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

Kristina Carter, Heather Wojton, Stephanie Lane, & Jonathan Snavely

Institute for Defense Analyses

DATAWorks 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

#### Determination

- **Systems Different**
  - Reality: $<$
  - Systems Equivalent: $=$

- **Power**
  - $1 - \beta = P(a \neq b | a \neq b)$

- **Type I Error**
  - $\alpha = P(a \neq b | a = b)$

- **Type II Error**
  - $\beta = P(a = b | a \neq b)$

- **Confidence**
  - $1 - \alpha = P(a = b | a = b)$
Power

Additional factors affecting power:

1. Acceptable risk level, \( \alpha \)
   \[ \alpha = P(a \neq b \mid a = b) \]
   risk of making a Type I Error

2. Magnitude of the effect (SNR), \( \frac{\delta}{\sigma} \)
   \[ \frac{\bar{X} - \bar{X}}{\sigma_{pooled}} \]

3. Size of the sample, \( N \)

Positively related to Power
Repeated Measures Design

Legacy System
- Speed
- Accuracy
- Perceived ease of use

New System
- Speed
- Accuracy
- Perceived ease of use

Veteran

Novice

$$\bar{X} - \bar{X}$$

$$\sigma_{pooled}$$
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**

\[
y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij}
\]

**Operator Model**

\[
\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + \zeta_{0j}
\]

\[
\beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + \zeta_{1j}
\]

**Mixed Model**

\[
y_{ij} = \gamma_{00} + \gamma_{01}Z_j + \gamma_{10}X_{ij} + \gamma_{11}Z_jX_{ij} + \zeta_{0j} + \zeta_{1j}X_{ij} + \epsilon_{ij}
\]
**Repeated Measures Mixed Model**

Level 1: Different observations of a single operator

Level 2: Different operators

**Intraclass Correlation (ICC)**

\[
ICC = \frac{\text{between operator variance}}{\text{between operator variance} + \text{within operator variance}}
\]
Benefits of Mixed Models

- Account for dependence within operators

  \[ r \neq 0 \]

- Account for varying dependency within operators

  \[ r = 0.50 \]
  \[ r = 0.80 \]
Benefits of Mixed Models

- Don’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
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
\[ N = 134 \]

….or even possible
Gaps in the Literature

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

\[
y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}z_j + \gamma_{11}z_jx_{ij} + \zeta_{0j} + \zeta_{1j}x_{ij} + \epsilon_{ij}
\]

\[
y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \zeta_{0j} + \epsilon_{ij}
\]

- **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

---

Overall Problem: Academic recommendations inconsistent with applied research realities
Current Study

• Even in small total sample conditions, bias of fixed effect estimates 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 level-2 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 \times 5 \times 4 = 4,860
Total mixed models* = 4,860 \times 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%
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

\( \alpha = .01 \)

\( \alpha = .05 \)
Results: Type I Error

$\alpha = .1$

$\alpha = .2$
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

\(\alpha = .01\)

\(\alpha = .05\)
Results: Power

\[ \alpha = .1 \]

\[ \alpha = .2 \]
**Results: Bias**

**Bias Descriptive Statistics**

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

Relative bias above 5% occurred only at combinations of low SNR and low sample size.

**Factor Impacts on Bias**

- **Level-1**
  - \( \beta = -0.07 \)
  - Bias: 0.01

- **Level-2**
  - \( \beta = -0.28 \)
  - Bias: 0.10

- **Effect**
  - \( \beta = -0.43 \)
  - Bias: 0.20

- **ICC**
  - \( \beta = 0.01 \)
  - Bias: 0.0001
Conclusions

• Smaller samples of participants can attain sufficient power in certain circumstances:
  – when a single fixed effect factor is of interest
  – when the minimum effect worth detecting is large (i.e., effect size = 1 or higher)
  – *when inflation in type I error is adjusted for by stricter standard*

• 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
  – Account for typically encountered challenges
  – Enable analysts to take advantage of data already available

• If you want to use mixed models with operators $\leq 10$, 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**
Selected References


