The ITEA Journal of Test and Evaluation 2019; 40: 16-29

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
Stephanie T. Lane
Institute for Defense Analyses, Alexandria, VA

Mixed models are ideally suited to analyze nested data from within-persons designs, designs that are advantageous in applied research. Mixed models enable the estimation of random effects, facilitating an accounting of the intra-cluster variation captured by multiple observations of the same participants. However, the sampling requirements for mixed models are prohibitive to areas of research which could greatly benefit from them. This simulation study examines the impact of small sample sizes in both observation and nesting levels of the model on the fixed effect bias, on type I error, and on the 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.

Keywords: mixed effect models, multi-level models, sample size, effect size, power

Kristina A. Carter

Heather M. Wojton

Stephanie T. Lane

In experimental design, exposing the same participants to all conditions allows researchers to control inter-person variability that would increase data noise and reduce detection of an effect. By accounting for individual-level variance, these within-person designs reduce overall error variance, granting statistical tests greater power than equivalently sized between-person designs. Additionally, repeated measures of each participant can capture longitudinal effects and reduce overall error by accounting for intra-person variability, which single measures of participants cannot (Hawkins et al., 2007; Frison and Pocock, 1992; Watson and Workman, 1981). A within-person design can be thought of as a hierarchical or nested design: repeated measurements are nested within the participants on which they are based.

Any experimental design includes both fixed effects, which are assumed constant across individuals, and random effects, which can vary (de Leeuw and Kreft, 1998). Gelman (2005) defines constant effects as identical for all groups in a population and varying effects as differing from group to group. Fixed effects can also be defined according to the researcher's interest; factors with explicitly chosen levels and effects that are interesting in themselves are fixed. Random effects are attributed to factors whose levels were randomly sampled from some underlying population of interest but whose effects are not themselves interesting to the researcher (Searle, Casella, and McCulloch, 1992).

For example, consider a defense researcher interested in the usability of a new system compared to an old system. The effect of system type on usability is of primary interest, but the researcher also documents what unit the system operators came from. In such a case, the usability of one system compared to the other is being investigated; that effect is constant. The unit is a random effect, unless groups are specifically selected from units of interest, in which case unit could be captured as a fixed effect. Differences in usability between systems can be allowed to vary across individuals within each grouping.

Mixed-effect analyses, which model both fixed and random effects, have been recommended for analyzing nested data (Gueorguieva and Krystal, 2004). However, the constraints of mixed-model analyses indicate that larger sample sizes than are readily available are necessary for fully effective models (Hox, 1998; Hox, Moerbeek, and van de Schoot, 2010; Dedrick et al., 2009; Snijders and Bosker, 1994). Recent research has examined the effect of small samples on mixed-effect models, but the impact of very small sample sizes—which can be unavoidable in some areas—has not been systematically examined for simple models relevant to applied research (McNeish and Stapleton, 2016). In this study, we used simulation methods to evaluate the impact of such sample sizes on the bias, type I error rate, and power of mixed-effect models to detect fixed effects in applied research.

Mixed Models Briefly Described

Mixed-effect models that can be described hierarchically (also called multilevel or hierarchical linear models) are a specific case of mixed-effect models (e.g., McNeish and Stapleton, 2016; Raudenbush and Byrk, 2002). For simplicity, we use the more general term mixed models. Mixed models are well-suited to analyzing within-person data because they account for dependencies across multiple observations of the same individual, whether across conditions or within conditions. In a within-persons design context, "level 2" refers to individual participants, and "level 1" refers to multiple measurements or observations of a single individual. Perhaps the most intuitive representation of mixed models is a linear regression structure (Raudenbush and Byrk, 2002). In keeping with our scenario of a defense researcher testing a new system's usability, a separate regression model is estimated on the level of the measurement (level-1 model) such that:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij}$$

where Y=usability rating, i=1,…,n_j observations for each person, j=1,…,j operators, X indicates the system rated, and ε represents observation-level error. In mixed models, variation among the regression coefficients (β_j) is modeled by a participant-level, or level-2 model:

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

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

a separate intercept (γ_{00}) and slope (γ_{10}) term is estimated for each operator, Z indicates an operator-level variable such as the unit the operator belongs to, and ζ represents operator-level error. The regression model can then be represented as a mixed model with fixed and random components grouped accordingly:

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

the first four terms on the right side of the equation contain all fixed effects, the subsequent two represent level-2 intercepts and slopes, and the final error term represents level-1 variability.

The error terms of the level-1 and level-2 models, ε_{ij} and ζ_{0j}, are assumed to be normally distributed, each having a mean of zero and variances of σ^2 and τ_{00} respectively. Intraclass correlation (ICC) is a measure of the extent to which the clustering of level-1 measures in level-2 accounts for the variance in Y. ICC is determined by the ratio of level-2 variance to the total variance:

$$\text{ICC} = \frac{\tau_{00}}{\tau_{00} + \sigma^2}$$

where τ_{00} is the level-2 variance (e.g., variance between operators) and σ^2 is the level-1 variance (e.g., variance between different measures of the same operators). The resulting ICC is the proportion of total variance that can be attributed to the nested nature of the data. Mixed-model analysis does not distinguish between data nested within people and data nested within groups of people, but ICC values tend to differ for the two designs, with ICC values for within-person data being substantially higher (e.g., Kwok et al., 2008; Rabe-Hesketh and Skrondal, 2008; Singer and Willet, 2003).

Mixed models commonly use maximum likelihood (ML) estimation. ML is generally robust and has an advantage in being more efficient and less biased than ordinary least squares methods, such as generalized least squares (Hox et al., 2010). Mixed models using ML have the added flexibility that residuals need not be normally distributed so long as the researcher's interests do not include inferential tests of the variance terms (Maas and Hox, 2004). Two likelihood functions are estimated in mixed models: full information maximum likelihood (FIML) and restricted maximum likelihood (REML). FIML treats regression coefficients as unknown fixed quantities when estimating variance components, failing to account for degrees of freedom lost by the estimation of fixed effects. The result is that FIML estimates are negatively biased; thus, its use is discouraged in favor of the less biased REML function, which includes variance components only for random effects and estimates fixed effects separately. REML fixed effect estimates are equivalent to those of FIML, and their variance effects may be improved over FIML. Therefore, use of REML estimation is preferable (Maas and Hox, 2005), particularly when sample sizes are small (Hox, 1998; de Leeuw and Kreft, 1998). An important characteristic of ML is that it is asymptotically consistent. That is, ML estimates converge on population values as sample sizes become large; therefore, large sample sizes are considered necessary to enable valid estimates (e.g., Hox et al., 2010). What constitutes a "large" sample size has been discussed, but uncharted territory remains. Two questions that have not been examined are the subject of the current study: (1) precisely how large the sample sizes should be for a within-study design (where ICC values are larger), and (2) the comparative impact of increases in level-1 and level-2 sample sizes on the power to detect a fixed effect in simple cases interested solely in that effect.

Previous Research on Mixed Model Sample Size

Previous simulation studies have examined how small sample sizes affect mixed-model outcomes. Because of the focus of the current study, we restrict our review to findings in three principal areas: parameter (fixed effect) bias, type I error, and power.

Fixed effect bias. Previous research indicates that the accuracy of parameter estimates varies based on whether the estimates are random or fixed, and on the level with which they are associated. Across simulation studies, fixed-effect estimates were found to be the least affected by sample size (McNeish and Stapleton, 2016). Fixed-effect estimates at both level 1 and level 2 were unbiased with level-2 sample sizes as low as 30. Level-1 fixed effect estimates continued to be unbiased even when level-2 sample sizes were less than 15. Level-2 fixed-effect estimates were more vulnerable: when level-2 sample sizes are below 15, these estimates, and cross-level estimates, tend to be positively biased. In contrast, some studies found fixed-effect bias is inconsequential even with level-1 and level-2 samples as small as five and 10 (e.g., Bell et al., 2010; Maas and Hox, 2005). Level-1 sample size did not influence fixed-effect estimate bias, nor did the ICC (notably, ICC values used were common for group nested designs, but low for within-person designs).

The extent to which random-effect estimates are affected by sample sizes varies according to the level. The bias of level-1 random-effect estimates is largely unaffected by sample size regardless of level, with level-2 sample sizes under 10 found to have bias under 1% (Bell et al., 2010; Browne and Draper, 2006; Maas and Hox, 2010). Findings are conflicted regarding level-2 random-effect estimate bias; some studies found that random-effect estimates have a positive bias as high as 25% with a level-2 sample size of 10 and a level-1 sample size of 5, whereas others found the estimates were minimally biased with a level-2 sample size of 6 and a level-1 sample size of 18 if REML was used (Bell et al., 2010; Maas and Hox, 2005; McNeish and Stapleton, 2016). McNeish and Stapleton (2016) posit that this may be because of differences in level-1 sample sizes, as other studies have found that small level-1 sample sizes result in positively biased level-2 variance estimates (Clarke, 2008; McNeish, 2014).

Type I Error. Traditional hypothesis testing is conducted by computing a test statistic: the quotient of the effect estimate over its standard error (SE). This procedure is problematic in mixed-model analysis because SE estimates for both fixed and random effects are particularly susceptible to level-2 sample size. SEs of fixed and random-effect estimates have been found to be unacceptably negatively-biased when level-2 sample sizes are small (Maas and Hox, 2005; McNeish and Stapleton, 2016). Underestimated SEs result in inflated test statistics, increasing the chances that the null hypothesis will be erroneously rejected (i.e., type I error).

Level-2 sample sizes as small as 10 have been found to underestimate the SEs of fixed effects, with simulation studies indicating a level-2 sample size of at least 30 is required to produce unbiased SE estimates when using REML (Maas and Hox, 2005; McNeish and Stapleton, 2016). McNeish and Stapleton (2016) conclude that a level-2 sample size of 10 is unacceptable with standard estimation procedures, as hypothesis tests have at least twice the typically accepted type I error rate (see Bradley, 1978). Random effect SEs were the most highly affected by sample size, with such elevated levels of negative bias among SEs that a minimum sample size of 50 in both levels is recommended, with some studies recommending higher numbers (Maas and Hox, 2005; McNeish and Stapleton, 2016).

The impact of biased SEs can be mitigated by using alternate inferential tests. Likelihood-ratio tests can be used to determine the significance of predictors, with FIML used to test fixed-effect hypotheses, and REML used for random-effect hypotheses (West, Welch, and Galecki, 2015). Methods such as the Kenward-Roger adjustment, bootstrapping, and a Bayesian approach can also be used to account for fixed-effect type I error; however, the computational complexity of these methods can make them prohibitive to use (Hox 2010; Maas and Hox, 2010; McNeish and Stapleton, 2016). Bayesian approaches additionally require selection of a prior distribution and they may not perform well when sample sizes are very small (Gelman 2006; Hox, 2010).

Fixed effect power. Previous studies have examined the effect of sample size on power. Mixed model power might refer to the hypothesis test of a level-1 or level-2 fixed effect, or to the hypothesis test of a level-1 or level-2 random factor. Results of simulations studies by Bell et al. (2010) indicated that a power level of .80 is not achieved when level-1 and 2 sample sizes are small. An average power of .80 or higher for detecting level-1 fixed effects was only achieved when level-1 and -2 sample sizes were 20-40 and 30 respectively. For level-2 fixed factors, an average power of .80 was not obtained even at the largest condition of 40 level-1 and 30 level-2 sample sizes. This latter finding casts doubt on the utility of the traditional 30-30 rule of thumb for detecting level-2 fixed effects, indicating that even higher sample sizes may be necessary if power of .80 is desired (Bell et al., 2010).

Power in Applied Research. Sampling standards and practices in applied settings vary by domain. In domains such as industrial-organizational psychology and education, sample sizes can be as large as thousands (Barlett, Kotrlik, and Higgins, 2001; Keselman et al., 1998; Dedrick et al., 2009). Studies in more novel areas tend to have fewer participants. For instance, recent research in usability studies recommends sample sizes of 10±2 or as high as 20, but earlier recommendations of five-person samples are still largely prevalent (e.g., Lewis, 1994; Nielsen and Molich, 1990; Nielsen, 1994; Virzi, 1992; Caulton, 2001; Faulkner, 2003; Hwang and Salvendy, 2010).

In areas such as biomedicine, psychiatry, and neuroscience, underpowered studies that lack sufficiently large samples are normal and often perceived as unavoidable (Button et al., 2013; Ioannidis, 2005, Ioannidis, 2011). Such small samples may be due to the rarity of conditions of interest or the prohibitive cost of obtaining and operating technological equipment such as neuroimaging scanners. Operational testing in the Department of Defense (DoD) is another area where adequate sample sizes can be challenging to procure, whether due to time-consuming and financially costly tests or because of scarcity in existing cases. For instance, defense operations involving novel systems might not only have a limited number of systems to test but a limited number of individuals equipped to operate them. In such cases, small samples may be unavoidable.

Bell et al.'s (2010) findings regarding power are problematic for applied research areas with available sample sizes well below the 30 level-2 and 20-40 level-1 samples that were found to yield the desirable standard of .80 power. However, in addition to the type II error rate of 20%, which is generally accepted as the standard, power is contingent upon other factors. These include effect size (represented by Cohen's d as the difference between means over their pooled standard deviation: $\frac{\bar{X}_1-\bar{X}_2}{s_p}$), whether the predictors' scale is binary or continuous, and the acceptable level of type I error. Bell et al.'s (2010) models had a maximum slope of .3 and were fairly complex; the simplest model included two level-1 predictors and two level-2 predictors. As the authors acknowledge, adequate power for small-sampled mixed models is likely contingent on very large effect sizes, which a slope of .3 does not begin to approximate. Effect sizes in behavioral research often range from d=.2 to d=.8. However, as outlined by Sawilowsky (2009), effect sizes surpassing 1 have been found in academic research areas such as education, with effect sizes greater than 2 being found in some instances. In many applied fields, the minimum effect sizes considered worth detecting surpass the values of d=.3 and d=.5 typically observed in academic behavioral research.

In addition to differences in effect size, differences in other standards affecting power exist across research domains. The type I error rate of .05, though common in academic research, is a standard born from convention and is not ubiquitous across other domains. For instance, the United States DoD deems a type I error rate as high as .2 to be an acceptable maximum limit (e.g., DOT&E, 2018; Rucker, 2014).

Current Study

Although the utility of mixed models has been well-documented, questions remain regarding the sampling conditions under which mixed models can perform sufficiently well to recommend their use to applied researchers. This study simulated the effect of small sample sizes on fixed effect parameter stability, type I error rate, and the power of mixed-models. Consistent with previous studies, we expected that *increasing level-2 sample size has a greater positive effect on power than increasing level-1 sample size* and that *even in small total sample conditions, fixed effect bias will be minimal*. As this study was interested in the extent that mixed-effect models can contribute to applied analysis, we investigated model conditions and standards uncommon in traditional academic areas but prevalent in applied areas such as operational testing. We expected that: *smaller sample sizes will have adequately high power and low type I error rate under the conditions and standards common in operational testing*. In short, we expected simulation results would validate the use of mixed-effect models in applied settings even when those settings are constrained by small sample sizes. The simulation study designed to test these expectations is detailed below.

Method

To compare the effect of level-1 and 2 sample sizes on parameter estimation and power, we used the simplest possible model, with a single predictor variable at level-1 and no level-2 predictors. This corresponds to a slightly simpler model than depicted above:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \zeta_{0j} + \zeta_{1j}X_{ij} + \varepsilon_{ij}$$

To further simplify the model, the slope error term ζ_{1j} was set to 0, resulting in the following four-term model:

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

In this simulation, four parameters were varied: level-1 sample size, level-2 sample size, effect size, and level-2 variance. For these simulations, level-2 groups were balanced–each individual had an equal number of observations. To determine the benefit of sample size to power, level-2 sample size was varied from 4 to 30 and level-1 sample size was varied from the lowest possible number of 2 to a maximum of 10. For simplicity and applicability, all simulation conditions included multiple level-1 units solely at the baseline measure (X=0), with a single follow-up (X=1) level-1 unit. This not only controlled number of baseline versus follow-up effects but increased the applicability of the study to applied domains such as operational testing, where archival data often provide a wealth of baseline measures, but follow-up measures are less easy to secure.

The regression intercept was specified as 1 and regression slope was varied among 0, 0.3, 0.5, 0.8, and 1 to account for both effect sizes commonly used in psychological literature and larger effect sizes seen in applied research (Cohen, 1988; Sawilowsky, 2009). The slope was set to zero to examine type I error rates. Level-2 variance was varied, resulting in alternate ICCs equal to .075, .25, .5, and .8, chosen to reflect ICC values common in mixed model research (e.g., Mass and Hox, 2005; Singer and Willet, 2003). Although ICC values typical in group mixed models are generally at or below .25, within-person designs often have much higher ICC values (e.g., Hedges and Hedberg, 2007; Kwok et al., 2008; Rabe-Hesketh and Skrondal, 2008; Singer and Willet, 2003). Residual variance was held constant and the covariance structure was assumed to be diagonal with covariance equal to zero.

Measures

Bias, seen in the numerator of the equation below, is a method of quantifying the error between estimated parameters and true parameters. Bias is more easily compared across modeled conditions when considered relative to the true parameter. The below equation of relative bias, where $\hat{\theta}$ represents the estimated fixed effect and θ represents the true fixed-effect parameter (the simulated slope), gives us the percentage parameter bias for a specific model's estimates.

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

Bias for every estimated parameter was computed using this equation, then average relative bias for each simulation condition was computed.

Type I error rate was measured by the proportion of models in which the fixed effect was found to be statistically significant despite having a slope equal to zero. Power for each simulation condition was determined by the proportion of models in which the fixed effect was significant at various alpha rates. Nominal alpha values of .01, .05, .10, and .20 were used to determine the proportion of statistically significant fixed effects for both type I error rate and power.

Analysis and Results

Nine possible level-1 sample sizes and 27 possible level-2 sample sizes resulted in 243 sampling conditions. Multiplying the sampling conditions by five effect sizes and four level-2 variances gives a total of 4,860 model conditions. A total of 1,000 simulated data sets with normally distributed errors were generated for each model condition. FIML was used to estimate fixed effects and, following recommendations of West et al. (2015), a likelihood-ratio test was used to compare the full model with the fixed effect to a reduced model without it. Convergence failure rates of the mixed model estimation varied according to ICC. In the lowest ICC conditions, failure rates ranged from 0.13% to 0.17%; among the highest ICCs, failure rates ranged from 0% to 0.008%. Reported results exclude estimates that failed to converge.

Linear regression was used to model the effect of simulation factors on fixed-effect estimate bias, type I error rate, and power. Effect sizes are highlighted to minimize reliance on *p*-values in interpreting factor effects.

Fixed Effect Bias

Across all conditions for which relative bias could be computed (i.e., effect size > 0), relative bias ranged from -16.27% to 14.32%. Due to negative values of fixed-effect bias, absolute relative bias was computed. Mean absolute relative bias was equal to 1.51. Absolute relative bias was regressed on effect size, ICC, level-1 sample-size, and level-2 sample size. All factors except ICC (p=.65, β=0.01, η^2_p=.00) were negatively related to bias and statistically significant at the $p < .001$ level. Partialling out the variance explained by other factors in the model, effect size (β=-0.43) most greatly impacted bias (η^2_p=.20). The partial eta-squared of level-2 sample size (β=-0.28) indicated that level-2 sample size accounted for 10% of variability in bias (η^2_p=.10) beyond that accounted for by other factors; level-1 sample size (β=-0.068) accounted for far less (η^2_p=.01).

Type I Error Rate

Type I error was inflated at all nominal alphas. Higher nominal alpha rates were associated with greater average inflation in Type I error rates, $r(5)$=.944, p=.016. Further investigation indicated this relationship was due to extreme values at the smallest total sample-size conditions. At the lowest nominal alpha ($p \leq .01$), type I error ranged from 0 to .07 (*M*=.02), indicating average inflation of 0.01 (*Mdn*=0.01). At the highest nominal alpha ($p \leq .2$), type I error ranged from .18 to .38 (*M*=.22), indicating average inflation of 0.02 (*Mdn*=0.01). This indicates that type I error inflation rate cannot be presumed constant across nominal alpha levels.

Type I error rate at the $p \leq .01$ level was regressed on ICC, level-1 sample-size, and level-2 sample size. Overall patterns presented in the regression results remained the same at higher alpha rates. All three factors were negatively related to type I error and statistically significant at the $p < .001$ level. After accounting for all other variables in the model, level-1 sample size (β=-0.36) had the greatest effect on type I error (η^2_p=.14). The partial eta-squared of level-2 sample size (β=-0.30) indicated that level-2 sample size accounted for approximately 11% of unique variability in bias (η^2_p=.11). ICC (β=-0.14) had the smallest effect (η^2_p=.02).

Fixed Effect Power

At the smallest sample conditions, mixed models were severely underpowered; therefore, medians are described in the descriptive statistics as they more accurately reflect central tendency. At effect sizes of 0.3 and 0.5, median power ranged from 0.09 (at slope=0.3, $p \leq .01$) to 0.71 (at slope=0.5, $p \leq .2$). At effect sizes of 0.8 and 1, median power ranged from 0.66 (at slope=.8, $p \leq .01$) to 0.99 (at slope=1, $p \leq .2$). With the exception of the minimum value, median power levels for both higher effect size values were above 0.80 for all *p*-values.

Power at the alpha ≤ .01 level was regressed on effect size, ICC, level-1 sample-size, and level-2 sample size. Regression result patterns described below remained the same when higher alpha rates were used. All factors except ICC were positively related to power and statistically significant at the $p < .001$ level. ICC (β=-0.01) accounted for less than 1% of variance in power (η^2_p=.002). Unsurprisingly, effect size had the greatest impact on power; after accounting for the other factors in the model, effect size (β=0.82) explained 88% of the variance in power (η^2_p=.88). After accounting for other factors' effects, level-2 sample size (β=0.46) accounted for approximately 70% of the unexplained variance in power (η^2_p=.70); in contrast, level-1 sample size (β=0.13) accounted for a far smaller proportion (η^2_p=.15). Follow-up investigation of the interaction between level-1 and level-2 sample size indicated a very small partial eta-squared (η^2_p=.005). Figure 1 depicts the effect of level-1 and level-2 sample size on power.

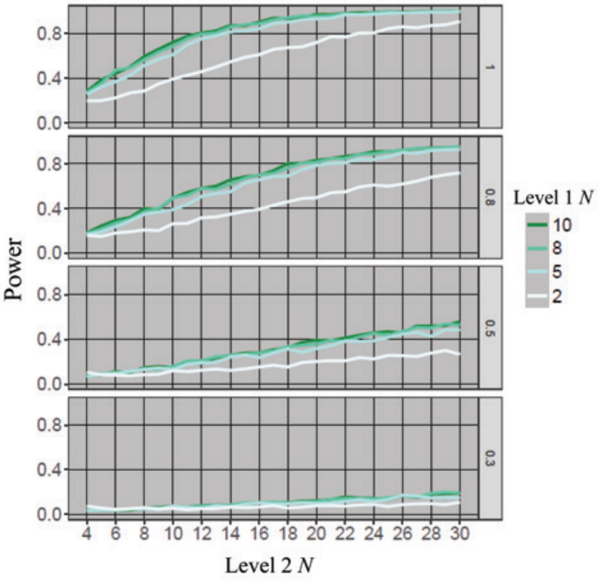

Figure 1: Power at Various Levels of Effect Size, p=.01, ICC=.5

Table 1: List of Symbols and Definitions for Equation Terms in This Paper

Symbol	Definition
d	Cohen's d: an effect size measure computed as the difference between two groups means over their pooled standard deviation
Mdn	Median value
M	Mean value
p	The probability of observing a result as extreme or more extreme than the observed result if the null hypothesis is true
s_p	Standard deviation pooled between two groups
X_{ij}	Level-1 or observation-level predictor variable in a multilevel model
X_i	Mean value of ith group
Y_{ij}	Estimate of the ith observation in a regression model of the jth participant
Z_j	Level-2 or person-level predictor variable in a multilevel model
β	Standardized estimate of coefficient in a regression model
β_{ij}	Estimator of the ith level-1 or observation-level coefficient in a regression model of the jth participant
ε_{ij}	Level-1 or observation-level error term in a multilevel model
ζ_{ij}	Level-2 or person-level error term in a multilevel model
η_p^2	Partial eta-squared, a measure of the effect of a given predictor in a linear regression model controlling for other predictors in the model
θ	The true value of the fixed effect slope, estimated by $\hat{\theta}$
$\hat{\theta}$	An estimate of some true parameter θ, representative of an estimate of the fixed effect slope
σ^2	Level-1 or observation-level variance in a multilevel model
τ_{00}	Level-2 or person-level variance in a multilevel model
γ_{oi}	Estimator of the ith level-2 or person-level intercept coefficient
γ_{1i}	Estimator of the ith level-2 or person-level slope coefficient

Conclusion

Previous research indicates that, as a rule of thumb, a minimum sample size of 30 at level-2 and 10 at level-1 is necessary to analyze data using mixed models. In some areas of applied research, such sample sizes may not be readily available or even possible. Our findings suggest that much smaller sample sizes, including sample sizes of 10 and under, can attain sufficient power in certain circumstances, specifically: (1) when a single fixed-effect factor is of interest, (2) when greater risk of type I error is acceptable and thus can be adjusted for, and (3) when the minimum effect worth detecting is large (i.e., effect size ≥1). This study demonstrates that under these conditions, fixed effect bias is low, inflation in type I error is manageable, and power is adequate despite small sample sizes.

This study also provides a clear demonstration of the comparative benefit of marginal increases in level-1 and level-2 sample size. Increasing the overall mixed model sample size will increase power, but increasing level-2 sample size, rather than level-1, has a greater impact on power. Within-person designs often have higher ICCs than other nested designs; our simulation found that ICC was negatively related to type I error rate but also negatively related to power (though the effect was very small).

The benefit of simulation studies is the systematic examination they facilitate, but they are constrained by the bounds they set. As with any simulation study, our findings are specific to the scope of this study and should not be generalized beyond its conditions. The model examined by these simulations was simpler than those generally undertaken by researchers interested in mixed models. Thus, although this study indicates implementing mixed models in applied research could be beneficial even when sample sizes are small, this may only be the case for researchers interested in a single fixed factor rather than more complex modeling.

Despite its constraints, this study contributes to a greater understanding of mixed models and offers important guidance to areas of research that could greatly benefit from using mixed models but face restraints in sampling. Furthermore, our findings indicate that more work in mixed models is needed to examine traditional "rules of thumb" and the boundary conditions under which such rules apply. ❑

KRISTINA CARTER is a graduate of Berea College and has been serving as an Adjunct Researcher with the Institute for Defense Analyses (IDA) in Alexandria, VA, since her completion of the IDA summer associate program in 2017. She is a Ph.D. candidate in Cognitive Psychology at Ohio University and her research focuses on investigating human judgment and information usage using individual-based modeling. She looks forward to completing her dissertation *comparing automatic variable selection methods in the Spring of 2019. Email: kcarter@ida.org.*

HEATHER WOJTON holds a B.A. in Psychology from Marietta College and earned a Ph.D. in Experimental Psychology from The University of Toledo in 2015. She has been a Research Staff Member at the Institute for Defense Analyses since 2015 and is currently serving as Technical Advisor to the Office of the Secretary of Defense/DOT&E. Email: hwojton@ida.org.

STEPHANIE LANE is a Research Staff Member at the Institute for Defense Analyses in Alexandria, VA. She completed her Ph.D. in Quantitative Psychology at the University of North Carolina at Chapel Hill, with a minor in Biostatistics from the Gillings School of Global Public Health. Her research spans latent variable and hierarchical models of change, with an emphasis on variable selection methodology. She is the author of multiple open source software packages that aim to disseminate novel methods to the broader research community. Her research has appeared in outlets such as Structural Equation Modeling, Psychological Methods, and Multivariate Behavioral Research. Email: slane@ida.org.

References

Barlett, J. E., J. W. Kotrlik, and C. C. Higgins. 2001. "Organizational Research: Determining Appropriate Sample Size in Survey Research." *Information Technology, Learning, and Performance Journal* 19 (1): 43-50.

Bell, B. A., G. B. Morgan, J. A. Schoeneberger, B. L. Loudermilk, J. D. Kromrey, and J. M. Ferron. 2010. "Dancing the Sample Size Limbo with Mixed Models: How Low Can You Go." In *Proceedings SAS Global Forum*, April 11-14, Seattle, WA, United States, chair L. Haworth, paper 197-2010.

Browne, W. J., and D. Draper. 2006. "A Comparison of Bayesian and Likelihood-Based Methods for Fitting Multilevel Models." *Bayesian Analysis* 1 (3): 473-514.

Button, K. S., J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò. 2013. "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience." *Nature Reviews Neuroscience* 14 (5): 365-376.

Caulton, D. A. 2001. "Relaxing the Homogeneity Assumption in Usability Testing." *Behaviour and Information Technology* 20 (1): 1-7.

Clarke, P. 2008. "When Can Group Level Clustering be Ignored? Multilevel Models versus Single-Level Models with Sparse Data." *Journal of Epidemiology and Community Health* 62 (8): 752-758.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. New Jersey: Lawrence Earlbaum Associates.

Dedrick, R. F., J. M. Ferron, M. R. Hess, K. Y. Hogarty, J. D. Kromrey, T. R. Lang, J. D. Niles, and R. S. Lee. 2009. "Multilevel Modeling: A Review of Methodological Issues and Applications." *Review of Educational Research* 79 (1): 69-102.

Director of Operational Test and Evaluation (DOT&E). 2018. *DOT&E FY 2018 Annual Report*. Department of Defense.

de Leeuw, J., and I. Kreft. 1986. "Random Coefficient Models for Multilevel Analysis." *Journal of Educational Statistics* 11 (1): 57-85.

Faulkner, L. (2003). "Beyond the Five-User Assumption: Benefits of Increased Sample Sizes in Usability Testing." *Behavior Research Methods* 35 (3): 379-383.

Frison, L., and S. J. Pocock. 1992. "Repeated Measures in Clinical Trials: Analysis using Mean Summary Statistics and its Implications for Design." *Statistics in Medicine* 11 (13): 1685-1704.

Gelman, A. 2005. "Analysis of Variance—Why it is More Important than Ever." *The Annals of Statistics* 33 (1): 1-53.

Gelman, A. (2006). "Prior Distributions for Variance Parameters in Hierarchical Models (comment on article by Browne and Draper)." *Bayesian Analysis* 1 (3): 515-534.

Gueorguieva, R., and J. H. Krystal. 2004. "Move Over Anova: Progress in Analyzing Repeated-Measures Data and its Reflection in Papers Published in the Archives of General Psychiatry." *Archives of General Psychiatry* 61 (3): 310-317.

Hawkins, N. G., R.W. Sanson-Fisher, A. Shakeshaft, C. D'Este, and L. W. Green. 2007. "The Multiple Baseline Design for Evaluating Population-Based Research." *American Journal of Preventive Medicine* 33 (2): 162-168.

Hedges, L. V., and E. C. Hedberg. 2007. "Intraclass Correlation Values for Planning Group-Randomized Trials in Education." *Educational Evaluation and Policy Analysis* 29 (1): 60-87.

Hox, J. 1998. "Multilevel Modeling: When and Why." In *Classification, Data Analysis, and Data Highways*, ed. Ingo Balderjahn, Rudolf Mathar and Martin Schader, 147-154. Berlin: Springer.

Hox, J. J., M. Moerbeek, and R. van de Schoot. 2010. *Multilevel Analysis: Techniques and Applications*. New York: Routledge.

Hwang, W., and G. Salvendy. (2010). "Number of People Required for Usability Evaluation: The 10±2 Rule." *Communications of the ACM* 53 (5): 130-133.

Ioannidis, J. P. 2005. "Why Most Published Research Findings are False." *PLoS Medicine* 2 (8): e124.

Ioannidis, J. P. 2011. "Excess Significance Bias in the

Literature on Brain Volume Abnormalities." *Archives of General Psychiatry* 68 (8): 773-780.

Keselman, H. J., C. J. Huberty, L. M. Lix,S. Olejnik, R.A. Cribbie, B. Donahue, R. K. Kowalchuck, L. L. Lowman, M. D. Petoskey, J.C. Keselman, and J. R. Levin. 1998. "Statistical Practices of Educational Researchers: An Analysis of Their ANOVA, MANOVA, and ANCOVA Analyses." *Review of Educational Research* 68 (3): 350-386.

Kreft, I. G., I. Kreft, I., and J. de Leeuw. 1998. *Introducing Multilevel Modeling*. London: Sage.

Kwok, O. M., A. T. Underhill, J. W. Berry, W. Luo, T. R. Elliott, and M. Yoon. 2008. "Analyzing Longitudinal Data with Multilevel Models: An Example with Individuals Living with Lower Extremity Intra-Articular Fractures." *Rehabilitation Psychology* 53 (3): 370-386.

Lewis, J. R. 1994. "Sample Sizes for Usability Studies: Additional Considerations." *Human Factors* 36 (2): 368-378.

Maas, C. J., and J. J. Hox. 2004. "The Influence of Violations of Assumptions on Multilevel Parameter Estimates and their Standard Errors." *Computational Statistics and Data Analysis* 46 (3): 427-440.

Maas, C. J., and J. J. Hox. 2005. "Sufficient Sample Sizes for Multilevel Modeling." *Methodology* 1 (3): 86-92.

McNeish, D. M., and L. M. Stapleton. 2016. "The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration." *Educational Psychology Review* 28 (2): 295-314.

Nielsen, J., and R. Molich. 1990. March. "Heuristic Evaluation of User Interfaces." In *Proceedings of the ACM CHI'90 Conference*, April 1-5, Seattle, WA, United States, 249-256.

Rabe-Hesketh, S., and A. Skrondal. 2008. *Multilevel and Longitudinal Modeling Using Stata*. Texas: Stata Press.

Raudenbush, S. W., and A. S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods.* California: Sage Publications, Inc.

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): 804-824.

Sawilowsky, S. S. 2009. "New Effect Size Rules of Thumb." *Journal of Modern Applied Statistical Methods* 8 (2): 597-599.

Searle, S. R., G. Casella, and C. E., McCulloch. 1992. *Variance Components.* New Jersey: John Wiley and Sons, Inc.

Singer, J. D., and J. B. Willett. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* New York: Oxford University Press, Inc.

Snijders, T. A., and R. J. Bosker. 1994. "Modeled Variance in Two-Level Models." *Sociological Methods and Research* 22 (3): 342-363.

Virzi, R. A. 1992. "Refining the Test Phase of Usability Evaluation: How Many Subjects is Enough?" *Human Factors* 34 (4), 457-468.

Watson, P. J., and E. A. Workman. 1981. "The Non-Concurrent Multiple Baseline Across-Individuals Design: An Extension of the Traditional Multiple Baseline Design." *Journal of Behavior Therapy and Experimental Psychiatry* 12 (3): 257-259.

West, B. T., K. B. Welch, and A. T. Galecki. 2014. *Linear Mixed Models: A Practical Guide using Statistical Software.* Florida: CRC Press.

Supplementary Material: Appendix A

```
Analysis R Code
#Generating Mixed Model Simulation Data
rm(list = ls())
set.seed(20170712)
install.packages("tcltk2")
install.packages("tidyverse")
library("tcltk2")
library(tidyverse)

############Simulating Repeated Measures Design######################
repeated_measures_design <- function(ns, nm){
  subjects <- list()
  measures <- list()
  ordermeasures <- list()
  conditions <- list()
  mescond <- list()
  subjectset <- list()
  subjectList <- list()
  index <- ns^2 + nm
  total <- ns*nm
  replicate <- total - ns

  subjects[[index]] <- list(ind = as.character(c(rep(1:i, j))))
  measures[[index]] <- tibble(mes = c(rep(1:j, i)))
  ordermeasures[[index]] <- with(measures[[index]], measures[[index]][order(mes) , ])
  conditions[[index]] <- tibble(cond = as.factor(c(rep(0, replicate), rep(1, i))))
  mescond[[index]] <- bind_cols(ordermeasures[[index]], conditions[[index]])
  subjectset[[index]] <- bind_cols(subjects[[index]], mescond[[index]])
  subjectList[[index]] <- with((subjectset[[index]]), subjectset[[index]][order(ind) , ])
  subjectList[[index]]

}

ns <- 4:30 #varying number of subjects per design
nm <- 2:10  #varying number of measures per design

datalist <- list() #creating list of data tables for each test design condition
for (i in ns){
  for (j in nm){
    index <- i^2 + j
    datalist[[index]] <- repeated_measures_design(i, j)
  }
}

fulldatalist <- datalist[vapply(datalist, Negate(is.null), NA)] #removing non-populated conditions
#^also renames lists according to their index position, names now more intuitive
#e.g., fulldatalist[[1]] is your first smallest subject-observation list sequenced by observation increases then subject ones

randomEffects <- "ind" #generating empty random effects matrix
fixedEffects <- tibble("ind", "cond") #generating empty fixed effects matrix

#function for populating fixed model matrix:
create_fixed_model_matrix <- function(tb, fixedEffects){

  myFormula <- formula(paste(c("~", paste(fixedEffects, collapse = "+")), collapse = ""))
  model.matrix(myFormula, tb)
}

#applying above function across all sampling conditions to create fixed model matrices for all designs
fixed_model_matrices <- lapply(fulldatalist, function(x) create_fixed_model_matrix(x, fixedEffects))

#function for populating random model matrix:
create_random_model_matrix <- function(tb, randomEffects){

  myFormula <- formula(paste(c("~ -1 + ", randomEffects, collapse = "")))
  model.matrix(myFormula, tb)

}

#applying above function across all sampling conditions to create random model matrice for all designs
random_model_matrices <- lapply(fulldatalist, function(x) create_random_model_matrix(x, randomEffects))

#####################Generating data across multiple conditions######################
#Depending on the number of data points you are trying to generate this process will need parallelized
#Alternately, you can run each iteration of whatever parameters you are altering below separately
#Without parallelizing:
#500 data points across the original 243 sampling conditions, 4 SNR conditions, and 3 devmod conditions takes about
20min
#10,000 datapoints across the above conditions will likely crash your computer

set.seed(20170712)#seed used for initial 500 runs
set.seed(714)#seed used for subsequent 500 runs

#function for generating dependent variables
```

```
generate_alt <- function(fulldatalist, fixed_model_matrices, random_model_matrices,
            n_sim_rep, SNR, stdev, devmod, rvName = "y"){
 dvlist <- list()
 for(i in 1:length(fulldatalist)){
  tb <- fulldatalist[i][[1]]
  X <- fixed_model_matrices[i][[1]]
  Z <- random_model_matrices[i][[1]]
  n <- nrow(tb)
    for(j in 1:n_sim_rep){
     beta <- c(0, rep(SNR*stdev, ncol(X) - 1))
        gamma <- rnorm(ncol(Z), 0, stdev * devmod) #devmod to increase or decrease second-level variance
      y <- X %*% beta + Z %*% gamma + rnorm(n, 0, stdev)
   tb[[paste(rvName, j, sep = "_")]] <- y
  }
  dvlist[[i]] <- tb
 }
 dvlist
}

#function for varying designated parameters for determined number of generations
model <- function(fulldatalist, fixed_model_matrices, random_model_matrices, n_sim_rep,
         SNRs, stdev, devmods) {
 t1 <- Sys.time()
 SNRlist <- list()
 devmodlist <- list()
 SDRats <- list()

 total <- (length(devmods)*length(SNRs))
 # create progress bar
 pb <- tkProgressBar(title = "progress bar", min = 0,
           max = total, width = 300)

 for (i in 1:length(SNRs)){
  SNR <- SNRs[[i]]
  for (j in 1:length(devmods)){
   k <- j + ((i - 1) * length(devmods))
    devmod <- devmods[[j]]
   SDRats[[j]] = round (stdev * devmods[[j]] / ((stdev * devmods[[j]]) + stdev), digits = 4)
      devmodlist[[j]] <- generate_alt(fulldatalist, fixed_model_matrices, random_model_matrices,
       n_sim_rep, SNR, stdev, devmod)
   save(devmodlist, file = #'Enter file location and name.Rda')
   Sys.sleep(0.1)
   setTkProgressBar(pb, k, label=paste(round(k/total*100, 0),"% done"))
  }

  SNRlist[[i]] <- devmodlist
  save(SNRlist, file = #'Enter file location and name.Rda')
  names(SNRlist[[i]]) <- paste0("SDRat_", SDRats)#
 }
 names(SNRlist) <- paste0("SNR_",SNRs)#assigns name of list
 print(Sys.time()-t1)
 SNRlist
}

#Generation commanded here
#Assigns the completed lists of tibbles to a final complete list
devlist <- list()
dvlist <- model(fulldatalist, fixed_model_matrices, random_model_matrices, n_sim_rep = 500,
        SNRs = c(0, .3, .5, .8, 1), stdev = 20, devmods = c(.0811, .3335, 1, 4))#should change devmods to desired modifiers
#saving complete dependent variable file
save(dvlist, file = '#Enter file name and location.Rda')
#if not parallelizing mixed modeling, may want to save separately as exemplified below for SNR = 0
#saving SNR 0 files
dvlist_SNR0_ICC0.075 <- dvlist$SNR_0$SDRat_0.075
save(dvlist_SNR0_ICC0.075, file = #'Enter file location and name.Rda ')
rm(dvlist_SNR0_ICC0.075)
```

```
dvlist_SNR0_ICC0.25 <- dvlist$SNR_0$SDRat_0.25
save(dvlist_SNR0_ICC0.25, file = #'Enter file location and name.Rda')
rm(dvlist_SNR0_ICC0.25)
dvlist_SNR0_ICC0.5 <- dvlist$SNR_0$SDRat_0.5
save(dvlist_SNR0_ICC0.5, file = #'Enter file location and name.Rda')

####################Running Mixed Model Analysis##############################
rm(list = ls())
library(tidyverse)
library(lme4)
library(nlme)

##Load your datalist###
datalist <- dvlist_SNR0.3_ICC0.075 #Assign your specific datalist to term

#Below is for the event you have not parallelized:
#Here you can easily designate your list terms from your different datalists
#Without paralellizing, each datalist of 243*500*1(SNR)*1(ICC) datapoints took approximately 4hrs
#this with each of 243 sampling conditions (500 data sets) taking approximately 1min each to model if they are taking longer
than 1min (time is printed) you can anticipate longer total time
############################################################################
#Designate your list terms, in this case, SNR and ICC
SNR <- .3
ICC <- .075
rm(dvlist_SNR0.3_ICC0.075)

#######################Mixed Model Analysis##################################
model_list <- list()

list_names <- list()

#Beginning your mixed model analysis
for (j in 1:length(datalist)){

  t1 <- Sys.time()

  df2 <- data_frame("SNR" = SNR, "ICC" = ICC)
  df2[1, "Sj"] <- length(unique(datalist[[j]]$ind))
  df2[1, "Mes"] <- length(unique(datalist[[j]]$mes))

  fit.model <- function(x){

    df3 <- data.frame(ind = datalist[[j]]$ind, cond = datalist[[j]]$cond)
    df3$y <- x

    #Models and tests
    m1 <- nlme::gls(y ~ 1, data = df3)#initial model, without intercept

    m2 <- nlme::lme(y~1, data=df3, random = ~ 1|ind)#model with intercept

    m1vm2  <- anova(m2, m1)#loglikelihood test of intercept-only and non-intercept models

    m3 <- lmer(y ~ cond + (1|ind), data = df3, REML = FALSE)#model with fixed factor, using FML

    m2vm3 <- anova(lmer(y ~ 1 + (1|ind), data = df3, REML = FALSE), m3)
    #^loglikelihood tests of reduced (intercept-only) and full(with fixed factor) models, using FML

    #Summary objects
    m2var <- as.data.frame(summary(lmer(y ~ 1 + (1|ind), data = df3, REML = FALSE))$varcor)
    m3var <- as.data.frame(summary(m3)$varcor)
    m3sum <- as.data.frame(summary(m3)$coefficients)

    #Results saved to table, we'll want to add more results here in future
    df4 <- data_frame()
    df4[1, "m2_ind_var"] <- m2var %>% filter(grp == "ind") %>% select(vcov) %>% .[[1]]
    df4[1, "m2_resid_var"] <- m2var %>% filter(grp == "Residual") %>% select(vcov) %>% .[[1]]
    df4[1, "m2_chsq"] <- m1vm2$L.Ratio[2]
    df4[1, "m2_df"] <- m1vm2$df[1]-m1vm2$df[2]
    df4[1, "m2_p"] <- m1vm2$`p-value`[2]
    df4[1, "m3_ind_var"] <- m3var %>% filter(grp == "ind") %>% select(vcov) %>% .[[1]]
    df4[1, "m3_resid_var"] <- m3var %>% filter(grp == "Residual") %>% select(vcov) %>% .[[1]]
    df4[1, "m3_cond_b"] <- m3sum["cond", "Estimate"]
    df4[1, "m3_cond_SEb"] <- m3sum["cond", "Std. Error"]
    df4[1, "m3_chsq_p"] <- m2vm3$`Pr(>Chisq)`[2]
    df4[1, "m3_df"] <- m2vm3$`Chi Df`[2]

    df5 <- bind_cols(df2, df4)
    df5
  }

  model_list[[j]] <- bind_rows(apply(datalist[[j]][,4:503], 2, fit.model))
#^^above^^in [,4:503], first value (here 4) is column number you wish your mixed models to begin
####second value (here 503) should be column number where mixed models end (i.e. 1st value + total data sets)
  list_names[[j]] <- paste0("SJs_", length(unique(datalist[[j]]$ind)), "; Mes_", max(datalist[[j]]$mes))

  print(Sys.time()-t1)
}

names(model_list) <- list_names
##################Summarizing Data Results###############################

rm(list=ls())
library(tidyverse)


#####In case you did run your models separately, function for combining model files######

combine_files_df <- function(directory_path){
  require(dplyr)
  require(stringr)
  file_list <- list.files(path = directory_path,
                          pattern = "\\.rda", ignore.case = TRUE)
  data_list <- list()
  for(i in 1:length(file_list)){
  dataname <- load(file = if_else(str_detect(directory_path, "/$"), paste0(directory_path, file_list[i]),
  paste0(directory_path, "/", file_list[i])))
      data_list[[i]] <- get(dataname)
  }
    df <- bind_rows(data_list[[1]])
    for(i in 2:length(data_list)){
      df <- bind_rows(data_list[[i]]) %>%
      bind_rows(df)
    }
  df
}

directory_path <- "#Assign directory path where files are located"

All_Models_df <-
  combine_files_df(directory_path) %>%
  arrange(Sj, Mes)

save(All_Models_df, file = "All_Models_df.rda")

########################Compute Model Descriptives#########################

Simulation_summary <-
  All_Models_df %>%
    mutate(icc_actual = m2_ind_var/(m2_ind_var + m2_resid_var),
          relbias_b = ((m3_cond_b - (SNR*20))/(SNR*20))*100,#relative bias in fixed effect estimates
          bias_b = (m3_cond_b - (SNR*20)),#bias in fixed effect estimates
          propbias_b = (m3_cond_b/(SNR*20)),#proportion bias in fixed effect estimates
```

```
        ratio = ifelse(ICC == 0.5, 1, ifelse(ICC == 0.25, 0.3335, 0.0811)),#
        indvar = (20^2)*ratio,
        relbias_indvar = ((m2_ind_var - indvar)/indvar)*100,#bias in subject variance estimates
        bias_indvar = (m2_ind_var - indvar), #bias in subject variance estimates
propbias_indvar = (m2_ind_var/indvar),#computing proportion bias in subject variance estimates
        relbias_resvar = ((m2_resid_var - (20^2))/(20^2))*100, #relative bias in subject variance estimates
        bias_resvar = (m2_resid_var - (20^2)),#bias in subject variance estimates
        propbias_resvar = (m2_resid_var/(20^2)),#proportion bias in subject variance estimates
        relbias_icc = ((icc_actual - ICC)/ICC) * 100,#relative bias in ICC estimates
        bias_icc = (icc_actual - ICC),#bias in ICC estimates
        propbias_icc = (icc_actual/ICC))%>% #relative bias in ICC estimates

 #summarizing by parameters of interest
  group_by(SNR, ICC, Sj, Mes) %>%
  summarise(m.b = mean(m3_cond_b),
            m.relbias_b = mean(relbias_b),
            m.bias_b = mean(bias_b),
            m.propbias_b = mean(propbias_b),
            m.relbias_indvar = mean(relbias_indvar),
            m.bias_indvar = mean(bias_indvar),
            m.propbias_indvar = mean(propbias_indvar),
            m.relbias_resvar = mean(relbias_resvar),
            m.bias_resvar = mean(bias_resvar),
            m.propbias_resvar = mean(propbias_resvar),
            m.icc_actual = mean(icc_actual),
            m.relbias_icc = mean(relbias_icc),
            m.bias_icc = mean(bias_icc),
            m.propbias_icc = mean(propbias_icc),
            power.b.05 = (sum(round(m3_chsq_p, 2) <= .05))/length(m3_chsq_p),
            power.b.1 = (sum(round(m3_chsq_p, 2) <= .1))/length(m3_chsq_p),
            power.b.15 = (sum(round(m3_chsq_p, 2) <= .15))/length(m3_chsq_p),
            power.b.2 = (sum(round(m3_chsq_p, 2) <= .2))/length(m3_chsq_p),
            power.icc.05 = (sum(round(m2_p, 2) <= .05))/length(m2_p),
            power.icc.2 = (sum(round(m2_p, 2) <= .2))/length(m2_p))

save(Simulation_summary, file = '#Enter File Name and Location Here.Rda')
```