Presented by
Sunthud Pornprasertmanit

Sample Size in Factor Analysis

MacCallum, Widaman, Zhang, & Hong (1999)

Outline

- Misconception in Sample Size Estimation of Factor Analysis (FA)
- Mathematical theory of sample size impacts
- Characteristics of FA that affect desired sample size
- Simulation Study
- Guideline of Sample Size Estimation
- Comments + Future Research

Rules of Thumb

- Minimum Number of Sample Size (*N*)
- Minimum ratio of *N* to Number of variables (*p*)
- Misconception: this rule is invariant across studies

Statistical Theory: Errors

- Model Error
  - Introduce lack of fit of the model in the population and the sample
  - Increase sample size does not help
- Sampling Error
  - Introduce inaccuracy and variability in parameter estimates
  - Increase sample size does help
- This study assumes no model error

Statistical Theory

$$\mathbf{y} = \mathbf{\Lambda x_c} + \mathbf{\Theta x_u}$$

$\mathbf{y}$	a random row vector of scores on *p* measured variables
$\mathbf{x_c}$	a row vector of scores on *r* common factors (variances = 1)
$\mathbf{x_u}$	a row vector of scores on *p* unique factors (variances = 1)
$\mathbf{\Lambda}$	population common factor loadings of *p* measured variables from *r* common factors
$\mathbf{\Theta}$	Diagonal matrix of unique factor loadings

Statistical Theory

$$\Sigma_{yy} = \Lambda\Sigma_{cc}\Lambda' + \Lambda\Sigma_{cu}\Theta' + \Theta\Sigma_{uc}\Lambda' + \Theta\Sigma_{uu}\Theta$$

Σ_{yy}	Covariance matrix of measured variables
Σ_{cc}	Covariance (correlation) matrix among common factor scores
Σ_{uu}	Covariance (correlation) matrix among unique factor scores
Σ_{cu}	Covariance (correlation) matrix between common and unique factor scores

Statistical Theory

- If hypothesized model is true,
  - True Sample Λ = Population Λ
  - True Sample Θ = Population Θ
  - $\mathbf{C}_{cc}$ not equal to $\mathbf{\Sigma}_{cc}$
  - $\mathbf{C}_{uc}$ not equal to $\mathbf{\Sigma}_{uc}$ (not zero matrix)
  - $\mathbf{C}_{uu}$ not equal to $\mathbf{\Sigma}_{uu}$ (not diagonal matrix)
  - $\mathbf{C}_{yy}$ not equal to $\mathbf{\Sigma}_{yy}$

Statistical Theory

- If hypothesized model is true,
  - True Sample Λ = Population Λ
  - True Sample Θ = Population Θ
  - C_{cc} not equal to Σ_{cc}
  - C_{uc} not equal to Σ_{uc}
  - C_{uu} not equal to Σ_{uu}
  - C_{yy} not equal to Σ_{yy}

No Sampling Error

Sampling Error

- Sampling error in Σ_{yy} is come from sampling error in Σ_{cc}, Σ_{uc}, and Σ_{uu}

Statistical Theory

Population

$$\Sigma_{yy} = \Lambda\Sigma_{cc}\Lambda' + \Lambda\Sigma_{cu}\Theta' + \Theta\Sigma_{uc}\Lambda' + \Theta\Sigma_{uu}\Theta$$

Sample

$$C_{yy} = \Lambda C_{cc}\Lambda' + \Lambda C_{cu}\Theta' + \Theta C_{uc}\Lambda' + \Theta C_{uu}\Theta'$$

Statistical Theory

Population

$$\Sigma_{yy} = \Lambda\Sigma_{cc}\Lambda' + \Lambda\Sigma_{cu}\Theta' + \Theta\Sigma_{uc}\Lambda' + \Theta\Sigma_{uu}\Theta$$

$$\Sigma_{cc} = \Phi \qquad \Sigma_{uu} = I \qquad \left(\Sigma_{uc} = \Sigma_{cu}\right) = 0$$

Sample

$$C_{yy} = \Lambda C_{cc}\Lambda' + \Lambda C_{cu}\Theta' + \Theta C_{uc}\Lambda' + \Theta C_{uu}\Theta'$$

$$C_{cc} \neq \Phi \qquad C_{uu} \neq I \qquad \left(C_{uc} = C_{cu}\right) \neq 0$$

Statistical Theory

Population

$$\Sigma_{yy} = \Lambda\Phi\Lambda' + \Theta^2$$

Sample

Contribute to Sampling Error

$$C_{yy} = \Lambda C_{cc}\Lambda' + \boxed{\Lambda C_{cu}\Theta' + \Theta C_{uc}\Lambda' + \Theta C_{uu}\Theta'}$$

$$C_{uu} \neq I \qquad \left(C_{uc} = C_{cu}\right) \neq 0$$

Statistical Theory

- How sampling error in Σ_{uu} and Σ_{uc}, make estimated sample Λ and true sample Λ different

$\mathbf{C}_{uu}$ and $\mathbf{C}_{uc}$ are not $\mathbf{I}$ and $\mathbf{0}$.

+

Model is constrained Σ_{uu} and Σ_{uc} equal to $\mathbf{I}$ and $\mathbf{0}$.

Introduce biases in Λ

Factors Affecting Λ Estimation

- Sample Size
- Communalities
- Overdetermination: Ratio of Number of Indicators to Number of Factors (p:r ratio)

Factors Affecting Λ Estimation

Smaller Sample Size

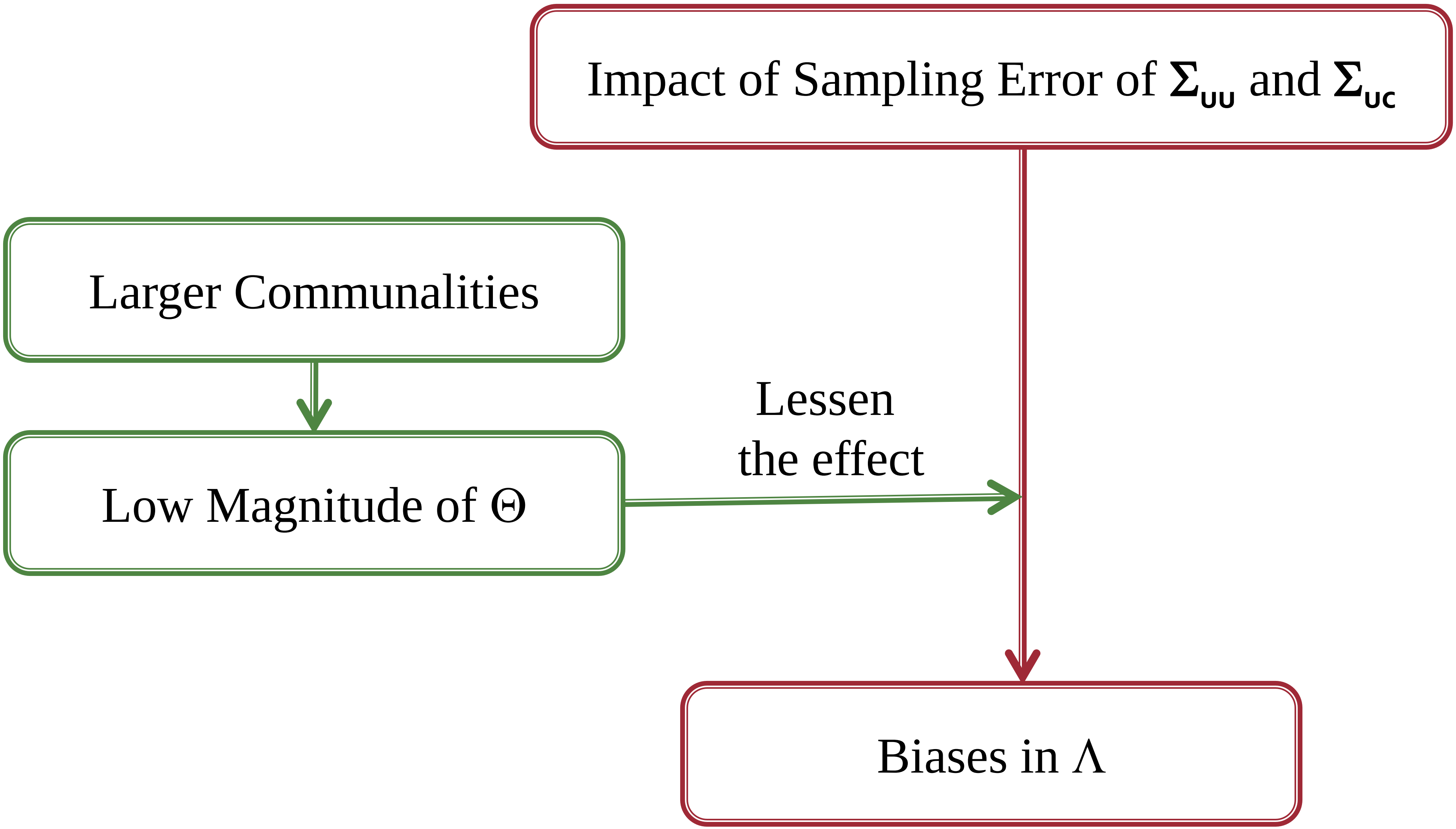
↓

$\mathbf{C}_{uu}$ and $\mathbf{C}_{uc}$ are more varied from **I** and **0**.

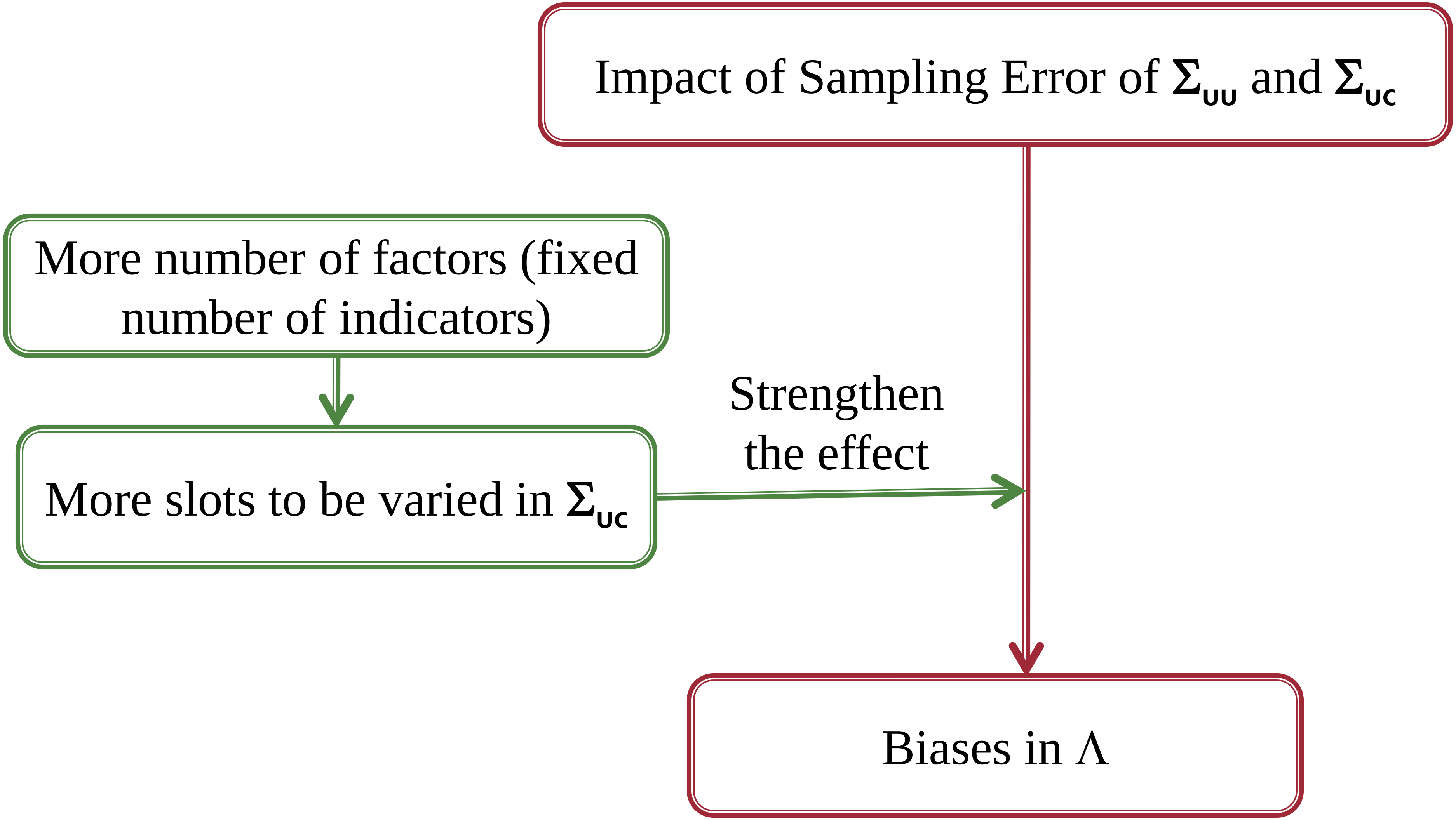
↓

Larger biases in Λ

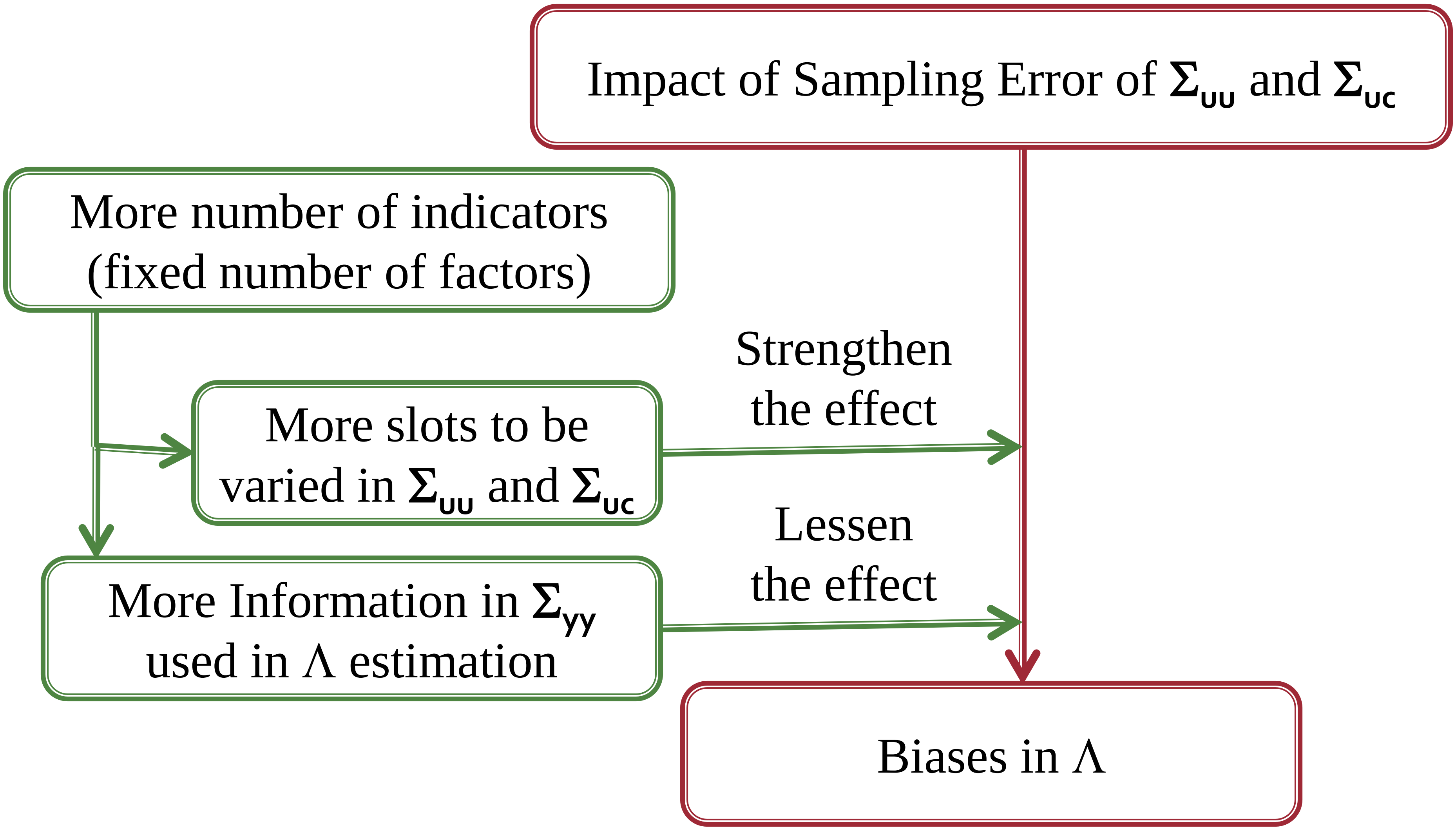
Factors Affecting Λ Estimation



Factors Affecting Λ Estimation



Factors Affecting Λ Estimation



Summary of Hypotheses

- Common factor loadings will be more accurately recovered when
  - *N* increases
  - Communalities increases
  - Overdetermination improves
- Large communalities will reduce impact of both *N* and overdetermination.

Simulation Study

- Investigate impact of *N*, communalities, and overdetermination of common factor loading recovery.

Simulation Study

- 36 Conditions
  - Sample Size: 60, 100, 200, and 400
  - Communalities
    - High: .6, .7, and.8
    - Wide: range of .2 to .8
    - Low .2, .3, and .4
  - # of indicators to # of factors: 10/3, 20/3, 20/7
- 100 replications on each condition

Simulation Study

Population Correlation Matrices

↓

Sample Correlation Matrices

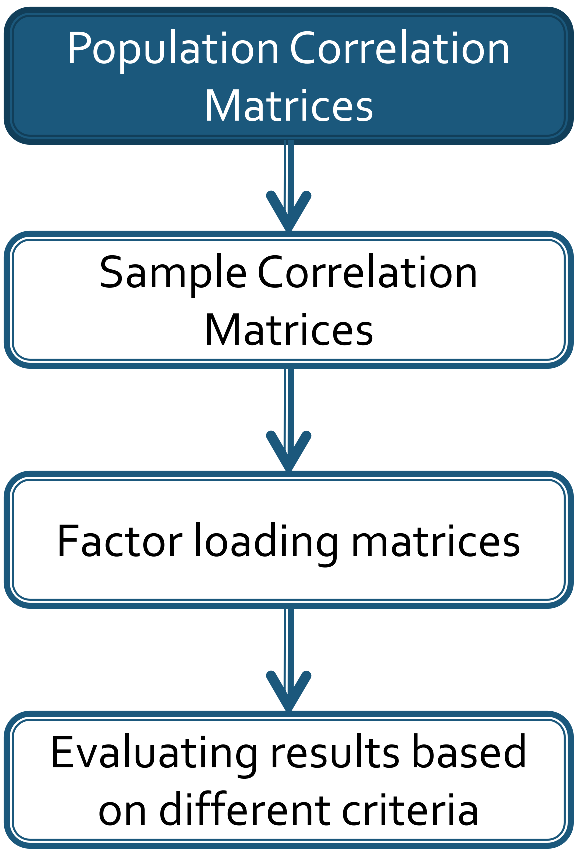
↓

Factor loading matrices

↓

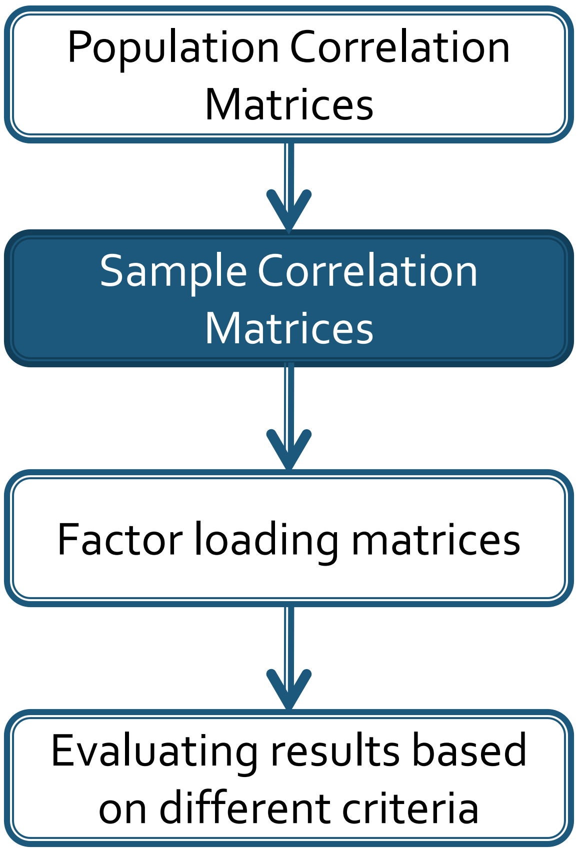
Evaluating results based on different criteria

Simulation Study



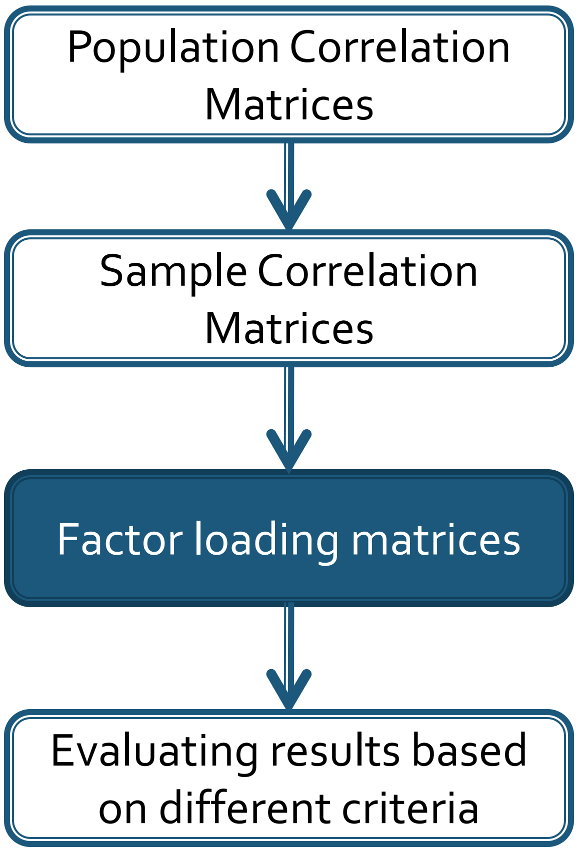
- Population correlation matrices are created based on
  - Communalities
  - Indicators to factor ratio
- Thus, nine population correlation matrices were used
- Clear Simple Structure
- Equal # of indicators per factor

Simulation Study



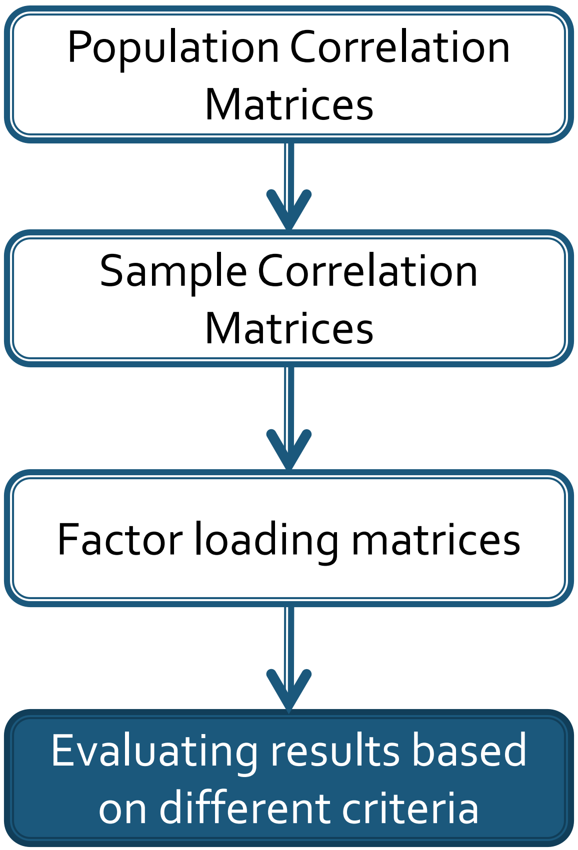
- Multivariate normal data with *N* observations were created
- Find sample correlation matrices from the data
- 100 replications per *N* and population correlation matrix

Simulation Study



- Sample correlation matrices were analyzed by ML with pre-specified number of factors
- Negative variance result will be dropped
- Quartimin Rotation
- Population correlation matrices were also analyzed similarly

Simulation Study



- Average congruence between sample and population factor loadings (average correlation)
  - The closer to 1, the better
- Variability of sample factor loadings across replications
  - The smaller, the better

Simulation Study

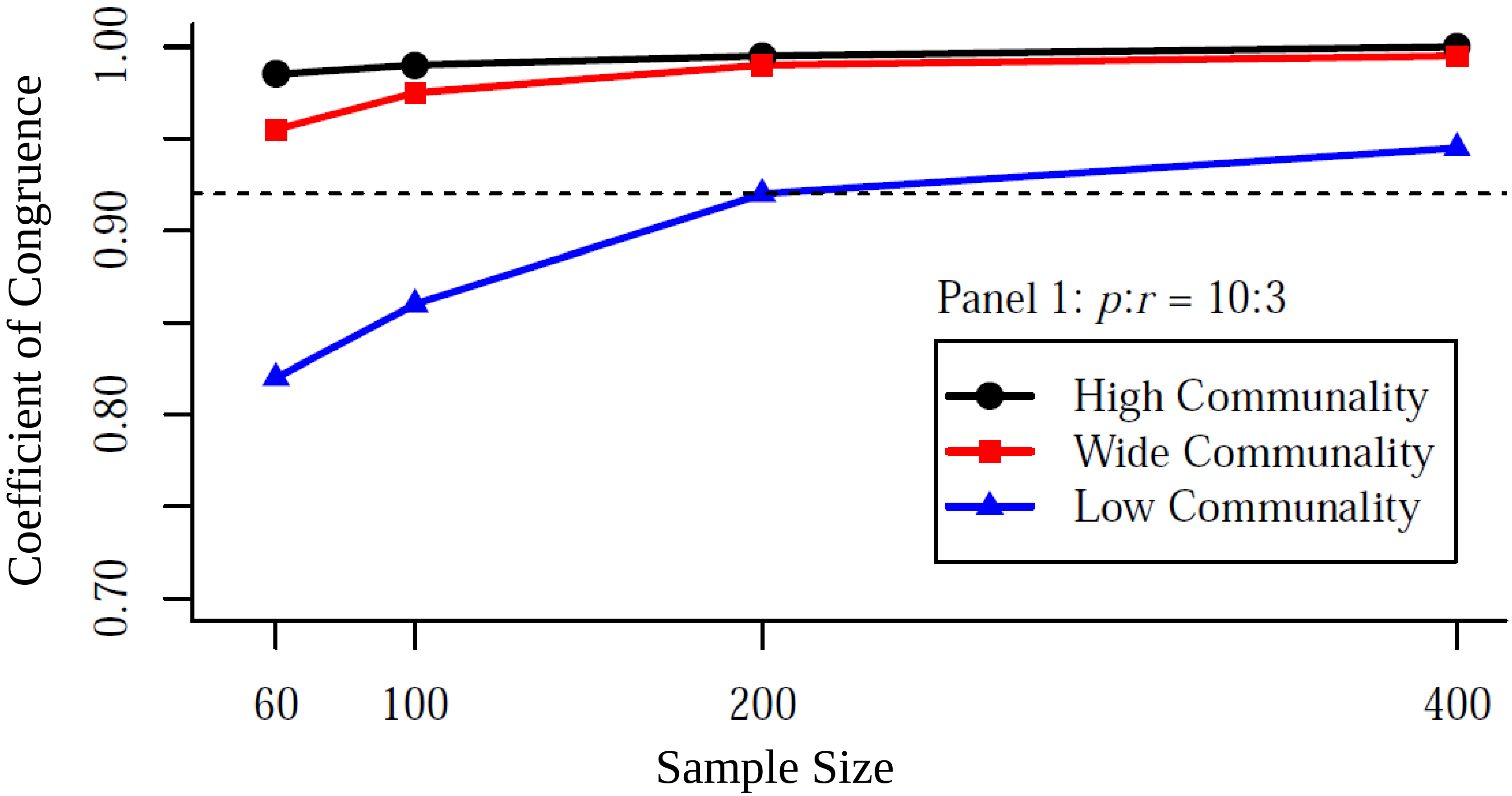
- ANOVA Results: Coefficient of congruence

Source	ω^2
Sample Size (*N*)	.15
Communality (*h*)	.41
Overdetermination (*d*)	.11
N x *h*	.08
N x *d*	.01
H x *d*	.05
N x *h* x *d*	.00

All sources were significant at .001

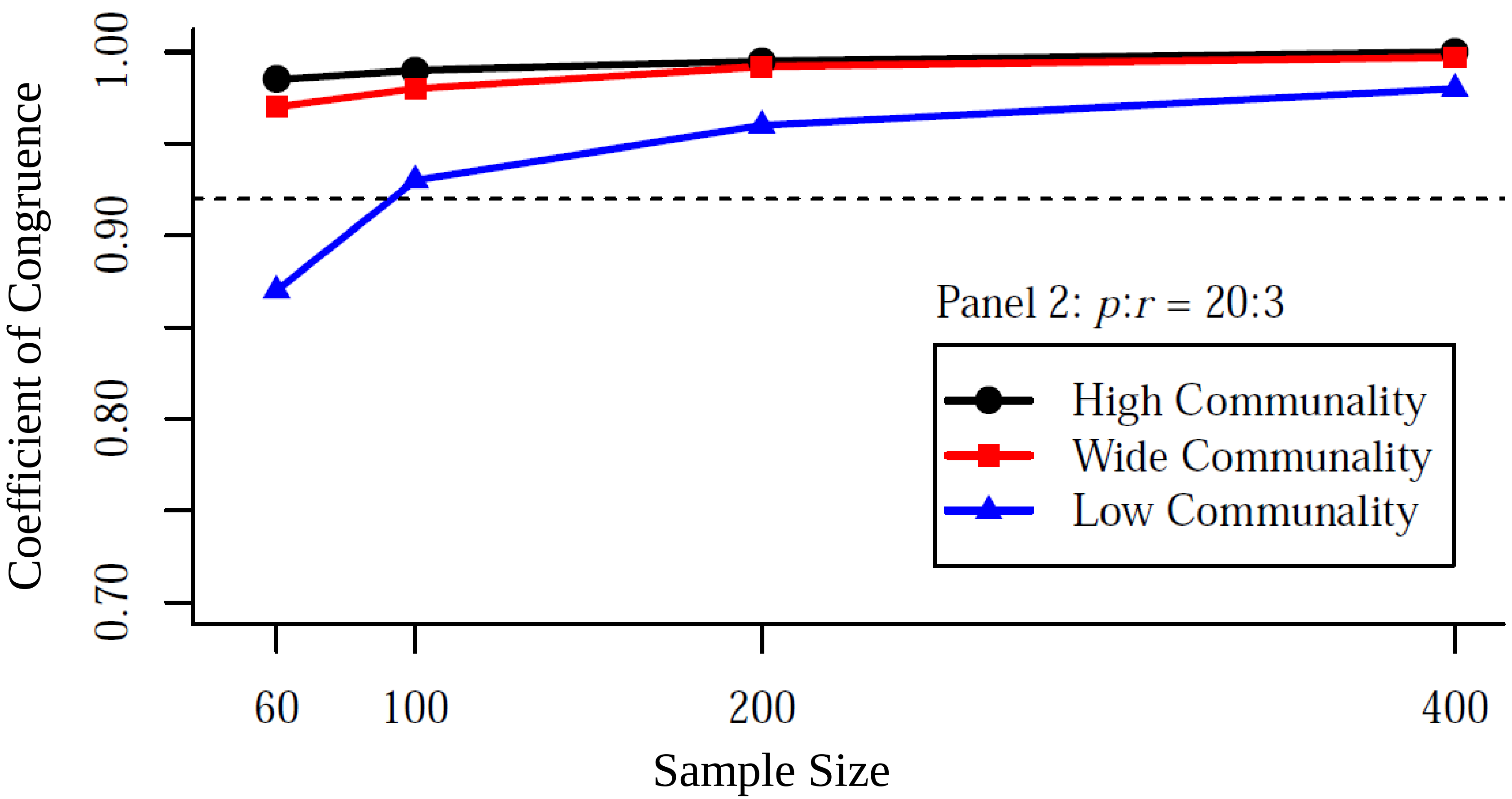
Simulation Study

- 10 indicators with 3 factors



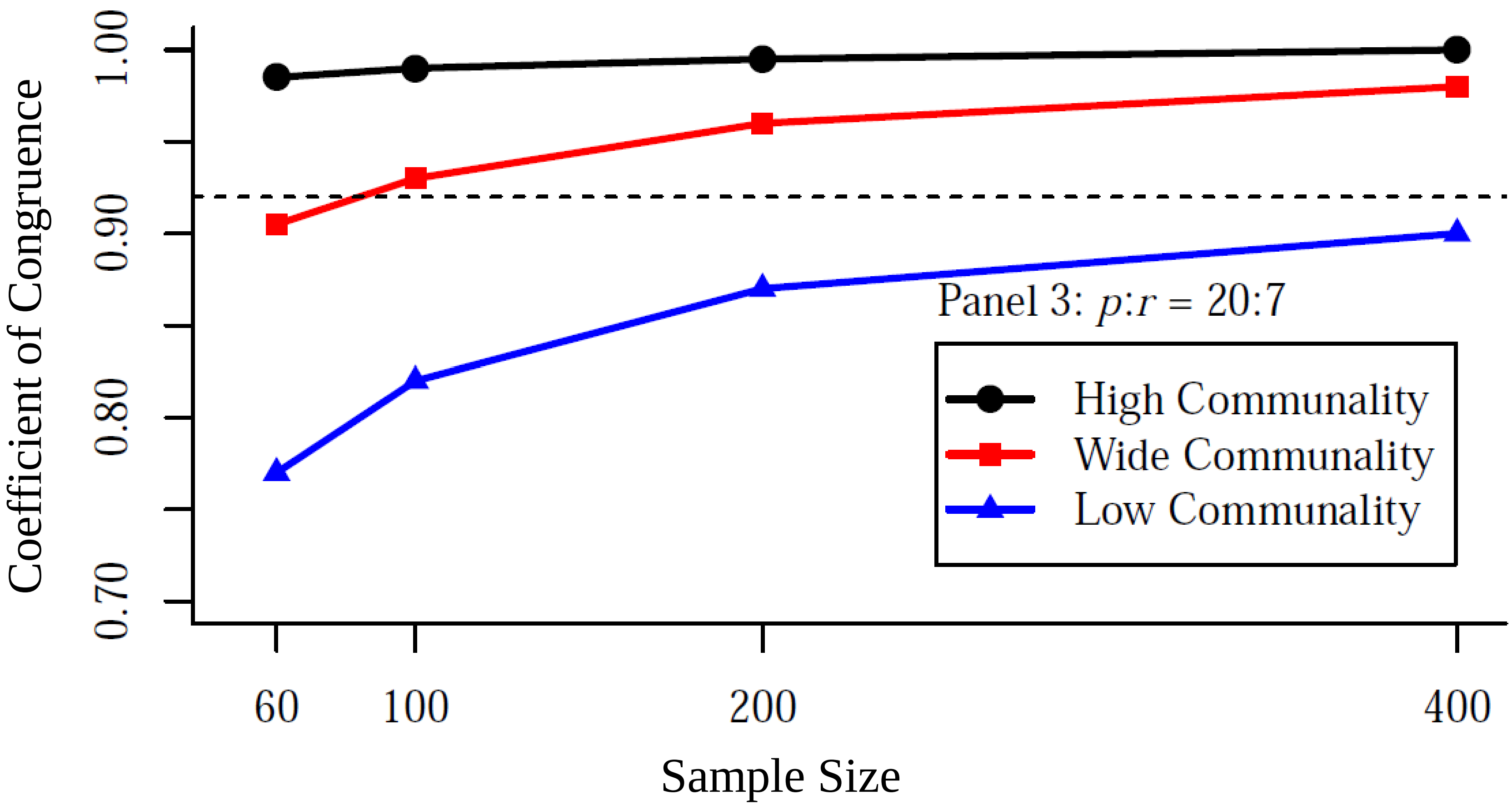
Simulation Study

- 20 indicators with 3 factors



Simulation Study

- 20 indicators with 7 factors



Simulation Study

- High communalities: Sample size really does not matter
- Low communalities: Sample size is crucial
- Low communalities + Low *p* to *r* ratio: Large sample size still have bad results
- The graphs of variability were similar to those of coefficient of congruence.

Conclusion

- Rules of thumb are not valid
- Sample size determination should consider from expected results (communalities, number of factors)
- High communalities → 100 is enough
- Low communalities → Large number
- Write more than three items per factor or write very good items

Comments

- Parameter recovery and variability are not the only desired properties in determining sample size
- Parameter model with or without model error provide the same results (MacCallum, Widaman, Preacher, & Hong, 2001)
- Sample size guideline is not useful for categorical indicators

Future Research

- Categorical indicators
- Other criteria
  - Accuracy of determining number of factors
  - Accuracy in high loading of each indicator