INSTITUTE FOR DEFENSE ANALYSES

The Effect of Extremes in Small Sample Size on Simple Mixed Models: A Comparison of Level-1 and Level-2 Size

Jane Pinelis, *Project Leader*

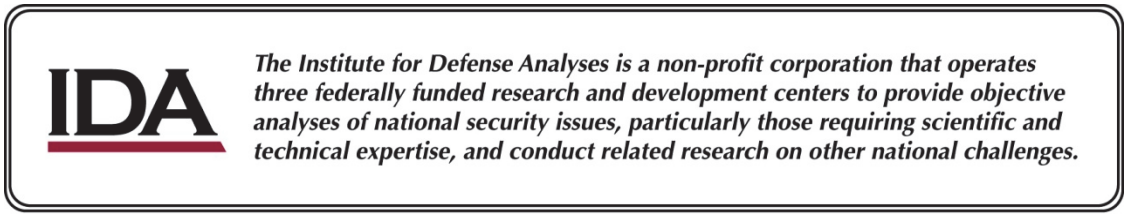
Kristina A. Carter
Heather M. Wojton

February 26, 2018

IDA Non-Standard Document
NS D-8965

Log: H 2018-000065

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882



About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under Central Research Project C9082, Statistics Working Group. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
Technical review was performed by Stephanie Lane of the Operational Evaluation Division.

For more information:
Jane Pinelis, Project Leader
ypinelis@ida.org • 703-845-6899

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

INSTITUTE FOR DEFENSE ANALYSES

IDA Non-Standard Document NS D-8965

The Effect of Extremes in Small Sample Size on Simple Mixed Models: A Comparison of Level-1 and Level-2 Size

Jane Pinelis, *Project Leader*

Kristina A. Carter
Heather M. Wojton

Executive Summary

Mixed models are ideally suited to analyzing nested data from within-person designs – designs that are advantageous in applied research. Mixed models have the advantage of enabling the modeling of random effects, facilitating an accounting of the intra-person variation captured by multiple observations of the same participants, and suggesting further lines of control to the researcher. However, the sampling requirements for mixed models are prohibitive for other areas that could greatly benefit from them.

This simulation study examines the impact of small sample sizes (in both levels of the model) on the fixed effects bias, type I error, and power of a simple mixed-model analysis. Despite the need for adjustments to control for type I error inflation, findings indicate that smaller samples than previously recognized can be used for mixed models under certain conditions prevalent in applied research. Examination of the marginal benefit of increases in sample subject and observation size provides applied researchers with guidance for developing mixed-model repeated measure designs that maximize power.

The Effect of Extremes in Small Sample Size on Simple Mixed Models: A Comparison of Level-1 and Level-2 Size

Kristina A. Carter

Heather M. Wojton

Institute for Defense Analyses

March 2018

Background & Overview

- **Operational testing**
- **Operational performance = *f*(operator, system)**
- **Mixed model analysis**
  - Addresses some challenges
  - Raises others
- **Quantify challenges**
  - Can mixed models be used in operational settings where sample sizes are small?
- **Provide recommendations**

Scenario

- **Testing: Gathering information**
- **Evaluating: Drawing conclusions**

Drawing Conclusions

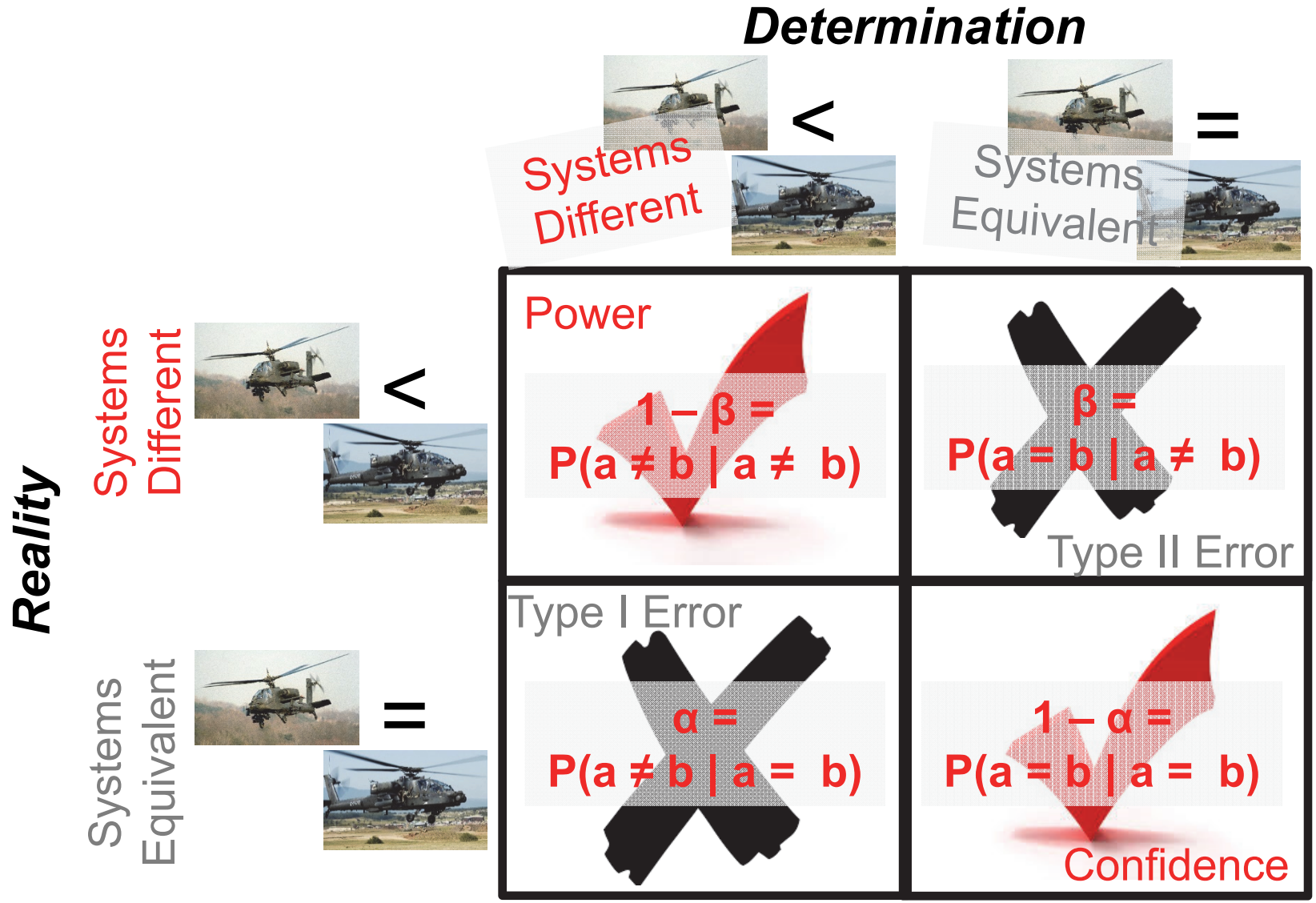
Systems Different

<

Systems Equivalent

=

Drawing Conclusions



Power

Additional factors affecting power:

Positively related to Power

- **Acceptable risk level, α**
  α = P(a ≠ b | a = b)
  risk of making a Type I Error
- **Magnitude of the effect (SNR), δ/σ**

$$\frac{\overline{X}_{\text{  }} - \overline{X}_{\text{  }}}{\sigma_{pooled}}$$

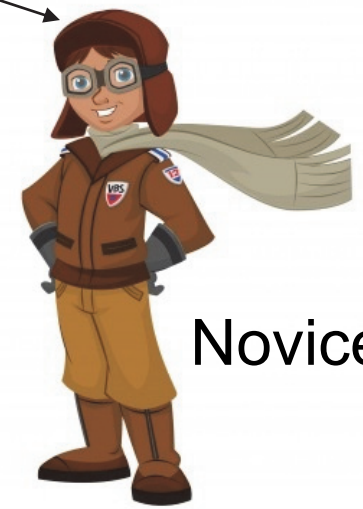
- **Size of the sample, *N***

Repeated Measures Design

Legacy System
- Speed
- Accuracy
- Perceived ease of use

$$\frac{\overline{X}_{\text{  }} - \overline{X}_{\text{  }}}{\sigma_{pooled}}$$

Veteran



Novice

New System
- Speed
- Accuracy
- Perceived ease of use

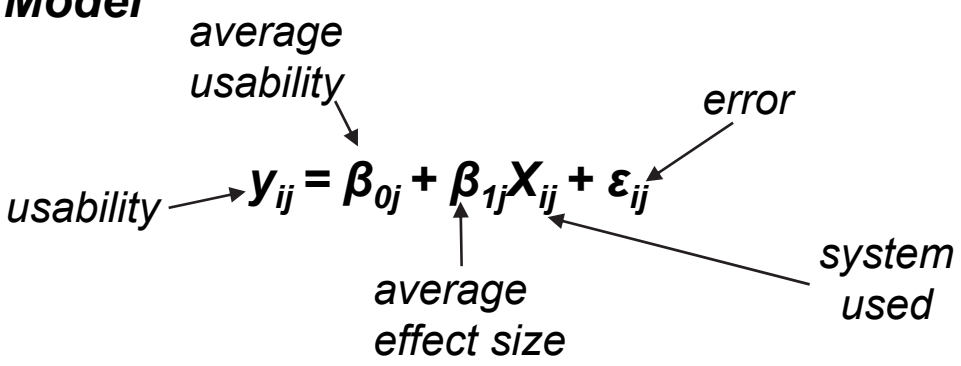
Repeated Observations

Legacy System
- Speed (5 times)
- Accuracy (5 times)
- Perceived ease of use (5 times)

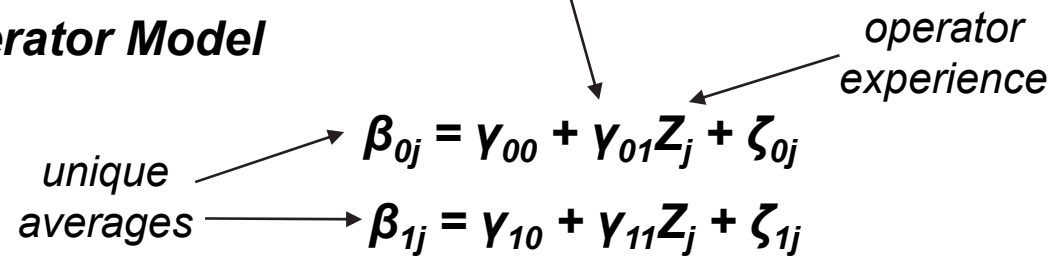
New System
- Speed
- Accuracy
- Perceived ease of use

Mixed Models for Repeated Measures

System Model



Operator Model



Mixed Model

$$y_{ij} = \gamma_{00} + \gamma_{01}Z_j + \gamma_{10}X_{ij} + \gamma_{11}Z_jX_{ij} + \zeta_{0j} + \zeta_{1j}X_{ij} + \varepsilon_{ij}$$

Mixed Models for Repeated Measures

Mixed Model

$$y_{ij} = \gamma_{00} + \gamma_{01} Z_j + \gamma_{10} X_{ij} + \gamma_{11} Z_j X_{ij} + \zeta_{0j} + \zeta_{1j} X_{ij} + \varepsilon_{ij}$$

Intraclass Correlation (ICC)

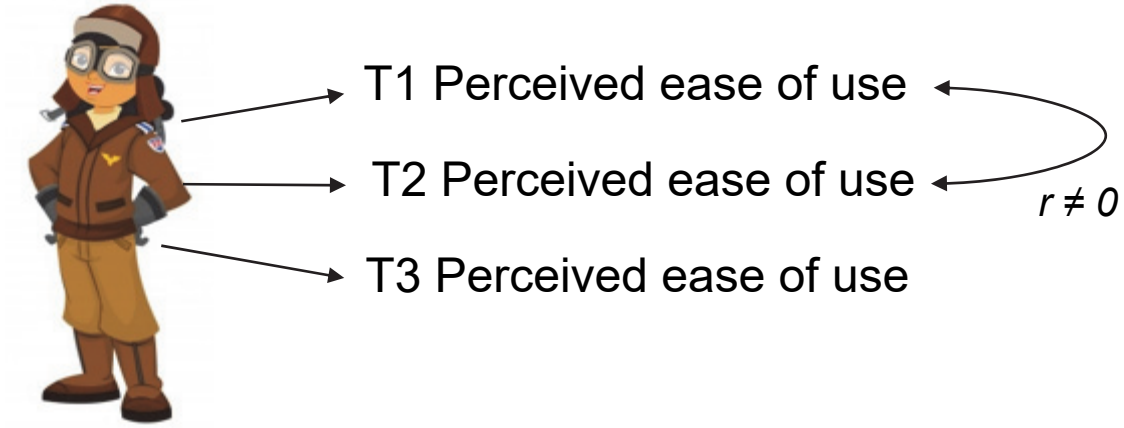
$$ICC = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}$$

level-1 variance

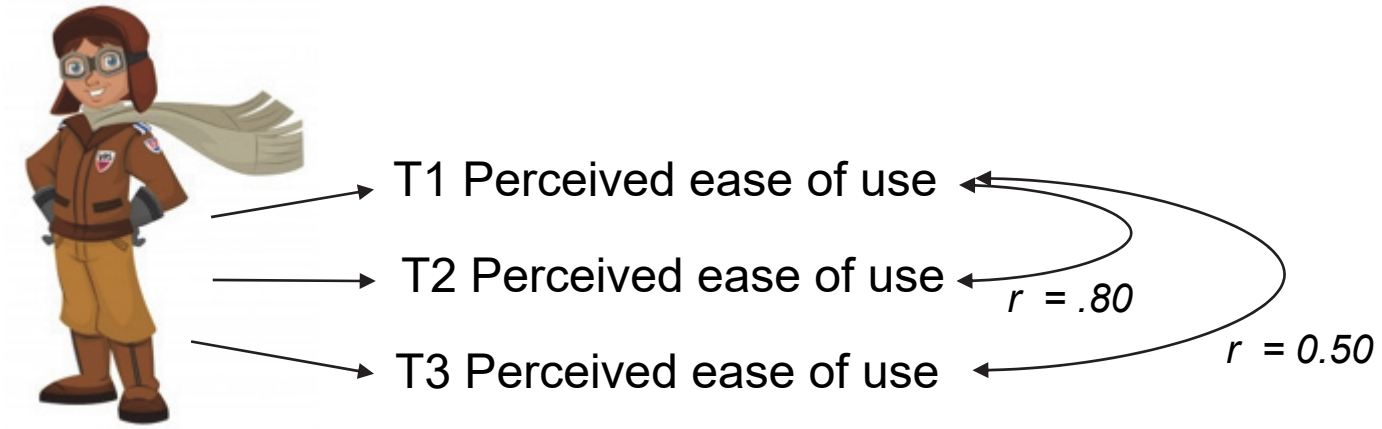
level-2 variance

Benefits of Mixed Models

- **Accounts for dependence within pilots**



- **Accounts for varying dependency within pilots**



Benefits of Mixed Models

- **Doesn't require complete data**

T1 Perceived ease of use

T2 Perceived ease of use

T3 Perceived ease of use

T1 Perceived ease of use

~~T2 Perceived ease of use~~

T3 Perceived ease of use

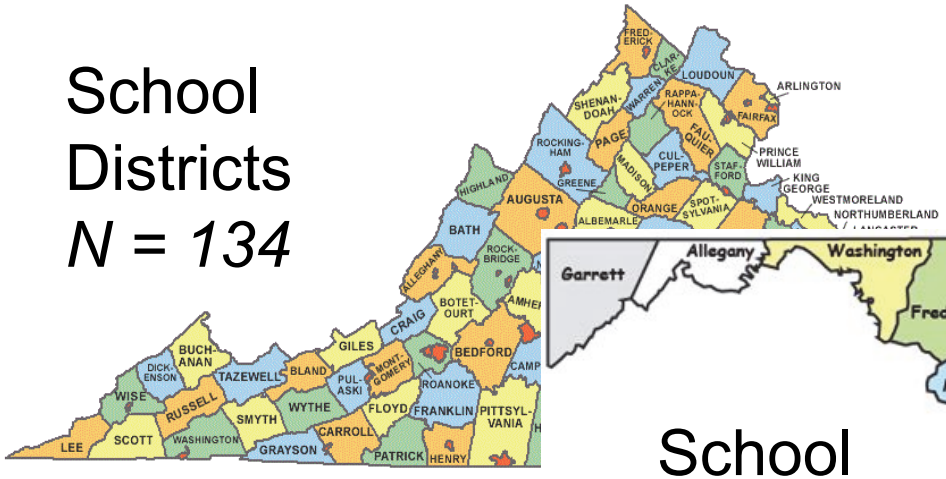
Looming Disadvantage

Previous research indicates sample sizes of at least 30 at the highest level should be used.

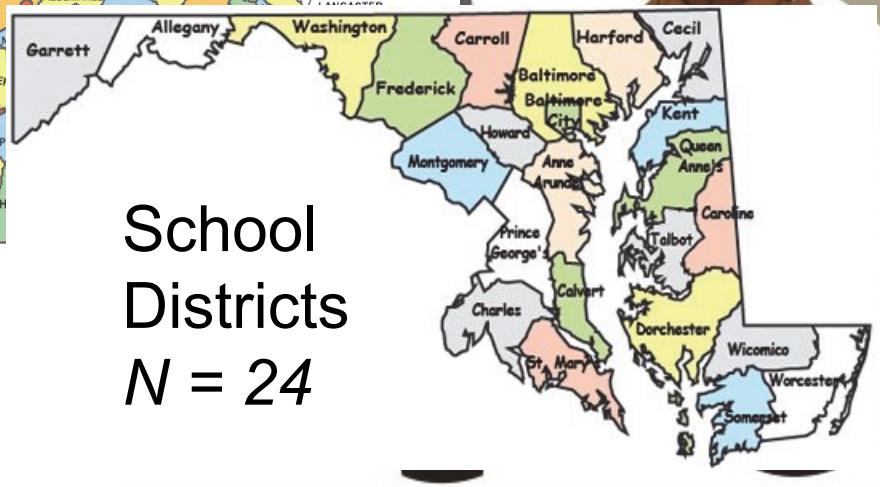
Scenario

Higher numbers needed
than are easily available….

School
Districts
N = 134



School
Districts
N = 24



….or even possible

Gaps in the Literature

- **How bad is "too small"?**
  - Lower limit of 10
    - » Simplest mixed model not explored:

$$y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}Z_jX_{ij} + \zeta_{0j} + \zeta_{1j}X_{ij} + \varepsilon_{ij}$$

$$y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \zeta_{0j} + \varepsilon_{ij}$$

- **Small**

Overall Problem:
Academic recommendations inconsistent with applied research realities

- **Small effect size**
  - Behavioral research often looking at tiny impacts
  - Impacts at that level not of interest to DOD

- **Small intraclass correlation**
  - Higher intraclass correlation exists in within-person designs

Current Study

- **Even in small total sample conditions, fixed effect bias will be minimal**
- **Increasing level-2 sample size has a greater positive effect on power than increasing level-1 sample size**
- **Smaller sample sizes will have adequately high power and low type I error rate under conditions and standards common in operational testing**
  - Higher type I error risk levels
    - » Power levels at DOD standard of $\alpha \leq .2$
  - Larger effect sizes
    - » Power at effect sizes relevant in applied research
  - Higher ICC levels
    - » ICC levels common to repeated measures designs

Method: Simulation Study

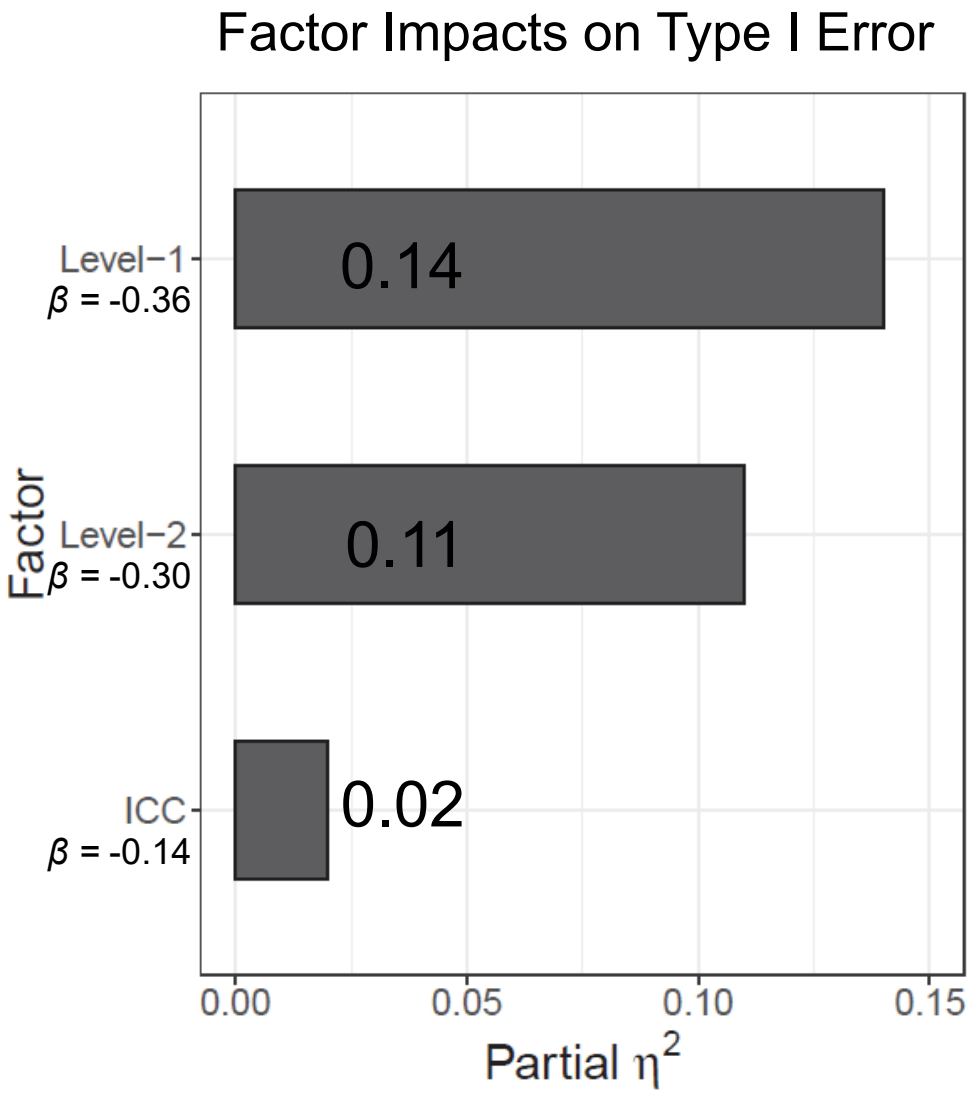
- **Continuous increases in sample size**
  - *N* = 4 to *N* = 30
- **Continuous increases in *baseline* observations**
  - *N* = 2 to *N* = 10
- **Varying levels of SNR**
  - SNR = 0, .3, .5, .8, 1
- **Varying level-2 variance**
  - ICC = .075, .25, .5, .8
- **1,000 datasets generated each**

Total sampling conditions: 243
Total conditions: 243 x 5 x 4 = 4,860
Total mixed models* = 4,860 x 1,000 = 4,860,000

Method: Simulation Study

- **FIML used to estimate fixed effects**
  - Fixed effect type I error, bias, and power of interest
- **Likelihood-ratio test used to compare full and reduced models**
  - Mitigates impact of downwardly biased standard error estimates
- **Convergence failure**
  - Negatively related to ICC
  - 0.13%-0.17% to 0.008% - 0%
- **Simulation factors impact modeled using linear regression**
  - Effect sizes highlighted to minimize reliance on *p*-values

Results: Type I Error

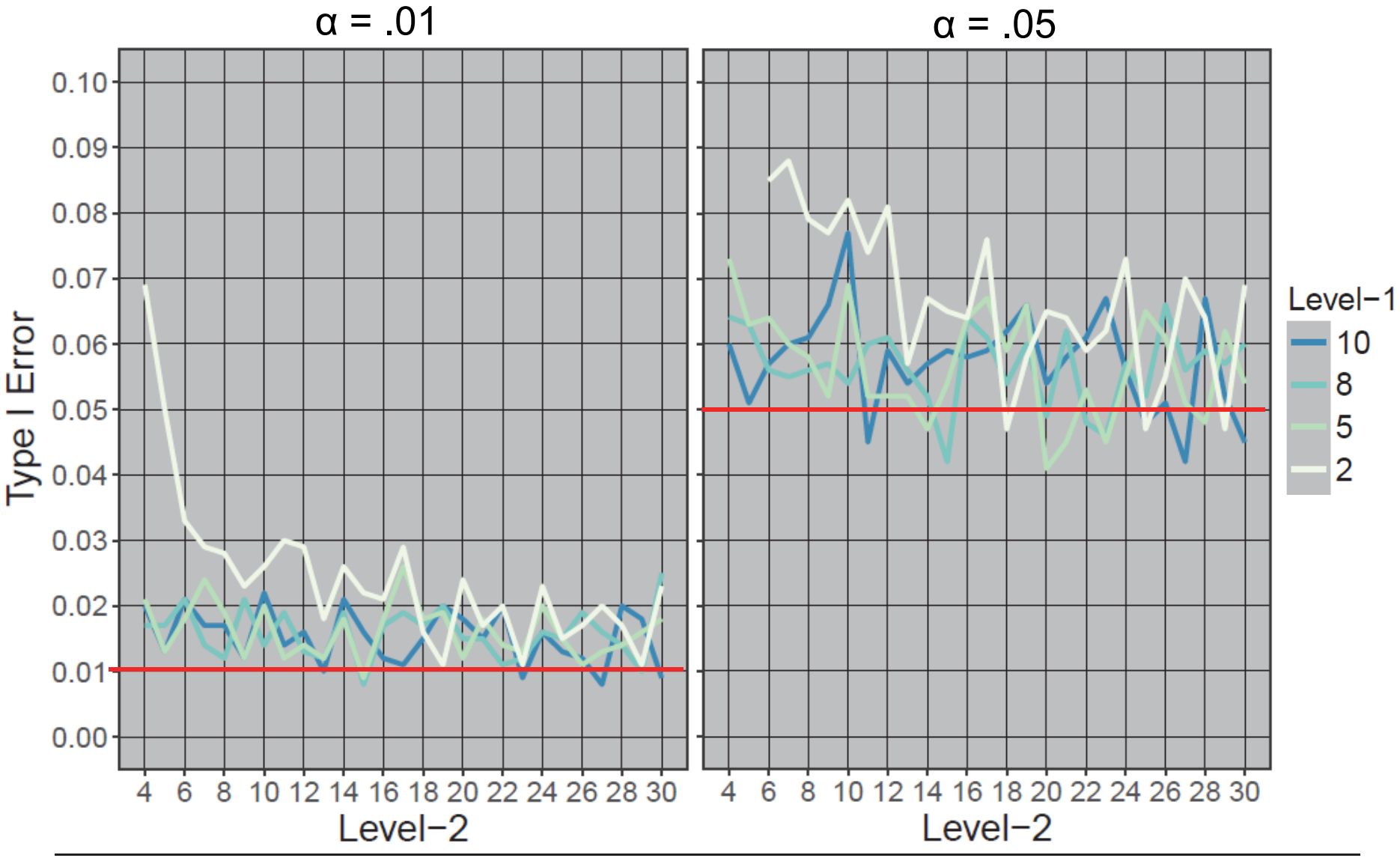


Type I Error:

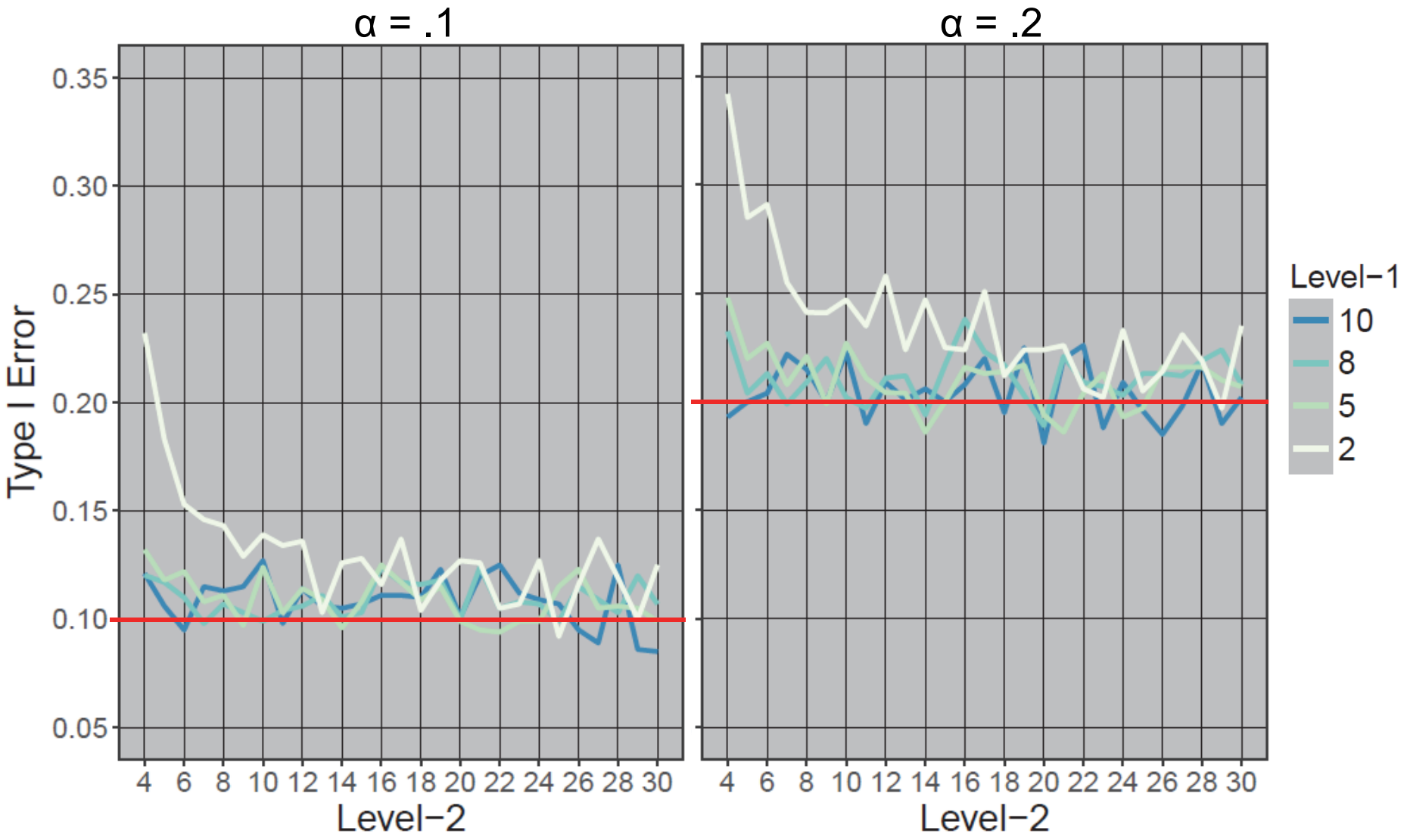
Proportion of models for which the fixed effect was found to be statistically significant despite having a slope equal to zero.

Type I error rate at the $p \leq .01$ level depicted, overall patterns present remained the same at higher alpha rates.

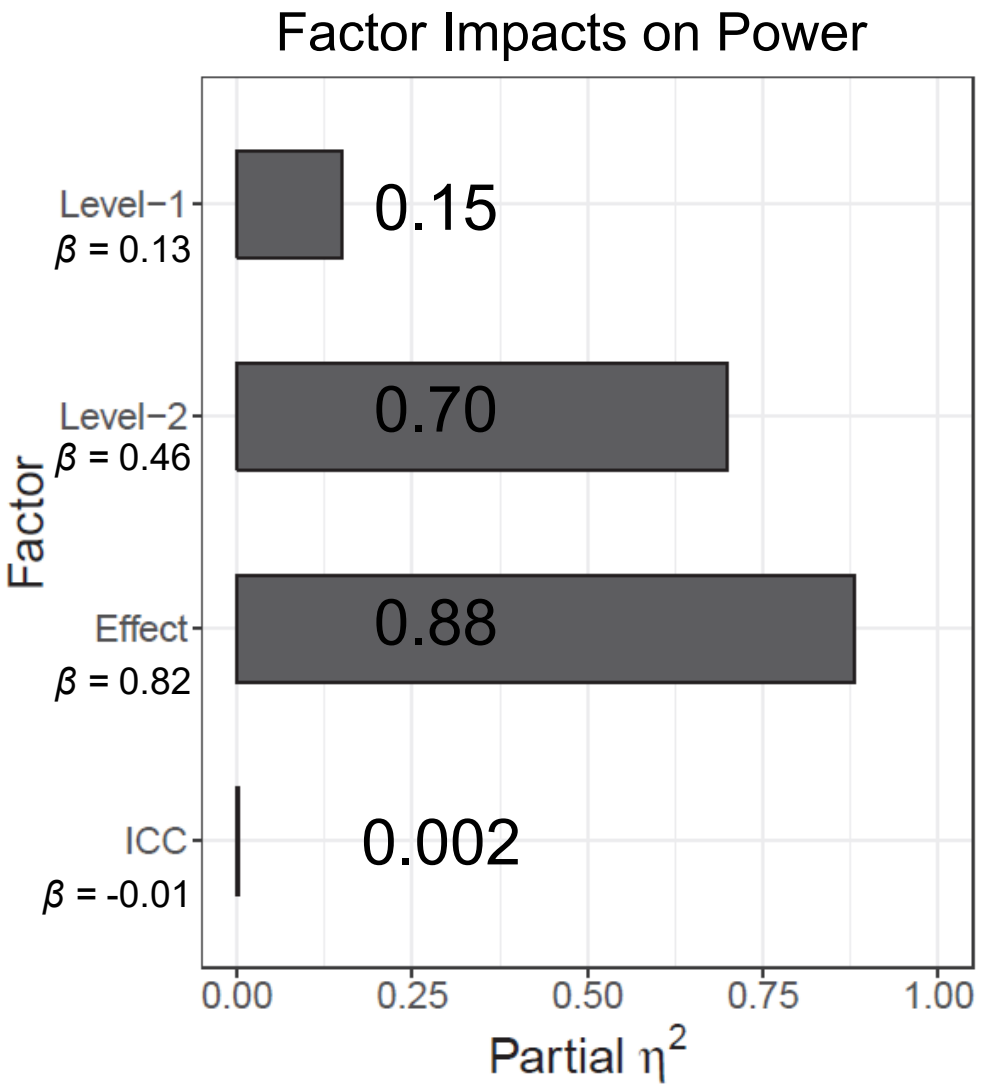
Results: Type I Error



Results: Type I Error



Results: Power

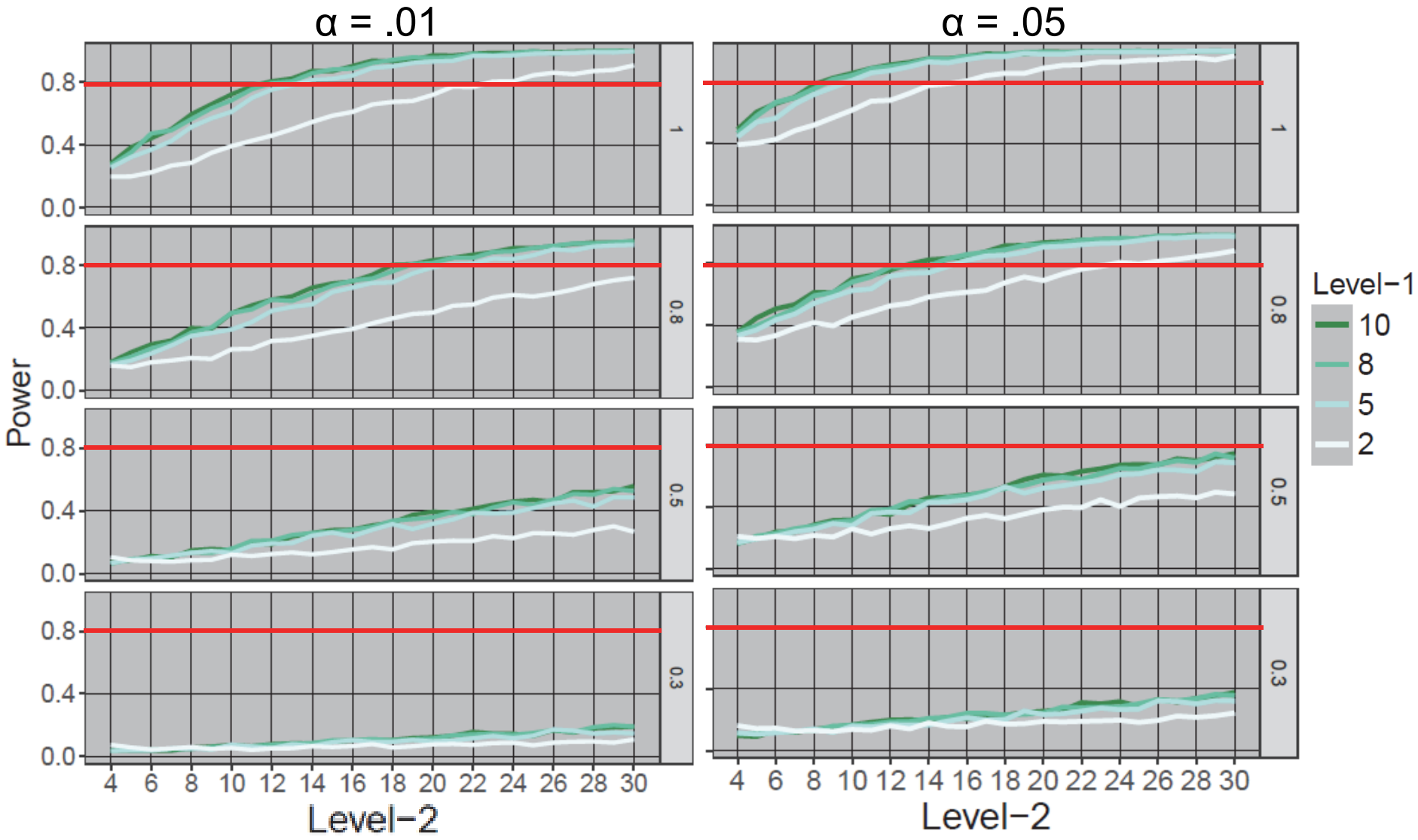


Power:

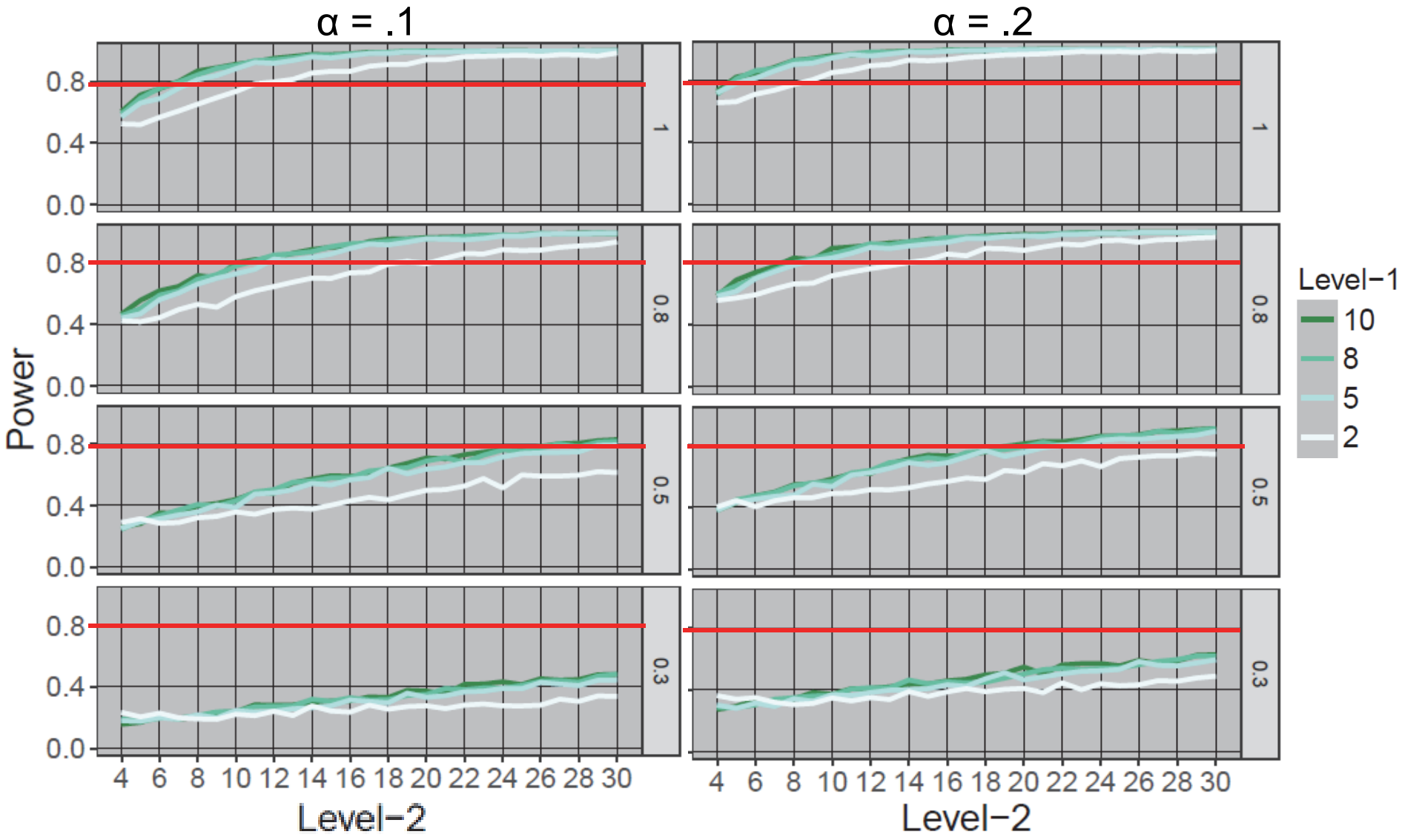
Proportion of models in which the fixed effect (slope > 0) was statistically significant.

Power at the $p \leq .01$ level depicted, overall patterns present remained the same at higher alpha rates.

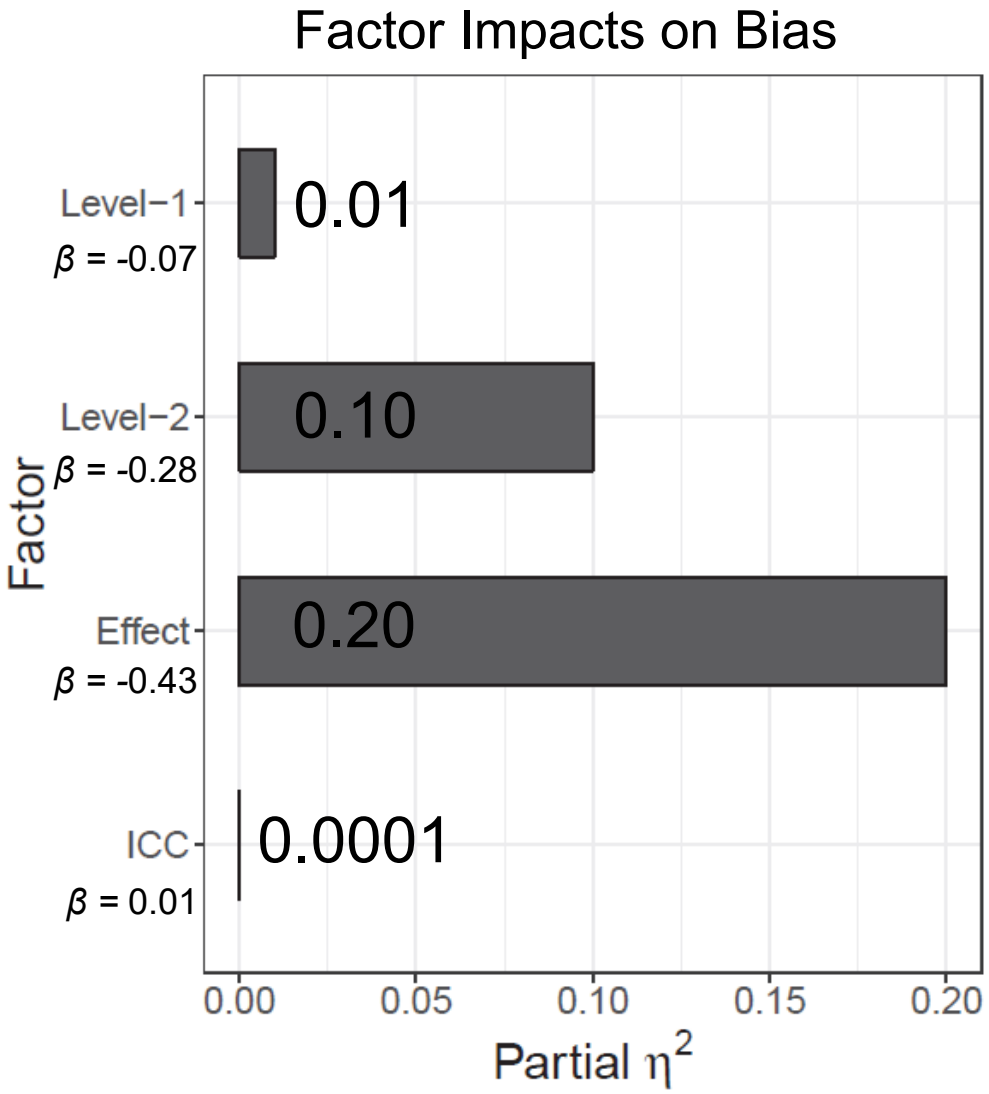
Results: Power



Results: Power



Results: Bias



Bias Descriptive Statistics

$$Relative\ Bias = \left(\frac{\hat{\theta} - \theta}{\theta}\right) \times 100$$

Across all conditions for which relative bias could be computed (i.e., effect size > 0), relative bias ranged from -16.27% to 14.32%. Relative bias above 5% occurred only at combinations of low SNR and low sample size.

IDA

Conclusions

- **Sample sizes of 10 participants and under can attain sufficient power in certain circumstances:**
  - when a single fixed effect factor is of interest
  - when greater risk of type I error is acceptable
  - when the minimum effect worth detecting is large (i.e., effect size = 1 or higher)

- **Under these conditions, fixed effect bias is low, inflations in type I error are manageable, and power is adequate despite small sample sizes.**

Implications

- **For operational research….**
  - Mixed models are a viable alternative, with minor adjustments
  - Accounts for typically encountered challenges
  - Enables analysts to take advantage of data already available
- **If you want to use mixed models with operators ≤ 10, you will only be able to detect large effect sizes**
  - Sampling numbers recommended here not unreasonable
  - Higher numbers available, mixed models can detect lower effect sizes

Further Research

- **Only simplest model examined here**
  - Binary vs. continuous predictors
  - Adding in fixed parameters, e.g., time of day
  - Cross level interactions, e.g., system-pilot experience interaction
  - Variance components, e.g., pilot unit
- **Impact of missing data**
  - Previous research indicates not problematic
  - Not tested on sample sizes this small
- **Using mixed models with empirical operator data**

Questions

Selected References

- Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Loudermilk, B. L., Kromrey, J. D., & Ferron, J. M. (2010). Dancing the sample size limbo with mixed models: How low can you go. *SAS Global Forum*, *4*, 11-14.
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Lawrence Earlbaum Associates. Hillsdale, NJ, 20-26.
- de Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 11*(1), 57-85.
- Hox, J. J., Moerbeek, M., & van de Schoot, R. (2010). Multilevel analysis: Techniques and applications. Routledge.
- Kreft, I. G., Kreft, I., & de Leeuw, J. (1998). Introducing multilevel modeling. Sage.
- Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92.
- McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28*(2), 295-314.
- Rucker, A. (2014). Improving statistical rigor in defense test and evaluation: Use of tolerance intervals in designed experiments. *Defense Acquisition Research Journal: A Publication of the Defense Acquisition University*, *21*(4).
- Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical methods, 8* (2), 597-599.
- Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods, 12*(2), 347-367.
- Snijders, T. A., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological methods & research, 22*(3), 342-363.
- West, B. T., Welch, K. B., & Galecki, A. T. (2014). *Linear mixed models: A practical guide using statistical software*. Boca Raton: CRC Press.

Recommendations

- **With SNR of 1**
  - if α = .2, power of .8 achievable with *N* = 6, observations ≥ 4
    - » At precise sample size, type I error of .18 to .27
- **With SNR of .8**
  - if α = .2, power of .8 achievable with N = 10, observations ≥ 4
    - » At precise sample size, type I error of .20 to .24
- **Desired rate of Type I error risk is .2**
  - If SNR = 1, type I error rates of .2 achievable with *N* = 9 & 10, observations 7+
  - ***Adjust for inflation in Type I error by stricter standard***