second
language
research

Second Language Research
28(3) 369–382

DOI: 10.1177/0267658312443651
slr.sagepub.com
SAGE

An overview of mixed-effects statistical models for second language researchers

Ian Cunnings
University of Edinburgh, UK

Abstract

As in any field of scientific inquiry, advancements in the field of second language acquisition (SLA) rely in part on the interpretation and generalizability of study findings using quantitative data analysis and inferential statistics. While statistical techniques such as ANOVA and *t*-tests are widely used in second language research, this review article provides a review of a class of newer statistical models that have not yet been widely adopted in the field, but have garnered interest in other fields of language research. The class of statistical models called mixed-effects models are introduced, and the potential benefits of these models for the second language researcher are discussed. A simple example of mixed-effects data analysis using the statistical software package R (R Development Core Team, 2011) is provided as an introduction to the use of these statistical techniques, and to exemplify how such analyses can be reported in research articles. It is concluded that mixed-effects models provide the second language researcher with a powerful tool for the analysis of a variety of types of second language acquisition data.

Keywords

statistics, mixed-effects models, fixed effect, random effects, second language acquisition

1 Introduction

Techniques of statistical analysis used in second language acquisition (SLA) research have increased in sophistication since the field's inception (Loewen and Gass, 2009). In the 1980s, Henning (1986) noted the growing use of quantitative analysis and inferential statistics in second language (L2) research, and agreement emerged within the field that increased literacy in the use of such methods was desired (Lazaraton et al., 1987). By the end of the 1990s, surveys reported that close to 90% of studies in applied linguistics were quantitative in nature, with parametric statistics such as ANOVA and *t*-tests prevailing

Corresponding author:
Ian Cunnings, Department of Psychology, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, EH8 7JZ, UK
Email: ian.cunnings@ed.ac.uk

(Lazaraton, 2000; Norris and Ortega, 2000). More recent surveys indicate reliance on means-based parametric statistical techniques has continued (Plonsky, 2011; Plonsky and Gass, 2011).

As statistical techniques have advanced in SLA, the field of statistics itself has not remained static. Advances in statistical techniques should, naturally, influence research practices in other fields to help ensure that the most insightful analysis of a given dataset is used. In the field of psycholinguistics, a 2008 special issue of the *Journal of Memory and Language* discussed emerging data analysis techniques, and several authors contributed articles on one class of statistical model in particular, namely linear mixed-effects regression modelling, or simply mixed-effects models (e.g. Baayen et al., 2008; Jaeger, 2008; Quene and van den Bergh, 2008).

Assume a study in which two second language (L2) teaching strategies are tested on a sample of nonnative English speakers, and performance on a language test is measured as in indicator of which strategy greater improved proficiency. In this study, the independent variable of interest, 'teaching strategy', is said to be a fixed effect, while the participants randomly sampled from a larger population of L2 learners are a random factor. If we were to replicate the study, the fixed effect ('teaching strategy') would be repeated while new participants would be recruited (see Baayen, 2008: 241–42). As such, fixed effects model the influence that independent variables have on a dataset, whereas random effects model variance arising from random population sampling. A mixed-effects model thus contains both fixed and random effects.[1]

Mixed-effects models have been used in certain subfields of SLA. Given trends in psycholinguistics, it is perhaps not surprising that recent research on nonnative language processing has begun to adopt these analyses (e.g. Baten et al., 2011; Bowden et al., 2010; Tremblay et al., 2011), and mixed-effects models have also been used in research on language testing (e.g. Barkaoui, 2010; Ferne and Rupp, 2007). The aim of this article is to provide an overview of mixed-effects models for L2 researchers more generally.

This review is outlined as follows. Below I first review some of the benefits that mixed-effects models can offer SLA researchers. I will then provide a practical example of a mixed-effects analysis using the statistical software package R (R development core team, 2011). I will continue by providing some guidelines in good practice when reporting the results of mixed-effects analyses, before concluding that mixed-effects models can provide L2 researchers with a powerful new statistical tool.

II Why mixed models?

Mixed-effects models can provide a number of benefits to the L2 researcher. The fixed effects component of a mixed-effects model can contain multiple independent variables of interest to the researcher, including categorical factors (e.g. native vs. nonnative, high proficiency vs. low proficiency etc.), continuous predictors (e.g. age or proficiency, if measured on a continuous scale), or a mixture of the two. Dependent variables in a mixed model can also be continuous, as in the case of reaction times, or categorical, as in a forced-choice grammaticality judgment (see Jaeger, 2008). While there are existing techniques to handle such data, mixed-effects models provide an additional general framework for their analysis.

Another property of mixed-effects models that could be of use in L2 research is that they can model change over time. Such growth curve analyses (Goldstein, 1987, 1995), which do not necessarily have to assume change over time to be linear, could be adopted to examine a number of issues in SLA, such as how L2 acquisition develops over time, or in assessing the success of different teaching strategies in longitudinal studies. Although there is some tradition for longitudinal designs in L2 research (Ortega and Byrnes, 2008), Ortega and Iberri-Shea (2005) noted that it would be beneficial if more advanced statistical techniques were used in the analysis of such data. Mixed-effects models could be used in such cases.

Random effects in mixed-effects analyses can model different types of random effects structures that arise during random population sampling. Consider again a study investigating how teaching strategy influences learner proficiency. To examine this question, a sample of students may be taken from a number of different language classes from a number of different schools. In this case, the sampled students are nested in a hierarchical fashion within classes within schools (Goldstein, 1987). It could be that performance correlates between students within the same class (and school) in a way that is not observed between different classes (and schools), and it would be beneficial to take such variance and covariance into account statistically. Hierarchical mixed-effects models with nested random effects were developed to account for precisely such situations (Goldstein, 1987, 1995; Raudenbush and Bryk, 2002; Snijders and Bosker, 1999). The ability to model such sampling could be of general use to any SLA research that involves sampling of populations within educational establishments.

Samples drawn from classes within schools can also vary in a non-nested fashion. Some of the students within a class may come from the same family, and students from the same family could be in different classes. In this case, families and classes are not nested hierarchically but are instead crossed at the same level of sampling, and mixed-effects analyses can model such crossed random effects as well (Raudenbush, 1993). The ability to model crossed random effects has recently been discussed in psycholinguistics in particular as a solution to Clark's (1973) 'language-as-fixed-effect fallacy'. Clark highlighted that in language research not only are participants randomly sampled from a population of speakers, but so too are language stimuli sampled from a much larger population of linguistic materials. Indeed, that in language a finite number of signs can be combined to make an infinite number of utterances ensures that often the linguistic stimuli used in language research are only a sample of the relevant materials rather than a complete set (Baayen et al., 2008).

To overcome this problem, Clark (1973) suggested two separate analyses be conducted, one in which data are averaged over participants (F_1), and another in which data are averaged over items (F_2). Clark originally proposed that these separate analyses should be combined into a single analysis that takes both participant and item variance into account in a single model. This second step has however not routinely been taken and instead researchers usually assume significance if an effect is reliable by both participants and items (dubbed the $F_1 \times F_2$ criterion; Raaijmakers et al., 1999). There are however problems with this approach. On a practical level, it becomes difficult to interpret results that are significant in one analysis but not the other. This practice also does not really provide a solution to Clark's (1973) objection, as while the participant analysis

takes into account random variance arising from participant sampling and the item analysis random variance arising from item sampling, neither takes both into account simultaneously.

The alternative offered by mixed-effects models is to treat participants and items as crossed random effects. As families of students in classes were considered crossed random effects in our earlier example, so too are the participants and items tested in language research crossed, in that participants in a study are tested on a series of items, and the same items are tested on a series of participants. The ability to include crossed random effects for participants and items in a single mixed-effect model thus provides, in comparison to the $F_1 \times F_2$ criterion, a more satisfactory solution to Clark's (1973) 'language-as-fixed-effect fallacy' (Baayen et al., 2008; Locker et al., 2007; Quene and van den Bergh, 2008).

The reporting of whether data meet the assumptions of statistical tests (e.g. that data are normally distributed) is rare in SLA (Plonsky, 2011; Plonsky and Gass, 2011). Linear mixed-effects models do, like ANOVA, assume a normal distribution, but models for other distributions are available. For example, logit mixed-effects models can be used to analyse data with a binomial distribution (Jaeger, 2008; Quene and van den Bergh, 2008). Mixed-effects models are also robust against violations of sphericity and homoscedasticity (Quene and van den Bergh, 2004, 2008). Finally, SLA researchers are sometimes confronted with unbalanced datasets, as L2 learners can provide high numbers of missing responses in experimental studies. Mixed-effects models are robust against missing data, assuming that the data are missing at random, obviating the need to replace missing values using debatable imputation techniques (Quene and van den Bergh, 2004, 2008).

III An example of mixed-effects models in R

Recent versions of software packages such as SAS, STATA and SPSS can run mixed-effects analyses. Another option is the open-source statistical package R (R development core team, 2011), which is free to download from http://www.r-project.org. Although researchers trained in packages such as SPSS, which provide a graphical user interface, might initially find daunting the fact that R is command-line rather than menu driven, it becomes a highly flexible tool after an initial learning curve. It is beyond the scope of this article to provide an introduction to R syntax. The interested reader is directed to textbooks such as Baayen (2008), Crawley (2007), Dalgaard (2008) or Vasishth and Brow (2011).

The standard installation of R can perform a variety of statistical analyses, and users can download additional packages that offer specialized tools. The package of current interest is lme4 (Bates, 2005), which provides up-to-date tools for running mixed-effects models. The following example is provided both as a demonstration of how mixed-effects models can be used in R, and to more generally highlight some of the benefits of such models.

Consider a fictional study designed to assess L2 acquisition of subject–verb agreement in English. Twenty-four nonnative English speakers, sampled from a population of university students, rated a series of sentences on a scale from 1 (unacceptable) to 10

(acceptable). Twenty pairs of grammatical and ungrammatical sentences were constructed (e.g. 'The books that John bought in town were boring' vs. 'The books that John bought in town was boring'). Two versions of the questionnaire were constructed so that participants only rated one version of each sentence, such that each participant rated 10 grammatical and 10 ungrammatical sentences. Although we are mainly interested in assessing the acquisition of subject–verb agreement by comparing scores in the grammatical and ungrammatical conditions, we may also want to control for other variables that are not of primary interest but which nonetheless vary in our sample and as such could influence the data. In this experiment, these are participant age in years and sentence length in words.

The supplementary file 'ratings.RData' contains this data as an R dataframe object named 'ratings'. This dataframe includes rows for each observation in the experiment, and seven columns. One column (subject) identifies each unique participant (1–24), another (item) identifies each of the 20 experimental sentences, and a third (condition) codes the main independent variable of theoretical interest as either 'g' (grammatical) or 'ug' (ungrammatical). There are also columns for the control variables age and length. Two final columns rating and zrating contain the (untransformed and *z*-score transformed) acceptability ratings.

If we were to analyse this data using traditional methods, such as paired *t*-test, we would first aggregate the data, averaging first over participants (t_1) and second over items (t_2). Rather than being averaged down to the 24 participants and 20 items, mixed-effects analyses require no prior aggregation and are run on the unaveraged data, in this case 480 datapoints.

Overall, the grammatical sentences received a higher mean acceptability rating (5.55, SD 1.83) than the ungrammatical sentences (4.16, SD 1.68). The acceptability ratings, bounded between 1 and 10, were *z*-score transformed (see Schütze and Sprouse, to appear) before further statistical analysis. Mean *z*-scores for the grammatical and ungrammatical sentences were 0.12 (SD 0.71) and –0.45 (SD 0.65) respectively.[2] The lmer() function in R is used to construct mixed-effects models to test the significance of this difference. We can create a model using the following command in R:

```
> model1 = lmer(zrating ~ condition + (1|subject) +
(1|item), data = ratings)
```

This syntax uses the lmer() function to create a mixed-effects model (model1) in which the dependent variable zrating is being analysed in terms of our independent variable, the fixed effect condition. The next part of the formula (1|subject) + (1|item) specifies crossed random effects for participants (i.e. subjects) and items, while the final part specifies which dataframe is being analysed. The summary() function provides a model summary:

```
> summary(model1)
Linear mixed model fit by REML
Formula: zrating ~ condition + (1|subject) + (1|item)
   Data: ratings
```

```
   AIC BIC logLik deviance   REMLdev
 996.1 1017  -493       978    986.1
Random effects:
 Groups    Name         Variance    Std.Dev.
 subject    (Intercept) 0.02748419  0.165784
 item       (Intercept) 0.00070097  0.026476
 Residual               0.43219709  0.657417
 Number of obs: 480, groups: subject, 24; item, 20
 Fixed effects:
             Estimate Std. Error t  value
 (Intercept)   0.11919    0.05460    2.183
 conditionug  -0.56898    0.06001   -9.481
 Correlation of Fixed Effects:
            (Intr)
 conditionug   -0.550
```

The summary first specifies the structure of the model and then provides various measures of model fit, which tell us how much of the variance in the data is being explained by the model. AIC, for example, measures of how much variance is left unexplained by the model, and thus lower scores mean that more variance is being accounted for. Next, the random effects of the model are specified, specifically random intercepts for participants and items, before providing information about the number of observations, participants and items. Finally, the fixed effect condition is provided including the model estimate, standard error and *t*-statistic.

The model above, which contains random intercepts, allows mean values for each participant and each item to vary. For example, some participants may generally rate sentences higher on the scale than others, and some items may generally receive lower ratings than others. The random intercepts allow the model to take such variance into account. We also however need to include random slopes, which account for the fact that different participants and different items may vary with regards to how sensitive they are to the manipulation at hand. For example, some participants may rate the grammatical sentences much higher than the ungrammatical ones, other participants may show this same trend but with a smaller difference, while others may not rate grammatical sentences any higher than ungrammatical sentences (and some may even rate the ungrammatical ones higher). The same can also be said of the different items, and not including random slopes in a model where there is considerable by participant and/or by item variance can lead to drastically increased Type I error rates (Barr et al., submitted; Schielzeth, and Forstmeier, 2009). Random slopes can be included for repeated measures fixed effects. In this experiment, participants were repeatedly measured on grammatical and ungrammatical sentences, and so a participant random slope for condition can be included. To include a participant random slope, we create a second model as follows:

```
> model2 = lmer(zrating ~ condition + (1+condition|subject) +
(1|item), data = ratings)
```

By typing summary(model2) at the R command line and examining the summary of model2, several differences between model1 and model2 become apparent. The AIC score for model2 (984.6) is smaller than for model1 (996.1), indicating that model2 is explaining more of the variance in the data. The *t* value for the fixed effect of condition is also lower in model2 (–6.372) than model1 (–9.481), suggesting that model1, containing random intercepts only, was indeed providing an overconfident estimate of this effect.

In this experiment, grammaticality was manipulated within minimal pairs of sentences, such that each sentence appeared in grammatical and ungrammatical conditions. Assuming grammaticality was thus repeatedly measured within sentences, we can also include an item random slope of condition as below:

```
> model3 = lmer(zrating ~ condition + (1+condition|subject) +
(1+condition|item), data = ratings)
```

This third model has an again lower AIC (978.9), and a smaller *t* value for condition (–5.107). We can formally test whether a particular model provides a significantly improved fit to the data over another using likelihood ratio tests with the anova() function. This function assesses whether the inclusion of a given model parameter results in an improved fit (i.e. whether it explains more of the variance in the data). With this function, we can incrementally compare each increasingly complex model. Below, we compare model1 with model2, and model2 with model3:

```
> anova(model1,model2,model3)
       Df   AIC      BIC  logLik  Chisq  Chi  Df Pr(>Chisq)
model1 5 987.95 1008.8 -488.98
model2 7 977.09 1006.3 -481.55 14.861    2    0.000593 ***
model3 9 971.75 1009.3 -476.88 9.343     2    0.009358 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These results, expressed as a chi-squared statistic, show that model2 provides a significantly better fit to the data than model1 ($\chi^2(2) = 14.861$, $p < .001$), as does model3 in comparison to model2 ($\chi^2(2) = 9.343$, $p = .009$), indicating that the random slopes have improved model fit, and thus need to be included in the model.

A benefit of mixed-effects models is that properties of both the participants and/or the items tested can be included in the analysis. In the current example, these are participant age and sentence length. Under traditional methods, the inclusion of such control predictors would involve various additional analyses. For age, one could include it as a continuous predictor in an analysis of covariance of the participant data (ANCOVA). However, this analysis by participants would not take into account any potential variance between the different experimental items (it could be that age effects are restricted to certain items). For sentence length, an additional ANCOVA could be run on the mean item ratings with length as a covariate to examine whether sentence length affected ratings. However in this case, any variance arising from the participants

(it could be that not all participants were affected by sentence length) would not be taken into account. Including both participant age and item length in a single analysis would be tricky, unless the continuous variables are turned into factors such as in a 2 (condition: grammatical/ungrammatical) by 2 (age: young/old) by 2 (length: short/long) ANOVA. As unaveraged data are analysed with mixed-effects models, properties of both the participants and the items can be included as continuous (or categorical) predictors in a single model.

When including continuous predictors in a mixed-effect model, it is often useful to centre each predictor around its mean value. This involves subtracting from each individual value of a predictor the predictor's overall mean, and is done to help reduce collinearity within the model (e.g. between main effects and interactions; see Jaeger, 2010). The commands below add new columns (clength and cage) to the ratings dataframe containing centred data for length and age:

```
> ratings$clength = ratings$length - mean(ratings$length)
> ratings$cage = ratings$age - mean(ratings$age)
```

We could simply create a new model with these two predictors included. Alternatively, we can again use the anova() function to verify whether the inclusion of additional fixed effects in the model are warranted by the data. In what follows, we incrementally add fixed effects of length and age into increasingly complex models, and then use anova() to examine whether the addition of each fixed effect improves how well the model explains the data. A fourth model includes a fixed effect of (centred) length, while a fifth additionally includes age:

```
> model4 = lmer(zrating ~ condition + clength +
(1+condition|subject) + (1+condition|item), data = ratings)
> model5 = lmer(zrating ~ condition + clength + cage +
(1+condition|subject) + (1+condition|item), data = ratings)
```

The models thus far only include main effects of condition, length and age, but it could be that these variables interact. Sentence length may affect ratings for grammatical but not ungrammatical sentences, and participant age could only be influencing ratings for ungrammatical sentences (indeed, there could even be a three-way interaction). As with the main effects, we can incrementally add these interactions into increasingly complex models (in this syntax ':' denotes specific interactions and '*' is shorthand for 'main effects and all possible interactions'):

```
> model6 = lmer(zrating ~ condition + clength + cage +
condition:clength + (1+condition|subject) + (1+condition|item),
data = ratings)
> model7 = lmer(zrating ~ condition + clength + cage +
condition:clength + condition:cage + (1+condition|subject) +
(1+condition|item), data = ratings)
```

```
> model8 = lmer(zrating ~ condition + clength + cage +
condition:clength + condition:cage + clength:cage +
(1+condition|subject) + (1+condition|item), data = ratings)
> model9 = lmer(zrating ~ condition * clength * cage +
(1+condition|subject) + (1+condition|item), data =
ratings)
```

To examine whether any of these additional main effects and interactions actually do improve model fit to the data, we successively compare each increasingly complex model and then report the most complex model that provides a significantly improved fit to the data:

```
> anova(model3,model4,model5,model6,model7,model8,model9)

       Df    AIC     BIC  logLik   Chisq Chi Df Pr(>Chisq)
model3  9 971.75 1009.3 -476.88
model4 10 969.56 1011.3 -474.78 4.1878       1   0.04072 *
model5 11 971.38 1017.3 -474.69 0.1813       1   0.67026
model6 12 972.42 1022.5 -474.21 0.9586       1   0.32755
model7 13 974.47 1028.7 -474.23 0.0000       1   1.00000
model8 14 976.23 1034.7 -474.11 0.2400       1   0.62423
model9 15 977.87 1040.5 -473.94 0.3536       1   0.55210
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that model4, which includes the main effect of length, provides a significantly improved fit over model3 containing the main effect of condition only. Neither the main effect of age nor any of the interactions provide any better fit. It thus seems that model4, containing fixed effects of condition and length (but no interaction), plus participant and item random intercepts and random slopes for condition, is the most complex model justified by the data. A final check examines whether a random slope for length is warranted. It could be that only certain participants are affected by sentence length. Note that we cannot include an item random slope for length as, while this was a repeated measure for participants, it is a property that varies between rather than within the experimental sentences.

```
> model10 = lmer(zrating ~ condition + clength +
(1+condition+clength|subject) + (1+condition|item), data =
ratings)
```

The anova(model4,model10) function, however, indicates that the addition of this random slope provides no better model fit ($\chi^2(3) = 2.4806$, $p = .479$). As such, model4 is the most complex model warranted by the data.

One thing immediately noticeable if one uses the summary(model4) command (as in the earlier example for model1) is that *t* values for fixed effects are provided but not accompanying *p* values. The calculation of exact *p* values is not straight forward for

mixed-effects models where, for example, it is not yet understood how the degrees of freedom should appropriately be calculated (see Baayen et al., 2008: 396; Bates, 2006). There are however ways in which *p* values can be estimated. Baayen (2008: 248) describes that *p* values can be estimated based on the *t* distribution using the following formula:

```
2 * (1 - pt(abs(X), Y - Z))
```

Where X is the *t* value, Y is the number of observations, and Z is the number of fixed effect parameters. In the case of the fixed effect of condition, *t* is –4.980, the number of observations 480 and the number of fixed effects parameters 3 (condition, length and the intercept), resulting in $p < .001$. For the fixed effect of length, $p = .032$.

Baayen (2008) and Baayen et al. (2008) explain that this estimated *p* value can be anticonservative (i.e. has an increased risk of Type I error) for small datasets. They advocate another way of estimating *p* values using the pvals.fnc() function from the LanguageR package (for discussion of how this *p* value is calculated, see Baayen, 2008: 247–48; Baayen et al., 2008: 396–99). However, using this function on model4 with the command pvals.fnc(model4) results in an error, as currently the pvals.fnc() function is not yet implemented for the types of model discussed here that contain random slopes. Where possible then, it is advised that pvals.fnc() should be used to estimate *p* values, otherwise they can be estimated as above, with the caveat that this method is anticonservative for small datasets.

We can thus now report the results of our simulated experiment, where we constructed a mixed-effects model containing a fixed effect of condition (grammatical vs. ungrammatical), random intercepts for participants and items and participant and item random slopes for condition. To control for any effects of participant age and sentence length, likelihood ratio tests examined whether the inclusion of these fixed effects improved model fit. These indicated that only a main effect for sentence length significantly improved model fit to the data ($\chi^2(1) = 4.19$, $p = .041$). Analysis of this model indicated that ungrammatical sentences were rated less acceptable than grammatical sentences (estimate = –0.57, SE = 0.11, $t(480) = –4.98$, $p < .001$). Additionally, longer sentences received lower ratings than shorter sentences (estimate = 0.04, SE = 0.02, $t(480) = –2.15$, $p = .032$).

It should be clear from this example that researchers need to be explicit when reporting the structure of a mixed-effects model. Unfortunately, this has not always been the case in language research thus far (Barr et al., submitted). For example, some studies have reported that random effects for participants and items were included without specifying whether this included random intercepts or random slopes. Given that random intercept only models can drastically inflate Type I error rates (Barr et al., submitted; Schielzeth and Forstmeier, 2009), it is imperative that precise information about the random effects structure of mixed-effects models be reported.

As a relatively new tool in language research, the conventions for best practice in the use of mixed-effects models are still under debate. Baayen (2008) and Baayen et al. (2008) use a model selection approach to analysis, similar to the example here, where successively complex models are compared to each other using the anova() function, and

the model that provides the best fit to the data is reported. This type of data-driven approach is similar to (often exploratory) analysis of large corpora containing multiple dependent and independent variables, where model selection can help in providing an understanding of precisely which statistical model provides the best explanation of the data. While model selection may well be suited to such cases, Barr et al. (submitted) argue that it is not appropriate for the analysis of carefully designed experiments. In experimental research, the researcher is testing specific hypotheses, and the statistical model should reflect these hypotheses on theoretical rather than data-driven grounds. They argue that experimental designs should instead use what they call 'maximal' random effects structures; that is, models containing random intercepts and slopes for every fixed effect of theoretical interest.

It is thus suggested that researchers follow Barr et al.'s recommendations and use mixed-effects models with 'maximal' random effects structures when analysing experimental data. Consider the example in this article. Of main theoretical interest was the fixed effect of 'condition', indeed the study was designed to test L2 acquisition of English subject–verb agreement. As such the 'maximal' model, containing participant and item random intercepts and participant and item random slopes for 'condition', should be used. In addition to this independent variable, two additional control variables were included that were not of prime theoretical interest (the study did not set out to examine how age and sentence length affect acquisition of subject–verb agreement), but their potential influence needed to be controlled. For such control variables, it may be that model selection is an appropriate tool for examining which need to be included in the statistical model.

IV Conclusions

In this review article I have discussed some of the benefits that mixed-effects models can provide to SLA researchers. Baayen et al. (2008) emphasize the key benefit of mixed-effects models as their ability to simultaneously include participant-level and item-level factors in a single analysis, as this facilitates a level of understanding of a dataset that is not possible with techniques that require averaging over participants or items.

Although many statistical tools are available to L2 researchers, Plonsky (2011) suggests that the default use of ANOVA in SLA may have shaped research practices in favour of factorial designs even when other methods of analysis would be more appropriate. In particular, Plonsky notes that the practice of converting continuous data (e.g. proficiency) into factorial categories (low vs. high) trades a loss of variance, and thus statistical power for what appears to be a cleaner statistical approach. It is hoped that the ability to model multiple categorical and continuous predictors in mixed-effects analyses can help steer researchers away from such practices, and instead promote use of the most appropriate and powerful analysis of a given dataset.

As with all analysis techniques, mixed-effects models have their limitations. As a new technique in language research, one current potential problem is the lack of rigid standards in the correct use and reporting of such analyses. It is hoped that the suggestions outlined in Section III can help L2 researchers in this regard, or at least make them aware of such issues, and I again emphasize the importance of transparency in reporting the

fixed and random effects structures used in mixed-effects analyses in research articles. Another potential issue that is frequent in language research relates to how populations are sampled. Samples in language research are sometimes not entirely random but are instead simply samples of those participants who were available at the time. As such, precisely what populations these samples are meant to be representative of is not always entirely clear (e.g. specific university students, learners of a specific age range, learners of a particular language, or indeed L2 learners in general?).[3] However, this issue of how research findings generalize is an interpretive problem whichever type of analysis is adopted. The ability to model different random effects structures – both nested (e.g. students within classes within schools) and crossed (e.g. students from the same families in different classes) – that arise during sampling is one property of mixed-effects models that could potentially help address this issue, assuming one's dataset is varied and large enough to model such effects.

It is hoped that the example experiment provided in this review article highlights how L2 researchers can apply this approach to experimental designs. Of course, there are many other designs that could have been exemplified but which were not possible given space limitations. One related design is the use of categorical responses (e.g. grammatical vs. ungrammatical) rather than scales of acceptability. Standard practice in such designs involves computing mean proportions of grammatical responses and then running statistical analyses on the averaged data. The use of ANOVA, which assumes a normal distribution, is notoriously problematic in such cases. Binomial logit mixed-effects models, which require no prior averaging of data, offer a more appropriate solution to the analysis of such data (Jaeger, 2008). In addition to experimental designs, mixed-effects models can also be used in the analysis of large corpora and also, as briefly touched upon earlier, in longitudinal studies, where they could be used to examine developmental change in L2 acquisition over time.

To conclude, mixed-effects models provide a flexible and powerful tool for the analysis of a variety of data types. For the interested L2 researcher, such models are implemented in recent versions of many statistical packages, including R, and commercial programs such as SPSS. Although SLA researchers already have a wide variety of analysis techniques at their disposal, mixed-effects models should be considered as a powerful addition to the existing arsenal of statistical tools.

Acknowledgements

I would like to thank Martin Corley, Ian Finlayson, Patrick Sturt and João Veríssimo for various insightful discussions on the use of mixed-effects models. I would also like to thank the Developmental Linguistics Research Group at the University of Edinburgh, an anonymous *Second Language Research* reviewer and the editor for comments on earlier drafts of this review. All remaining errors are my own. This work was completed whilst funded by a British Academy Postdoctoral Fellowship (PF 100026).

Notes

1. Analyses such as ANOVA are 'mixed' in the sense of including fixed and random effects. However, in recent psycholinguistic research, the term 'mixed-effect model' has been reserved for the class of hierarchical linear mixed effects models reviewed in this article.

2. In reality, participants would rate both experimental items and fillers, and *z*-scores would be calculated using all items. In this example, *z*-scores are calculated using the 20 experimental items, and 20 randomly generated fillers not included in the ratings dataframe.
3. I thank an anonymous *Second Language Research* reviewer for raising this issue.

References

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

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

Barkaoui K (2010) Explaining ESL essay holistic scores: A multilevel modelling approach. *Language Testing* 27: 515–35.

Barr D, Levy R, Scheepers C, and Tily H (submitted) Random effects structure in mixed-effects models: Keep it maximal. Unpublished manuscript, University of Glasgow, UK.

Baten K, Hofman F, and Loeys T (2011) Cross-linguistic activation in bilingual sentence processing: The role of word class meaning. *Bilingualism: Language and Cognition* 14: 351–59.

Bates D (2005) Fitting linear models in R: Using the lme4 package. *R News* 5: 27–30.

Bates D (2006) [R] lmer, p-values and all that. Post to the R-help mailing list, 19 May 2006. Retrieved from: https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html (May 2012).

Bowden H, Gelfand M, Sanz C, and Ullman M (2010) Verbal inflectional morphology in L1 and L2 Spanish: A frequency effects study examining storage verses composition. *Language Learning* 60: 44–87.

Clark H (1973) The language-as-fixed-effect fallacy: A critique of language statistics in psychology research. *Journal of Verbal Learning and Verbal Behavior* 12: 335–59.

Crawley M (2007) *The R book*. Chichester: John Wiley and Sons.

Dalgaard P (2008) *Introductory statistics with R*. 2nd edition. New York: Springer.

Ferne T and Rupp A (2007) A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly* 4: 113–48.

Goldstein H (1987) *Multilevel models in educational and social research*. London: Griffin.

Goldstein H (1995) *Multilevel statistical models*. London: Arnold.

Henning G (1986) Quantitative methods in language acquisition research. *TESOL Quarterly* 20: 701–08.

Jaeger F (2008) Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59: 434–46.

Jaeger F (2010) Common issues and solutions in regression modelling (mixed or not). Presentation at Brain and Cognitive Sciences (BCS), University of Rochester, UK, 4 May 2010. Retrieved from: http://wiki.bcs.rochester.edu/HlpLab/StatsCourses?action=AttachFile&do=view&target=lecture2McGill.pdf (May 2012).

Lazaraton A (2000) Current trends in research methodology and statistics in applied linguistics. *TESOL Quarterly* 34: 175–81.

Lazaraton A, Riggenbach H, and Ediger A (1987) Forming a discipline: Applied linguists' literacy in research methodology and statistics. *TESOL Quarterly* 21: 263–77.

Locker L, Hoffman L and Bovaird J (2007) On the use of multilevel modelling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods* 39: 723–30.

Loewen S and Gass S (2009) The use of statistics in L2 acquisition research. *Language Teaching* 42: 181–96.

Norris J and Ortega L (2000) Effectiveness in L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning* 50: 417–528.

Ortega L and Byrnes H (2008) *The longitudinal study of advanced L2 capacities*. New York: Routledge.

Ortega L and Iberri-Shea G (2005) Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics* 25: 26–45.

Plonsky L (2011) Study quality in SLA: A cumulative and developmental assessment of designs, analyses, reporting practices, and outcomes in quantitative L2 research. Unpublished doctoral thesis, Michigan State University, MI, USA.

Plonsky L and Gass S (2011) Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning* 61: 325–66.

Quene H and van den Bergh H (2004) On multi-level modelling of data from repeated measures designs: A tutorial. *Speech Communication* 43: 103–21.

Quene H and van den Bergh H (2008) Examples of mixed-effects modelling with crossed random effects and with binomial data. *Journal of Memory and Language* 59: 413–25.

R development core team (2011) R: *A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Raaijmakers J, Schrijnemakers J, and Gremmen G (1999) How to deal with the 'the language as fixed effect fallacy': Common misconceptions and alternative solutions. *Journal of Memory and Language* 41: 413–26.

Raudenbush S (1993) A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics* 18: 321–49.

Raudenbush S and Bryk A (2002) *Hierarchical linear models: Applications and data analysis methods*. 2nd edition. Thousand Oaks, CA: Sage.

Schielzeth H and Forstmeier W (2009) Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology* 20: 416–20.

Schütze C and Sprouse J (to appear) Judgment data. In: Podesva R and Sharma D (eds) *Research methods in linguistics*. Cambridge: Cambridge University Press.

Snijders T and Bosker R (1999) *Multilevel analysis*. London: Sage.

Tremblay A, Derwing B, Libben G, and Westbury C (2011) Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning* 61: 569–613.

Vasishth S and Broe M (2010) *The foundations of statistics: A simulation-based approach*. Heidelberg: Springer.