2007

On the Use of Multilevel Modeling as an Alternative to Items Analysis in Psycholinguistic Research

Lawrence Locker Jr.
Georgis Southern University, kdbodily@georgiasouthern.edu

Lesa Hoffman
University of Nebraska-Lincoln, lhoffman2@unl.edu

James A. Bovaird
University of Nebraska-Lincoln, jbovaird2@unl.edu

Behavior Research Methods
2007, 39 (4), 723-730

On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research

Lawrence Locker, Jr.
Georgia Southern University, Statesboro, Georgia

and

Lesa Hoffman and James A. Bovaird
University of Nebraska, Lincoln, Nebraska

The use of multilevel modeling is presented as an alternative to separate item and subject ANOVAs ($F_1 \times F_2$) in psycholinguistic research. Multilevel modeling is commonly utilized to model variability arising from the nesting of lower level observations within higher level units (e.g., students within schools, repeated measures within individuals). However, multilevel models can also be used when two random factors are crossed at the same level, rather than nested. The current work illustrates the use of the multilevel model for crossed random effects within the context of a psycholinguistic experimental study, in which both subjects and items are modeled as random effects within the same analysis, thus avoiding some of the problems plaguing current approaches.

A great deal of research in cognitive psychology has been devoted to the study of word recognition in skilled readers (e.g., Locker, Simpson, & Yates, 2003; Pexman & Lupker, 1999; Rubenstein, Garfield, & Millikan, 1970; Strain, Patterson, & Seidenberg, 1995; Yates, Locker, & Simpson, 2004). These studies typically involve the selection of a set of words that vary on some lexical dimension(s). The word list is then presented to a group of participants for the purpose of recording relevant dependent measures (e.g., response time) that serve as the basis of inferences drawn in regard to the processes and structure of the language system. For example, Yates et al. (2004) were interested in how phonological neighborhood density (i.e., the number of words that differ from a target word by one phoneme) influenced visual word recognition. In each of the experiments reported in their study, a list was constructed composed of an equal number of words with many phonological neighbors (high-density words) and words with few neighbors (low-density words). The words were presented along with pronounceable pseudowords in a lexical decision task (i.e., word/nonword discrimination task). Yates et al. found that words with many phonological neighbors were responded to more rapidly on average than words with few neighbors, supporting the notion that phonology is an important component in word processing.

Although such studies are relatively simple in design, issues concerning data analysis in this area can be quite contentious. Typically, two analyses are conducted for response times. In the subjects analysis, or F_1, condition means are obtained for each subject and submitted to an ANOVA. In the items analysis, or F_2, condition means are obtained for each item and also submitted to an ANOVA. Obtaining significant treatment effects in both analyses is referred to as meeting the $F_1 \times F_2$ criterion (Raaijmakers, Schrijnemakers, & Gremmen, 1999). It is commonly believed that if both F_1 and F_2 analyses yield significant findings, then the effects will generalize to different samples of subjects *and* items, assuming that the subjects and items in the experiment can each be considered random samples from larger populations (Raaijmakers et al., 1999). Simply put, under this belief, one can be confident of a given result if both F_1 and F_2 analyses are significant.

Although the $F_1 \times F_2$ criterion is by far the most common approach to data analysis in psycholinguistic studies, there has been some resistance. For example, it is not uncommon to find studies that report F_1 and F_2 analyses, but in which conclusions are based primarily on significant F_1 analyses, ignoring nonsignificant F_2 analyses (e.g., Locker, Simpson, et al., 2003; Siakaluk, Sears, & Lupker, 2002). Although such an approach may be justified under certain conditions (i.e., that item variability has been experimentally controlled; Raaijmakers, 2003; Raaijmakers et al., 1999), Clark (1973) argued in a classic paper that such an approach implicitly assumes that the materials used in an experiment can be treated as fixed factors (i.e., the "language-as-fixed-effect fallacy"). In reality, a given stimulus set of words may constitute only a subset of items that could be utilized in a given experi-

L. Locker, Jr., llocker@georgiasouthern.edu

ment. Random or pseudorandom selection of items results in random variance that could lead to a positive bias in the F_1 test, increasing the likelihood of Type I error. That is, unaccounted item variability (nested within treatments) could contribute to differences between treatment conditions when, in reality, there is a null effect of the experimental manipulation. As a consequence, Clark advocated the use of a quasi-*F* ratio, or F', which is a random effects model that takes into account both item and subject variability, as shown in Equation 1:

$$F' = (MS_{\mathrm{T}} + MS_{\mathrm{S\times I\times T}}) \div (MS_{\mathrm{T\times S}} + MS_{\mathrm{I\times T}}), \quad (1)$$

in which MS_{T} represents the mean square for the treatment effect, $MS_{\mathrm{S\times I\times T}}$ represents the error term of the subjects by items by treatment interaction, $MS_{\mathrm{T\times S}}$ is the error term of the treatment by subjects interaction, and $\mathrm{MS}_{\mathrm{I\times T}}$ is the error term of the items by treatment interaction. As illustrated by Equation 1, rather than constituting separate tests, F' involves the simultaneous treatment of subjects and items as random factors. Unfortunately, F' cannot be computed when the data are *unbalanced* or when responses are missing for certain item/subject combinations. Because response times are almost never included for incorrect trials, and subjects almost never exhibit perfect accuracy when response time is emphasized, this approach is nearly impossible to use in practice. In contrast, the minimum bound of the F' (Clark, 1973) can be computed quite easily from the results of separate $F_1 \times F_2$ analyses, as shown in Equation 2:

$$F'_{\min}(i,j) = (F_1 \times F_2) \div (F_1 + F_2), \quad (2)$$

in which *i* represents the numerator degrees of freedom in each analysis, and *j* represents the denominator degrees of freedom. If $F'_{\min}$ is significant, then F' is assumed to be significant as well (Raaijmakers et al., 1999).

However, despite its initial adoption, researchers have largely abandoned $F'_{\min}$ and have instead utilized the $F_1 \times F_2$ criterion (Raaijmakers et al., 1999; Raaijmakers, 2003). Although compelling arguments have been made for both sides (Forster & Dickinson, 1976; Raaijmakers, 2003; Wike & Church, 1976), the zeitgeist in the field of psycholinguistic research is apparently to continue with the use of the $F_1 \times F_2$ criterion, even though the flaws associated with this approach are well known (see Raaijmakers, 2003). More specifically, because F_1 ignores systematic variability due to the individual items, and F_2 ignores systematic variability due to the individual subjects, neither is truly an appropriate description of all sources of systematic variance within the outcome (e.g., response time or accuracy). Therefore, it is necessary to explore alternative methods for analyzing psycholinguistic data that do not erroneously treat items as fixed effects, while at the same time providing a means by which to treat both subjects and items as random factors within a single analysis, as originally advocated by Clark (1973).

The alternative proposed in the present article is the multilevel model. Multilevel modeling is a tool for the analysis of data with nested sources of variability (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). That is, there are lower level observations nested within higher level observations, such as students sampled from multiple classrooms. One might be interested in the effect of a student-level predictor on academic performance (e.g., socioeconomic status on math achievement). However, because the residuals of students within the same classroom are likely to be correlated, a typical regression analysis is inappropriate. Multilevel models properly account for variability at each level of analysis and permit the examination of predictors of that variability at each level of analysis. For example, the effect of student characteristics on variability across students in math achievement could be assessed, as well as the effect of school characteristics on variability in mean math achievement across schools, as well as any cross-level interactions (e.g., Singer, 1998). Both classrooms and students are assumed to be random samples from their respective populations, although students are nested within classrooms. This approach has also been used by Wright (1998), who demonstrated how the multilevel modeling of autobiographical memories as nested within individuals may be more appropriate than ANOVAs.

In contrast with the above examples, however, a different scenario may arise when sources of variability are not strictly nested within one another, such as when sampling students who live in different neighborhoods and who attend different schools. Because students from the same school may not live in the same neighborhood and vice versa, these two random factors are *crossed* instead of nested (Raudenbush, 1993). Random factors that are crossed at the same level of analysis can also be included within the multilevel modeling framework, however, and predictors of each type of variability (i.e., characteristics of schools and of neighborhoods) may still be evaluated, as well as predictors of student-level variability.

The crossed random factors multilevel model directly meets the recommendations of Clark (1973), in that both subjects and items can be considered as random effects simultaneously within a single model. In this way, effects of experimental manipulations (i.e., treatment effects) can be assessed without falsely reducing the observed variance in the outcome (e.g., response time or accuracy), such as by collapsing across items to form cell means for a subjects analysis (F_1), or by collapsing across subjects to form cell means for an items analysis (F_2). Treatment effects can be modeled as fixed effects or as random effects (i.e., in which the magnitude of the treatment effect is specified as varying over subjects or items). Critically, the inclusion of both subjects and items as random factors provides a more complete description of all systematic sources of variance in the outcome, whereas the $F_1 \times F_2$ criterion does not. Another important advantage is that the multilevel model also uses full information maximum likelihood as a means of directly addressing unbalanced or incomplete data, and thus complete cases are not required.

The purpose of the present article is to further demonstrate the application of multilevel modeling to the analysis of psycholinguistic data. Although prior research has shown the applicability of this approach when treating subjects and items as nested (Baayen, Tweedie, &

Schreuder, 2002), the first aim of the present study is to demonstrate the validity of this approach when random subject and item variability are treated as *crossed* at the same level, rather than as nested. As argued above and elsewhere (Ghisletta & Renaud, 2005), there is reason to believe that this constitutes a more appropriate treatment of the data in the present context, and fulfills the requirements originally advocated by Clark (1973). Furthermore, such a demonstration will add to the growing body of literature demonstrating that multilevel modeling is a viable alternative to data analysis in cognitive research (e.g., Baayen et al., 2002; Ghisletta & Renaud, 2005; Hoffman & Rovine, 2007; Wright, 1998).

A second aim of the present work is to demonstrate the ease with which multilevel modeling can be applied in data analysis and to serve as a reference for investigators who may wish to apply this approach in their own research. To this end, the data and analysis syntax for both SAS and SPSS, used for the examples below have been included in an electronic appendix. Given the capabilities of SAS and SPSS, as well as other packages that can estimate these types of models (e.g., HLM, MLwiN, R, Mplus), a multilevel approach is a reasonable option for the analysis of data in which both subjects and items constitute random factors.

To demonstrate the viability of this approach in terms of analysis of data from psycholinguistic research, we present F_1 and F_2 analyses from an experiment conducted by Locker, Yates, and Simpson (2003), followed by a crossed random factors multilevel analysis. This facilitates a direct comparison of the $F_1 \times F_2$ and multilevel outcomes, as well as demonstrates that multilevel modeling is indeed a tenable approach in this context.

ILLUSTRATIVE EXAMPLE

$F_1 \times F_2$ Subjects and Items ANOVAs

The purpose of the experiment originally conducted by Locker, Yates, et al. (2003) was to assess the interaction of phonological neighborhood frequency and semantic neighborhood in a visual lexical decision task.[1] *Neighborhood frequency* in this example refers to the average frequency of a word's phonological neighbors (phonological neighborhood frequency values were obtained from the Wordmine database; Buchanan & Westbury, 2000). *Semantic neighborhood* refers to the number of words that are meaningfully related to a given target word (Nelson, McEvoy, & Schreiber, 1998). A 39-word list constructed by crossing neighborhood frequency (high vs. low) and semantic neighborhood (large vs. small)[2] was administered to 38 undergraduate students. Stimuli were presented on an IBM-compatible PC with E-Prime software (Schneider, Eschman, & Zuccolotto, 2002). Participants were instructed that a series of letter strings that formed words and pronounceable pseudowords would be presented on the computer screen one at a time. Participants were asked to respond as quickly and accurately as possible by pressing buttons on the keyboard designated "word" and "nonword."

Response times for the word responses were analyzed with both F_1 and F_2 ANOVAs. Observed means for each condition for each analysis are provided in Table 1, and results from the ANOVAs are provided in Table 2. Both analyses revealed significant main effects of neighborhood frequency and semantic neighborhood size, as well as a significant interaction. The calculation of $F'_{\min}$ for each effect, however, suggests a different pattern of results. As shown in Table 2, although the $F'_{\min}$ interaction effect was significant ($p < .05$), the main effects of neighborhood frequency and semantic neighborhood size were not significant (ps = .053 and .071, respectively). Thus, according to the logic of the $F_1 \times F_2$ criterion, the main effects and interaction of neighborhood frequency and semantic neighborhood size could be generalized to both subjects and items; but according to the $F'_{\min}$ criterion, only the interaction effect could be. As discussed above, however, both F_1 and F_2 analyses are each potentially biased, such that neither is an appropriate model for the multiple sources of variability within these response times. Importantly, the calculation of $F'_{\min}$ does not overcome this limitation, and has been suggested to be unnecessarily conservative (e.g., Wike & Church, 1976).

Accordingly, we will now utilize a crossed random effects multilevel model for the same data in order to further assess these findings. Although multilevel models are often presented hierarchically (i.e., as separate equations for each level), in the present example it is more straightforward to specify a combined equation, in which the higher level effects are inserted directly into the level-1 equation. This also parallels how these models were estimated within a general linear mixed model (i.e., as used by SAS and SPSS). It is important to note that a different data structure is required for multilevel analysis than is typically used for ANOVAs. Specifically, the data need to be in a "stacked" or "long" format, in which each case contains the independent and dependent variables for a single subject and a single item (see Hoffman & Rovine, 2007, for more information).

Crossed Random Effects Multilevel Analysis

The first step in the analysis should be to examine the extent to which subjects and items both exhibit systematic effects, and thus the extent to which subjects and items each need to be considered as random factors. One can

Table 1
Mean Response Times (RTs, in Milliseconds) and Standard Errors (*SE*s) per Condition × Model

Model and Condition	Neighborhood			
	Small		Large	
	RT	*SE*	RT	*SE*
Subject ANOVA (F_1)				
Low frequency	615.0	11.8	620.3	12.5
High frequency	676.0	15.3	617.7	14.0
Item ANOVA (F_2)				
Low frequency	616.7	14.7	621.0	10.1
High frequency	689.7	23.7	619.5	6.9
Crossed random effects multilevel model				
Low frequency	615.8	18.6	620.2	18.5
High frequency	685.8	18.7	618.2	18.6

Table 2
ANOVA Approximate *F* Test Results

Effect	*F*	*df*	MS_e	*p* value
Subjects ANOVA (F_1)				
Phonological neighborhood frequency	16.1	1,37	2,012.9	.0003
Semantic neighborhood size	14.9	1,37	1,793.3	.0004
Interaction	38.2	1,37	1,007.1	<.0001
Items ANOVA (F_2)				
Phonological neighborhood frequency	5.3	1,35	2,361.9	.0278
Semantic neighborhood size	4.5	1,35	2,361.9	.0415
Interaction	5.7	1,35	2,361.9	.0225
F'_{min}				
Phonological neighborhood frequency	4.0	1,56.09		.053
Semantic neighborhood size	3.5	1,54.62		.071
Interaction	5.0	1,45.27		.031
Crossed random effects multilevel model				
Phonological neighborhood frequency	5.4	1,31.8*	–	.0272
Semantic neighborhood size	4.6	1,31.8*	–	.0393
Interaction	6.0	1,31.8*	–	.0199

*Estimated denominator degrees of freedom using the Satterthwaite method; no mean squares estimated.

estimate an "empty" model with no random effects (i.e., only one error term) as a baseline for comparison, as shown in Equation 3:

$$Y_{si} = \gamma_0 + e_{si} \quad (3)$$

in which Y_{si} is the observed response time for subject *s* and item *i*, γ_0 is the intercept, or expected mean response time for the overall sample, and e_{si} is the residual deviation from the sample mean response time for subject *s* and item *i*. This model further specifies that residuals (the e_{si}s) are uncorrelated; that is, that no systematic effects of subjects or items are present. This assumption is not likely to be tenable, but provides a baseline for comparison with more complex models.

A random effect for subjects is added next, as seen in Equation 4:

$$Y_{si} = \gamma_0 + U_{0s} + e_{si} \quad (4)$$

in which Y_{si} is now also predicted from U_{0s}, a random effect for subject *s*, which is the deviation of that subject's mean response time from the grand mean response time. The residuals are now assumed to be uncorrelated across observations after considering from which subject the observation was taken. Because the empty model is nested within the random subject model, the improvement over the empty model from adding a random effect for subjects can be assessed by comparing the model deviance values from each. The difference of the model deviances is distributed as a χ^2, with degrees of freedom equal to the difference in the number of parameters estimated within each model, or in this case, $df = 1$. The difference in the model deviances is 280, which is highly significant ($p < .001$), as is expected.

A random effect for items is then added in order to estimate a *crossed random effects model*, as seen in Equation 5:

$$Y_{si} = \gamma_0 + U_{0s} + V_{0i} + e_{si} \quad (5)$$

in which Y_{si} is the observed response time for subject *s* and item *i*, γ_0 is the intercept, or expected mean response time for the overall sample, U_{0s} is the random effect of subject *s*, V_{0i} is the random effect of item *i*, and e_{si} is the residual deviation from the expected value for subject *s* and item *i*. The residuals are now assumed to be uncorrelated after considering from which item *and* from which subject the observation was taken. The difference in the model deviances from adding a random effect for items ($df = 1$) is 100, which is again highly significant ($p < .001$).

One way of expressing the relative contribution of variance in response time due to items versus variance due to subjects is to calculate intraclass correlations (ICC) for each effect (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). The ICC is calculated as the proportion of variance of the random effect (i.e., subjects or items) over the total variance (i.e., subjects variance + items variance + residual variance). Given the variance components shown in Table 3, the proportion of total variance in response time due to *subjects* is 24% (5,167 ÷ 21,920), the proportion of total variance in response time due to *items* is 11% (2,409 ÷ 21,920), and the unexplained variance in response time, or *subject* × *item interaction*, is 65% (14,344 ÷ 21,920).

Upon identifying the proper error structure for the model (i.e., the presence of random effects of subjects and items), one can then examine the independent variables (i.e., predictors) of interest. In this example, we examine the effects of the item characteristics of neighborhood frequency (Freq) and semantic neighborhood size (Size) by adding to the model contrast codes representing low/high frequency (coded −.5 or .5, respectively), small/large neighborhood (coded −.5 or .5, respectively), as well as their interaction, as shown in Equation 6:

$$Y_{si} = \gamma_0 + \gamma_1(\text{Freq}) + \gamma_2(\text{Size}) + \gamma_3(\text{Freq})(\text{Size}) + U_{0s} + V_{0i} + e_{si} \quad (6)$$

in which all terms are defined as in Equation 5, except that γ_1 represents the main effect of frequency, or the mean difference between items of low versus high frequency averaged across semantic neighborhood size, γ_2 represents the main effect of size, or the mean difference between items with small versus large neighborhoods averaged across

Table 3
Mean Variance Components With Standard Errors and Proportion Reduction × Model

Model	Unexplained Residual Variance		Random Subject Variance		Random Item Variance		Total Variance
	M	*SE*	*M*	*SE*	*M*	*SE*	
Subject ANOVA (F_1) of condition means							
Empty model	2,418	320	5,066	1,321			7,484
Predictor model	1,604	215	5,269	1,319			6,873
Proportion reduction in variance							.082
Item ANOVA (F_2) of condition means							
Empty model	3,154	724					3,154
Predictor model	2,362	565					2,362
Proportion reduction in variance							.251
Crossed random effects multilevel model of all responses							
Empty model	14,344	560	5,167	1,293	2,409	678	21,920
Predictor model	14,341	560	5,168	1,293	1,692	527	21,201
Proportion reduction in variance							.033

neighborhood frequency, and γ_3 represents the two-way interaction of frequency and size, or the additional difference in response time for items that are of high frequency and which have large neighborhoods.

The condition means as estimated via maximum likelihood are given in the bottom of Table 1. In contrast to the observed condition means, the means estimated via maximum likelihood take into account the unbalanced nature of the data (i.e., that response times for individual trials may be missing if the response was not correct). Approximate *F* tests from the multilevel analysis for the two main effects and for their interaction are given in Table 2. Each effect is significant, with the same general pattern of results as was seen in the ANOVAs, although the significance levels obtained from the crossed random effects model are more comparable to those obtained from the items analysis than that of the subjects analysis.

The proportion reduction in total variance (Snijders & Bosker, 1999) due to the three predictor effects was calculated as .033 [i.e., as (21,920 − 21,201) ÷ 21,920]. However, although only approximately 3% of the total variance was accounted for by the predictors, this estimate does not take into consideration *which variance* could be accounted for by item characteristics, given that there are three variances estimated: subjects variance, items variance, and subject × item residual variance. Thus, a more appropriate comparison is to consider the *proportion of random item variance* accounted for, which was calculated as approximately 30% (1,692 ÷ 2,409). Further, the random effect for items remained significant even after including the predictors, suggesting that item variance was not sufficiently accounted for in terms of the experimental control variables or the model predictor variables. Therefore, an analysis treating items as a random sample from a larger population of words is appropriate in this example.

In order to illustrate the difference between the crossed random effects solution and that from a typical analysis, the F_1 and F_2 ANOVA models using condition means were also estimated as multilevel models. The variance components and the proportion of total variance accounted for from F_1 and F_2 analysis can then be compared with those of the crossed random effects model. Two things are readily apparent in comparing these values, as shown in Table 3: The overall amount of variance within the F_1 and F_2 analyses is smaller, and the proportion of variance accounted for is larger. These discrepancies occur because of the difference in the unit of analysis across methods. In the ANOVAs, trials are first averaged into condition means (either across items for F_1 or across subjects for F_2), and those condition means are then subjected to analysis. Accordingly, a significant portion of the observed variability in response times is never included in each analysis, which can result in overestimates of the size of the effects of the predictors. In contrast, the crossed random effects model considers all sources of variance simultaneously, without lost information, and thus is likely to be a more accurate depiction of the total observed variance in response time across subjects and items.

SUMMARY AND CONCLUSIONS

The purpose of the present work was to illustrate how a crossed random effects multilevel model can be used within psycholinguistic research as an alternative to separate subjects and items ANOVAs (i.e., the $F_1 \times F_2$ criterion or $F'_{\min}$). There are many advantages of a multilevel modeling approach within this context. The primary advantage is that it allows one to generalize to both populations of subjects and items on the basis of a single analysis. A second advantage is that because multilevel models can be estimated with incomplete data, they do not suffer from the same drawbacks as the F' test originally advocated by Clark (1973).

A third advantage of the multilevel model is that any combination of continuous or categorical independent variables that pertain to subjects, items, or their interaction may be included as predictors, and the reduction in each source of variance can be considered. For example, although the predictors of neighborhood frequency and semantic neighborhood size were dichotomous, this need not be the case; continuous predictors may be included as needed (see Hoffman & Rovine, 2007, or Quené & van den

Bergh, 2004, for more discussion). Furthermore, in Latin square designs, order of presentation can also be included as a covariate. Finally, the extent to which the effects of item characteristics also vary systematically across subjects can also be examined. In other words, are there systematic individual differences in the effects of the independent variables? This notion can be formalized and tested statistically in the form of a random effect across subjects of the independent variables, and individual differences in mean performance as well as in the effects of the independent variables can then be related to subject-level predictors (e.g., reading ability, phonemic awareness).

Given these advantages and the relative ease by which multilevel analyses can be conducted in standard statistical packages (see the electronic appendix), crossed random effects multilevel models may provide a viable approach to the "language-as-fixed-effect fallacy" (Clark, 1973; Raaijmaker et al., 1999; Wike & Church, 1976). This conclusion has also been supported by simulation studies in which the multilevel model was shown to perform significantly better than the $F_1 \times F_2$ criterion in terms of both Type I error rates and statistical power (Ghisletta & Renaud, 2005).

An important issue to consider in future research is the extent to which differential levels of experimental control upon selecting the items may necessitate different analytic strategies. As discussed by Raaijmakers et al. (1999), it is possible that if sufficient control of item variability can be achieved through matching of items on relevant control variables, then the issues raised by Clark (1973) may not be serious points of concern. Indeed, within the present example, the same conclusions about the experimental manipulations would have been drawn in the $F_1 \times F_2$ analysis as in the multilevel model (although a slightly different conclusion might have been reached by considering $F'_{\min}$ instead). However, it is easy to envision scenarios in which the inferences might change across analytic models, as well as experiments in which the matching of items across a wide range of extraneous variables is not feasible. In these scenarios a crossed random effects multilevel model could potentially be used as a diagnostic tool in order to assess whether item variance is indeed a systematic effect that should be modeled. In the absence of significant random item variance, a more traditional ANOVA may be sufficient.

In summary, the present work represents an attempt to resolve the difficulties surrounding Clark's (1973) "language-as-fixed-effect fallacy" through the use of multilevel models for crossed random effects. Although the current $F_1 \times F_2$ criterion is the standard approach within the study of psycholinguistics, the problems surrounding this method are well known. Thus, it is necessary as a field to investigate viable alternatives that ensure the quality of our inferences.

AUTHOR NOTE

The electronic appendix and the accompanying data are available from the second author at psych.unl.edu/hoffman/HomePage.htm. The authors thank Marc Brysbaert and an anonymous reviewer for helpful comments on an earlier draft of this article. Correspondence concerning this article should be addressed to L. Locker, Jr., Department of Psychology, P. O. Box 8041, Statesboro, GA 30460 (e-mail: llocker@georgiasouthern.edu).

REFERENCES

BAAYEN, R. H., TWEEDIE, F. J., & SCHREUDER, R. (2002). The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon. *Brain & Language*, **81**, 55-65.

BUCHANAN, L., & WESTBURY, C. (2000). *Wordmine database: Probabilistic values for all four to seven letter words in the English Language.* www.wordmine.org.

CLARK, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning & Verbal Behavior*, **12**, 335-359.

FORSTER, K. I., & DICKINSON, R. G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F_1, F_2, F', min F'. *Journal of Verbal Learning & Verbal Behavior*, **15**, 135-142.

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.

HOFFMAN, L., & ROVINE, M. J. (2007). Multilevel models for experimental psychologists: Foundations and illustrative examples. *Behavior Research Methods*, **39**, 107-117.

LOCKER, L., SIMPSON, G. B., & YATES, M. (2003). Semantic neighborhood effects on the recognition of ambiguous words. *Memory & Cognition*, **31**, 505-515.

LOCKER, L., YATES, M., & SIMPSON, G. B. (2003, November). *The influence of phonological neighborhood frequency in visual lexical decision.* Poster presented at the 44th Annual Meeting of the Psychonomic Society, Vancouver, BC.

NELSON, D. L., MCEVOY, C. L., & SCHRIEBER, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms.* Available at www.usf.edu/FreeAssociation.

PEXMAN, P. M., & LUPKER, S. J. (1999). Ambiguity and visual word recognition: Can feedback explain both homophone and polysemy effects? *Canadian Journal of Experimental Psychology*, **53**, 323-334.

QUENÉ, H., & VAN DEN BERGH, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication*, **43**, 103-121.

RUBENSTEIN, H., GARFIELD, L., & MILLIKAN, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning & Verbal Behavior*, **9**, 487-494.

RAAIJMAKERS, J. G. W. (2003). A further look at the "language-as-fixed-effect fallacy." *Canadian Journal of Experimental Psychology*, **57**, 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 & Language*, **41**, 416-426.

RAUDENBUSH, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational & Behavioral Statistics, 18*(4), 321-349.

RAUDENBUSH, S. W., & BRYK, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

STRAIN, E., PATTERSON, K., & SEIDENBERG, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1140-1154.

SINGER, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models and individual growth models. *Journal of Educational & Behavioral Statistics*, **24**, 323-355.

SNIJDERS, T. A. B., & BOSKER, R. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.

SCHNEIDER, W., ESCHMAN, A., & ZUCCOLOTTO, A. (2002). *E-Prime user's guide*. Pittsburgh, PA: Psychology Software Tools.

SIAKALUK, P. D., SEARS, C. R., & LUPKER, S. J. (2002). Orthographic neighborhood effects in lexical decision: The effects of nonword orthographic neighborhood size. *Journal of Experimental Psychology: Human Perception & Performance*, **28**, 661-681.

WIKE, E. L., & CHURCH, J. D. (1976). Comments on Clark's "The

language-as-fixed-effect fallacy." *Journal of Verbal Learning & Verbal Behavior*, **15**, 249-255.

WRIGHT, D. B. (1998). Modeling clustered data in autobiographical memory research: The multilevel approach. *Applied Cognitive Psychology*, **12**, 339-357.

YATES, M., LOCKER, L., & SIMPSON, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review*, **11**, 452-452.

NOTES

1. The data utilized in the present article are presented *only* as a means of illustrating the relevant statistical analyses.

2. The word list originally included 40 items. However, following data collection, it was observed that one item had been miscoded and incorrectly included in the stimulus list. The removal of this item did not affect the results.

APPENDIX
SAS Syntax for Estimating Crossed Random Effects Multilevel Models

```
*Library for data files;
*Replace path with location of .sasb7sat file;
LIBNAME folder 'F:\folder';
/***  Note: These models assume a stacked data structure in which each
row provides the response time for a single subject and a single item.
***/
*SAS: Bringing in data from folder to work library;
DATA Example; SET folder.Example; run;
TITLE 'Eq3: SAS Empty Model: No Random Effects';
PROC MIXED DATA = Example COVTEST NOCLPRINT NOITPRINT METHOD = REML;
        *Observations for subjects and items are considered categorical;
                CLASS Subject Item;
        *RT predicted from intercept only;
                MODEL rt = / SOLUTION DDFM = Satterthwaite; run;
TITLE 'Eq4: SAS Random Effects of Subjects Model';
PROC MIXED DATA = Example COVTEST NOCLPRINT NOITPRINT METHOD = REML;
        *Observations for subjects and items are considered categorical;
                CLASS Subject Item;
        *RT predicted from intercept only;
                MODEL rt = / SOLUTION DDFM = Satterthwaite;
        *Level 2 variance for subjects;
                RANDOM INTERCEPT / SUBJECT = Subject TYPE = UN; run;
TITLE 'Eq5: SAS Random Subjects by Random Items Crossed Model';
PROC MIXED DATA = Example COVTEST NOCLPRINT NOITPRINT METHOD = REML;
        *Observations for subjects and items are considered categorical;
                CLASS Subject Item;
        *RT predicted from intercept only;
                MODEL rt =  / SOLUTION DDFM = Satterthwaite;
        *Level 2 variance for items;
                RANDOM INTERCEPT / SUBJECT = Item TYPE = UN;
        *Level 2 variance for subjects;
                RANDOM INTERCEPT / SUBJECT = Subject TYPE = UN; run;
TITLE 'Eq6: SAS Random Subjects by Random Items Crossed Model';
PROC MIXED DATA = Example COVTEST NOCLPRINT NOITPRINT METHOD = REML;
        *Observations for subjects and items are considered categorical;
        *Item predictors are also categorical;
                CLASS Subject Item freq size;
        *RT predicted from freq, size, and freq*size;
                MODEL rt = freq|size / SOLUTION DDFM = Satterthwaite;
        *Level 2 variance for items;
                RANDOM INTERCEPT / SUBJECT = Item TYPE = UN;
        *Level 2 variance for subjects;
                RANDOM INTERCEPT / SUBJECT = Subject TYPE = UN;
        *Requesting means per condition;
                LSMEANS freq*size; run;
```

(*Continued on next page*)

APPENDIX (Continued)

SPSS Syntax for Estimating Crossed Random Effects Multilevel Models

```
* Note: SPSS v. 11.5 or higher is required to estimate these models.
* Results reported are from SAS Proc Mixed – SPSS estimates differ slightly.
* In SPSS, BY is equivalent to CLASS in SAS.
* WITH denotes continuous variables.
* FIXED is equivalent to MODEL in SAS.
* EMMEANS is equivalent to LSMEANS in SAS.
* Replace path with location of .sav file.
GET FILE = 'F:\folder\example.sav'.
TITLE 'Eq3: SPSS Empty Model: No Random Effects'.
MIXED rt BY Subject Item
        /FIXED  =
        /METHOD = REML
        /PRINT  = SOLUTION TESTCOV.
TITLE 'Eq4: SPSS Random Effects of Subjects Model'.
MIXED rt BY Subject Item
        /FIXED  =
        /METHOD = REML
        /PRINT  = SOLUTION TESTCOV
        /RANDOM = INTERCEPT | SUBJECT(Subject) COVTYPE(UN).
TITLE 'Eq5: SPSS Random Subjects by Random Items Crossed Model'.
MIXED rt BY Subject Item
        /FIXED  =
        /METHOD = REML
        /PRINT  = SOLUTION TESTCOV
        /RANDOM = INTERCEPT | SUBJECT(Item) COVTYPE(UN)
        /RANDOM = INTERCEPT | SUBJECT(Subject) COVTYPE(UN).
TITLE 'Eq6: SPSS Crossed Predictor Model'.
MIXED rt BY Subject Item Freq Size
        /FIXED  = Freq Size Freq*Size
        /METHOD = REML
        /PRINT  = SOLUTION TESTCOV
        /RANDOM = INTERCEPT | SUBJECT(Item) COVTYPE(UN)
        /RANDOM = INTERCEPT | SUBJECT(Subject) COVTYPE(UN)
        /EMMEANS TABLES (Freq*Size).
```

(Manuscript received January 16, 2006;
revision accepted for publication September 11, 2006.)