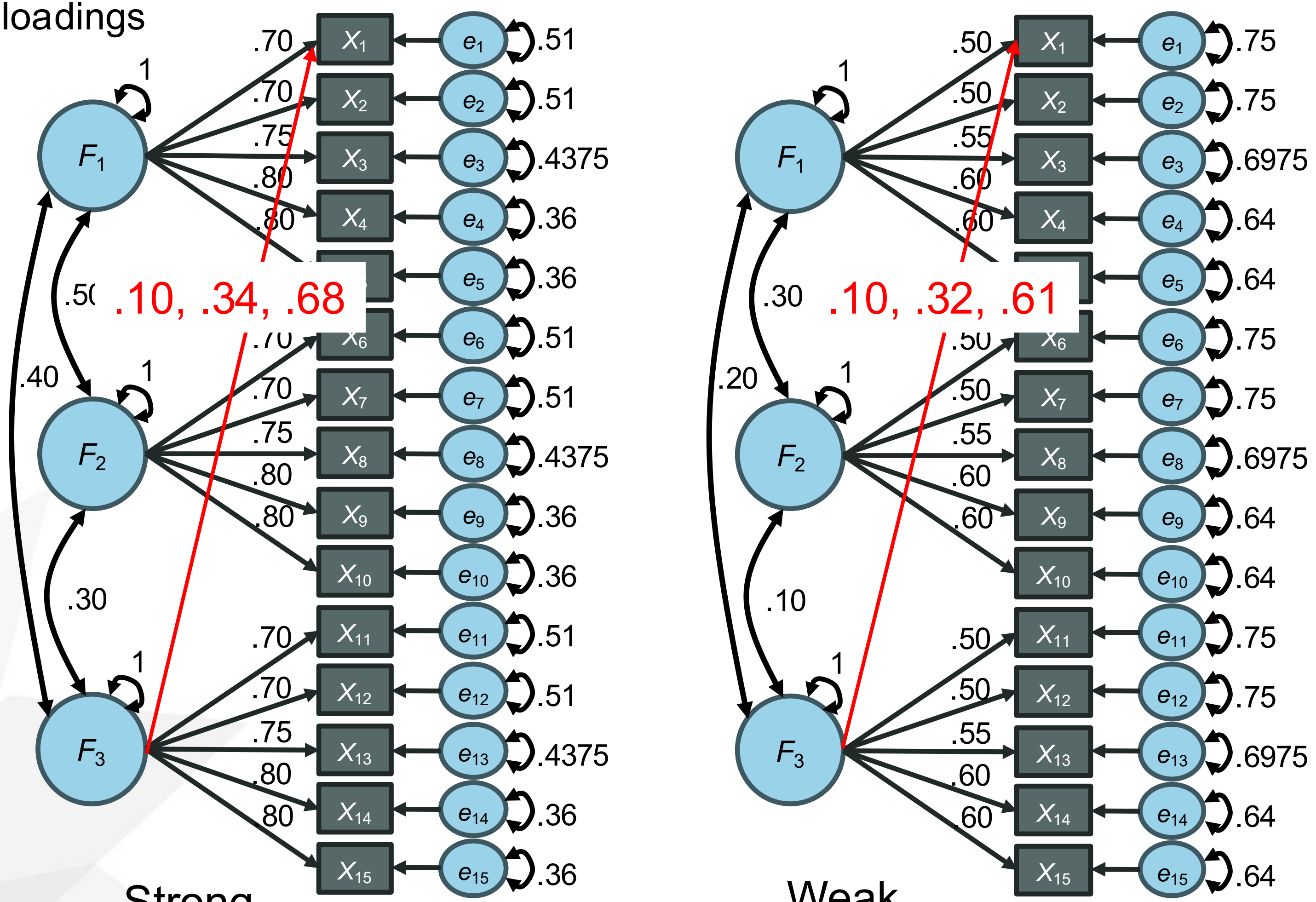
Simulation Study

- Study 1
  - Data: Three-factor CFA, 3 indicators per factor (similar to Hu & Bentler, 1999)
  - Sample size: N = 100 to 640000 (15 levels)
  - Weak vs. strong parameter values
  - Type of misspecification
    - No misspecification
    - Standardized cross loadings of .10, .30, and .50
    - Error correlations of .025, .075, and .125



Weak

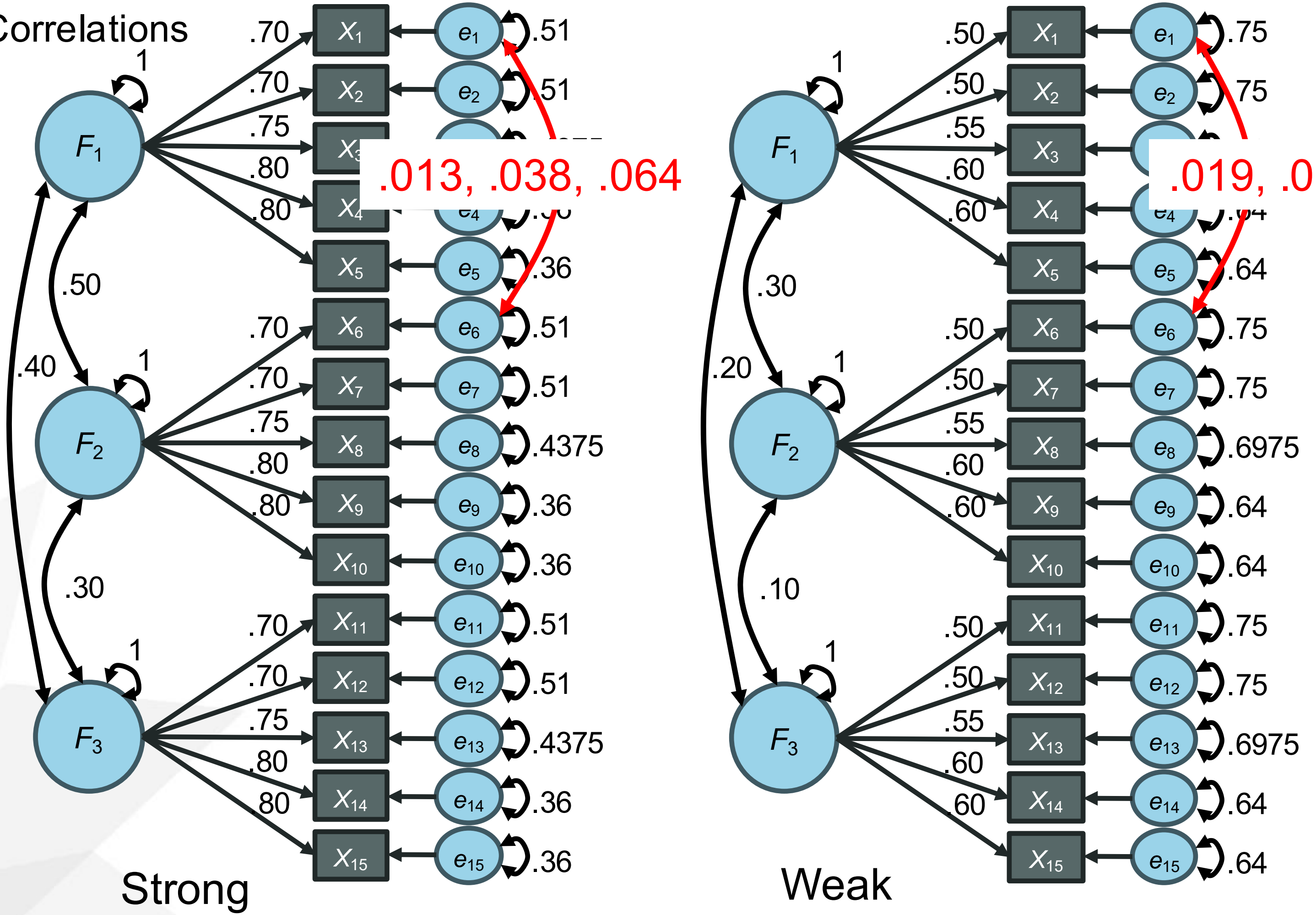
- Cross loadings



Strong

Weak

- Error Correlations



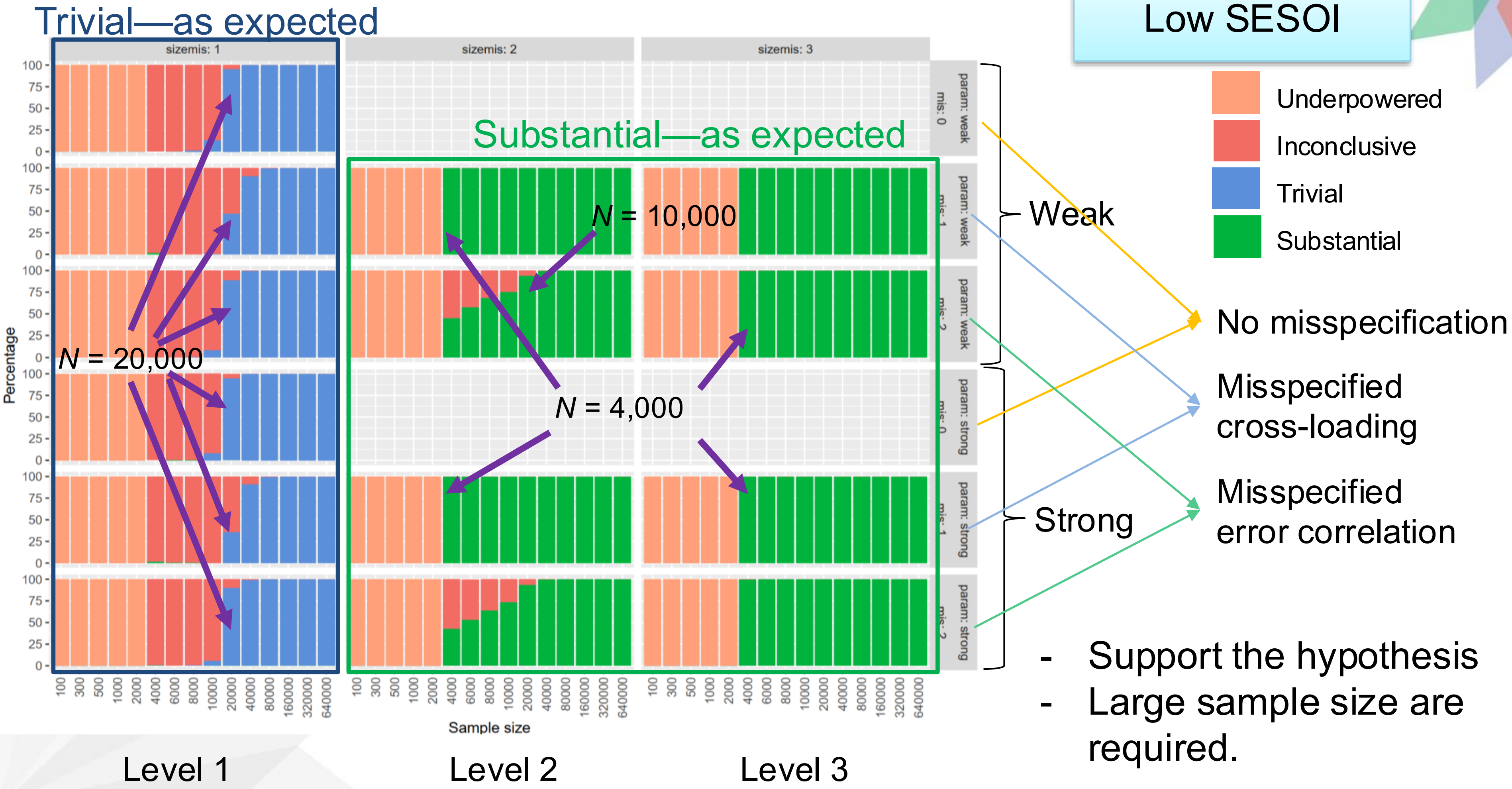
Simulation Study

- SESOI:
  - **Low**: cross-loading = .20, error correlation = .05
  - **High**: cross-loading = .40, error correlation = .10
  - Expected classification

SESOI \ Misfit	None	Level 1	Level 2	Level 3
Low	Trivial	Trivial	Substantial	Substantial
High	Trivial	Trivial	Trivial	Substantial

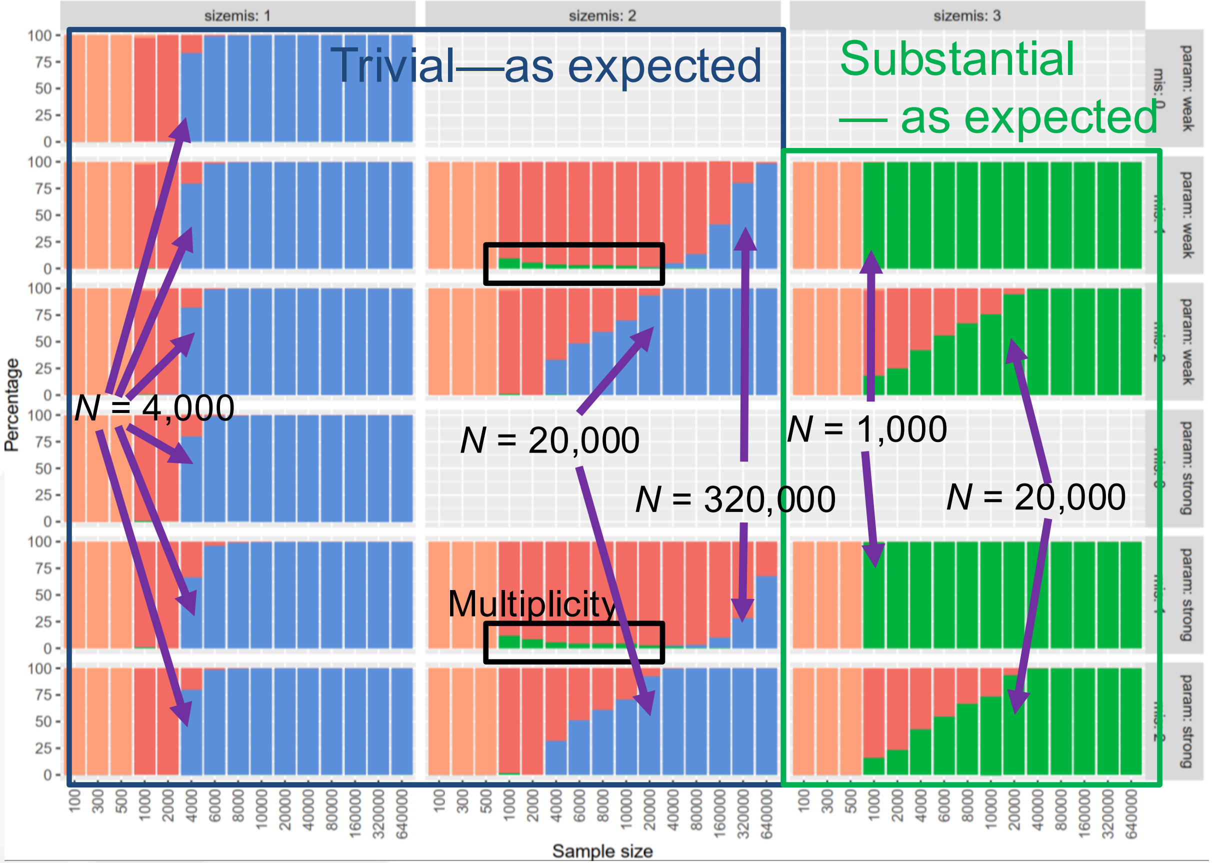
- 1,000 replications per condition

Low SESOI



- Support the hypothesis
- Large sample size are required.

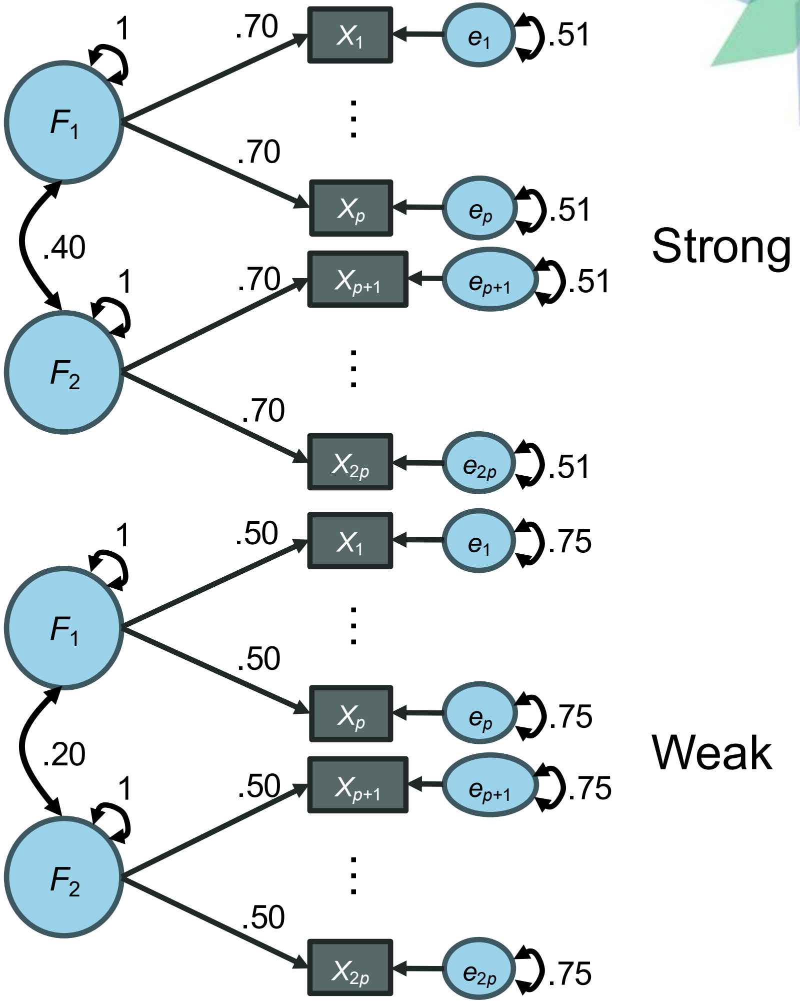
High SESOI

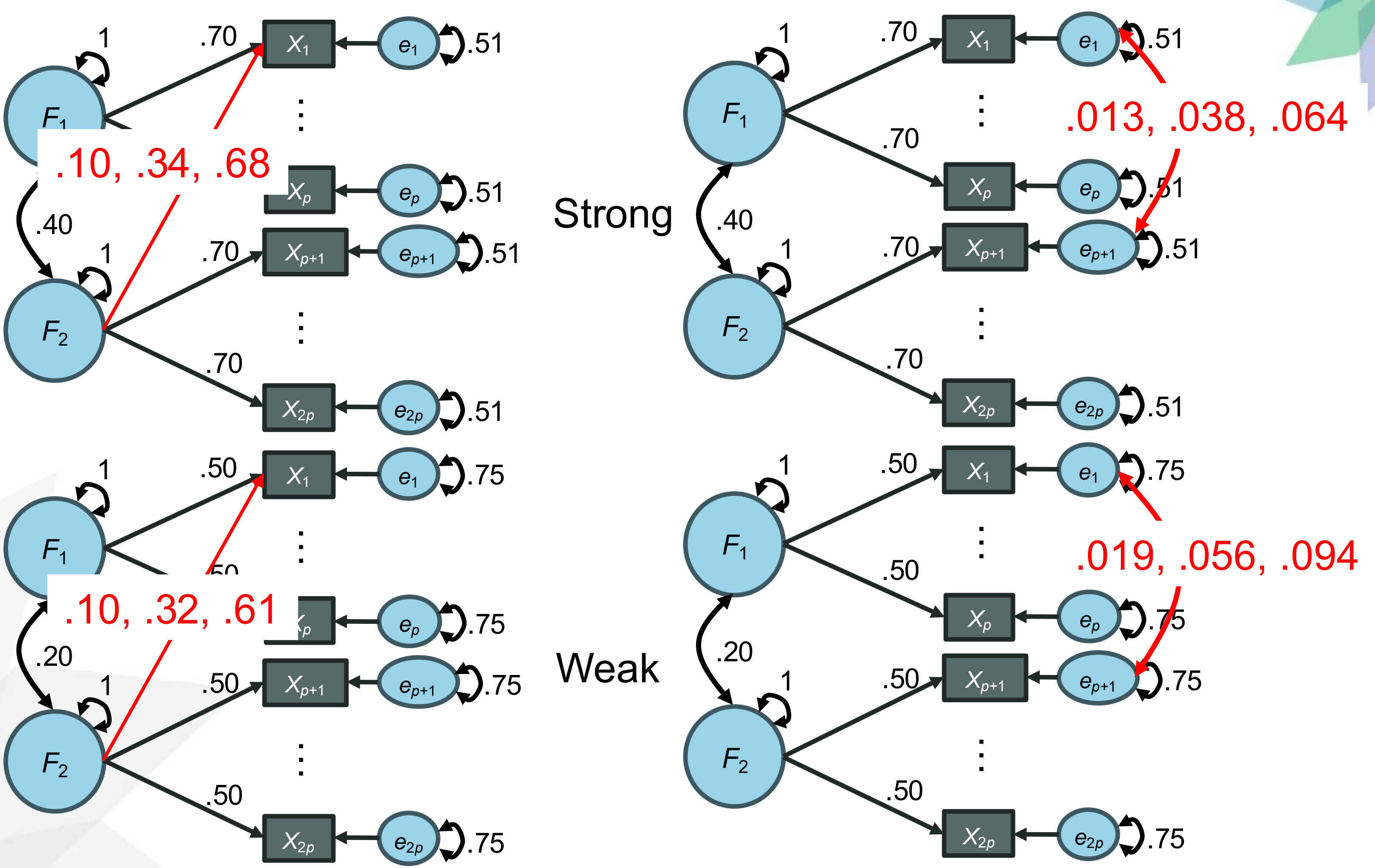


- Support the hypothesis
- No misspecification: smaller but still large, *N* required
- Level 1: smaller, but still large, *N*
- Level 2: requires very large *N*
- Level 3
  - larger *N* required for error misspecification
- Misspecification > Level 3: would require smaller *N*
- Increased probability of at least one substantial classification due to multiplicity

Simulation Study

- Study 2
  - Data: Two-factor CFA
  - 3, 4, 5, 10, or 15 items per factor
  - Sample size: N = 100 to 640000 (15 Levels)
  - Weak vs. strong parameter values
  - Type of misspecification
    - No misspecification
    - Standardized cross loadings of .10, .30, and .50
    - Error correlations of .025, .075, and .125





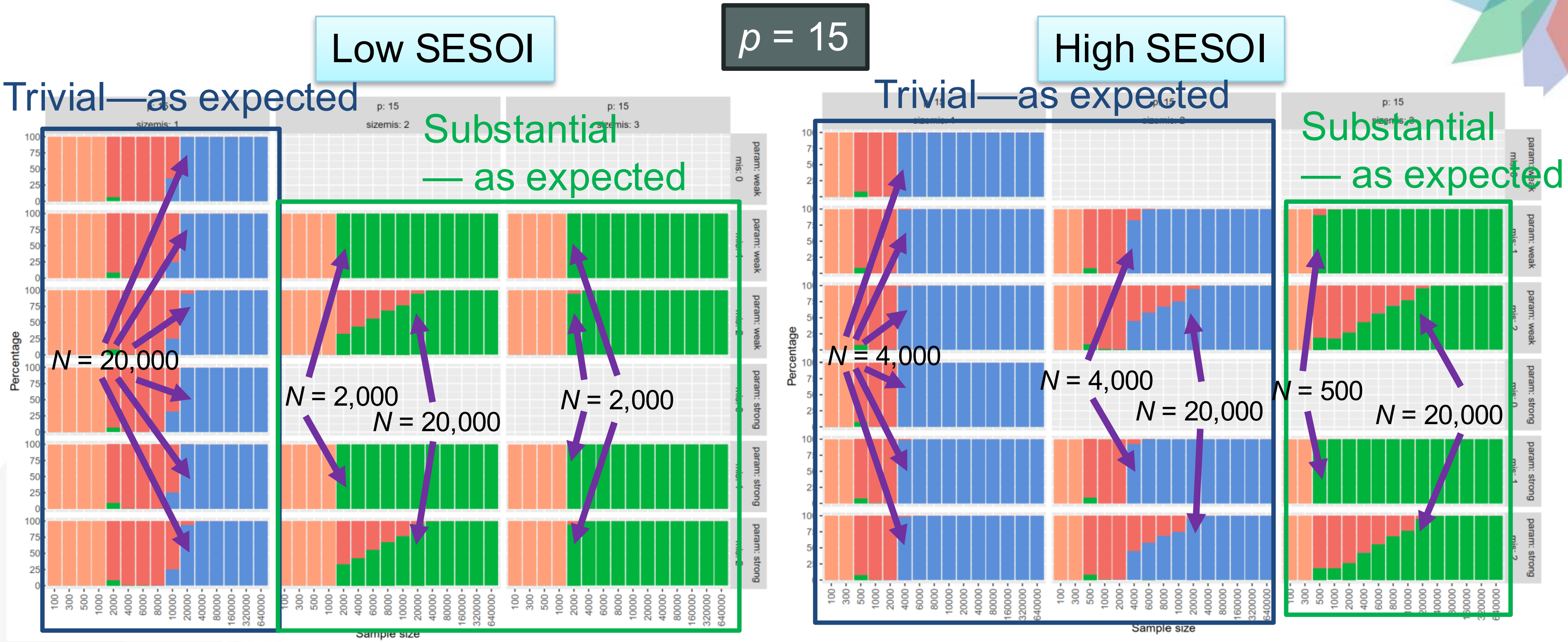
Strong
Weak
.10, .34, .68
.013, .038, .064
.10, .32, .61
.019, .056, .094
F1
F2
X1
Xp
Xp+1
X2p
e1
ep
ep+1
e2p
1
.70
.51
.40
.50
.75
.20

Simulation Study

- SESOI:
  - **Low**: cross-loading = .20, error correlation = .05
  - **High**: cross-loading = .40, error correlation = .10
  - Expected classification

SESOI \ Misfit	None	Level 1	Level 2	Level 3
Low	Trivial	Trivial	Substantial	Substantial
High	Trivial	Trivial	Trivial	Substantial

- 1,000 replications per condition

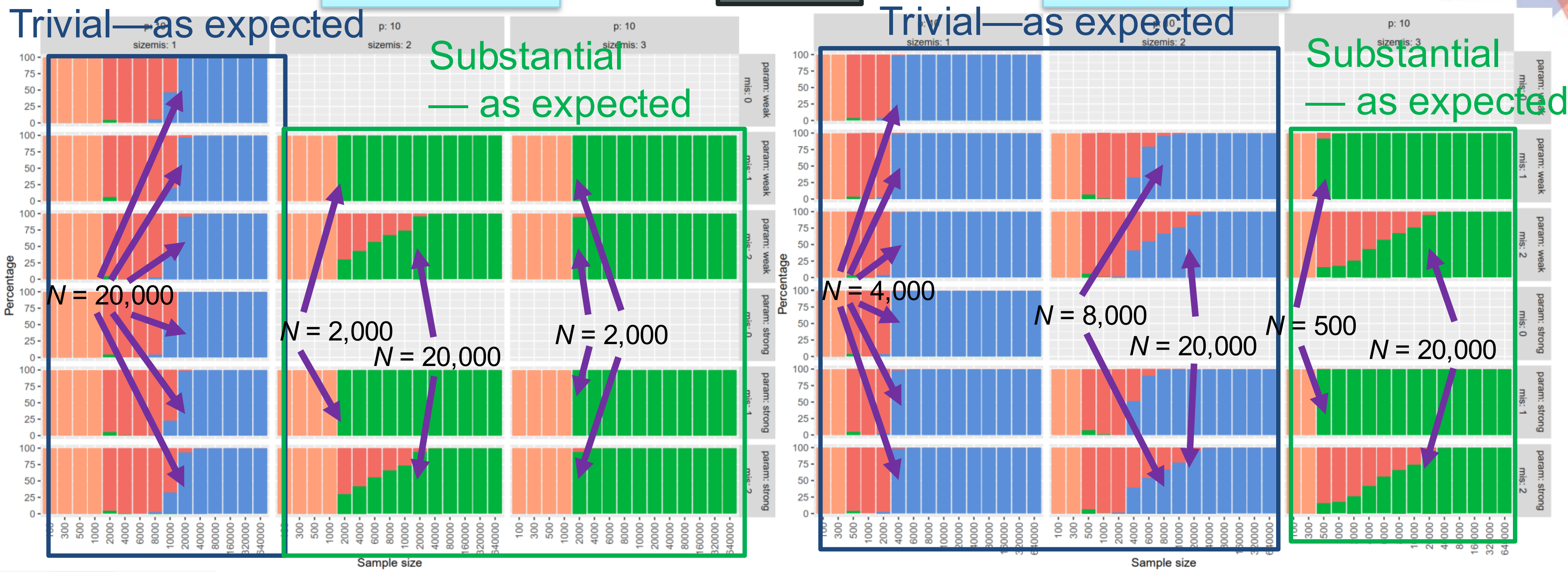


- Support the hypothesis
- Requires smaller, but still large, N

Low SESOI

p = 10

High SESOI

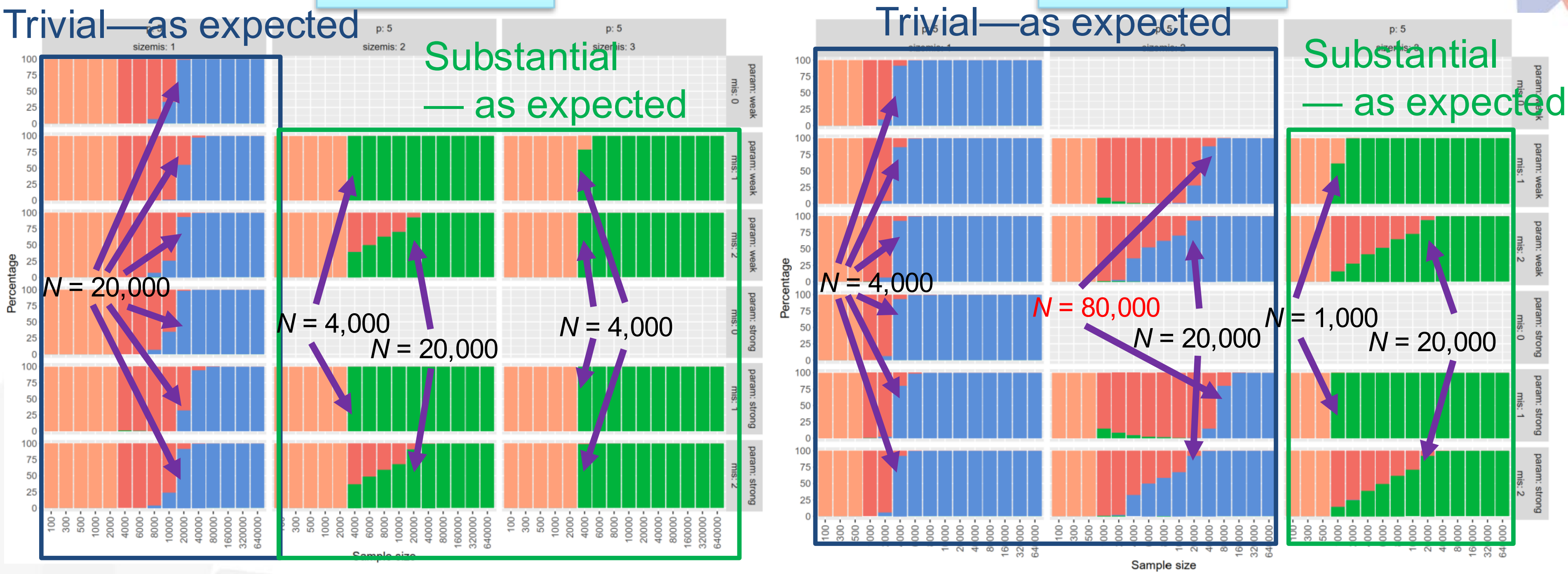


- Support the hypothesis
- Sample size requirement are comparable to those for p = 15

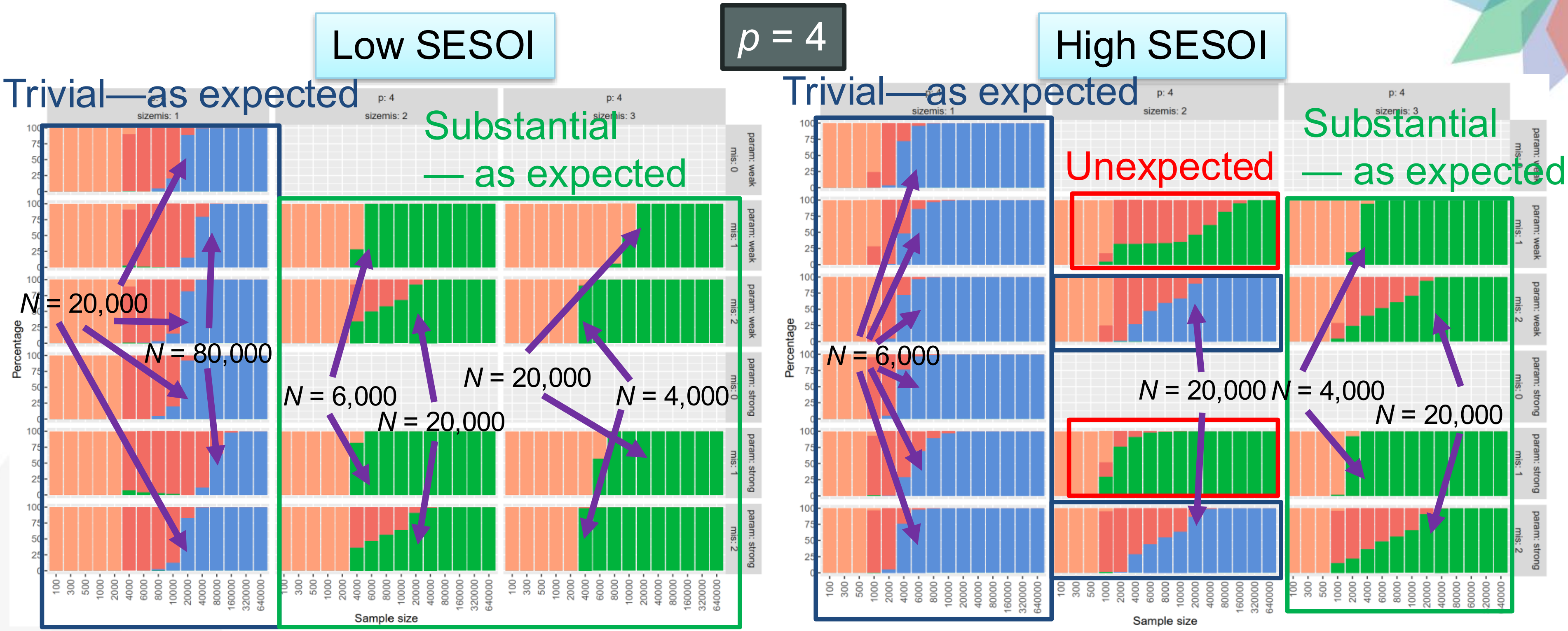
Low SESOI

p = 5

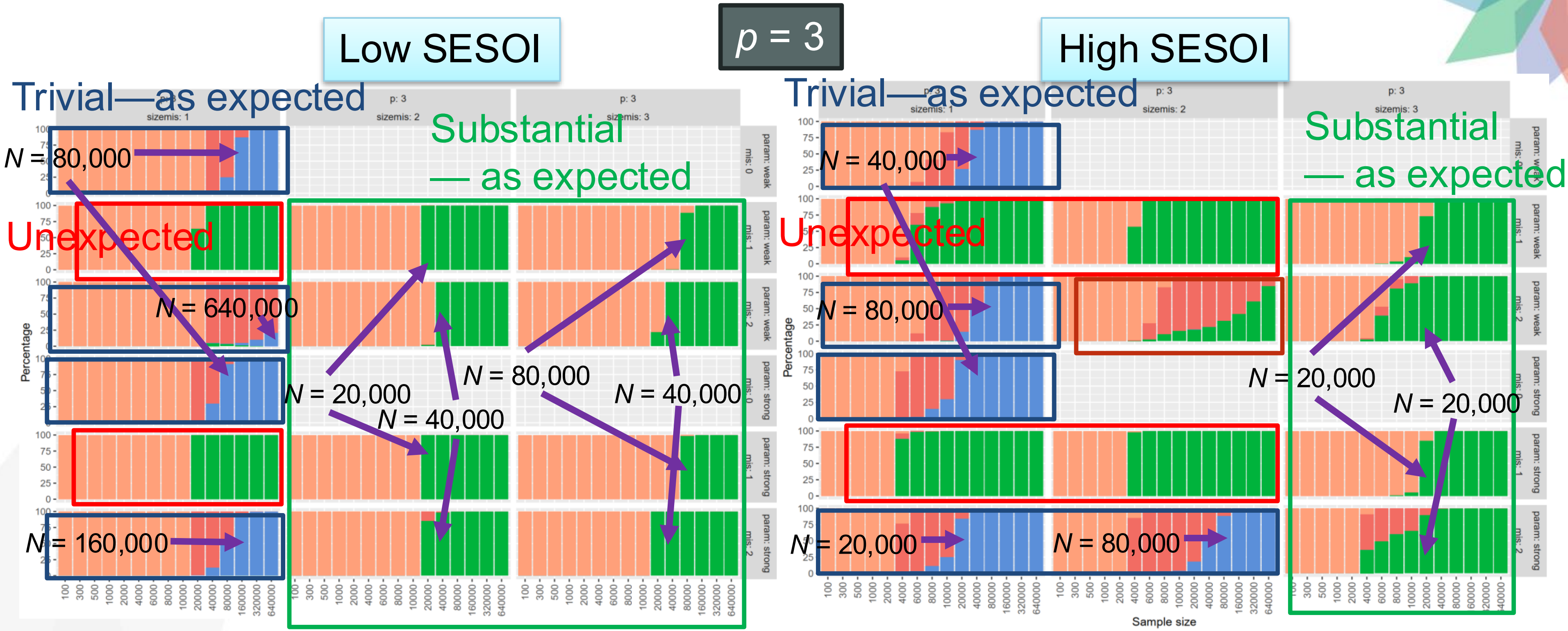
High SESOI



- Support the hypothesis
- Sample size requirements are comparable to those for p = 10 and p = 15 except for Level 2 cross-loading misspecification under high SESOI



- Partially support the hypothesis
- Sample size requirements are higher.
- Standardized cross-loadings of .30 led to SEPC > .40.
- SEPC inflation due to weak identification



- Partially support the hypothesis
- Sample size requirements are even higher.
- Standardized cross-loadings of .10 led to SEPC > .40.
- With low factor correlation, a misspecified error correlation of .075 is between Items 1 and 4 resulted in SEPCs for Items 2—3 and 5—6 exceeding .10.
- Instability arises from a weakly identified two-factor structure.

Simulation Study

- Study 3
  - Data: One-factor CFA with design features similar to Study 2
  - The absence of cross-factor contamination stabilizes SEPC behavior
  - Results are fully consistent with expectations

SESOI \ Misfit	None	Level 1	Level 2	Level 3
Low	Trivial	Trivial	Substantial	Substantial
High	Trivial	Trivial	Trivial	Substantial

Illustrative Example (Christopher et al., 2012)

- Study of reading and comprehension performance in children.
- **Data**: Twin study (Colorado Learning Disabilities Research Center)
  - One twin per pair included
  - Adjusted sample size = 4,000 (from original *N* = 265, ages 8–10).
- **SESOI**: Cross-loading = .40; Error correlation = .10.
- Implementation available in **semTools**
  - Function: `epcEquivFit()`

Christopher, M. E., Miyake, A., Keenan, J. M., Pennington, B., DeFries, J. C., Wadsworth, S. J., ... & Olson, R. K. (2012). Predicting word reading and comprehension with executive function and speed measures across development: a latent variable analysis. *Journal of Experimental Psychology: General*, *141*(3), 470-488.

Table 2

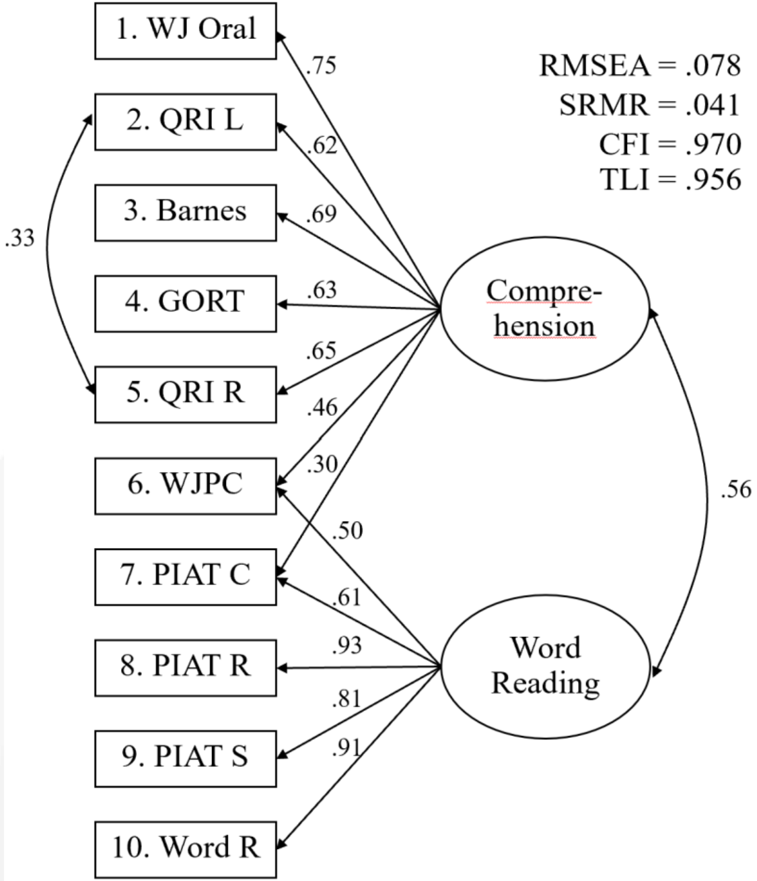
Summary of Zero-Order Correlations for Scores on the Reading and Listening Variables by Age Group With Estimates of Reliability on Diagonal

Variable	1	2	3	4	5	6	7	8	9	10
1. WJ Oral	**.62**[a]	.58	.40	.59	.56	.46	.56	.43	.32	.47
2. QRI L	.44	**.67**[a]	.56	.62	.74	.52	.61	.47	.35	.47
3. Barnes	.54	.44	**.59**[a]	.42	.49	.35	.36	.24	.15	.29
4. WJ PC	.58	.47	.45	**.71**[a]	.57	.52	.63	.59	.51	.57
5. QRI R	.44	.60	.48	.52	**.59**[a]	.53	.50	.39	.26	.38
6. GORT	.48	.40	.45	.47	.34	**.48**[a]	.50	.44	.30	.39
7. PIAT C	.49	.40	.39	.66	.51	.42	**.79**[a]	.63	.50	.62
8. PIAT R	.42	.32	.27	.68	.42	.34	.74	**.82**[a]	.66	.83
9. PIAT S	.36	.27	.17	.61	.36	.33	.66	.75	**.67**[a]	.67
10. Word R	.42	.33	.25	.72	.40	.33	.69	.85	.73	**.85**[a]

Note. All correlations signficant at $p < .05$, with correlations greater than .15 significant at $p < .01$. Variables standardized within age groups; Ages 8–10 located below diagonal; Ages 11–16 above diagonal; WJ Oral = Woodcock-Johnson (Woodcock et al., 2001) oral comprehension; QRI L = Qualitative Reading Inventory (Leslie & Caldwell, 2001) mean listening question score (standardized within QRI level); Barnes = Barnes KNOW-IT (Barnes & Dennis, 1996; Barnes et al., 1996) average of coherence inference, elaborative inference, and literal proportions; WJ PC = Woodcock-Johnson passage comprehension; QRI R = Qualitative Reading Inventory mean reading question score (standardized within QRI level); GORT = Gray Oral Reading Test-3 (Wiederholt & Bryant, 1992); PIAT C = Peabody Individual Achievement Test (Markwardt, 1970) comprehension; PIAT R = PIAT reading recognition; PIAT S = PIAT spelling; Word R = time-limited oral reading of single words (Olson et al., 1994).

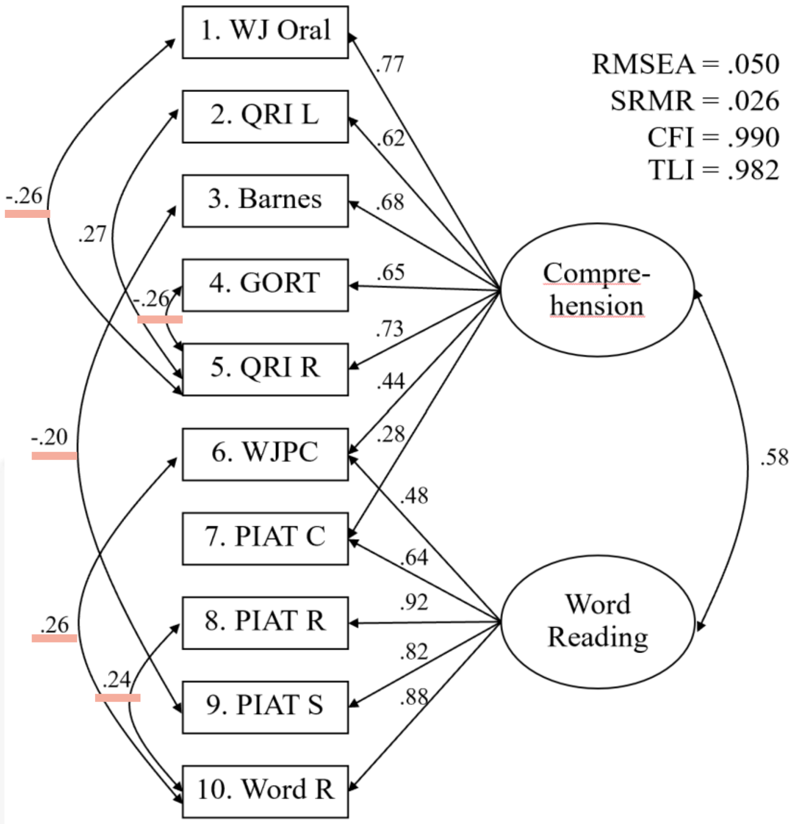
[a] Reliabilties estimated from monozygotic twin partial correlations (n = 144 twin pairs, controlling for age). Monozygotic twin correlations can be used as low-bound estimates of reliablitiy. The twins share their genes and their family environment, meaning any within-pair differences in performance are due to nonshared environmental influences including measurement error.

The lower-diagonal elements (ages 8-10) were used. The sample size is adjusted to N = 4,000.



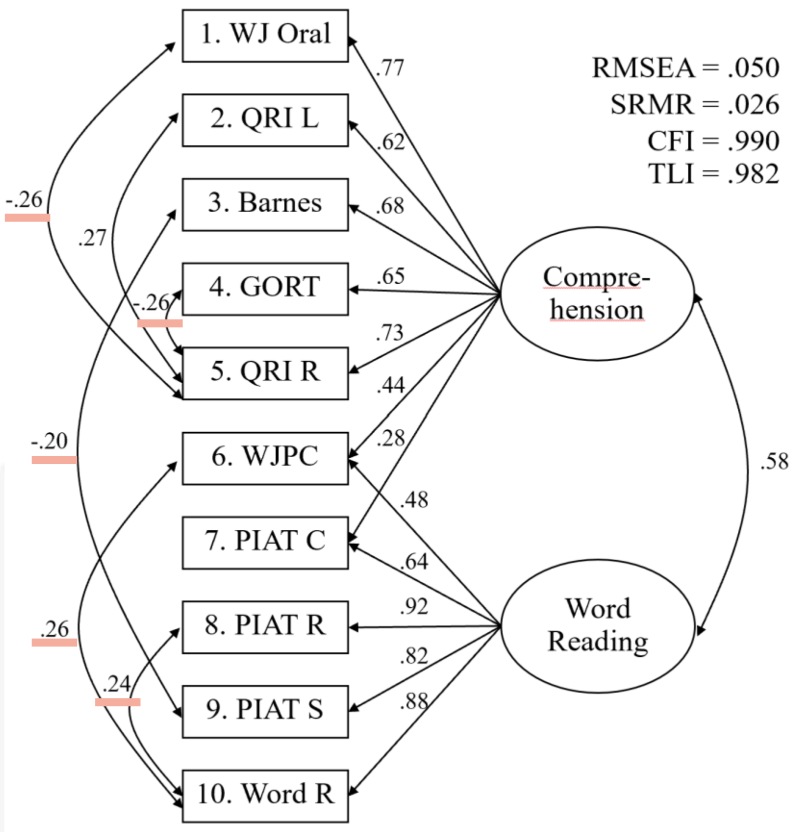
Substantially Misspecified

Inconclusive	Substantial	Trivial	Underpowered
10	7	35	0



Inconclusive

Inconclusive	Substantial	Trivial	Underpowered
11	0	36	0



lhs	op	rhs	std.epc	decision.ci
WJORAL	~~	QRLL	-.118	I
WJORAL	~~	Barnes	.082	I
WJORAL	~~	GORT	-.064	I
QRLL	~~	WJPC	.073	I
GORT	~~	PIATS	.076	I
QRIR	~~	WJPC	-.126	I
QRIR	~~	PIATC	.077	I
PIATC	~~	PIATR	.106	I
PIATC	~~	WordR	-.077	I
PIATR	~~	PIATS	-.130	I
PIATS	~~	WordR	.072	I

SESOI: Cross-Loading = .40;
Error Correlation = .10

Inconclusive	Trivial
11	36

SESOI: Cross-Loading = .40;
Error Correlation = .19

Inconclusive	Trivial
0	47

- The model fits the data if cross-loadings of .40 and error correlations of .19 are considered trivial.
- With a stricter SESOI (e.g., error correlation = .10), model fit remains inconclusive.

Conclusion

- Equivalence testing of EPCs provides a more consistent framework for evaluating model fit.
- Performs well in well-identified CFA models but becomes **unstable in weakly identified models** (e.g., three indicators per factor).
- **Requires large sample sizes**, especially under strict SESOI
- **Main advantage: Results are interpretable in substantive effect-size terms** (e.g., cross-loadings, error correlations) rather than abstract fit indices.
- Illustrative example demonstrates how equivalence testing clarifies the practical meaning of misfit.

Thank you.

Questions?

Slides & Analysis Code

https://github.com/psunthud/epc-equivalence-sem