PSY 503: Foundations of Psychological Methods
Lecture 20: Multilevel Models

Robin Gomila

Princeton

November 16-18, 2020

The independence assumption

- So far, we have operated in settings that do not violate the independence assumption
- What does this mean?
  - No form of connection between data points (within a variable)
  - e.g., rolling a die, flipping a coin
- Examples of study designs that violate the independence assumption?
  - Repeated measures: more than one observation per participant
  - Data scraping: Tweets (observations) may come from the same twitter account
  - Corpus linguistics: multiple data points from the same text or author
  - Cluster-randomized experiments: random assignment of communities, schools, classroom to experimental conditions

Consequences of violation of independence assumption

- If the regression analysis does not explicitly account for non-independence:
  - Standard errors are biased downwards! i.e., they are unrealistically small.
  - Therefore, p-values are also biased downwards! i.e., they are unrealistically small too.
- As a result, **non-independence increases the probability of false positives**

Misconceptions about non-independence

- Non-independence **DOES NOT** bias estimates of treatment effects
  - i.e., regression coefficients will not be biased if you don't include random intercepts or random slopes in your model (more soon!)
- **UNLESS:**
  - small number of clusters AND unequal cluster sizes AND cluster size covaries with potential outcomes (Gerber & Green, 2012, p.83; see also Green & Vavrek, 2008)
  - If you are in this situation, check out Middleton & Aronow (2011)

Consequences of violation of independence assumption

- In large enough samples, violations of independence assumption is about standard errors, just like heteroskedasticity
- But non-independence is a much much more important issue than heteroskedasticity because the impact on the standard error is consequential
- Why is that?

Consequences of violation of independence assumption

- Suppose an experimental design with “repeated measures”
- Observations from the same participant are more similar to each other than observations from different participants
- As a result, residuals become clustered
  - All of the residuals of each participant act as a group
- This “misleads” statistical inference
  - Equation for classic standard errors not appropriate. i.e., it does not produce a realistic estimate of the standard deviation of the sampling distribution of the parameter!
  - Sample size is artificially inflated
  - Estimates of the parameters “seem” more precise than they actually are
- This is why standard errors are smaller than they should be to reflect “empirical standard errors”

When the independence assumption is violated. . .

- Aggregation? Could we average observations from the same cluster?
  - Resolves the non-independence issue because we end up with one data point per participant
  - This used to be the main way to deal with non-independence in many fields
  - Not optimal: We lose some information when we aggregate data —the variation across the non-independent cases is not retained in the final analysis

When the independence assumption is violated...

- Inform your analysis about non-independence in your data
- Use analytic strategy that allows you to incorporate non-independent clusters of data into your regression analysis
- Objective: draw **appropriate** statisitical inferences

When the independence assumption is violated...

- Two possible ways to go:
  - Specify **clustered standard errors** in your usual lm_robust() function
  - Mixed models / random effects using the `lmer` from the lme4 package
- How to decide what to do?
  - It depends on your study design and what sources of variation you need to account for

Clustered standard errors

- Use clustered SEs usually used in designs in which a treatment is assigned in clusters
    - assignment of classrooms, schools, groups to an experimental condition
    - i.e., inference at the cluster level but you are getting multiple data points per cluster
- In these cases, use the argument `clusters =` in the following way:

```r
lm_robust(Y ~ Z,
          clusters = classroom,
          data = dat)
```

Clustered standard errors

- Using clustered SEs basically indicates to the analysis that different clusters may have different intercepts
- In the mixed models framework, this is equivalent to models with "random intercepts" or "varying intercepts"
  - A mixed model with random intercepts for clusters in lmer will produce the exact same output!
  - Similarly, clustered SEs for participants in a "repeated measures" design will produce the exact same result as a mixed model with lmer with "random intercepts for participants"

Mixed models: “fixed effects” vs. “random effects”

- Differentiate “fixed effects” and “random effects”
- We are very familiar with fixed effects
  - In the past few weeks, we have been fitting “fixed-effects-only” models
  - All the coefficients that we have looked at so far were “fixed effects” (e.g., donation, income, Z, gender)
- When we turn to mixed models, we can specify “random effects”
  - Random effects constitute different **sources and forms** of non-independence in the data

Examples of random effects

- Different participants may have different intercepts (“random intercepts”)
- Different items may have different intercepts (“random intercepts”)
- The relationship between Y and a fixed effect predictor (e.g., Z) may vary by participant (“random slopes”)
- The relationship between Y and a fixed effect predictor (e.g., Z) may vary by item (“random slopes”)

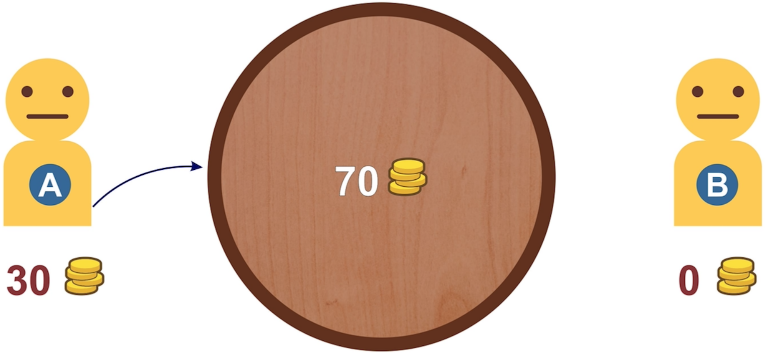
Let's try to understand what this means

Random effects: Illustration with a trust game study

- Imagine a study testing the effect of looking trustworthy vs. neutral on trust decisions in a behavioral game called the trust game
- This game involves two players: Player A (first mover) and Player B (second mover)
- At the beginning of the game, Player A is endowed with 100 Monetary Units (MUs)

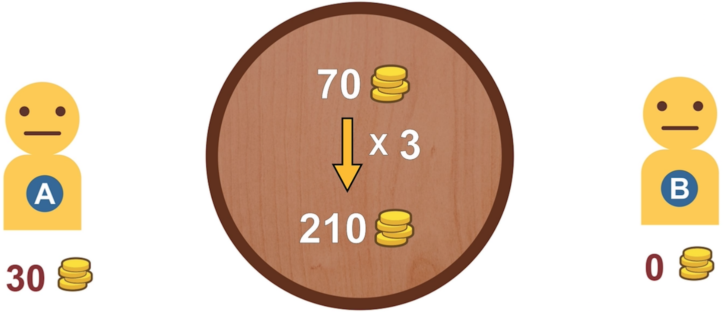
Random effects: Illustration with a trust game study

- Player A is the first mover and decides how much of their endowment to send to Player B



Random effects: Illustration with a trust game study

- Player A's contribution is multiplied by 3 before arriving in the hands of Player B

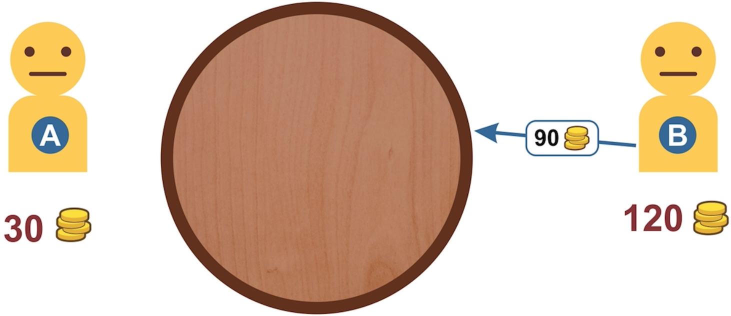


Random effects: Illustration with a trust game study

- Player A's contribution is multiplied by 3 before arriving in the hands of Player B

Random effects: Illustration with a trust game study

- Finally, Player B decides how much to send to Player A



Random effects: Illustration with a trust game study

- In the present case, both players would end the round with the same amount of MUs

Random effects: Illustration with a trust game study

- In this hypothetical study:
    - All participants play with a bot
    - All participants are assigned to the role of Player A
    - Participants believe that they see the picture of their game partner

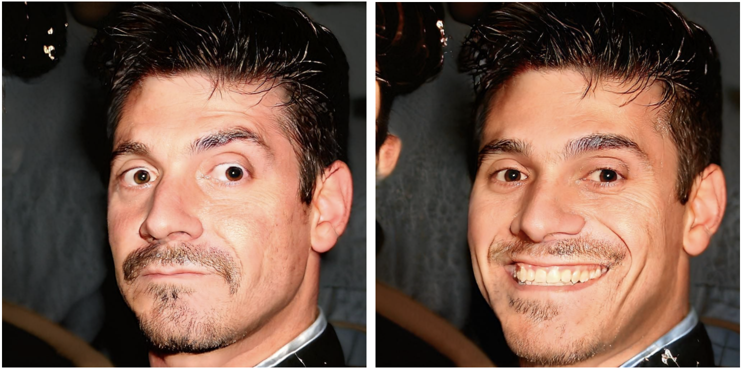
Random effects: Illustration with a trust game study

- Researchers randomly assign participants to one of two experimental conditions:
  - Control condition: Play with a neutral looking face
  - Treatment condition: Play with a trustworthy looking face
- Participants randomly assigned to play with 5 different faces from a pool of 10 faces
  - Either 5 out of 10 neutral faces
  - Or 5 out of 10 trustworthy faces

Random effects: Illustration with a trust game study

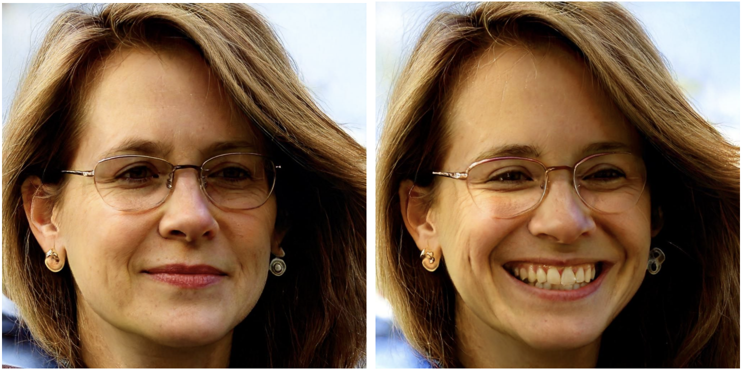
Random effects: Illustration with a trust game study

Random effects: Illustration with a trust game study



Random effects: Illustration with a trust game study

Random effects: Illustration with a trust game study



Random effects: Illustration with a trust game study

- Quantity of interest: Average Treatment Effect
- Outcome variable Y: Amount of MUs that participants “invest” in Player B
- Random effects:
  - Random intercept for participants
  - Random intercepts for items
  - Random slopes for items

Random effects: Illustration with a trust game study

- Our plan:
  - Open R Studio
  - Generate population data—including these “random effects”
  - Study how different analytic strategies “perform” with regard to estimating the ATE
- Let’s do it!

Mixed models in R

- Most widely used R package for random effects is `lme4`
- Syntax:

```r
lmer(Y ~ Z + (1 | id),
     data = dat)
```

Mixed models in R

```r
lmer(Y ~ Z + (1 | id),
     data = dat)
```

- This model estimates the (fixed) effect of Z on Y, allowing intercepts to vary by participants
- `Y ~ Z` looks familiar: estimates treatment effect
- `(1 | id)` allows for random intercepts “conditional on” / “with respect to” participants

Random effects vs. Clustered standard errors

- When robust and classic SEs agree, the following two functions yield identical inferences:

```r
lmer(Y ~ Z + (1 | id),
     data = dat)
```

```r
lm_robust(Y ~ Z,
          clusters = id,
          data = dat)
```

Mixed models in R

```r
lmer(Y ~ Z + (1 | id) + (1 | item),
     data = dat)
```

- This model estimates the (fixed) effect of Z on Y, allowing intercepts to vary with respect to participants and items
- For the ongoing hypothetical trust study, the mixed model could represent the population data that we generated even better
  - We could include varying slopes with respect to items

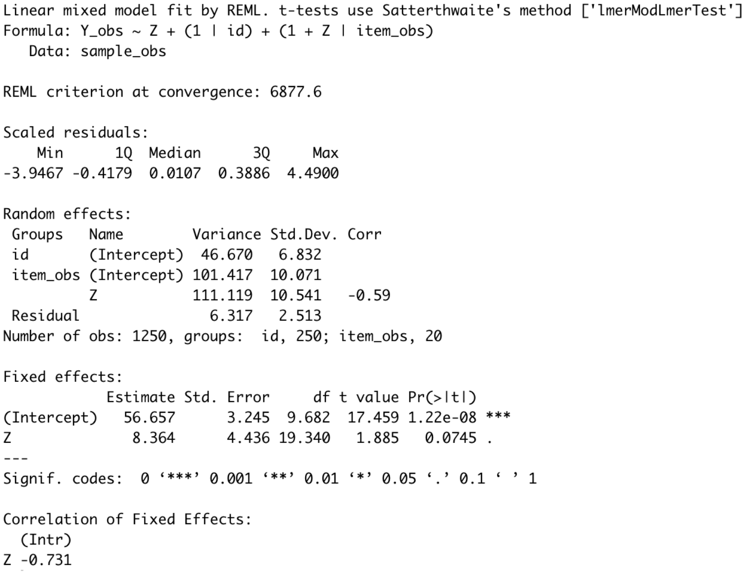
Mixed models in R

- The model that best represents the population data is

```
lmer(Y ~ Z + (1 | id) + (1 + Z | item),
     data = dat)
```

- Let's look at the main elements of this regression output
  - Note that the output produced on the next page uses `summary()` and requires loading the `lmerTest` package

Mixed models in R



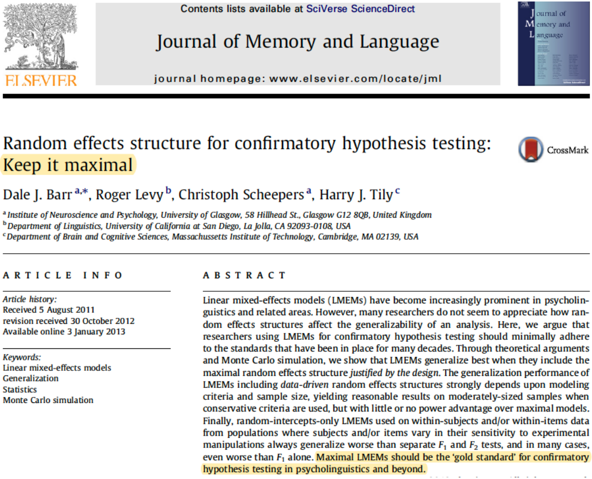
```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: Y_obs ~ Z + (1 | id) + (1 + Z | item_obs)
   Data: sample_obs

REML criterion at convergence: 6877.6

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-3.9467 -0.4179  0.0107  0.3886  4.4900 

Random effects:
 Groups   Name        Variance Std.Dev. Corr 
 id       (Intercept)  46.670   6.832        
 item_obs (Intercept) 101.417  10.071        
          Z           111.119  10.541   -0.59
 Residual               6.317   2.513        
Number of obs: 1250, groups:  id, 250; item_obs, 20

Fixed effects:
            Estimate Std. Error     df t value Pr(>|t|)    
(Intercept)   56.657      3.245  9.682  17.459 1.22e-08 ***
Z              8.364      4.436 19.340   1.885   0.0745 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
  (Intr)
Z -0.731
```

Should you “keep it maximal”?

- Fully specified (a.k.a. “maximal”) mixed models often result in “combinatorial explosions” (Winter, 2019) and often lead to so called “convergence issues”
- Example: The “maximal model” previously displayed for the present trust study led to convergence issues for about 90% of the simulations with 250 participants
  - That is, even though the model actually perfectly represents the underlying structure of the data
- Yet, you’ll often hear: “keep it maximal”
  - Let’s try to understand where this idea comes from!

Should you “keep it maximal”?

- Keeping it maximal became the norm / “the right thing to do” in 2013 after this article was published



Contents lists available at SciVerse ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Random effects structure for confirmatory hypothesis testing: Keep it maximal

Dale J. Barr[a,*], Roger Levy[b], Christoph Scheepers[a], Harry J. Tily[c]

[a] *Institute of Neuroscience and Psychology, University of Glasgow, 58 Hillhead St., Glasgow G12 8QB, United Kingdom*
[b] *Department of Linguistics, University of California at San Diego, La Jolla, CA 92093-0108, USA*
[c] *Department of Brain and Cognitive Sciences, Massachussetts Institute of Technology, Cambridge, MA 02139, USA*

ARTICLE INFO

Article history:
Received 5 August 2011
revision received 30 October 2012
Available online 3 January 2013

Keywords:
Linear mixed-effects models
Generalization
Statistics
Monte Carlo simulation

ABSTRACT

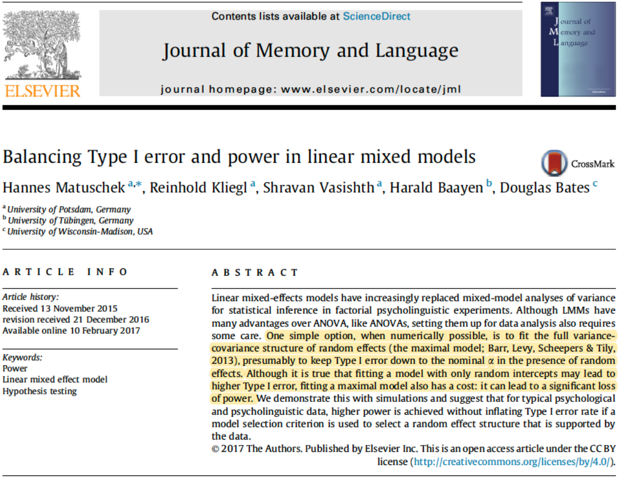
Linear mixed-effects models (LMEMs) have become increasingly prominent in psycholinguistics and related areas. However, many researchers do not seem to appreciate how random effects structures affect the generalizability of an analysis. Here, we argue that researchers using LMEMs for confirmatory hypothesis testing should minimally adhere to the standards that have been in place for many decades. Through theoretical arguments and Monte Carlo simulation, we show that LMEMs generalize best when they include the maximal random effects structure *justified by the design*. The generalization performance of LMEMs including *data-driven* random effects structures strongly depends upon modeling criteria and sample size, yielding reasonable results on moderately-sized samples when conservative criteria are used, but with little or no power advantage over maximal models. Finally, random-intercepts-only LMEMs used on within-subjects and/or within-items data from populations where subjects and/or items vary in their sensitivity to experimental manipulations always generalize worse than separate F_1 and F_2 tests, and in many cases, even worse than F_1 alone. Maximal LMEMs should be the 'gold standard' for confirmatory hypothesis testing in psycholinguistics and beyond.

Should you “keep it maximal”?

- Authors argue that the gold standard is to include all possible random effects in the model
- The idea is to prevent false positives
- OK but we saw that preventing false positives by incurring a penalty on SEs / p-value impacts the probability of false negatives
  - Remember that false positives can only occur when the null is true
  - What happens when the null is not true? Penalty on standard errors leads to larger p-values and therefore, lower statistical power
    - Increase in false negatives when null is true

Should you “keep it maximal”?

- That’s why 5 years later, same journal:



Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Balancing Type I error and power in linear mixed models

Hannes Matuschek[a,*], Reinhold Kliegl[a], Shravan Vasishth[a], Harald Baayen[b], Douglas Bates[c]

[a] *University of Potsdam, Germany*
[b] *University of Tübingen, Germany*
[c] *University of Wisconsin-Madison, USA*

ARTICLE INFO

Article history:
Received 13 November 2015
revision received 21 December 2016
Available online 10 February 2017

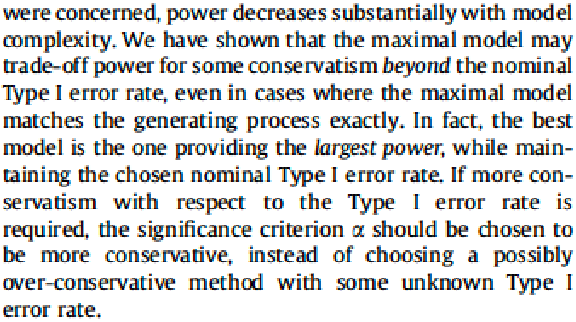
Keywords:
Power
Linear mixed effect model
Hypothesis testing

ABSTRACT

Linear mixed-effects models have increasingly replaced mixed-model analyses of variance for statistical inference in factorial psycholinguistic experiments. Although LMMs have many advantages over ANOVA, like ANOVAs, setting them up for data analysis also requires some care. One simple option, when numerically possible, is to fit the full variance-covariance structure of random effects (the maximal model; Barr, Levy, Scheepers & Tily, 2013), presumably to keep Type I error down to the nominal α in the presence of random effects. Although it is true that fitting a model with only random intercepts may lead to higher Type I error, fitting a maximal model also has a cost: it can lead to a significant loss of power. We demonstrate this with simulations and suggest that for typical psychological and psycholinguistic data, higher power is achieved without inflating Type I error rate if a model selection criterion is used to select a random effect structure that is supported by the data.

Should you “keep it maximal”?

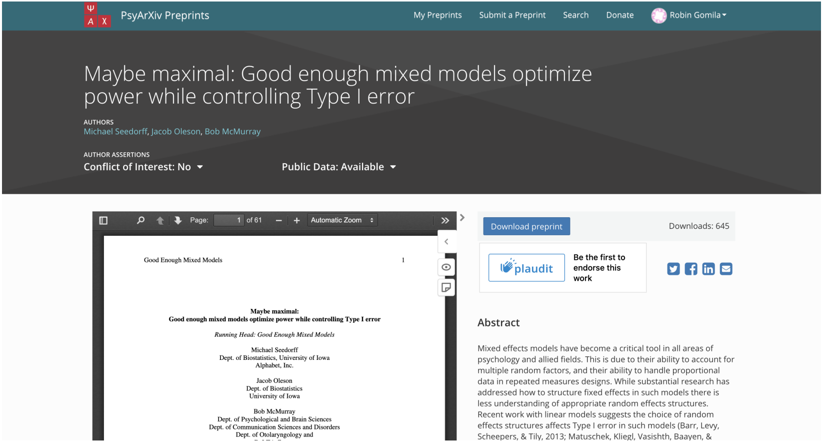
- The authors conclude:



were concerned, power decreases substantially with model complexity. We have shown that the maximal model may trade-off power for some conservatism *beyond* the nominal Type I error rate, even in cases where the maximal model matches the generating process exactly. In fact, the best model is the one providing the *largest power*, while maintaining the chosen nominal Type I error rate. If more conservatism with respect to the Type I error rate is required, the significance criterion α should be chosen to be more conservative, instead of choosing a possibly over-conservative method with some unknown Type I error rate.

Should you “keep it maximal”?

- Finally, check out this new preprint posted in August 2020:



Maybe maximal: Good enough mixed models optimize power while controlling Type I error

AUTHORS
Michael Seedorff, Jacob Oleson, Bob McMurray

AUTHOR ASSERTIONS
Conflict of Interest: No
Public Data: Available

Good Enough Mixed Models 1

Maybe maximal:
Good enough mixed models optimize power while controlling Type I error

Running Head: Good Enough Mixed Models

Michael Seedorff
Dept. of Biostatistics, University of Iowa
Alphabet, Inc.

Jacob Oleson
Dept. of Biostatistics
University of Iowa

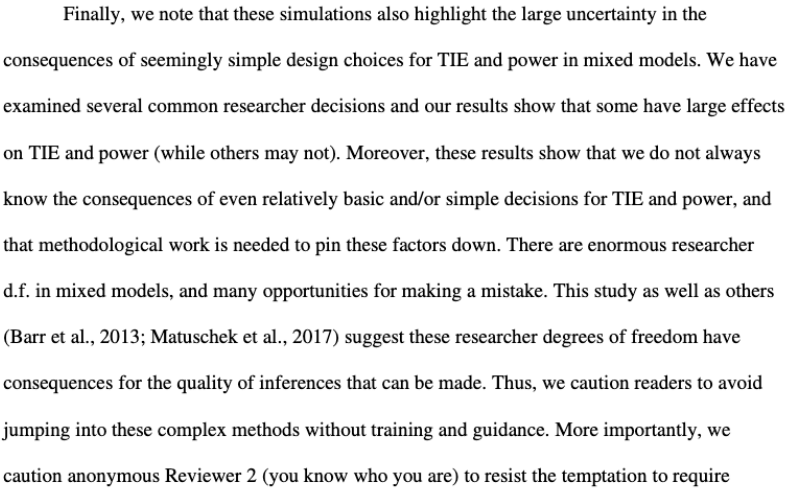
Bob McMurray
Dept. of Psychological and Brain Sciences
Dept. of Communication Sciences and Disorders
Dept. of Otolaryngology and

Abstract

Mixed effects models have become a critical tool in all areas of psychology and allied fields. This is due to their ability to account for multiple random factors, and their ability to handle proportional data in repeated measures designs. While substantial research has addressed how to structure fixed effects in such models there is less understanding of appropriate random effects structures. Recent work with linear models suggests the choice of random effects structures affects Type I error in such models (Barr, Levy, Scheepers, & Tily, 2013; Matuschek, Kliegl, Vasishth, Baayen, &

Should you “keep it maximal”?

- The answer is: it depends! From this last preprint’s discussion



Finally, we note that these simulations also highlight the large uncertainty in the consequences of seemingly simple design choices for TIE and power in mixed models. We have examined several common researcher decisions and our results show that some have large effects on TIE and power (while others may not). Moreover, these results show that we do not always know the consequences of even relatively basic and/or simple decisions for TIE and power, and that methodological work is needed to pin these factors down. There are enormous researcher d.f. in mixed models, and many opportunities for making a mistake. This study as well as others (Barr et al., 2013; Matuschek et al., 2017) suggest these researcher degrees of freedom have consequences for the quality of inferences that can be made. Thus, we caution readers to avoid jumping into these complex methods without training and guidance. More importantly, we caution anonymous Reviewer 2 (you know who you are) to resist the temptation to require

Should you “keep it maximal”?

- Last part of R code compares the estimates and standard errors of different models
    - Keep in mind that the in the present simulated population, there exist a treatment effect that we are trying to figure out
    - So we are looking at one side of the coin: Power to detect an existing effect and probability of false negatives
    - And our setup does not include a lot of repeated measures!