The utility and application of mixed-effects models in second language research

Article

Accepted Version

The Utility and Application of Mixed Effects Models in Second Language Research

Jared A. Linck [a] and Ian Cunnings [b]

[a] University of Maryland Center for Advanced Study of Language

[b] School of Psychology and Clinical Language Sciences, University of Reading

Introduction

Researchers in second language acquisition (SLA) face particular challenges when attempting to generalise study findings to a wider population. Study participants in SLA are often taken from convenience samples that are not truly random. For example, in a sample of students from local schools, the observations from individual students may be 'clustered' into classes, which in turn may be clustered by schools. It could be that performance within classes (and schools) may correlate in a way that is not observed between classes (and schools). The researcher would want to take this random variation within and between classes and schools into account to be sure that study findings generalise across the wider population. Additionally, second language (L2) learners constitute a heterogeneous group, in which participants within and across studies may differ in several non-trivial ways, such as language background, proficiency, length and type of language exposure, amongst many factors. It is often the case that researchers will average data across samples in an attempt to neutralise the effects of these many possible individual differences. While such averaging may lead to a 'cleaner' analysis, important individual differences between participants may be overlooked.

In this paper, we provide an overview of a statistical analysis technique that applied linguists and L2 researchers might find useful in dealing with these and related problems, namely mixed-effects models. Hierarchical mixed-effects models were devised to deal with precisely the types of clusterings of observations that are often found in SLA settings (see e.g. Goldstein, 1987; Raudenbush & Bryk, 2002). Mixed-effects models also easily allow for the inclusion of multiple participant-level and stimulus-level independent variables in a single analysis, potentially offering a fruitful way of examining how individual differences may affect L2 acquisition. We begin by providing an overview of mixed-effects models and their potential

benefits for L2 researchers, before providing a practical example of how such analyses can be conducted using the statistical software package R (R Development Core Team, 2014). As a relatively new advancement in statistical analysis in the language sciences, standards of best practice in the use of mixed-effects models are still being developed. We will thus conclude by offering some advice on how researchers using mixed-effects models might best report such analyses. R code and an example dataset are provided as online supplemental materials.

Mixed-effects models

Consider a study investigating the processing of English agreement morphology in German and Chinese learners of L2 English. The researcher may construct a series of sentences with grammatical and ungrammatical agreement morphology, and then have participants read the sentences on a computer and press a button to measure the time taken to read each sentence. In this study, the researcher would want to examine how the independent variables of interest, L1 background (German vs. Chinese) and sentence grammaticality (grammatical vs. ungrammatical), influence the dependent variable, the reaction times for the sentences. One hypothesis that could be tested would be that adequate acquisition of English agreement morphology should lead to longer reaction times for ungrammatical than grammatical sentences. In a statistical analysis of such data, the independent variables are modelled with *fixed* effects, while random variation in the sample is modelled using *random* effects. A model with both fixed and random effects is a *mixed effects* model.

It could be that the researcher tested German learners from three different classes in the same language school, and Chinese learners from three classes as well. As mentioned above, in this type of design the students can be thought of as being hierarchically 'clustered' into different classes (Goldstein, 1987). The statistical analysis of the reaction time data will obviously need to

take into account random variance across the students sampled, but the different classes may also add random variance that should also be taken into account. For example, assume that *on average* the Chinese learners had slower reaction times than the German learners. From such a finding one might conclude that L1 language background influences the processing of English agreement morphology. However, this conclusion might be premature if the clustering of students into classes is ignored. An assumption of parametric statistical tests is that individual observations are independent of each other. However, observations within classes are not truly independent, as the performance of students within the same class may correlate in a way that is not observed between students in different classes. For example, one class may have a particularly good teacher that makes performance in that specific class different to other classes. In the current example, it could be that only one particular class of Chinese learners performed slowly, while the other two classes behaved more similarly to the German learners. In this case, once the random variation between classes is taken into account, it would be premature to make any strong conclusions about the role of L1 background in the processing of English agreement morphology. Mixed-effects models with *hierarchical* or *nested* random effects were developed to account for this type of nested random variation (Goldstein, 1987, 1995; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999).

Students in classes can also vary in a non-nested fashion. For example, students in the same class may come from different families, and students from the same family might be in different classes. In this case, while students are nested in both classes and families, classes are not nested under families and neither are families nested under classes. Rather, classes and families are *crossed* at the same level of sampling. Mixed-effects models can model both nested

and crossed sources of random variation using nested and crossed random effects (Raudenbush, 1993).

In this example, L1 background is tested *between* two groups of learners. Performance on the reaction time task may however be tested *within* participants. For example, the researcher may have adopted a repeated measures design in which all the participants rated a series of ten grammatical and ten ungrammatical sentences. The data in this type of repeated measures study may vary randomly in different ways. For example, individual participants may differ randomly in their overall reaction times. Some particularly alert participants may *on average* have relatively faster reaction times than other participants (irrespective of grammaticality), while other participants may *on average* have slower reaction times (a slow participant may have randomly had a particularly bad nights' sleep for example). Additionally, participants may differ in their sensitivity to the grammaticality manipulation. Some participants may have much slower reaction times for ungrammatical than grammatical sentences, while other participants may show a similar trend but with a smaller difference. Some participants may have similar reaction times for grammatical and ungrammatical sentences, while some may even have slower reaction times for grammatical than ungrammatical sentences. Different types of random effects are required to model these different types of random variation. A random *intercept* takes into account how each participant's *average* reaction times (irrespective of grammaticality) may differ, while random *slopes* are required to take into account any variability in sensitivity to the repeated measures grammaticality manipulation (see Barr, Levy, Scheepers & Tily, 2013 for further discussion). In an entirely between-groups design, only random *intercepts* are required to account for the random variation in the data. In our example, participants are either in the Chinese or the German group, so we cannot estimate a by-participant slope for 'language group' (a participant cannot be

repeatedly tested on *both* the values of 'language group' – they are in either the Chinese or German group). For within group variables, random *slopes* are required to account for the random variation in the repeated measures. It is imperative to stress the importance of random *slopes* in a repeated measures design, as not including random slopes in the presence of considerable random slope variance can lead to drastically increased Type I error rates (for further discussion, see Barr 2013; Barr et al. 2013; Schielzeth & Forstmeier, 2009).

A standard analysis of this type of study may involve calculating an average reaction time for the grammatical and ungrammatical conditions for each participant and then submitting these averages to a 2x2 ANOVA with the factors L1 background (German vs. Chinese) and grammaticality (grammatical vs. ungrammatical). This analysis would test whether the results generalise from the L2 learners sampled to the wider learner population. As highlighted by Clark (1973) however, not only can we consider the participants in the study as being sampled from a wider population, the same too can also be argued for the linguistic materials. That is, the ten grammatical and ungrammatical sentences tested are a sample of the *possible* English sentences with these properties. This leads to the possibility that it could be, for example, that any results in the analysis averaged over participants are carried largely by a few of the experimental items rather than the item set as a whole. To overcome this 'language-as-fixed-effect' fallacy, Clark originally suggested that a single analysis (dubbed min F') be performed that takes into account random variation arising from both the participants and the materials tested. In practice however, researchers have tended to conduct an analysis with the data averaged over participants (the F_1 analysis) and a second analysis averaged over the linguistic items (F_2). A result is then considered reliable if it is significant by both participants and items. However, while the F_1 analysis takes into account random subject variance and the F_2 analysis random item variance,

neither analysis provides a true solution to Clark's problem, as they do not take *both* sources of variation into account at the same time. Mixed-effects models offer a solution to this problem. Just as classes and families can be considered *crossed* random effects, so too the participants and items in a language experiment are *crossed* at the same level of sampling. As such, mixed-effects models with *crossed* random effects for subjects and items offer a better solution to the 'language-as-fixed-effect' fallacy than separate subjects and items analyses (Baayen, Davidson & Bates, 2008; Locker, Hoffman & Bovaird, 2007).[1]

Traditional ANOVA requires a balanced dataset with no missing cells. One reason data is averaged over participants (and items) is to ensure this assumption is met. While averaging obviates problems with regards to missing data, it leads to the possibility that the averages for each subject and/or item are not based on the same amount of data (if individual data points are missing, the participant/item averages will not all be based on the same number of observations). Mixed-effects models do not require balanced datasets and can be conducted on the raw data with no prior averaging. It is this non-averaging of data that allows for the simultaneous estimation of crossed random effects. In the case of subjects and items, this also means that it is possible for the researcher to include any number of participant-level and item-level covariates in a single analysis, assuming there is sufficient data to model such effects, allowing for a level of analysis not possible in procedures that require prior averaging (Baayen et al. 2008).

Parametric statistics should only be used if assumptions about the data are met. The reporting of whether assumptions are met is rare in SLA (Plonsky, 2011; Plonsky & Gass, 2011). Mixed-effects models with a continuous dependent variable make similar assumptions regarding

[1] See Barr et al. (2013) for in-depth discussion and comparison of how traditional ANOVA analyses and mixed-effects models perform, particularly with regards to protection from Type I error rates, based on a series of simulations.

normality as ANOVA. However, models for other distributions are available. Logit mixed-effects models for example can be used to analyse data with a binomial dependent variable, such as a binary grammaticality judgement (see Jaeger, 2008 for review). Mixed-effects models do not make assumptions of homoscedasticity or sphericity and are robust against missing data, assuming that it is missing completely at random (Quene & van den Burgh, 2008). These properties make mixed-effects models not only suitable for the analysis of standard experimental paradigms that researchers may typically analyse with ANOVA, but also other types of unbalanced paradigms with missing data, such as corpus analyses and longitudinal studies (see e.g. Collins, 2006; Goldstein, 1987, 1995; Raudenbush, 2001; Singer, 1998).

In the following section, we provide a practical example of how such analyses can be carried out in R (R Core Team, 2014). R is an open source, command-line driven statistical software package. It is beyond the scope of the current chapter to provide an in-depth introduction into R syntax. We direct the interested reader to Baayen (2008), Gries (2013) and Vasishth and Broe (2010) for accessible introductions, which also include chapters on mixed effects models. See also Cunnings (2012) and Cunnings and Finlayson (under review) for further worked examples, including longitudinal analysis.

Sample data: Linck et al. (2009) immersion study

In this section, we reanalyze data from a previously published study as a worked out demonstration of how to fit and interpret mixed effects models in R. In the original study, the authors examined the impact of the context of L2 learning on L1 and L2 lexical processing for adult learners (for a more complete description, see Linck, Kroll, & Sunderman, 2009). Two participant groups were included in the analysis: a group of immersed learners studying abroad, and a comparison group of classroom learners at their home university who had no immersion

experience. For this demonstration, we focus on the data from a translation recognition task. In this task, participants were presented with a sequence of word pairs – an L2 word followed immediately by an L1 word. Participants were instructed to indicate, with a button press, whether the two words were correct translations of one another. The materials included correct translations requiring a 'yes' button response (e.g., cara—face), and a variety of incorrect word pairs requiring a 'no' button response. The distractor ('no') trials included critical item word pairs that were related in form (e.g., cara—card) or meaning (e.g., cara—head), and control item word pairs that were unrelated to one another (e.g., cara—lake). To the extent that participants were affected by the form or semantic relationship of the critical distractor pair, the authors expected to see slower responses on those items compared to the unrelated control distractor items, i.e. they predicted a *relatedness effect*. The critical research question that we focus on here was: does this relatedness effect vary by group (immersed vs. classroom) and by distractor type (form vs. semantic)?[2]

This research question reflects a 2 (group) x 2 (distractor condition) x 2 (relatedness) factorial design. For the traditional by-subjects (or F_1) ANOVA approach, we would model the aggregated data using a mixed model ANOVA, with group being a between-subjects factor and both distractor condition and relatedness being within-subjects factors. The distinction between between- and within-subjects factors is important in appropriately partitioning the variance in this dataset. In particular, the within-subject repeated measures ANOVA accounts for the structure inherent to the dataset by explicitly modeling the relationships between data points (e.g., different condition means coming from the same participant). This is precisely what we

[2] The original analysis also included the factor grammatical class (i.e., whether the two words came from the same or different classes). However, here we exclude this factor to simplify the analyses and exposition, and thus the results will not exactly match those initially reported by Linck et al. (2009).

aim to do when fitting mixed effects models – appropriately account for the structure in the data by explicitly modeling these relationships. Consider the mixed effects analog to the F_1 ANOVA. For the fixed effects, we would include the factorial combination of the three factors of group, distractor condition, and relatedness. The random components account for the fact that we have tested multiple participants and importantly have multiple observations per participant (i.e., participants have been measured repeatedly). For this, we include random *intercepts* to account for overall mean differences between subjects, and random *slopes* to allow sensitivity of the repeated measures factors (distractor condition, relatedness, and their interaction) to vary by subject.

Setting up the dataset

To fit a mixed effects model in R, the first step is to ensure that the dataset is setup in the "long format" with a separate row for each unique observation (i.e., trial, with multiple rows per subject) and columns indicating grouping factors. The example below demonstrates this format. The data from this example are available in the "rt_data.txt" supplementary file available on the journal's website. Here, the `head` function in R is used to show the top six rows of the R dataframe `rtdata`, which we will subsequently analyse using mixed effects models.

```
> head(rtdata)

  Subject item     z.acc related  type  group    RT
1     301   59 -1.599547    -0.5  -0.5   -0.5  1159
2     301   63 -1.599547    -0.5  -0.5   -0.5  1449
3     301   57 -1.599547    -0.5  -0.5   -0.5   482
4     301   58 -1.599547    -0.5  -0.5   -0.5   558
5     301   61 -1.599547    -0.5  -0.5   -0.5   817
6     301   62 -1.599547    -0.5  -0.5   -0.5   544
```

In this example, the `RT` column contains the reaction time data from each individual trial. The column `Subject` identifies each participant, while `item` identifies the different linguistic stimuli tested. The `related` and `type` columns identify the experimental manipulation, the

`group` column denotes the participant's group membership, and the `z.acc` column is a subject-level covariate (standardized z-score of each participant's overall percent correct on the translation recognition task). Note that subject-level covariates are merged with the trial-level data, leading to each subject's `z.acc` value being repeated on all trials; this is how subject-level predictors should be coded for use in `lmer`. Once in this format, mixed effects models can be fit to the data.

As with standard regression, it is important to consider the coding scheme used with categorical factors (e.g., treatment coding, contrast coding), as they impact the interpretation of the model coefficients (see Gelman & Hill, 2007; Pedhazur 1997; Raudenbush & Bryk, 2002).[3] By default, R applies treatment coding to all character vectors and factors. Using treatment coding, a reference level is defined for the categorical factor and all other levels are compared to that reference level. For the present analyses however, the three categorical factors were recoded to use contrast coding in order to more closely match the inferences drawn from ANOVA. This was done by converting each predictor variable into a numeric variable with the values of -0.5 and 0.5 (e.g., for `related`, unrelated = -0.5, related = 0.5; see sample code in Online Supplemental Materials). The result can be seen above in the output from the `head` function. Contrast coding is recommended with two-level factors, as it can prevent some convergence issues by reducing multicollinearity among the predictors.

Mixed effects models in R using 'lme4'

The `lme4` package (Bates, Maechler, Bolker, & Walker, 2013) in R is used to fit mixed effects models by calling the `lmer` function. We explore the syntax and results of mixed effects

[3] For a comparison of various coding schemes, see http://www.ats.ucla.edu/stat/r/library/contrast_coding.htm

models fit with the `lmer` function by building up from simple to more complex statistical models fit to the example dataset.

Model 1: Factorial effects varying by subject

The primary research question in our working example is whether the effects of lexical and semantic relatedness differ between the two groups of learners (classroom vs. immersed). The target inferences are captured by the factorial combination of the within-subjects factors of `related` and `type` and the between-subjects factor of `group`. Because different individuals might respond differently to the within-subjects factors, we want to allow the two repeated measures main effects and their interaction effect to vary randomly by subjects. This is achieved by including by-subject random slopes for these effects. Note that this specification follows precisely from the traditional F_1 repeated measures ANOVA, except that we can fit the model to the raw observed data without need for aggregation within conditions. This factorial model would be fit to the `rtdata` dataframe with the following code:

```
> m.factors.Rsub <- lmer(data = rtdata,
    formula = RT ~ related*type*group + (1+related*type|Subject))
```

R is an object-oriented language. That is, when fitting a model, rather than calling a process and waiting for the results to print to the screen, you instead create a data object where the output from the analysis is stored for later examination. The left-pointing arrow `<-` is an assignment operator and instructs R to do whatever is on the right side of the arrow and save it in the object on the left. In this case, we create a mixed effects model object called `m.factors.Rsub` using the `lmer` function (note the object name is arbitrary). The first argument of the `lmer` syntax,

`data = rtdata`, specifies the dataframe being analysed.[4] Then, the formula for the model is specified with the dependent variable `RT` on the left side of the tilde (`~`) and the fixed and random effects on the right. We first specify the factorial fixed effects with the code `related * type * group`, while `(1+related*type|Subject)` denotes the by-subject random effects. Specifically, the code specifies a random intercept for subjects and by-subject random slopes for the two repeated measures main effects and their interaction. Note that the code `(1|Subject)` could be used to specify random intercepts only (but no slopes) for each subject. Here however, as the main effects and interaction for `related` and `type` are repeated measures manipulations, random slopes for these effects are included.[5] However, as `group` is a between-subjects factor, a by-subject random slope for `group` cannot be included in the model (i.e., subjects cannot vary on the effect of group, because they are either classroom or immersed learners). Note also, that we cannot include random slope interactions when a repeated measure interacts with a between subjects variable. Although a random slope for the `related` by `type` repeated measures interaction is included, we do not include a random slope for the `related` by `group` interaction, or any of the other possible interactions that involve the between groups manipulation (see Barr, 2013). Once the model is fit, we can display the results using the `summary` command on the resulting model object as below.

```
> summary(m.factors.Rsub)
Linear mixed model fit by REML ['lmerMod']
```

[4] By explicitly specifying each argument (e.g., `data = rtdata`), the order of arguments in the call to `lmer` is flexible; that is, we could rearrange them to first specify the formula followed by the data and the model be the same.

[5] Note that the asterisks (*) in this formula are a shorthand for the factorial combination of mains effects and interactions. Specific interactions can be specified with a colon (`:`). For example, the factorial combination of random slopes here could also be specified with the code `related + type + related:type`.

```
Formula: RT ~ related * type * group + (1 + related * type | Subject)
   Data: rtdata

REML criterion at convergence: 36741.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4030 -0.6418 -0.2238  0.4117  4.4114

Random effects:
 Groups   Name         Variance Std.Dev. Corr
 Subject  (Intercept)   41690.3 204.18
          related         754.0  27.46   -0.73
          type            760.7  27.58   -0.80  0.99
          related:type   2687.2  51.84   -0.06  0.73  0.65
 Residual              112885.5 335.98
Number of obs: 2534, groups:  Subject, 45

Fixed effects:
                   Estimate Std. Error t value
(Intercept)         937.839     31.365  29.901
related              51.644     14.112   3.660
type                  1.178     14.125   0.083
group               -31.411     62.730  -0.501
related:type         52.149     28.104   1.856
related:group        18.906     28.224   0.670
type:group          -21.843     28.249  -0.773
related:type:group  108.375     56.209   1.928

Correlation of Fixed Effects:
            (Intr) relatd type   group  rltd:t rltd:g typ:gr
related     -0.193
type        -0.219  0.124
group       -0.111  0.023  0.028
related:typ -0.006  0.097  0.115  0.004
related:grp  0.023 -0.100  0.000 -0.193  0.005
type:group   0.028  0.000 -0.100 -0.219 -0.005  0.124
rltd:typ:gr  0.004  0.005 -0.005 -0.006 -0.100  0.097  0.115
```

The resulting output contains three main sections: the model summary, the random effects components, and the fixed effects (and associated correlation matrix). The first few lines contain the model summary, including a statement of the type of model (here fit with restricted maximum likelihood, or REML) and the model formula.

The second output section provides information on the random components and the number of observations, including the number of unique levels of the grouping factor(s) – here, Subjects. In the table of random effects, any varying intercept(s) and slope(s) are grouped by the

grouping factor over which they vary. For each random component, a variance and standard deviation are provided. When multiple correlated random effects are included in the model, their correlation matrix appears on the right side of this section. The final line of the table provides the residual variance.

The third output section provides the parameter estimates for the fixed effects, along with their standard errors and *t*-values. Note that no *p*-values are provided, as there is still debate regarding how the appropriate degrees of freedom for such *t*-statistics with linear mixed effects models should be calculated (see Baayen 2008: 247-248; Baayen et al. 2008: 396-399). As such, there are different ways to test significance in a mixed effects model. One rule of thumb is to take *t* values above 2.0 as being statistically significant (Gelman & Hill, 2007). Another way is to estimate *p* values from the *t* distribution as below (from Baayen 2008: 248).

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

Here, X is the *t* value, Y the number of observations and Z the number of fixed effects parameters. One can also perform a model comparison between a model with the fixed effect of interest and a reduced model that excludes this fixed effect (for details on the model comparison approach, see Barr et al., 2013: 276-277; Gelman & Hill, 2007). If excluding the fixed effect leads to a significant decrease in goodness of fit of the model (i.e., if the model comparison χ^2 test is significant), this suggests that the fixed effect is significantly contributing to the model. For our purposes, we rely on the '$|t| \geq 2.0$' rule of thumb for evaluating significance. Looking at our example, we see that the main effect of relatedness is significant, and the 3-way interaction is nearly significant ($t = 1.93$). Indeed, using the formula above, `2 * (1 - pt(abs(1.928), 2534 - 8))`, reveals that the 3-way interaction is marginally significant ($p = .054$). With regard to our research question, this suggests that participants were significantly slowed by the

relatedness of the critical distractor items, and that the two groups of participants differed (marginally) in the patterns of interference across the two distractor conditions.

Extension 1: Allowing effects to vary by Subjects *and* Items

Depending on the nature of the dataset at hand, there may be other factors that still need to be controlled for. In the example, because the dataset was produced by sampling a subset of items from a population of possible word pairs, the standard approach in psycholinguistics would be to also model `item` as a random effect. This would typically involve a separate by-items (F_2) analysis. As mentioned above, one advantage of mixed effects models is that it is possible to simultaneously include *crossed* random effects of both subjects *and* items in the same analysis. In our working example, we can take the previous `lmer` call and simply add an additional parenthetical term `(group|item)` to specify effects varying by `item`:

```
> m.factors.Rsubit <- lmer(data = rtdata,
   formula = RT ~ related*type*group + (1+related*type|Subject) +
   (group|item))
```

Note that even though we have not explicitly specified a by-item random intercept, the model includes a by-item random *intercept* for items (R automatically includes this even without specifying it with "`1 +`") and also a by-item random slope for `group`. Although the factor `group` was manipulated *between* participants, it is manipulated *within* items because the same items were presented to both groups of participants. As such, a random *slope* allows the `group` effect to vary by `item`, whereas `related` and `type` were both manipulated between items and therefore cannot vary by `item` and thus do not require by-item random slopes. The output below (edited for space) provides a summary of this model.

```
> summary(m.factors.Rsubit)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: RT ~ related * type * group + (1 + related * type | Subject) +
(group | item)
   Data: rtdata

Random effects:
 Groups   Name         Variance Std.Dev. Corr
 item     (Intercept)   11615.9 107.777
          group            18.2   4.266  1.00
 Subject  (Intercept)   41994.4 204.925
          related         777.5  27.884  -0.92
          type            832.3  28.850  -0.93  1.00
          related:type   2019.0  44.933  -0.20  0.57  0.54
 Residual              101676.0 318.867
Number of obs: 2534, groups:  item, 375; Subject, 45

Fixed effects:
                   Estimate Std. Error t value
(Intercept)         940.182     31.911  29.462
related              53.856     17.624   3.056
type                  3.676     17.667   0.208
group               -31.040     62.829  -0.494
related:type         57.309     34.905   1.642
related:group        17.942     27.100   0.662
type:group          -21.085     27.205  -0.775
related:type:group  115.491     53.315   2.166
```

Here, the first section of the output shows our expanded model formula that now includes random components varying by both `subject` and `item`. The second section now provides the variance and SD for the random effects varying by both `subject` and `item`. Note that separate correlation matrices are provided for the random subject effects and random item effects. Finally, the fixed effects table shows us that the `related` effect is still significant and positive; moreover, the 3-way interaction is now significant.

Extension 2: Controlling for potential confounds (covariates)

Another benefit of employing mixed effects models is that it is relatively easy to control for potential confounds by simply adding predictors to the model.[6] In our example, one potential

[6] Covariates can also be incorporated into ANOVAs. However, because no prior aggregation is required with mixed effects models, continuous predictors can simultaneously be incorporated into any level of analysis (i.e. subjects and items), which is difficult in analyses requiring prior aggregation (Baayen et al., 2008).

concern is that any group differences may not be due to the `group` factor (i.e., immersion vs. classroom-only learning context), but instead may simply reflect differences in L2 proficiency. To address this concern, we can include a measure of L2 proficiency as a covariate in the model, so that any effect of group then reflects differences due to group membership above and beyond any individual differences in L2 proficiency.

In the sample dataset, to control for L2 proficiency, we include an additional predictor - overall accuracy in the translation recognition task (`z.acc`). Proficiency was a between-subjects variable but a repeated measures variable within items. Therefore we allow `z.acc` to vary by `item` but not `Subject`. To incorporate this covariate into our previous model, we add it to the right-hand side of the formula equation in two places: outside of the parentheses as a main effect that does not interact with the other variables, and within the parentheses to indicate it varies by item. We then examine the results with a call to `summary`.

```
> m.factors.covariates <- lmer(data = rtdata,
   formula = RT ~ z.acc + related*type*group + (related*type|Subject) +
   (z.acc + group|item))

> summary(m.factors.covariates)

Linear mixed model fit by REML ['lmerMod']
Formula: RT ~ z.acc + related * type * group + (related * type | Subject) +
(z.acc + group | item)
   Data: rtdata

Random effects:
 Groups   Name         Variance Std.Dev. Corr
 item     (Intercept)  11948.5  109.31
          z.acc         6583.5   81.14   -0.87
          group         2683.9   51.81    0.51 -0.86
 Subject  (Intercept)  29511.3  171.79
          related        523.7   22.88   -0.58
          type           833.9   28.88   -0.76  0.97
          related:type  3214.5   56.70    0.04  0.79  0.62
 Residual              95602.1  309.20
Number of obs: 2534, groups:  item, 375; Subject, 45

Fixed effects:
                  Estimate Std. Error t value
(Intercept)         938.865     27.154   34.58
```

```
z.acc              -115.835     27.612   -4.20
related              60.754     16.376    3.71
type                  6.225     16.621    0.37
group                 9.379     54.067    0.17
related:type         58.287     33.193    1.76
related:group        21.011     26.515    0.79
type:group          -20.645     27.054   -0.76
related:type:group  108.317     54.039    2.00
```

Examining the results, we see in the second section of the output that `z.acc` has been added to the random effects structure for `item`. In the third section, we now see a parameter estimate for the covariate with an absolute *t*-value greater than 2 indicating that variation in RTs were significantly accounted for by L2 proficiency. Note, however, that both the relatedness effect and the three-way interaction still remained significant, suggesting that group differences could not solely be accounted for by L2 proficiency.

Extension 3: Modeling binary outcomes

So far, we have focused on fitting linear mixed effects models to data with a continuous dependent variable. However, in L2 research, it is not uncommon to have binary outcomes (e.g., correct/incorrect, grammatical/ungrammatical). We can analyze such binary outcomes by using the mixed effects implementation of logistic regression, or mixed logit models. Recent work indicates that mixed effects modeling of binary outcomes is superior to simply computing the average scores by subjects (e.g., proportion correct) for each condition and then analyzing those results with ANOVA (see Jaeger, 2008). That approach is problematic, in part, because the outcome is not on a continuous scale, but rather is bounded at zero and one, violating one of the assumptions of ANOVA. This issue does not affect logistic mixed effects models, although problems can arise when there is no variability in responses for a cell within the research design.

The supplementary file "acc_data.txt" contains data with a binary dependent variable. This file contains accuracy data (rather than reaction times) under the variable `acc` (coded as 1 =

correct and 0 = error). With this coding scheme, we are now modeling the probability of making a correct response, without need of prior aggregation. The analysis of binary dependent variables is similar to the method we used before, except that we now use the function `glmer` (*generalized* linear mixed model) and the dependent variable is specified as a binary outcome (rather than continuous) with the `family = binomial` argument.

```
> m.acc.Rsubj.items <- glmer(data = accdata,
   formula = acc ~ related*type*group + (related*type|Subject) +
   (group|item), family = binomial)
```

The `glmer` call above produced a message warning of failed convergence ("`Model failed to converge with max|grad| = 0.226126 (tol = 0.001, component 18)`" and "`Model failed to converge: degenerate Hessian with 4 negative eigenvalues`"), indicating that the model did not produce stable results. To remedy this, we chose a different optimizing function using the `control` argument and refit the model (see Recommendations section below for discussion of this and other steps to follow when encountering convergence issues):

```
> m.acc.Rsubj.items_opt2 <- glmer(data = accdata,
      formula = acc ~ related*type*group + (related*type|Subject) +
(group|item),
      family = binomial,
      control = glmerControl(optimizer = "bobyqa"))
```

This successfully produced a stable model. When we examine the output using the `summary` function (see below), the first difference relative to the linear models we fit before can be seen in the first section of the output, where it identifies the model as a generalized linear mixed model - a useful confirmation that the binary data were treated as binary outcomes, rather than zeroes and ones on a continuous scale. As before, in the second output section we can

obtain variance and correlation values for the random components. In the third section of the output, however, the output has changed slightly - in addition to the fixed effect parameter estimates and *SE*s, we now have *z*-values in place of *t*-values, along with their associated *p*-values. With logistic mixed effects models, the underlying distribution of the parameter estimates is assumed to be normal and therefore probability values can be computed from the normal distribution. In contrast, parameters in linear mixed effects models are assumed to follow the *t*-distribution, and it is unclear how to best determine the degrees of freedom for computing their probability value. The interpretation of the parameter values also changes for logistic regression. In brief, the scale of the model coefficients is now the log-odds of a correct response rather than the scale of the dependent variable (see Jaeger, 2008). Note that this is true of any logistic regression analysis, whether involving mixed effects or not (for a more detailed discussion, see Pedhazur, 1997).

```
> summary(m.acc.Rsubj.items_opt2)

Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: acc ~ related * type * group + (related * type | Subject) + (group |
item)
   Data: accdata
Control: glmerControl(optimizer = "bobyqa")

Random effects:
 Groups  Name         Variance Std.Dev. Corr
 item    (Intercept)  5.7523   2.3984
         group        0.1868   0.4322   -1.00
 Subject (Intercept)  0.8481   0.9209
         related      1.2933   1.1372   -0.80
         type         0.1237   0.3517   -0.77  0.90
         related:type 0.5165   0.7187    0.53 -0.74 -0.37
Number of obs: 2840, groups:  item, 376; Subject, 45

Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         5.651737   0.602725   9.377  < 2e-16 ***
related            -2.255631   0.614562  -3.670 0.000242 ***
type               -1.132694   0.585548  -1.934 0.053062 .
group              -0.883673   0.819298  -1.079 0.280778
```

```
related:type       -0.180539   1.149877  -0.157 0.875239
related:group      -0.003816   0.691384  -0.006 0.995596
type:group          0.074080   0.584255   0.127 0.899103
related:type:group -0.815222   1.163087  -0.701 0.483358
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Recommendations for reporting results

The use of mixed effects models in the language sciences is a relatively recent analytical advancement, and unfortunately the standards in best practice of conducting and reporting mixed effects analyses are still maturing (though see Barr et al. 2013). We note the following recommendations that are based on ongoing discussions among scholars as well as our own experience working with mixed effects models and communicating their results. In particular, we emphasise the importance of explicitly stating the structure of your statistical model when it is reported.

It is particularly important to describe the random effects structure of your analysis. Firstly, explain whether your model included *crossed* or *nested* random effects, which random *intercepts* and random *slopes* were included in the analysis, and the grouping factor(s) over which the intercepts and slopes varied. Deciding on what random *slopes* to include in a mixed effects model is a matter of contention. Barr et al. (2013) argue that in cases of confirmatory hypothesis testing, when a researcher has designed a study to test a specific set of hypotheses, the structure of the random effects should reflect the hypotheses being tested. In such cases, Barr et al. recommend that researchers should adopt the ‘maximal’ random effects structure possible based on the design of the study. That is, random slopes should be included for any repeated measures fixed effect that is of prime theoretical interest. In the final reaction time example above, we included random *intercepts* for subjects and items, by-subject random *slopes* for `related`, `type` and their interaction, and by-item random *slopes* for `group` and `z.acc`. For

exploratory analyses, for example of corpus data, which may include a multitude of fixed effects, it may not be practical to include random slopes for all of the fixed effects being tested. In such cases, a possibility might be to only include random slopes if they provide a significantly improved model fit to the data compared to a model without them (see Baayen et al., 2008; Baayen 2008). For confirmatory hypothesis testing (i.e. the vast majority of experimental research conducted in the L2 literature), we suggest researchers adopt 'maximal' models (Barr et al., 2013). For exploratory research, it remains imperative to justify the random effects structure that was ultimately adopted. Regardless of the approach, a clear and explicit description of the random effects structure should be provided.

Researchers should describe the software package used to conduct the analysis, including version number (e.g., `lme4` version __ in R version __), as the underlying computations can vary between different analysis packages or even versions of the same package. The version numbers for R and each package can be obtained by using the `sessionInfo` function, as demonstrated in the sample code in the Supplemental Materials. The estimation method should be identified (e.g., maximum likelihood, restricted maximum likelihood), and it is also important to describe your dependent variable (e.g., continuous, binary outcome) and explain whether any transformations (e.g., log-transformation, *z* scores) were applied before analysis. For the fixed effects components, describe how you decided on what fixed effects to include – whether based on *a priori* theoretical motivations or empirically determined via exploratory analysis. For fixed effects that are categorical factors, explain the coding scheme that was used. For continuous fixed effects, describe any adjustments or transformations that were applied before analysis. When reporting the results of the fixed effects, include model estimates, standard errors and the

test statistic (e.g. *t, z*) for each fixed effect parameter. Explain how you assessed significance (e.g., *t* above 2).

Occasionally a mixed effects model can fail to converge as we saw with the accuracy analysis above. When this happens in R, the `lmer` function will display an error such as `singular convergence`, `false convergence` or `iteration limit reached without convergence`. This usually occurs when an overly complex model is fit to a dataset that is too sparse to accurately estimate one or more of the parameters. This can often be the case in complex designs with multiple random *slope* parameters. Although the `summary` command will provide a summary of the statistical model in such cases, the model estimates should not be interpreted or reported. Instead, a first step could be to see if the model will fit using other optimizing functions. For example, in the accuracy analysis above, we specified the "bobyqa" optimizer, which produced a stable model (see Supplemental Materials for details). If this does not resolve the issue, the model should then be simplified until convergence is achieved. Unfortunately there is currently no consensus on how this issue should be tackled (though see Barr et al. 2013: 275-276 for some discussion). When working with factorial combinations of factors (as in our example analysis), alternative approaches include (a) removing the correlations among random effects, (b) removing the random effect that is contributing the least amount of variance, or (c) removing the random slope for the highest-order interaction term, as the lower-level factors may already capture most relevant variability between the levels of the grouping factor. Whichever option is chosen, we again emphasise the importance of explicitness in reporting what criteria were used to overcome this issue.

As a concrete example, if we were to report the reaction time example above involving a covariate (model `m.factors.covariates`, see Extension 2), we would state the following. An example format for reporting mixed effects modelling results is provided below in Table 1.

"Analyses were conducted using mixed effects models with crossed random effects for subjects and items using the lme4 package (version 1.1-7) of R (version 3.1.1). The analysis included contrast coded fixed effects for relatedness (-.5 = unrelated, .5 = related), distractor type (-.5 = lexical, .5 = semantic) and group (-.5 = classroom, .5 = immersed) in a 2x2x2 factorial design. Participant proficiency was assessed by inclusion of a continuous fixed effect predictor of overall accuracy in the translation recognition task (standardized as *z*-scores). Random effects were fit using a 'maximal' random effects structure (Barr et al. 2013). This included random intercepts for subjects and items, by-subject random slopes for relatedness, distractor type, and their interaction, and by-item random slopes for group and overall task accuracy. Models were fit using a maximum likelihood technique. A fixed effect was considered significant if the absolute value of the *t* statistic was greater than or equal to 2.0 (Gelman & Hill, 2007). Results indicated that the covariate of overall task accuracy was significantly related to performance (estimate = -116, SD = 27, t = -4.20). There was also a significant main effect of relatedness (estimate = 61, SD = 16, t = 3.71), which was qualified by a significant relatedness x distractor type x group interaction (estimate = 108, SD = 54, t = 2.00). No other main effects or interactions were significant (all other ts < 1.76). The results of the final best-fitting model are reported in Table 1."

Table 1. Example format for presenting results from a mixed effects model.

Parameters	Fixed effects	Random effects	
		By Subject	By Items

	Estimate	*SE*	*t*		*SD*	*SD*
Intercept	938.9	27.2	34.58	*	171.8	109.3
z.acc	-115.8	27.6	-4.20	*	--	81.1
Related	60.7	16.4	3.71	*	22.9	--
Type	6.2	16.6	0.37		28.9	--
Group	9.4	54.1	0.17		--	51.8
Related x Type	58.3	33.2	1.76		56.7	--
Related x Group	21.0	26.5	0.79		--	--
Type x Group	-20.6	27.1	-0.76		--	--
Related x Type x Group	108.3	54.0	2.00	*	--	--

Note. z.acc = overall accuracy on the translation recognition task, standardized as *z*-scores. All factors were coded using contrast coding, as follows: Related (-.5 = unrelated, .5 = related), Type (-.5 = lexical, .5 = semantic), Group (-.5 = classroom, .5 = immersed). Model formula: RT ~ z.acc + related * type * group + (related * type | Subject) + (z.acc + group | item).
* $|t| > 2.0$, indicating a significant effect (Gelman & Hill, 2007)

If you employed a model comparison procedure to determine the best-fitting model (see Barr et al., 2013; Gelman & Hill, 2007), we recommend you consider reporting all preliminary models as supporting materials (e.g., in an appendix or online supplemental material). If taking this approach, be sure to include the model comparison statistics with associated *p* values. If the model comparisons are theoretically relevant or provide useful information for other scholars, you should consider including those relevant preliminary models in the main manuscript results table (for an example, see Hoffman & Rovine, 2007, Tables 3 and 4).

Finally, it may benefit the readers to note any model checking steps you took to examine the goodness of the fit of the model. These could include examination of residual plots (using the `resid` function in R) or the distributions of the random effects. R syntax for some example diagnostic checks is provided in the Supplemental Materials.

Conclusions

L2 researchers face a number of analytical challenges when attempting to generalise study findings from a sample of language learners to the wider population. Despite the varied

nature of SLA research, surveys of statistical analysis techniques used in L2 acquisition have noted a near ubiquitous use of ANOVA and *t*-test (Lazaraton, 2000; Norris & Ortega, 2000; Plonsky, 2011; Plonsky & Gass, 2011). Admittedly, some initial time investment on the part of the analyst is required to learn to fit and interpret mixed effects models appropriately, but this is true for any analytic technique. Nonetheless, with their flexibility and their ability to relax assumptions of traditional models, mixed effects models provide a single framework for a range of analyses while simultaneously providing advantages over more traditional methods (e.g., ANOVA). Mixed effects models constitute a powerful additional statistical tool that can aid researchers in SLA and applied linguistics in the analysis of a wide variety of data types from different experimental and non-experimental paradigms.

References

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

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

Barr, D. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, 4: 328. doi: 10.3389/fpsyg.2013.00328

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-5. http://CRAN.R-project.org/package=lme4

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, 335-359.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, 28, 369-382.

Cunnings, I. & Finalyson, I. (under review). 'Mixed effects modelling and longitudinal data analysis'. To appear in Plonsky, L. (ed.) *Advancing Quantitative Methods in Second Language Research*.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press.

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

- (1995). *Multilevel statistical models.* London: Arnold.

Gries, S. (2013). *Statistics for linguistics with R. A Practical Introduction, 2nd Edition*. Berlin: De Gruyter Mouton.

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods, 39,* 101-117.

Jaeger, F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59, 434-446.

Lazaraton, A. (2000). Current trends in research methodology and statistics in applied linguistics. *TESOL Quarterly*, 34, 175-181.

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

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

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction (3rd ed.)*. Orlando, FL: Harcourt Brace.

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.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: the case of interaction research. *Language Learning*, 61, 325-366.

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

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

- (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, 52, 501-525.

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

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, 20, 416-420.

Singer, J. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and residual growth curve models. *Journal of Educational and Behavioral Statistics*, 23, 323-355.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis.* London: Sage Publications.

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