Avoiding the language-as-a-fixed-effect fallacy: How to estimate outcomes of linear mixed models

Sterling Hutchinson (S.C.Hutchinson@tilburguniversity.nl)
Tilburg Centre for Cognition and Communication (TiCC), Tilburg University
PO Box 90153, 5000 LE, Tilburg, The Netherlands

Lei Wei (Lei.Wei@roswellpark.org)
Department of Biostatistics and Bioinformatics, Roswell Park Cancer Institute
Buffalo, NY 14263 USA

Max M. Louwerse (mlouwerse@tilburguniversity.nl)
Tilburg Centre for Cognition and Communication (TiCC), Tilburg University
PO Box 90153, 5000 LE, Tilburg, The Netherlands

Abstract

Since the 1970s, researchers in psycholinguistics and the cognitive sciences have been aware of the language-as-fixed-effect fallacy, or the importance in statistical analyses to not only average across participants (F_1) but also across items (F_2). Originally, the language-as-fixed-effect fallacy was countered by proposing a combined measure (*minF'*) calculated by participant (F_1) and item (F_2) analyses. The scientific community, however, reported separate participant and item (F_1 and F_2) regression analyses instead. More recently, researchers have started using linear mixed models, a more robust statistical methodology that considers both random participant and item factors together in the same analysis. There are various benefits to using mixed models, including being more robust to missing values and unequal cell sizes than other linear models, such as ANOVAs. Yet it is unclear how conservative or liberal mixed methods are in comparison to the traditional methods. Moreover, reanalyzing previously completed work with linear mixed models seems cumbersome. It is therefore desirable to understand the benefits of linear mixed models and to know under what conditions results that are significant for one model might beget significant results for other models, in order to estimate the outcome of a mixed effect model based on traditional F_1, F_2, and *minF'* analyses. The current paper demonstrates that it is possible, at least for the most simplistic model, for an F or p value from a linear mixed model to be estimated from the same values from more traditional analyses.

Keywords: statistics; parametric statistics; linear mixed models; Analysis of Variance, language-as-a-fixed-effect fallacy.

Introduction

Researchers in cognitive science, and in psycholinguistics specifically, have often incorrectly analyzed their experimental data simply by failing to use the proper statistical methods (Raaijmakers, Schrijnemakers, & Gremmen, 1999). This paper aims to answer the question whether the results of a proper statistical analysis can be estimated on the basis of the traditional, but improper, statistical analysis.

Many experimental studies in psycholinguistics consist of a generic simple reading time (RT) experiment whereby participants are asked to make semantic judgments about a word (or sentence, or paragraph). The time it takes for each participant to respond to an item (RT) is typically used as the dependent variable. Most of the time, participants are drawn from a convenience sample of university undergraduate students. However, to generalize findings to a larger population, participants are treated as a random factor in a regression analysis. Consequently, if the experiment were to be repeated with a different group of participants, the same effects are assumed to hold. In other words, any variation in RT specific to an individual participant (e.g., if one participant overall tends to respond faster than another) should be disregarded as random error. This allows for the generalization to a greater population than those participants included in the experiment. For the most part, researchers correctly identify when it is necessary to do this, and they accurately treat participants as random factors, keeping the Type I (and Type II) error rate low.

However, this method is not always used for the item stimuli in an experiment. Coleman (1964) and Clark (1973) recognized that although researchers in psycholinguistics correctly specified participants as random factors, variance in items (words, sentences, and paragraphs) was all but ignored. Like generalizing over participants, Clark (1973) argued that in most cases, researchers would like to be able to run their experiment with a different set of stimuli and find the same effects. He therefore argued that not only participants should be treated as random factors, but items as well. Just as participants in an experiment do not represent an entire population, items in an experiment are by no means representative of all the possibilities of language (Baayen, Davison, & Bates, 2008; Barr, Levy, Scheepers, & Tily, 2013).

The failure to also indicate items as being a random factor, and thereby also failing to generalize past the specific items included in a particular experiment, is known as the language-as-a-fixed-effect fallacy (Clark, 1973). Thankfully, in addition to pointing out this fallacy, Clark (1973) also proposed a simple solution to this problem. He recommended calculating an estimation of a combined F value representing a combined model, one with a random participant factor (F_1) and the other with a random item factor (F_2). This estimate of a combined F value is referred to as *minF'*.

minF'

MinF' is calculated from the familiar F_1 and F_2 values and is computed as (F_1 x F_2) / ($F_1 + F_2$), where F_1 is the F value of the by-participant ANOVA analysis and F_2 is the F value of the by-item ANOVA analysis. However, the *minF'* value suggested by Clark (1973) is only an approximation of another value, namely *F'*. *F'* is derived from the formula (MS_T + MS_{SxIxT}) + (MS_{TxS} + MS_{IxT}), whereby MS_T is the mean square of the treatment effect, MS_S is the error term of the participants, and MS_I is the error term of the items. *F'* is often too difficult to calculate due to a variety of reasons, such as when dealing with a large dataset or missing data (Raaijmakers, Schrijnemakers, & Gremmen, 1999).

The situation becomes more complicated too, as *F'* itself is an approximation of a combined F value, and like *F'*, the combined F it approximates is also difficult to compute when data are missing. Furthermore, because *minF'* is an approximation of an approximate value (*F'*), it is important to note that *minF'* is a conservative (minimum lower bound) approximation of *F'*. *F'*, in turn, is also a conservative approximation of the combined F it approximates. Therefore, the significance for *minF'* must be calculated independently from F_1 and F_2 because *minF'* does not automatically inherit significance simply because F_1 and F_2 are significant.

F$_1$ and F$_2$

Most studies report the less conservative (and therefore more often significant) F_1 and F_2 values instead of *minF'* values, despite the fact that they thereby might be making a Type I error. Raaijmakers (2003) and Raaijmakers, Schrijnemakers, and Gremmen (1999) suggest that researchers simply may have misunderstood that they are supposed to report *minF'* and not the components used to calculate *minF'*. There are two reasons for the incorrect practice of reporting F_1 and F_2. First, as suggested by Raaijmakers, Schrijnemakers, and Gremmen (1999), there might be a lack of understanding on the part of the researcher. Second, and equally problematic, is the fact that researchers regard *minF'* as too conservative and rather than reporting an insignificant *minF'* value, they would rather report significant F_1 and F_2 values, or worse, a single significant F_1 or F_2 value,.

F_1 and F_2 were intended as intermediate steps used to calculate minF'and not as a replacement for *minF'*. Yet the components of the formula to compute *minF'* (F_1 and F_2) have now become standard values to report in and of themselves. The correct *minF'* all but disappeared from the literature, only to be replaced with the F_1 (by-participant) and F_2 (by-item) analyses, incorrect when considered separate. Raaijmakers, Schrijnemakers, & Gremmen (1999) reported that the use of *minF'*, since introduced, has steadily declined in use till it is virtually unseen in published articles. In fact, Raaijmakers, Schrijnemakers, & Gremmen (1999) report that out of 220 that mention F_1 and F_2, published in the *Journal of Memory and Language* between 1993 - 1997, a total of 120 papers only report F_1 and F_2 values, ignoring *minF'* altogether.

The reporting of the correct statistics further degraded when it not only became more or less acceptable to report F_1 and F_2 values, but also to report F_1 and F_2 values, of which only one value was significant, while still concluding significant results. Locker, Hoffman, and Bovaird (2007) reported that it is not uncommon to find studies only reporting F_1 values, ignoring insignificant F_2 values. Reporting F_1 and F_2 is better than only reporting the by-participants analysis (F_1) and committing the language-as-fixed fallacy but there is still a glaring problem. The problem with either of these approaches is that Clark's (1973) advice is ignored altogether and *minF'* is not calculated at all.

Linear mixed models

A solution to the F_1 and F_2 problem that lies as the heart of the language-as-fixed fallacy is the use of linear mixed models. Linear mixed models, first seen in biomedical research, are also known as multilevel models, hierarchical linear models, mixed effects models, or variance component models (Baayen, Davidson, & Bates, 2008; Brysbaert, 2007; Locker, Hoffman, & Bovaird, 2007; Pinherio & Bates, 2000; Richter, 2006).

Linear mixed models are more powerful than linear regressions because they allow for considering both participant and item error simultaneously in the one model and thereby increase model fit by driving down random error. In essence, linear mixed models do not treat language as a fixed effect, thereby offering an alternative to the infrequently used *minF'*. In addition to solving the language-as-a-fixed-effect fallacy, these models also have several additional advantages compared to traditional models, such as ANOVAs and *minF'* analyses. First, they can accommodate more complicated nested and crossed designs (Quené & van den Bergh, 2008). In addition, linear mixed models allow for missing data at random and do not need to perform listwise deletion. Mixed models can be further extended to allow for time-varying covariates and they accurately present the relationships between variables over time. They easily allow for clustering, longitudinal, or repeated measures as well as specific covariate structures. Finally, linear mixed models generalize non-normal data and do not assume independent observations, thereby being more applicable to a wide range of datasets.

Recent work by Baayen, Davidson, and Bates (2008) demonstrated the outcomes of different models applied to the same datasets, encouraging researchers to recognize the benefits of linear mixed models. Raaijmakers (2003) and Raaijmakers, Schrijnemakers, and Gremmen (1999) similarly encouraged cognitive scientists to avoid only reporting F_1 and F_2 values by addressing concerns about minF' and proposing alternative solutions. Although software is readily and sometimes freely available in *R*, *SPSS*, *SAS, MLwiN* and other packages, and despite the convincing demonstrations of the benefits of linear mixed models (Baayen, 2008a; Brysbaert, 2007; West, Welch, & Gałecki, 2006; Winter, 2013) the use of mixed models is still not widespread. Out of 56 published articles mentioning F_1, F_2, *minF'*, or mixed models, in the *Journal of Memory and Language* between 2012 - Jan 2014, 30 still report F_1 and F_2 values, three of which also report *minF'*. At the same time, almost half of the papers (*n* = 26) do correctly report results from linear mixed models, suggesting that at least some researchers are starting to recognize that reporting F_1 and F_2 is not correct.

The advice of the current paper for researchers still reporting F_1 and F_2 values is to correctly reanalyze data that was originally reported as F_1, F_2 and *minF'*. But such advice would likely not be received enthusiastically, particularly because it is unclear whether the conclusions drawn from the results would in fact still hold, despite the incorrect analysis. Ideally, it would be desirable to estimate, on the basis of F_1 and F_2 values, whether mixed effect models would generate significant results and vice versa. Such an estimate would not replace a reanalysis of the data with mixed models, but could serve as an estimate of the effect of a proper statistical method on the findings. Hopefully, this would subsequently motivate a mixed model analysis of the original data, or a replication of the experiment with new data using the proper statistical model. By manipulating the effect of treatment in a variety of datasets this paper sheds light on the conditions under which results that are obviously significant for one model might beget insignificant results for other models.

Method

Following the principle of parsimony, we started by selecting a simple design with only one independent variable, one dependent variable, and normally distributed errors. We reasoned that if significance can indeed be estimated, it is more than likely to first be estimated with a simple model. Potentially, more factors would add to a model's complexity, making it more difficult to make accurate estimates. In addition, to stay close to designs reported in cognitive science literature, typically not so simple models, we also selected four variations of our design such that we included both within-participant and between-participant designs, and cases where there were different items in each treatment condition or cases where there were the same items in each treatment condition (see Table 1). The number of subjects for each condition ranged from 10 - 40 and there with 40 items in each experiment. Data for each of the four designs was simulated 100 times with different values, as calculated below. Next, these 400 simulations were repeated between six to ten times each contingent upon how long it took to vary the effect of treatment (E_T) from no effect ($p > .99$) to a strong effect ($p < .01$). In total there were 3400 different simulations of a dataset, as explained below.

Table 1: The four different designs used to simulate data

	Within Participants	Between Participants
Repeated items in each condition or "Within word"	Ranging from no effect to completely significant (100 simulations of each)	Ranging from no effect to completely significant (100 simulations of each)
Different items in each condition or "Between word"	Ranging from no effect to completely significant (100 simulations of each)	Ranging from no effect to completely significant (100 simulations of each)

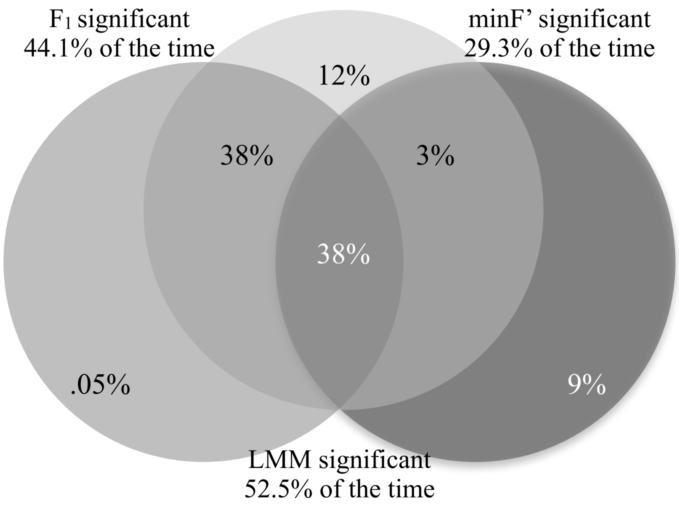

Figure 1: Venn diagram representing overlap of number of *p* values meeting the $p < .05$ criteria for each type of analysis. The total percentage of significant *p* values for each model is also included.

A linear model has the following structure: $Y=Y_0+E_T+E_S+E_I+E$. The base value, or the expected mean response time with no treatment (Y_0) for each response was set to 400ms. All normally distributed errors (by-participant error (E_S), by-item error (E_I), and by-observation error (E)) were set to be normally distributed randomly generated numbers centered at 0, where the *SD* of the error was a random number ranging between 0 and 20. Again, the strength of the effect (E_T) was manipulated such that each design was simulated between six and ten times, ranging from no effect of the independent variable, to all 100 cases resulting in highly significant effects at $p < .01$. Linear mixed models were computed using the `lme4` package (Bates & Sarkarin, 2007). Significance was estimated from the two tailed MCMC probability as calculated from the `pvals.fnc` function found in the `languageR` package (Baayen, 2008b).

Simulations

Four models, a mixed effect model, an F_1 model, an F_2 model, and a *minF'* model were conducted on the data for each of the 3400 simulations. The number of significant cases out of 3400 for each model is represented in Figure 1. In the figure, to increase legibility we focused on comparing F_1, *minF'*, and linear mixed models, as F_2 is rarely reported alone, however F_2 is certainly considered independently in all of the analyses. As is evident from Figure 1, linear mixed models were the least conservative, resulting in significant *p* values of $p < .05$ for 52.5% of the time with a large overlap with F_1 models (which were significant 44.1% of the time). In other words, linear mixed models and F_1 results were the most similar. *MinF'* values were the most conservative, with significance at $p < .05$ in 29.3% of the data. These findings are in line with the fact that F_1 and F_2 analyses have less power than linear mixed models (Ghisletta & Renaud, 2005), and that *minF'* has reduced power compared to both other models (Wickens & Keppel, 1983). Keep in mind, however, that for all four sets of simulations, the effect of treatment (E_T) varied from no effect to always significant. Although linear mixed models always detected more significant results than did other models, it is important to note that when E_T was barely significant, linear mixed models detected more significant effects than did F_1, F_2, or *minF'* (see Figure 2). These findings suggest that findings reported with significant F_1 results, are likely significant when data is analyzed with linear mixed models. Moreover, findings that have not been reported because results were not significant, should perhaps be reanalyzed and reported because significant results might be found with linear mixed models.

We next aimed to determine if the results from linear mixed models could be estimated from the output of the other models. There are several possible factors that might impact whether significance can be estimated in one model based on the results from another. For example, the experimental design, the size of the effect, the number of factors, and the degrees of freedom must be taken into account when making such estimates. Nevertheless, we decided to try estimate the outcome of linear mixed models (in this simple model) from respectively very little information (i.e. *p* and *F* values).

First, to see if F_1, F_2, and *minF' F* values estimate *F* values in linear mixed models, we entered F_1, F_2, and *minF' F* values in a regression. We found that F_1, $F(1, 3396) = 5931.156$, $p < .001$, F_2, $F(1, 3396) = 198.73$, $p < .001$, and *minF'*, $F(1, 3396) = 267.49$, $p < .001$, all estimated *F* values in linear mixed models.

To see if the same factors were able to estimate significance, as degrees of freedom were calculated differently in linear mixed models than for standard regressions, we entered all p values into a regression model and found that both F_1 *p* values, $F(1, 3396) = 28537$, $p < .001$, and *minF' p* values, $F(1, 3396) = 42.47$, $p < .001$, significantly estimated *p* values in linear mixed models. At the same time F_2 *p* values failed to estimate *p* values in linear mixed models, $F(1,3396) = .0001$, $p = .10$, possibly due to the fact that F_1 accounts for the majority of the variance in the model.

However, when using these simulations to estimate *F* and *p* values, it is likely that researchers only have one type of *F*

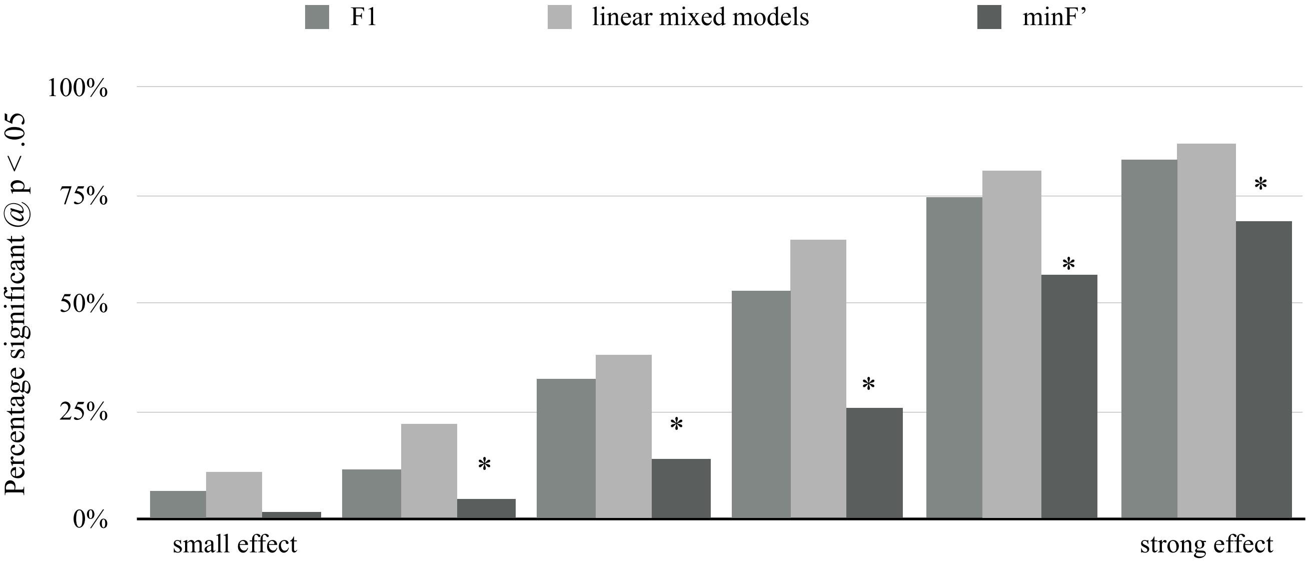

Figure 2: The total percentage of significant p values for each model. The E_T is split into six bins, from a small effect (first bin) to a strong effect (last bin).
* denotes a significant difference between other groups at $p < .05$

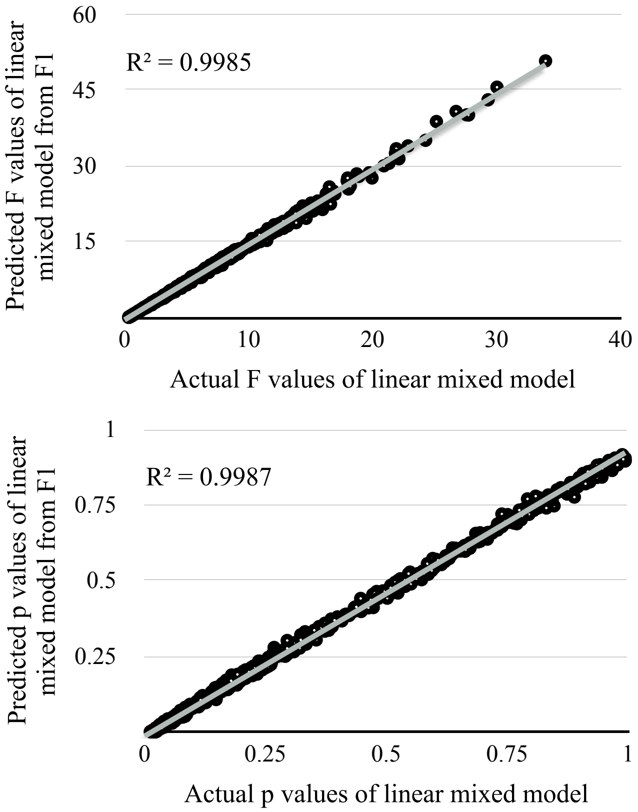

Figure 3: *F* and *p* values estimated from F_1 and F_2 plotted against actual *F* and *p* values for the dataset `splitplot`.

value (either F_1/F_2 or *minF'*, not both). If only one *F* value is used to estimate the likely output of a linear mixed model we decided to run additional analyses where F_1/F_2 and *minF'* were entered into separate analyses. Again, we found significance for F_1/F_2 models estimating mixed effect *F* values for F_1, *F* (1, 3397) = 40089, $p < .001$, and for F_2, *F* (1, 3397) = 2339, $p < .001$. We found that for *p* values only F_1 *p* values contributed to mixed effect *p* values, *F* (1, 3397) = 29338, $p < .001$, where F_2 did not, *F* (1, 3397) = 1.84, $p = .17$. Again, F_1 accounts for the majority of the variance in the model, perhaps explaining the insignificant effects of the *p* value from F_2.

For *minF'*, *F* values were also significant, *F* (1, 3398) = 73089, $p < .001$, as were p values, *F* (1, 3398) = 545.22, $p < .001$. Despite the fact that many factors might contribute to whether or not *F* and *p* values can be estimated from other *F* and *p* values, we find here that with a simple model, this seems quite possible.

Estimates

We next aimed to see if we would be able to estimate the significance of linear mixed models on a different dataset using the formulas derived from our simple design simulations above. We tested our formulas from these simulations on the dataset `splitplot` in the `languageR` package (Baayen, 2008b). We selected this dataset because this dataset is freely available, ensuring replicability, and also because Baayen, Davidson, and Bates (2008) previously analyzed the same dataset using a variety of methodologies. The experimental design for this dataset involved two counterbalanced lists of words, each with 40 words. Each list consisted of related prime words and unrelated prime words. Twenty participants were tested on one list, or the other.

One thousand simulations of linear mixed models predicting RT with the priming condition as a fixed factor and participant and item as random factors were conducted on the `splitplot` dataset. Regressions were also conducted for F_1 and F_2 values and for *minF'* values (Clark, 1973). This resulted in a total of 4000 outcomes. To ensure 1000 different datasets, RT values for the `splitplot` dataset were calculated using the parameters of the original data such that all simulated data were generated from the distribution of the original mean and *SD* for each parameter. The effect of the IV (E_T) was set randomly so that models would vary from a weak effect of treatment at $p > .999$ to a strong effect of treatment at $p < .001$.

We then estimated values of significance for each dataset from our previous formulas and compared these values to the actual output from 1000 simulations of the dataset provided in `splitplot` (see figure 3). As can be seen from Figure 3, predicting F and p values for linear mixed models from the F_1/F_2 analyses is almost perfect for simple designs with one independent variable, one dependent variable, and normally distributed errors.

Discussion and Conclusion

This paper demonstrates that it is possible, at least for the most simplistic models, for an *F* or *p* value from a linear mixed model to be estimated from the same values from more traditional analyses. It is important to recognize that this paper only demonstrates this for the most simple of designs, and that with more complexity, it is likely that it becomes more difficult to so accurately estimate *F* and *p* values for linear mixed models. Nevertheless, the strength of the relationship between F_1/F_2 or *minF'* and the *F* value from a linear mixed model is not unexpected, as all of these *F* values are calculated from the same dataset in a similar way. The same logic stands for the *p* values. This at least suggests that it might be possible to estimate *F* and *p* values of linear mixed models from more complex designs. In future work, we intend to explore such factors. In addition, it would be interesting to include varying random effect structures, as the generalizability and the performance of linear mixed models are influenced by the assumptions of the random structures of the models (Barr, Levy, Scheepers, & Tily, 2013). Furthermore, including random slopes by treatment would increase applicability for real datasets and such factors might further impact the resulting values.

This paper has also elaborated upon some of the benefits of linear mixed models, and suggested its' use over alternative traditional methodologies such as F_1 and F_2 analyses. Although, sometimes F_1 is the proper analysis to use, this can be the case when items are nested in participants, and participants are nested in treatments (Clark, 2008, p. 348), or when items are properly counterbalanced or matched. It is nevertheless important for researchers to understand when particular analyses are appropriate to use and when they are not. Even more practically, linear mixed models provide some benefits to researchers with regards to the flexibility and robust nature of the analysis.

In this paper, we suggest that researchers analyze current data and reanalyze past data that was originally reported as F_1, F_2 or *minF'* using linear mixed models. We realized such a suggestion might not be eagerly considered, therefore we demonstrated that it is possible to estimate, on the basis of F_1 and F_2 values and *minF'* values, whether linear mixed effect models would generate significant results. Indeed we not only estimated F values, but also p values. These estimates are not intended to replace a reanalysis of the data, but rather they are intended to motivate researchers to analyze and properly reanalyze data using linear mixed models.

References

Baayen, R. H. (2008a). *Analyzing linguistic data: A practical introduction to statistics.* Cambridge: Cambridge University Press.

Baayen, R. H. (2008b). languageR: Data sets and functions with “Analyzing Linguistic Data: A practical introduction to statistics”. R package version 0.953.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal, *Journal of Memory and Language, 68(3)*, 255–278.

Bates, D. M., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes, R package version 0.99875-6.

Brysbaert, The language-as-fixed-effect fallacy”: Some simple SPSS solutions to a complex problem. (2007). The language-as-fixed-effect fallacy”: Some simple SPSS solutions to a complex problem. London: Royal Holloway.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological re- search. *Journal of Verbal Learning and Verbal Behavior, 12,* 335–359.

Ghisletta, P., & Renaud, O. (2005). *Multilevel models for cross- factors data to generalize across both subjects and items*. Paper presented at the 58th annual scientific meeting of the Gerontological Society of America, Orlando, FL.

Locker, L., Hoffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods, 39(4),* 723–730.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS* (Statistics and Computing). New York: Springer.

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59(4),* 413–425.

Raaijmakers, J. G. W. (2003). A further look at the "language-as-fixed-effect fallacy". *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale, 57(3),* 141–151.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with “the language-as-fixed-effect fallacy”: Common misconceptions and alternative solutions. *Journal of Memory and Language, 41(3),* 416–426.

West, B. T., Welch, K. B., & Gałecki, A. T. (2006). Linear mixed models: a practical guide using statistical software.

Wickens, T. D., & Keppel, G. (1983). On the choice of design and of test statistic in the analysis of experiments with sampled materials. *Journal of Verbal Learning and Verbal Behavior*, *22*, 296-309.

Winer, B. J. (1971). *Statistical principles in experimental design.* New York: McGraw–Hill.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv: 1308.5499.