Tutorials in Quantitative Methods for Psychology
2013, Vol. 9(1), p. 25-41.

DOI: 10.20982/tqmp.09.1.p025

Crossing Language Barriers: Using Crossed Random Effects Modelling in Psycholinguistics Research

Robyn J. Carson and Christina M. L. Beeson
University of Ottawa

The purpose of this paper is to provide a brief review of multilevel modelling (MLM), also called hierarchical linear modelling (HLM), and to present a step-by-step tutorial on how to perform a crossed random effects model (CREM) analysis. The first part provides an overview of how hierarchical data have been analyzed in the past and how they are being analyzed presently. It then focuses on how these types of data have been dealt with in psycholinguistic research. It concludes with an overview of the steps involved in CREM, a form of MLM used for psycholinguistics data. The second part includes a tutorial demonstrating how to conduct a CREM analysis in SPSS, using the following steps: 1) clarify your research question, 2) determine if CREM is necessary, 3) choose an estimation method, 4) build your model, and 5) estimate the model's effect size. A short example on how to report CREM results in a scholarly article is also included.

Many statistical techniques, such as analysis of variance (ANOVA) and ordinary least-squares (OLS) multiple regression, assume that observations are not correlated with one another. However, this is not always the case. Within many areas of psychology, measurements are not fully independent of one another, but are instead nested, or hierarchical, in nature.

Data from both cross-sectional and longitudinal studies can be fully nested. In a cross-sectional design, for example, students (level-1) can be nested within classrooms (level-2), which can be further nested within schools (level-3). One could expect that students in the same classroom are more similar to one another than they are to students in a different classroom, and classrooms within the same school should be more similar to each other than to classrooms in a different school (Field, 2009; Peugh, 2010). In a longitudinal design, depression scores (level-1) can be nested within patients (level-2), which can be nested within therapists (level-3; Field, 2009). Again, one could expect that depression scores within the same patient will be more similar to one another across time than to those of a different patient, and that patients of the same therapist will have scores more similar to one another than to those of a different therapist. Alternatively, cross-sectional and longitudinal data can be partially nested, where lower levels are still nested within higher levels, but higher levels are independent of one another, not nested one within the other. For example, students (level-1) can be nested within middle schools

Robyn J. Carson, School of Psychology, University of Ottawa; Christina M. L. Beeson, School of Psychology, University of Ottawa.

The authors would like to thank Dr. Alain Desrochers and Karine Côté for their permission to use the dataset reported in the tutorial portion of this paper. The data are from a larger study investigating how word frequency and word animacy impact response time during lexical and gender decision tasks. Please note that Robyn Carson and Christina Beeson contributed equally to this manuscript and should both be considered first authors.

Correspondence concerning this paper should be addressed to Robyn Carson, School of Psychology, University of Ottawa, 136 Jean Jacques Lussier, Vanier Hall, Room 3002, Ottawa, Ontario, Canada K1N 6N5. Tel: (613) 562-5800 ext. 4289 Email: rcars035@uottawa.ca.

(level-2), as well as within high schools (level-3; Beretvas, 2011). Students who attended a particular middle school will not all attend the same high school; similarly, students attending the same high school did not all attend the same middle school. Thus, although students are nested within the two higher levels, the higher levels are not nested within the other; they are independent of one another.

Hierarchical Data Analysis: Past and Present

Until recently, hierarchical data were often analyzed using aggregation or disaggregation, simple linear regression techniques wherein parameters are fixed and the hierarchical nature of the data is cast aside (Woltman, Feldstain, MacKay, & Rocchi, 2012). In aggregation, all of the variables are collapsed across a higher hierarchical level, where data from the lower level(s) are ignored and transformed into averages for the higher level variable(s) (Snijders & Bosker, 1999). This leads to the assumption that individuals within a group are one entity, resulting in a loss of individual, or within-group, variance (Woltman et al., 2012). Further, aggregated data can only be properly interpreted in the context of the higher level variable(s) of interest (Snijders & Bosker, 1999), which is often not ideal. In disaggregation, all of the variables are collapsed across the lowest hierarchical level, and the group data from the higher level(s) are ignored. This leads to the assumption individual results are not influenced by the group(s) within which the lower level data are nested and results in a loss of between-group variation (Woltman et al., 2012). Moreover, the risk of committing Type II and Type I errors, for aggregation and disaggregation respectively increases substantially (Bovaird & Shaw, 2012; Heck & Thomas, 2009; Peugh, 2010; Snijders & Bosker, 1999).

Another way hierarchical data were analyzed was by running separate analyses on each data level, known as the slopes-as-outcomes approach. In this approach, groups are analyzed one at a time for each level of data, and the estimates from each analysis are combined into a group level matrix (Hox & Roberts, 2011). The problem with analyzing hierarchical data in this manner is that levels are not considered simultaneously in relation to one another, but individually, which gives inaccurate results and leads to inferential errors (Bovaird & Shaw, 2012; Raudenbush & Bryk, 1986).

The more traditional ways in which hierarchical data have been analyzed are not adequate at reflecting the non-independence of the data, as well as the potential interactions between, or across, levels. Since the early 1980s, multilevel modeling (MLM) has been introduced as a solution to these problems (Janssen, 2012; Woltman et al., 2012). The theory behind MLM has developed simultaneously within a number of disciplines, resulting in many different, albeit synonymous, statistical terms. Specifically, multilevel models are also known as hierarchical linear models, mixed-effect models, mixed linear models, random coefficient models, and multilevel covariance structure models (Heck & Thomas, 2009; Woltman et al., 2012). Since MLM is the term predominantly used in psycholinguistics literature, which is our focus, this is the term we will employ.

Essentially, MLM is an extension of OLS multiple regression, except that instead of being confined to fixed coefficients, it allows for one or more random coefficients to exist within the same model (Field, 2009; Raudenbush & Bryk, 1986). Intuitively, MLM has the same assumptions as traditional OLS multiple regression. The only assumption that differs between OLS multiple regression and MLM is that MLM does not require observations to be independent of one another, which allows for the analysis of hierarchical data. Additionally, since there can be more than one random coefficient, a final assumption unique to MLM is that the random coefficients are normally distributed around the model (Field, 2009).

In OLS multiple regression, the parameters (i.e., slope and intercept) are fixed and are estimated based on the sample data. Because the coefficients are fixed, it is assumed that the regression model is accurate across all of the data. However, in MLM, these parameters can vary, resulting in three possible models. For the random intercept model, the assumption is that the intercepts vary across the higher level groups. That is, the relationship between the predictor and the outcome is the same across all groups, or has the same slope, but the groups have a different intercept. Alternatively, the random slope model assumes that the slopes vary across the groups. That is, the groups have the same intercept, but the relationship between the predictor and the outcome differs across the groups, or has a different slope. Finally, there is the random intercept and slope model, which is the most realistic, where both the slopes and the intercepts vary across the groups (Field, 2009).

Based on the arguments provided above, it should be evident that using MLM to analyze hierarchical data has a number of benefits. First, because slopes can be random, homogeneity of regression slopes does not need to be assumed. In the likely event that individuals (level-1) in a group (level-2) are more similar to one another than they are to those in another group (i.e., the slopes vary across groups), MLM can be used to account for this, whereas OLS multiple regression cannot. Second, because level-specific parameters can be incorporated into one model, independence does not need to be assumed (Field, 2009; Peugh, 2010; Woltman et al., 2012). This allows for the analysis of a sample where variables are related contextually, as is the case with hierarchical data. Finally, whereas OLS multiple regression will provide inaccurate results when there are missing data or when group sizes are not equal, MLM can accommodate missing values at the individual level, as well as discrepant group sizes, and still provide accurate results (Field, 2009; Woltman et al., 2012). Multilevel modelling does not have any limitations (Field, 2009); however, it does require large sample sizes for adequate power (Woltman et al., 2012).

Introduction to Multilevel Modelling in Psycholinguistics Research

In psycholinguistics research, experiments involving word recognition or lexical decision tasks are common (Locker, Hoffman, & Bovaird, 2007). In these tasks, participants are shown a list of words, and must decide whether or not each word is a true word or a non-word. Instead of focusing on participants' accuracy, which is usually near ceiling, researchers often focus on participants' reaction time (RT), or how long it takes them to identify the stimulus as a true or non-word. In this type of experiment, there are two random effects impacting the dependent variable, RT. Participants are randomly selected from the larger population, and words are also randomly selected from a larger list of total potential words.

In the past, researchers analyzed this type of data using an ANOVA, including the participants as the random unit of analysis while holding the items (i.e., words) constant. In doing so, however, they were ignoring the fact that words were also randomly selected from a larger population. They generalized their findings to all words, when they should have only been drawing conclusions based on the sample of words used (Field, 2009). This problem was coined the "language-as-fixed-effects fallacy" (Clark, 1973).

Although there is some debate in the literature as to whether items should be considered randomly selected (see Raaijmakers, 2003; Raaijmakers, Schrijnemakers, & Gremmen, 1999; Wike & Church, 1976), the majority of researchers no longer use one ANOVA to analyze their psycholinguistic data. Two alternative statistical approximations were developed to try and address the fallacy, the Quasi-F Ratio, denoted as F' (Clark, 1973), and, the more commonly used, F_1 x F_2 subjects and items repeated measures ANOVAs (Clark, 1973; Janssen, 2012; Locker et al., 2007). In this technique, two ANOVAs are performed. The first analyzes the data with participants as the random factor while holding the items constant. The second analyzes the data with items as the random factor while holding the participants constant. Only if both F_1 and F_2 reach significance can a researcher entertain generalizing the results to both the population of participants and the total items (Locker et al., 2007). Only when both ANOVAs are statistically significant can both samples be considered random, and the results generalizable. Whereas this technique has become the norm in psycholinguistics research, neither ANOVA treats the data properly, both ignore the second random factor and do not reflect the true results (Locker et al., 2007).

How to Perform a Crossed Random Effects Model Analysis

A crossed random effect model (CREM) is a type of MLM that can encompass one or more random factors within the same model, a requirement when analyzing psycholinguistic data. There are a number of steps to follow in order to perform a CREM analysis. In this section of our paper, we will briefly outline the five main steps, as well as their key theoretical considerations.

1. Clarify your research question. Although clarifying the research question seems like an obvious step, it is important because it will guide the decisions made in subsequent steps (Peugh, 2010). By specifying the research question, it clarifies at which hierarchical level the variable(s) of interest lie. Specifically, in a dataset with two levels, the question can focus on level-1 variables, on level-2 variables, or on the interaction between them. A question focusing on level-1 examines the relationship between lower level (individual) predictors and the outcome variable. A question focusing on level-2 examines the relationship between higher level (group) predictors on a higher level outcome variable. When focusing on an interaction, the research question examines whether the relationship between a lower level predictor and an outcome variable is moderated by a higher level variable. To illustrate these scenarios using an adapted example from above, suppose we have the math achievement scores of students (level-1) grouped within classrooms (level-2). If we were interested in looking at the impact of level-1, we would simply look at student differences to explain math achievement scores. If we were interested in looking at the impact of level-2, we would look at classroom differences to explain overall classroom math achievement scores. Finally, if we were interested in looking at the interaction between levels we would look at how classroom differences moderate, or interact with, student differences to explain math achievement scores (Peugh, 2010).

2. Determine if crossed random effects modelling is necessary. A dataset that is hierarchical does not automatically require MLM. Specifically, if no variation exists across higher level variables (i.e., if an individual's group association does not influence the outcome), a traditional OLS multiple regression could be sufficient. In order to quantify if the use of MLM is warranted, the intraclass correlation (ICC) is used. The ICC is defined as both the proportion of the outcome variation that is due to higher level variables, as well as the expected correlation between scores of individuals nested within the same group. It measures how much variance can be attributed to higher level variables (Field, 2009; Peugh, 2010). When the ICC is small, the higher level variable has little influence on the outcome, and most of the variation is due to lower level variables. In this case, traditional techniques can be used (Field, 2009; Peugh, 2010). As the ICC increases, the higher level variables are explaining more variability, with less variability being explained by the lower level variables (Field, 2009; Peugh, 2010). In this case, the use of MLM is warranted (see Hayes, 2006 for a debate on whether a small ICC negates the use of MLM). In addition to the ICC, some researchers also take the design effect, which evaluates the effect of independent violations on standard error estimates, into consideration when evaluating the need for MLM (Peugh, 2010).

There are two important additional questions to ask when deciding whether or not to conduct a CREM analysis: 1) Do you have more than one random effect in your dataset?, and 2) Is CREM supported by current theories or knowledge in your area of research? (Peugh, 2010; Snijder & Bosker, 1999). If the answer is yes to both of these questions, you should use CREM.

3. Choose an estimation method. There are two possible maximum likelihood (ML) estimations to choose from, full information maximum likelihood (FIML) and restricted maximum likelihood (REML; Peugh, 2010). In FIML, the assumption is that the MLM regression coefficients are known, so these parameters are fixed in the likelihood estimation. The resulting between group variance is often underestimated, however the difference becomes negligible when the sample size is large (Peugh, 2010, see also Maas & Hox, 2005 and Paccagnella, 2011 for a discussion on sample size and ML estimation). In REML, regression coefficients are treated as unknown quantities; therefore, the parameters are estimated based on sample data. For smaller sample sizes, REML is the preferred estimation method (Heck & Thomas, 2009).

In both ML estimation methods, a chi-square log-likelihood value is used to measure the probability that the estimated model adequately accounts for the data. To obtain the deviance value, which compares the fit of two successive models, you multiply the log-likelihood by -2 (-2LL). For FIML, the deviance calculates the fit of both the regression coefficients and the variance estimates, whereas for REML, the deviance calculates only the fit of the variance estimates (Peugh, 2010).

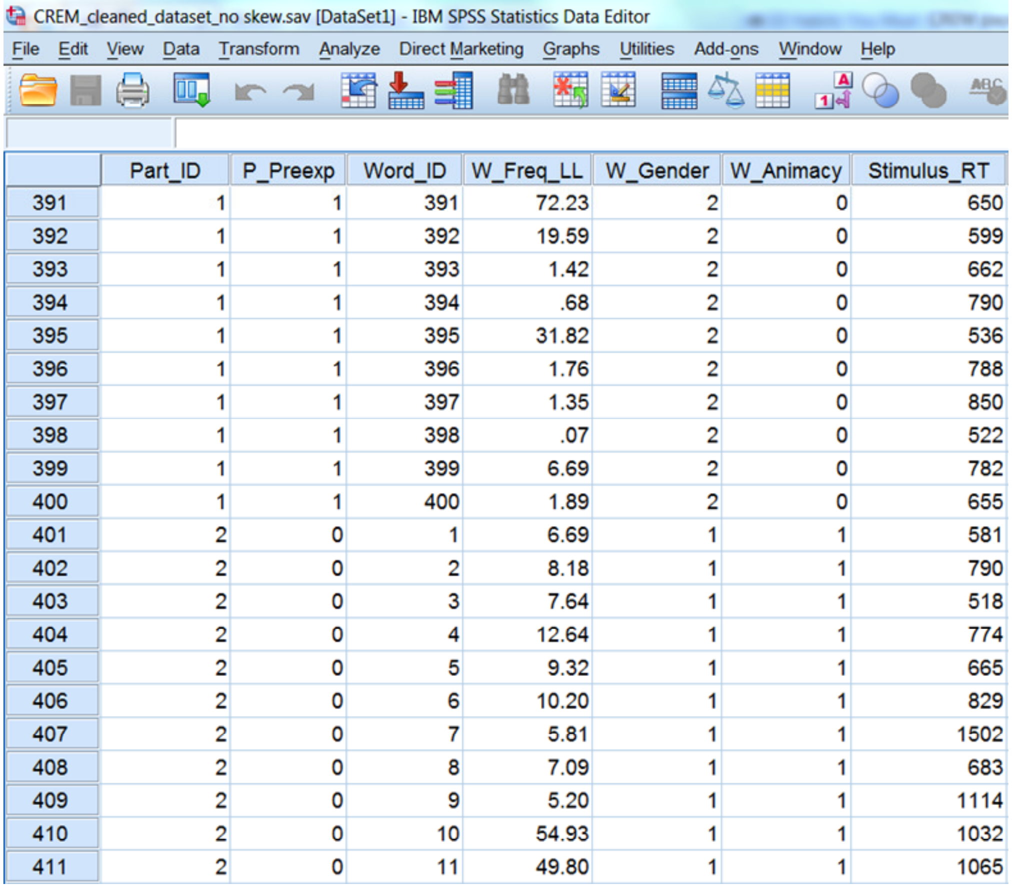
4. Build your model. Building a one-level CREM encompasses several steps: 1) testing an "empty" model, 2) adding and testing the random effects, and 3) adding and testing the fixed effects. There are additional steps for two-level models, where random and fixed effects need to be tested on both levels (see Raudenbush & Bryk, 2002 and Snijders & Bosker, 1999 for a thorough review).

It is helpful to begin by testing an "empty" model which is free of any random or fixed predictors. This model is also known as the "null," "baseline," or "unconditional" model and provides a baseline comparison for subsequent models being tested.

One way to account for the variation found in the empty model is to add random predictors one at a time and test the fit of each subsequent model. To compare models, the chi-square likelihood test is used. The -2LL of the new model is subtracted from the old one, with a positive difference indicating a better fit for the new model (Field, 2009). Once all of the random variables have been added and tested, fixed variables of interest can be added and tested. However, before you do this, you need to choose and apply a centring method to each of the fixed variables.

Centring involves rescaling variables around a fixed point, which allows for a meaningful interpretation of a score of zero (Field, 2009; Peugh, 2010). There are two methods of centring that can be used, grand mean centring and group mean centring. Grand mean centring, which is the most common method, takes an individual's score on the predictor variable and subtracts the grand mean for that variable (i.e., the mean across all groups) from their score (Field, 2009; Peugh, 2010). Alternatively, group mean centring takes an individual's score on the predictor variable and subtracts the group mean for that variable (i.e., the mean for the individual's specific group) from their score (Field, 2009). The centring method chosen should reflect the research question. If the research question is focused on a level-1 variable or if it is focused on an interaction, then grand mean centring should be used. However, if the research question focused on a level-2 variable, then group mean centring should be used (Heck & Thomas, 2009). Once all of the centred level-1 variables of interest have been added to the model, if your dataset includes level-2 variables, you can build a level-2 model. If your research question indicates an interest in an interaction, the level-2 variables must also be added to the level-1 model (Peugh, 2010).

5. Estimate the model's effect size. Since both fixed and random coefficients are estimates in MLM, determining a multilevel effect size is complex (Field, 2009). Consequently, there is currently no agreement as to which type of



	Part_ID	P_Preexp	Word_ID	W_Freq_LL	W_Gender	W_Animacy	Stimulus_RT
391	1	1	391	72.23	2	0	650
392	1	1	392	19.59	2	0	599
393	1	1	393	1.42	2	0	662
394	1	1	394	.68	2	0	790
395	1	1	395	31.82	2	0	536
396	1	1	396	1.76	2	0	788
397	1	1	397	1.35	2	0	850
398	1	1	398	.07	2	0	522
399	1	1	399	6.69	2	0	782
400	1	1	400	1.89	2	0	655
401	2	0	1	6.69	1	1	581
402	2	0	2	8.18	1	1	790
403	2	0	3	7.64	1	1	518
404	2	0	4	12.64	1	1	774
405	2	0	5	9.32	1	1	665
406	2	0	6	10.20	1	1	829
407	2	0	7	5.81	1	1	1502
408	2	0	8	7.09	1	1	683
409	2	0	9	5.20	1	1	1114
410	2	0	10	54.93	1	1	1032
411	2	0	11	49.80	1	1	1065

Figure 1. A snapshot of the tutorial dataset in SPSS.

estimated effect size is the most appropriate (Peugh, 2010). Currently, effect sizes can be defined as either global or local. Global effect sizes measure the outcome that can be explained by all of the predictors in the model. They resemble R^2, measuring the variance in the outcome variable explained by all of the predictors in the model. Local effect sizes resemble “change in R^2,” or ΔR^2, measuring the effect of level-1 variables on the outcome variable (Peugh, 2010). They also resemble a squared semi-partial correlation coefficient (Hayes, 2006; Radenbush & Bryk, 2002).While it is possible to compute both types of effect size, it is important to keep in mind that all MLM effect sizes are estimates (Snijders & Bosker, 1999).

A Tutorial on Crossed Random Effects Modelling in SPSS

The following tutorial section will demonstrate how to use CREM for psycholinguistic data in lieu of the standard F_1 x F_2 subjects and items repeated measures ANOVAs. All analyses are performed using SPSS, version 19.0.

Sample Dataset Content

The dataset for this tutorial contains results from an experiment involving 49 undergraduate students who completed a French lexical decision task (LDT). Specifically, we are interested in how participants’ pre-exposure to stimuli (0 = no pre-exposure, 1 = pre-exposure), word frequency, word gender (1 = masculine, 2 = feminine), and word animacy (0 = inanimate, 1 = animate) related to participants’ response times for the 400 real words presented.

The dataset was screened for invalid and impossible values. Several impossible values were found due to a computer error in registering response times. In addition, response time data were severely and positively skewed. Extreme outliers were removed from the dataset to reduce the skew to a more acceptable level. In all, 729 response time data points were removed, resulting in a total of 3.7% missing data for this variable. For this tutorial, we assume that all MLM assumptions (explained in detail above) have been met.

Data File Set-Up

To conduct multilevel analyses, you create a single SPSS data file containing all the possible variables of interest. Figure 1 provides a snapshot the tutorial dataset. Note that participant variables (part_ID and pre-exposure) are repeated across word variables (word_ID, frequency, gender, and animacy) and vice versa.

Table 1. Regression Coefficient Estimates and Variance-Covariance Estimates for CREMs Predicting Observed Response Time

Parameters	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Regression coefficients (fixed effects)							
Intercept (γ_0)	714.26 (1.92)***	717.25 (17.00)***	719.91 (18.23)***	719.92 (17.49)***	719.92 (17.44)***	719.91 (17.43)***	719.90 (17.40)***
Part. Pre-exposure (γ_1)				-71.93 (33.24)*	-71.94 (33.24)*	-71.94 (33.24)*	-71.94 (33.24)*
Word Frequency (γ_2)					-.34 (0.07)***	-.35 (0.07)***	-.34 (0.07)***
Word Gender (γ_3)						27.72 (10.96)*	27.69 (10.76)*
Word Animacy (γ_4)							-41.60 (10.77)***
Variance components (random effects)							
Residual (σ^2)	69884.36 (719.44)***	56090.33 (578.19)***	44552.59 (464.24)***	44552.61 (464.24)***	44552.48 (464.23)***	44552.30 (464.23)***	44552.21 (464.23)***
Participants (τ_{0s})		14012.23 (2860.86)***	14704.64 (2994.89)***	13410.46 (2733.58)***	13409.89 (2733.44)***	13410.40 (2733.53)***	13409.66 (2733.37)***
Words (τ_{0i})			11906.36 (912.60)***	11905.89 (912.55)***	11246.74 (865.67)***	11056.55 (852.03)***	10625.14 (821.35)***
Model summary							
Deviance statistic (-2LL)	264051.98	260126.89	256835.20	256830.73	256809.67	256803.32	256788.68
# of estimated parameters	2	3	4	5	6	7	8

Parameter estimate standard errors listed in parentheses.

* $p < 0.05$ *** $p < 0.001$

Performing a Crossed Random Effects Modelling Analysis

We will follow and apply each of the steps outlined above to our dataset. It is important to note that although we are presenting a one-level CREM, CREMs can include several levels, with the potential for random and fixed effects at each level (see Beretvas, 2011; Hoffman & Rovine; Raudenbush & Bryk, 2002; and Snijders & Bosker, 1999 for examples of two-level MLM analyses).

1. Specify your research question. Our research question is: Do participant pre-exposure, word frequency, word gender, and word animacy predict observed response times for real words in a lexical decision task?

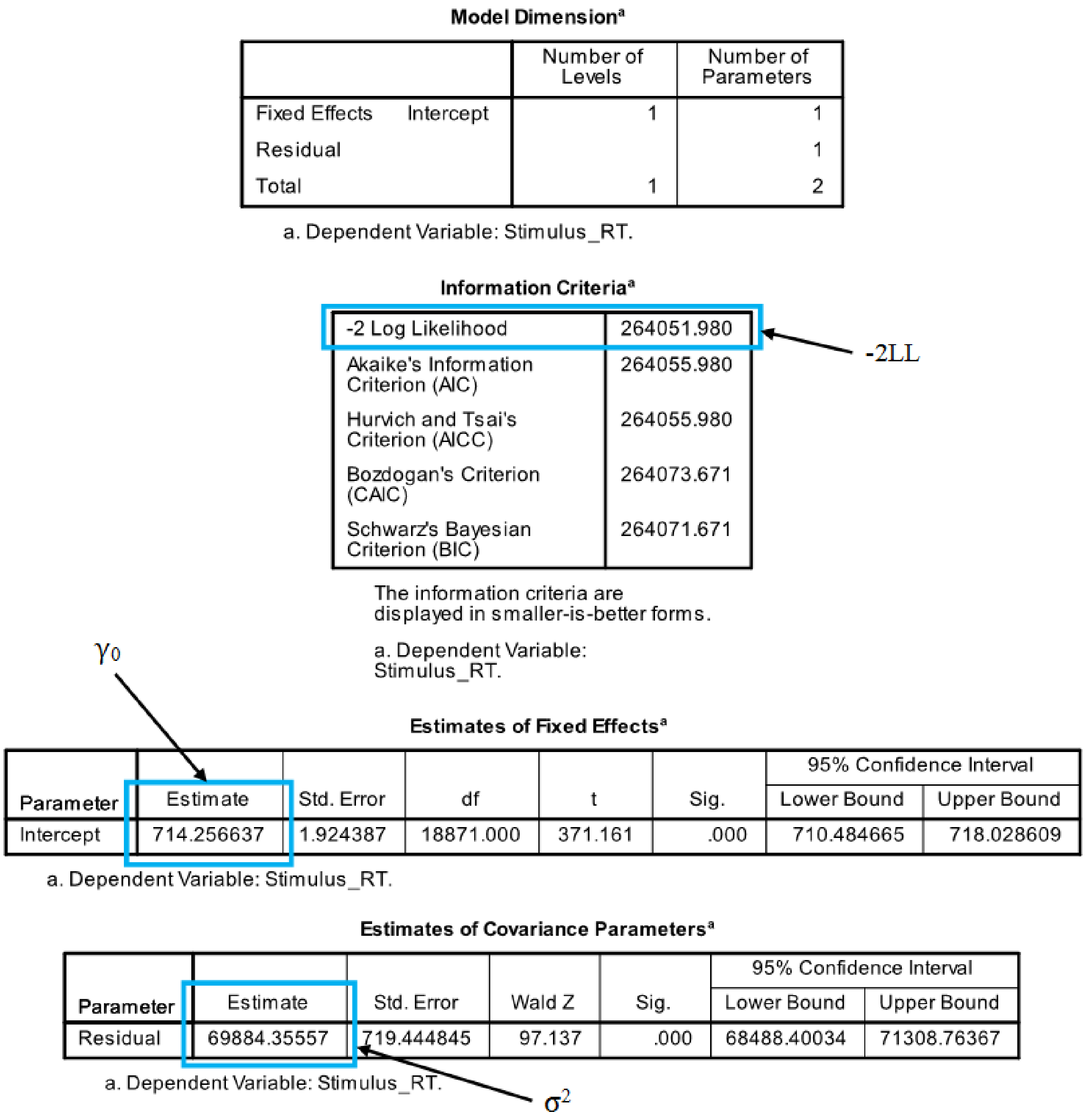
2. Determine if crossed random effects modelling is necessary. In our design there are two random effects: participants and words. Additionally, several researchers (Baayen, Davidson, & Bates, 2008; Baayen, Tweedie, & Schreuder, 2002; Janssen, 2012; Locker et al., 2007; Quené & van den Bergh, 2008) advocate that CREM is the best option for psycholinguistic data analyses. As a result, we will demonstrate a CREM with two random effects and four fixed effects.

For this tutorial, the ICC will be used to test the proportion of variance accounted for by our two random effects in Step 4 below.

3. Choose an estimation method. Our research question requires that we compare models with varying regression coefficients, which is not possible with REML. Our words sample size is large (n = 400) while our participant sample size is moderate (n = 49). Based on recent maximum likelihood simulation studies (Maas & Hox, 2005; Paccagnella, 2011), our sample sizes are large enough to use a FIML estimation without an unreasonable underestimation of the variance standard error. Thus, we will implement the FIML estimation method (referred to as ML in SPSS).

4. Build a crossed random effects model.[1]

[1] The syntax for all the CREM models discussed in this section can be found in the Appendix. Please note that if you are using SPSS version 11 or earlier, the provided syntax may not work. If you are using the SPSS menus, ensure that



Model Dimension[a]

		Number of Levels	Number of Parameters
Fixed Effects	Intercept	1	1
Residual			1
Total		1	2

a. Dependent Variable: Stimulus_RT.

Information Criteria[a]

-2 Log Likelihood	264051.980
Akaike's Information Criterion (AIC)	264055.980
Hurvich and Tsai's Criterion (AICC)	264055.980
Bozdogan's Criterion (CAIC)	264073.671
Schwarz's Bayesian Criterion (BIC)	264071.671

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: Stimulus_RT.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	714.256637	1.924387	18871.000	371.161	.000	710.484665	718.028609

a. Dependent Variable: Stimulus_RT.

Estimates of Covariance Parameters[a]

Parameter	Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Residual	69884.35557	719.444845	97.137	.000	68488.40034	71308.76367

a. Dependent Variable: Stimulus_RT.

Figure 2. SPSS output for Model 1.

Test "empty" model. The first model fit when estimating a CREM is the "empty" model. This model, shown in the equation below, does not include any random effects or predictors; it serves as a point of comparison for later models which will include parameters of interest.

$$Y_{si} = \gamma_0 + \varepsilon_{si} \tag{1}$$

where:

Y_{si} is the observed response time for subject s and item i [2]

γ_0 is the intercept, or expected mean response time for the overall sample, and

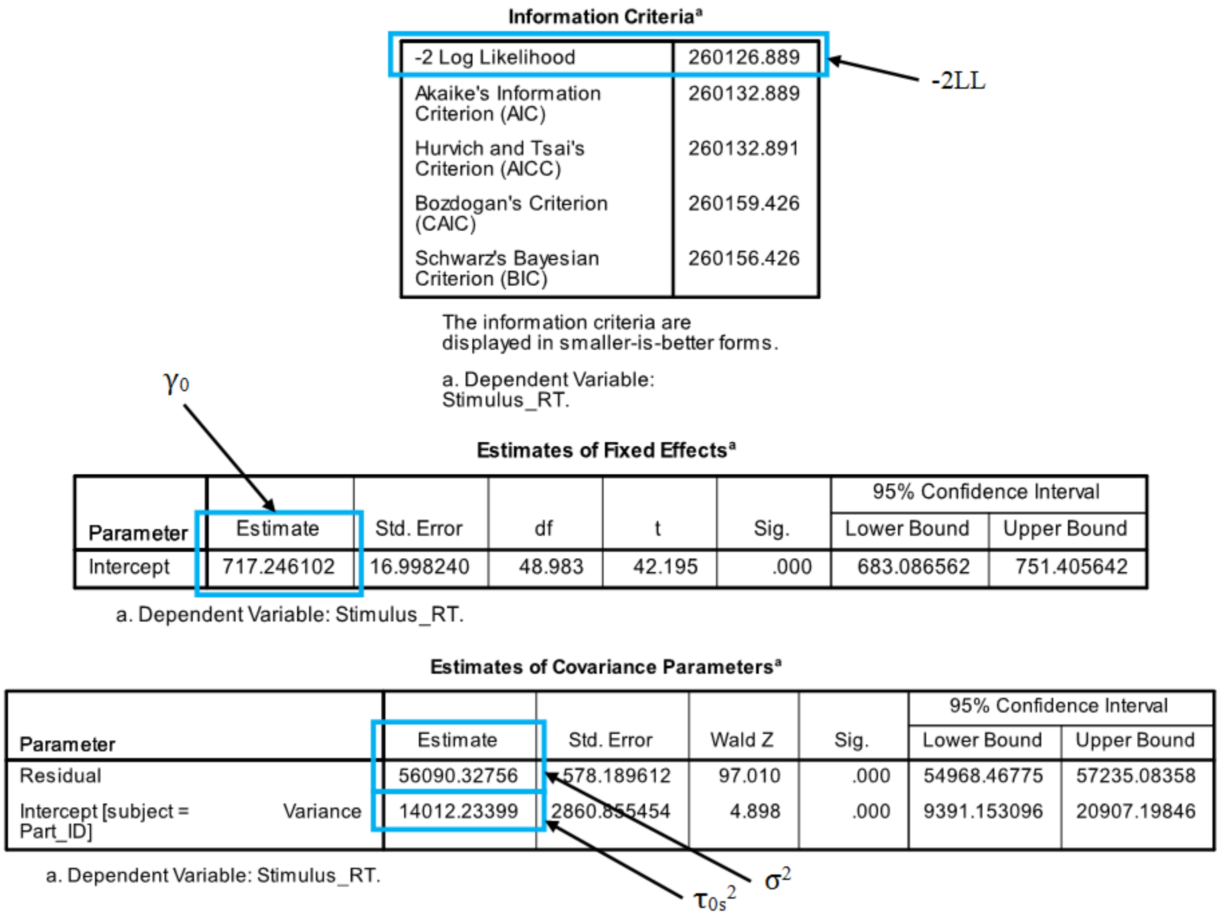
ε_{si} is the residual deviation from the sample mean response time for subject s and item i

This model assumes that the residuals (ε_{si}) are uncorrelated, meaning that no systematic effects of subjects or items are present (Beretas, 2011; Snijders & Bosker, 1999).

Running the analysis for Model 1, we generate an output with several tables (see Figure 2, see also Table 1, Model 1). The Model Dimension table displays which variables have been included in the model tested. For Model 1, no variables

you change the maximum number of iterations default of 100 to 150. This was done in to match the estimation values that would be obtained using *R* with the lme4 package.

[2] Subject is interchangeable with participant and item is interchangeable with word in our example. This is to keep the denotations of *s* and *i* in the CREM equations consistent with the recent literature on this topic.



Information Criteria[a]

-2 Log Likelihood	260126.889
Akaike's Information Criterion (AIC)	260132.889
Hurvich and Tsai's Criterion (AICC)	260132.891
Bozdogan's Criterion (CAIC)	260159.426
Schwarz's Bayesian Criterion (BIC)	260156.426

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: Stimulus_RT.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	717.246102	16.998240	48.983	42.195	.000	683.086562	751.405642

a. Dependent Variable: Stimulus_RT.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		56090.32756	578.189612	97.010	.000	54968.46775	57235.08358
Intercept [subject = Part_ID]	Variance	14012.23399	2860.855454	4.898	.000	9391.153096	20907.19846

a. Dependent Variable: Stimulus_RT.

Figure 3. SPSS output for Model 2.

were entered, so none appear in the table. The Information Criteria table provides deviance estimates that can be used to calculate how well the model fits the dataset using a chi-square likelihood ratio test (see Field, 2009 for information on the distinction between the different criteria). For the purposes of this tutorial we will use the -2LL results, where a smaller value indicates a better fit to the dataset. The Estimates of Fixed Effects table displays the estimated regression coefficient, or mean, for each of the model's fixed effects along with their associated standard error. The *t*-test indicates whether the estimated intercept is statistically different from zero. A significant grand mean response time score is observed, $\gamma_0 = 714.26$, $p < .001$. Finally, the Estimates of Covariance Parameters table displays the estimated variance for each of the model's random effects along with their associated standard error. The Wald Z test indicates whether the estimated variance is statistically different from zero (Hayes, 2006). A non-zero residual variance is observed, $\sigma^2 = 69884.36$, $p < .001$.

Add and test random effects. The next step is to add any random effect parameters to your model. Based on our research question and dataset, we will be adding two, the random effect for participants and the random effect for words.

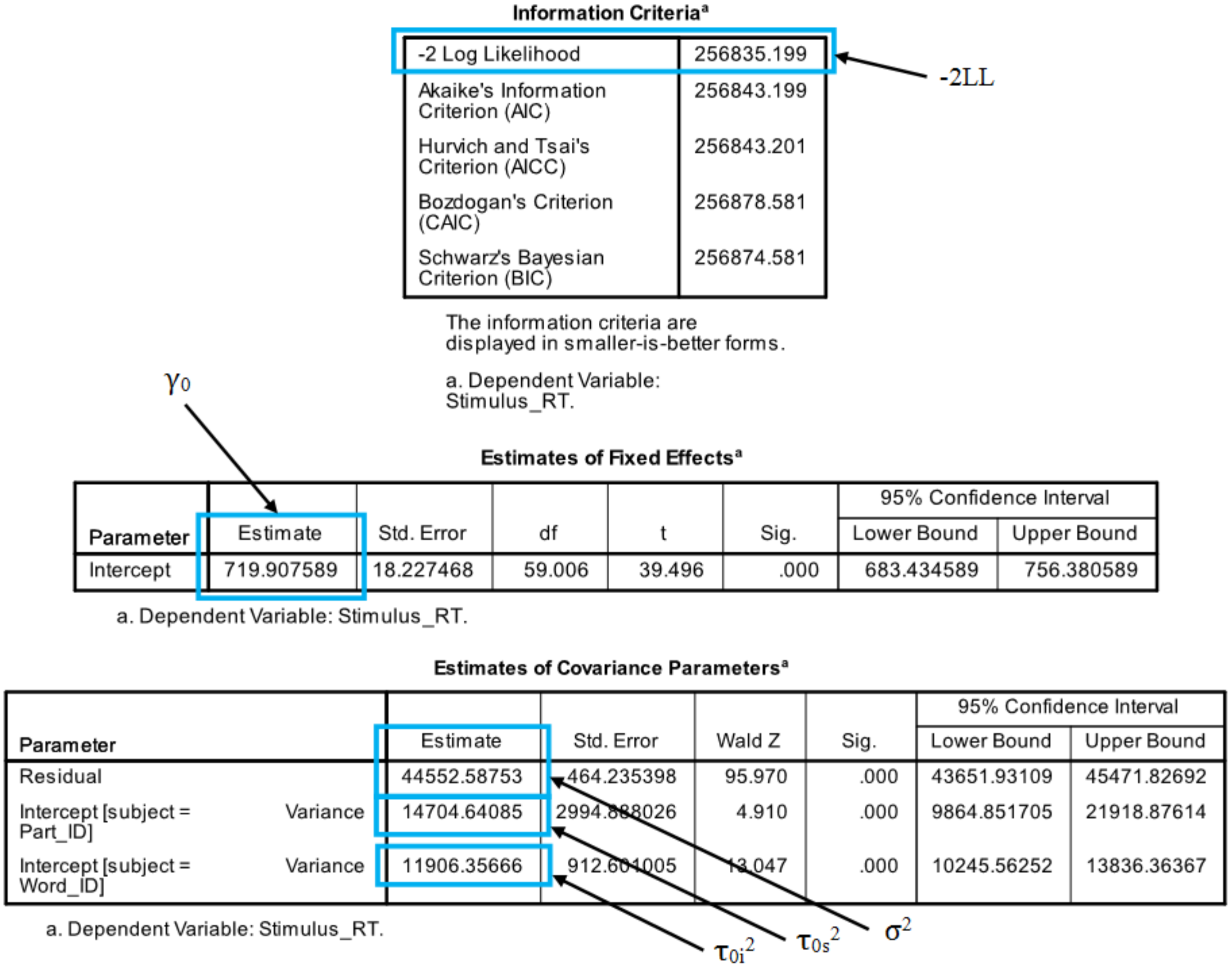
Random effect for participants (subjects). The equation below is equivalent to the "empty" model, with the addition of the random effect for participants.

$$Y_{si} = \gamma_0 + \mu_{0s} + \varepsilon_{si} \tag{2}$$

where u_{0s} is the random effect for subject *s*, or the deviation of subject *s*'s mean response time from the grand mean response time

This model assumes that the residuals (u_{0s} and ε_{si}) are uncorrelated across observations after taking into consideration which participant generated the observation (Beretas, 2011; Snijders & Bosker, 1999).

Running the analysis for Model 2, we generate a new output (see Figure 3) with four notable results (see also Table 1, Model 2). First, an adjusted, yet still significant, grand mean response time score is observed, $\gamma_0 = 717.25$, $p <$ .001. Second, an adjusted non-zero residual variance is observed, $\sigma^2 = 56090.33$, $p < .001$. Third, a new non-zero variance for the random effect of participants is observed, $\tau_{0s}^2 = 14012.23$, $p < .001$, indicating that the random effect for participants is significant. Fourth, we can test whether Model 2 fits the dataset better than Model 1 via the chi-



Information Criteria[a]

-2 Log Likelihood	256835.199
Akaike's Information Criterion (AIC)	256843.199
Hurvich and Tsai's Criterion (AICC)	256843.201
Bozdogan's Criterion (CAIC)	256878.581
Schwarz's Bayesian Criterion (BIC)	256874.581

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: Stimulus_RT.

Estimates of Fixed Effects[a]

						95% Confidence Interval	
Parameter	Estimate	Std. Error	df	t	Sig.	Lower Bound	Upper Bound
Intercept	719.907589	18.227468	59.006	39.496	.000	683.434589	756.380589

a. Dependent Variable: Stimulus_RT.

Estimates of Covariance Parameters[a]

						95% Confidence Interval	
Parameter		Estimate	Std. Error	Wald Z	Sig.	Lower Bound	Upper Bound
Residual		44552.58753	464.235398	95.970	.000	43651.93109	45471.82692
Intercept [subject = Part_ID]	Variance	14704.64085	2994.888026	4.910	.000	9864.851705	21918.87614
Intercept [subject = Word_ID]	Variance	11906.35666	912.601005	13.047	.000	10245.56252	13836.36367

a. Dependent Variable: Stimulus_RT.

Figure 4. SPSS output for Model 3.

square likelihood ratio test. The -2LL deviance of Model 2 is subtracted from that of Model 1.[3] The significance is determined by the chi-square distribution, with degrees of freedom calculated based on the difference in the number of parameters in each model (Locker et al., 2007). We find a difference of χ^2 (1) = 3925.09, $p < .001$, indicating that Model 2 fits the dataset significantly better than Model 1.

Random effect for words (items). The equation below adds the random effect for words.

$$Y_{si} = \gamma_0 + \mu_{0s} + v_{0i} + \varepsilon_{si} \qquad (3)$$

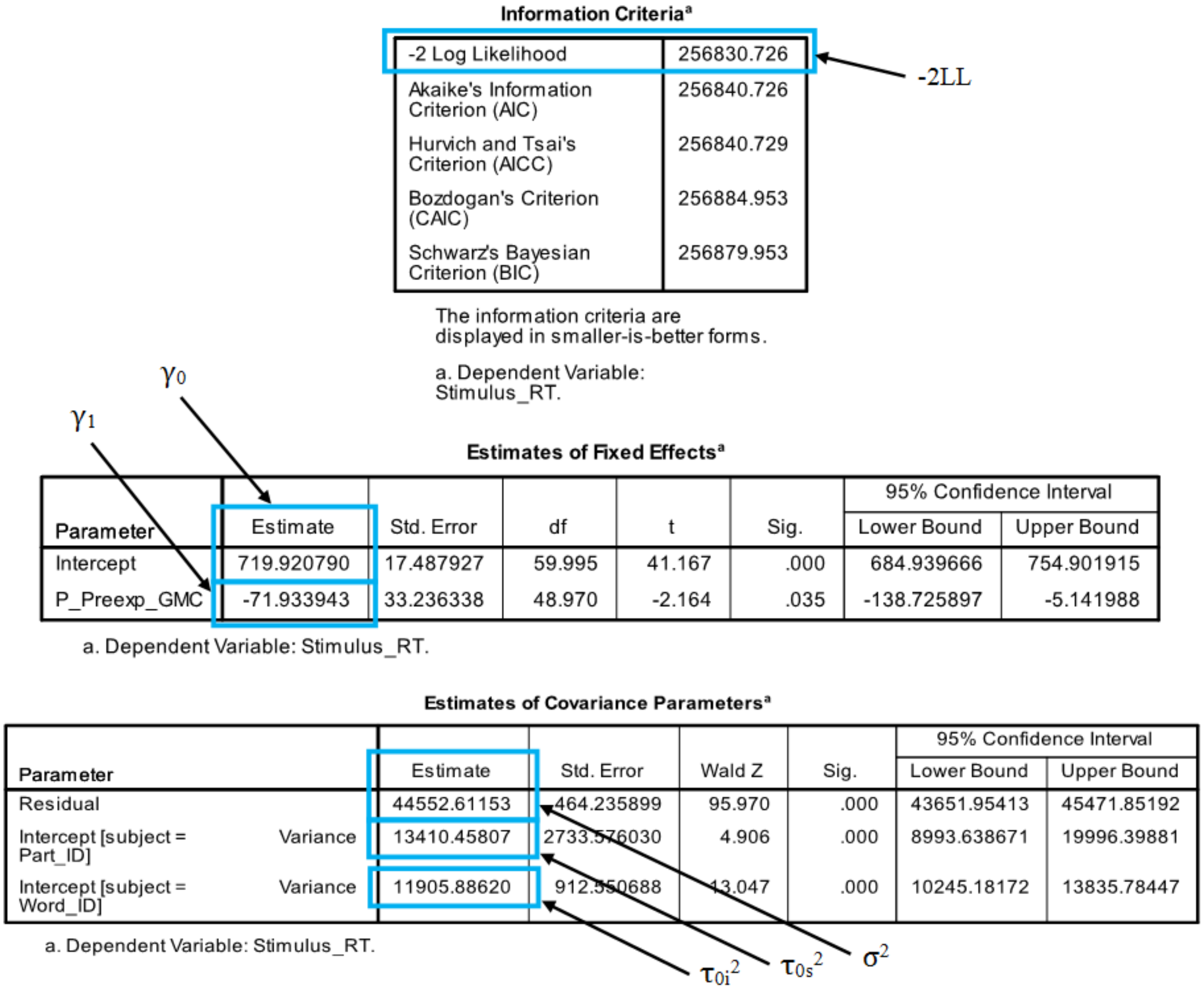
where v_{0i} is the random effect of item *i*

This model assumes that the residuals (u_{0s}, v_{0i}, and ε_{si}) are uncorrelated across observations after taking into consideration which participant and which word generated the observation (Beretas, 2011; Snijders & Bosker, 1999).

Running the analysis for Model 3, we generate a new output (see Figure 4) with five notable results (see also Table 1, Model 3). First, an adjusted, yet still significant, grand mean response time score is observed, γ_0 = 719.91, $p < .001$. Second, an adjusted non-zero residual variance is observed, σ^2 = 44552.59, $p < .001$. Third, an adjusted non-zero variance for the random effect of participants is observed, τ_{0s}^2 = 14704.64, $p < .001$. Fourth, a new non-zero variance for the random effect of words is observed, τ_{0i}^2 = 11906.36, $p < .001$, indicating that random effect for words is significant. Fifth, we find that Model 3 fits the dataset significantly better than Model 2, χ^2 (1) = 3291.69, $p < .001$.

Using the estimated parameter variances from Model 3 we can determine the proportion of response time variance explained by participants versus that explained by words through means of the ICC (Locker et al., 2007). The ICC is calculated as the proportion of variance of the random effects (participant variance or word variance) over the total variance (participant variance + word variance + residual variance). Using the variance parameters in Table 1, the total proportion of response time variance explained by participants is 20.7%, by words is 16.7%, and the remaining unexplained variance is 62.6%. Thus, the random effects

[3] SPSS does not calculate this difference for you, you need to do this calculation by hand.



Information Criteria[a]

-2 Log Likelihood	256830.726
Akaike's Information Criterion (AIC)	256840.726
Hurvich and Tsai's Criterion (AICC)	256840.729
Bozdogan's Criterion (CAIC)	256884.953
Schwarz's Bayesian Criterion (BIC)	256879.953

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: Stimulus_RT.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	719.920790	17.487927	59.995	41.167	.000	684.939666	754.901915
P_Preexp_GMC	-71.933943	33.236338	48.970	-2.164	.035	-138.725897	-5.141988

a. Dependent Variable: Stimulus_RT.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		44552.61153	464.235899	95.970	.000	43651.95413	45471.85192
Intercept [subject = Part_ID]	Variance	13410.45807	2733.576030	4.906	.000	8993.638671	19996.39881
Intercept [subject = Word_ID]	Variance	11905.88620	912.550688	13.047	.000	10245.18172	13835.78447

a. Dependent Variable: Stimulus_RT.

Figure 5. SPSS output for Model 4.

together explain 36% of the model variance.

Add and test fixed effects. Now that we have validated the inclusion of our random effects in a CREM, we need to test whether our predictors add anything. Based on our research question, we want to test how participant pre-exposure, word frequency, word gender, and word animacy predict observed response time.

Choose a centring method. Before predictors can be entered into the model as fixed effects, they need to be centred. For our tutorial we are interested in what is happening at level-1 so we will use grand-mean centring. We computed new centered variables to be used in our CREM analysis.

Fixed effect for participant pre-exposure. The equation below adds participant pre-exposure to the random effects model.

$$Y_{si} = \gamma_0 + \gamma_{1\ \text{preexp}_s} + u_{0s} + v_{0i} + \varepsilon_{si} \qquad (4)$$

where $\gamma_{1\ \text{preexp}_s}$ is the main effect of participant pre-exposure

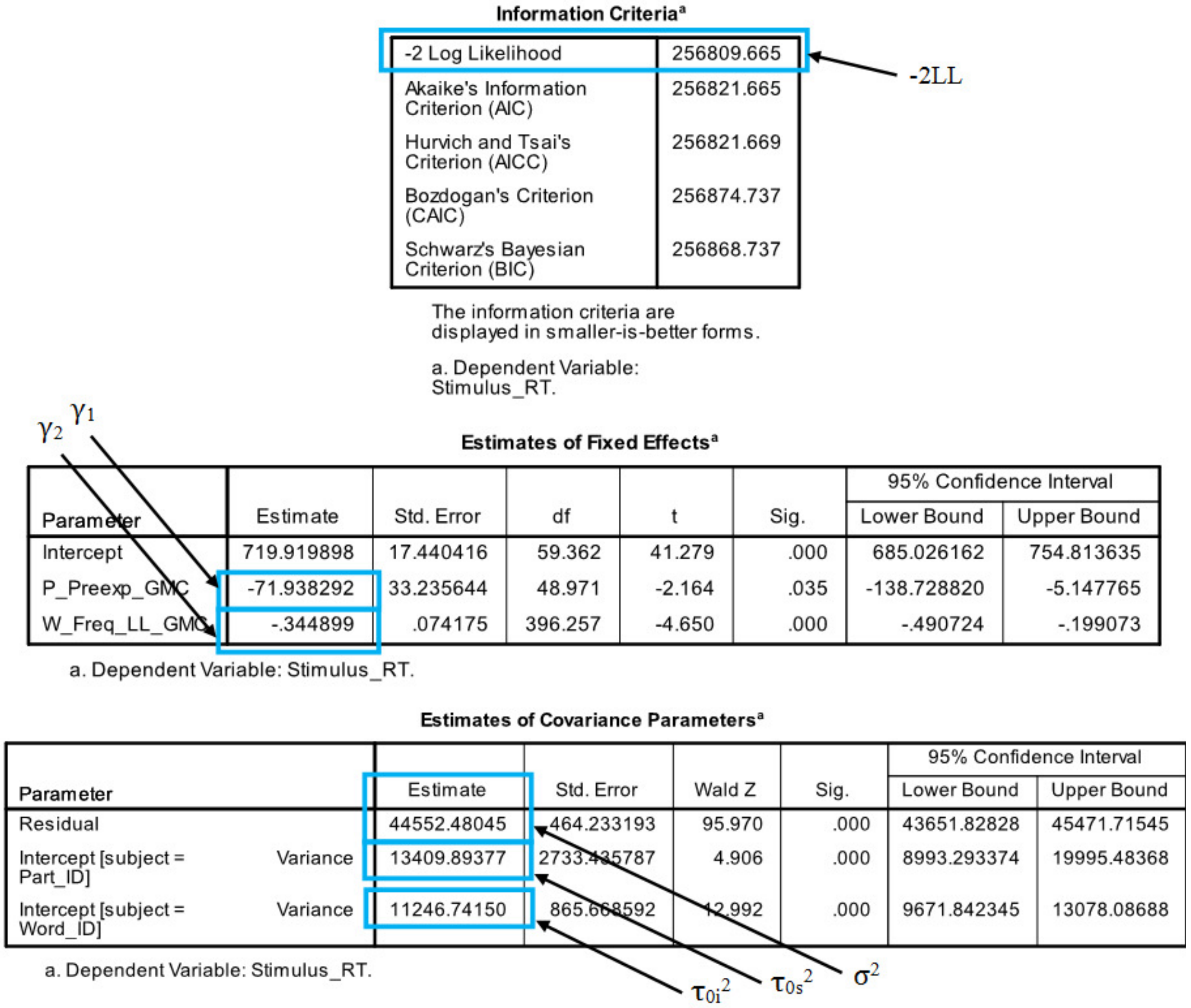
Running the analysis for Model 4, we generate a new output (see Figure 5) with six notable results (see also Table 1, Model 4). First, an adjusted, yet still significant, grand mean response time score is observed, γ_0 = 719.92, p < .001. Second, a new and significant regression coefficient for the main effect of participant pre-exposure is observed, γ_1 = -71.93, p < .05. Third, an adjusted non-zero residual variance is observed, σ^2 = 44552.61, p < .001. Fourth, an adjusted non-zero variance for the random effect of participants is observed, τ_{0s}^2 = 13410.46, p < .001. Fifth, an adjusted non-zero variance for the random effect of words is observed, τ_{0i}^2 = 11905.89, p < .001. Sixth, we find that Model 4 fits the dataset significantly better than Model 3, χ^2 (1) = 4.47, p < .05.

Fixed effect for word frequency. The equation below adds the second predictor of word frequency to the CREM.

$$Y_{si} = \gamma_0 + \gamma_{1\ \text{preexp}_s} + \gamma_{2\ \text{freq}_i} + u_{0s} + v_{0i} + \varepsilon_{si} \qquad (5)$$

where $\gamma_{2\ \text{freq}_i}$ is the main effect of word frequency

Running the analysis for Model 5 we generate a new output (see Figure 6) with six notable results (see also Table 1, Model 5). First, an adjusted, yet still significant, non-zero regression coefficient for the main effect of participant pre-exposure is observed, γ_1 = -71.94, p < .05. Second, a new and significant regression coefficient for the main effect of word



Information Criteria[a]

-2 Log Likelihood	256809.665
Akaike's Information Criterion (AIC)	256821.665
Hurvich and Tsai's Criterion (AICC)	256821.669
Bozdogan's Criterion (CAIC)	256874.737
Schwarz's Bayesian Criterion (BIC)	256868.737

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: Stimulus_RT.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	719.919898	17.440416	59.362	41.279	.000	685.026162	754.813635
P_Preexp_GMC	-71.938292	33.235644	48.971	-2.164	.035	-138.728820	-5.147765
W_Freq_LL_GMC	-.344899	.074175	396.257	-4.650	.000	-.490724	-.199073

a. Dependent Variable: Stimulus_RT.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		44552.48045	464.233193	95.970	.000	43651.82828	45471.71545
Intercept [subject = Part_ID]	Variance	13409.89377	2733.435787	4.906	.000	8993.293374	19995.48368
Intercept [subject = Word_ID]	Variance	11246.74150	865.668592	12.992	.000	9671.842345	13078.08688

a. Dependent Variable: Stimulus_RT.

Figure 6. SPSS output for Model 5.

frequency is observed, $\gamma_2 = -.34$, $p < .001$. Third, an adjusted non-zero residual variance is observed, $\sigma^2 = 44552.48$, $p < .001$. Fourth, an adjusted non-zero variance for the random effect of participants is observed, $\tau_{0s}^2 = 13409.89$, $p < .001$. Fifth, an adjusted non-zero variance for the random effect of words is observed, $\tau_{0i}^2 = 11246.74$, $p < .001$. Sixth, we find that Model 5 fits the dataset significantly better than Model 4, χ^2 (1) = 21.06, $p < .001$.

Fixed effect for word gender. The equation below adds the third predictor of word gender.

$$Y_{si} = \gamma_0 + \gamma_{1\ \mathrm{preexp}_s} + \gamma_{2\ \mathrm{freq}_i} + \gamma_{3\ \mathrm{gender}_i} + u_{0s} + v_{0i} + \varepsilon_{si} \tag{6}$$

where $\gamma_{3\ \mathrm{gender}_i}$ is the main effect of word gender

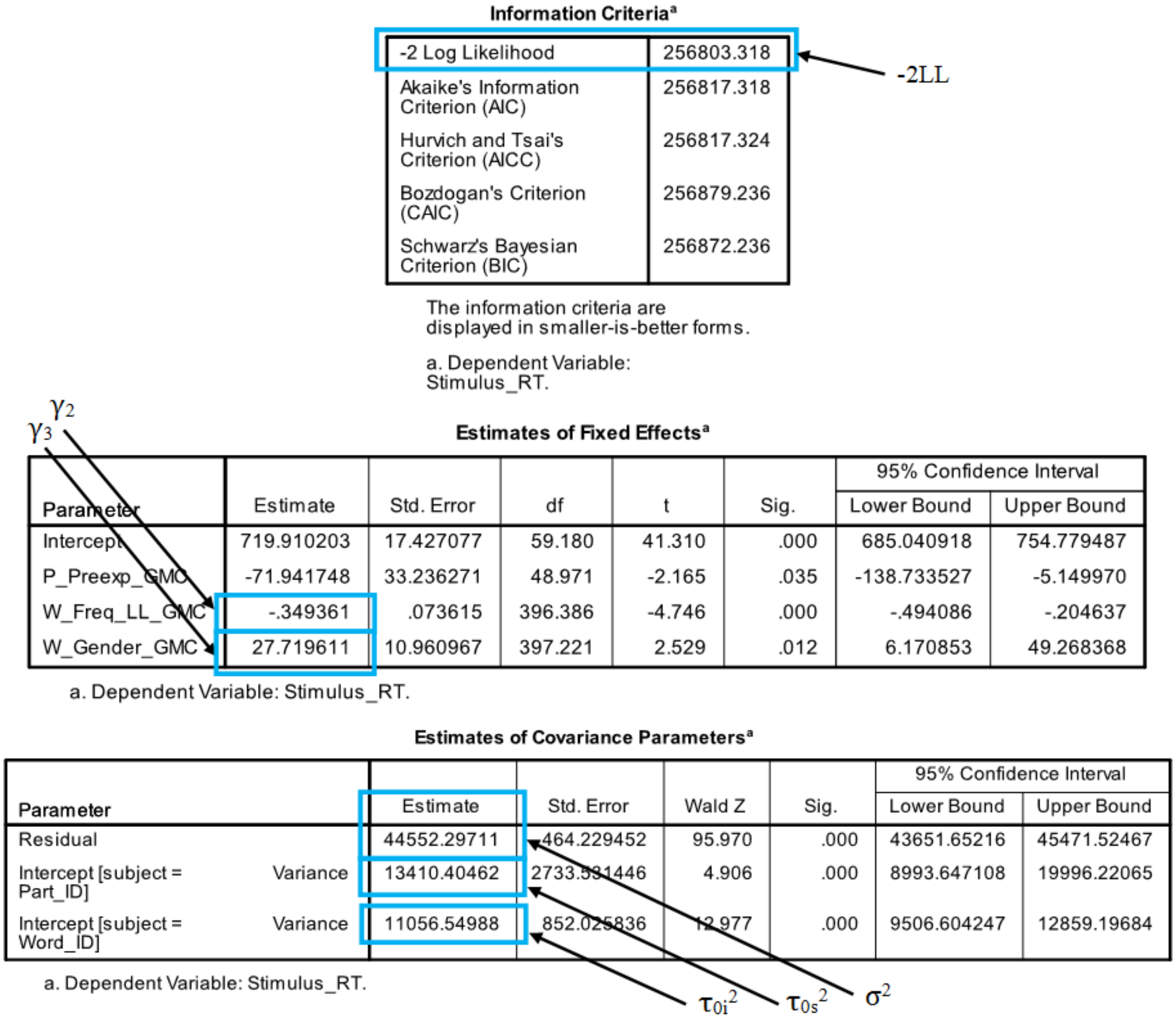
Running the analysis for Model 6, we generate new output (see Figure 7) with six notable results (see also Table 1, Model 6). First, an adjusted, yet still significant, regression coefficient for the main effect of word frequency is observed, $\gamma_2 = -.35$, $p < .001$. Second, a new and significant regression coefficient for the main effect of word gender is observed, $\gamma_3 = 27.72$, $p < .05$. Third, an adjusted non-zero residual variance is observed, $\sigma^2 = 44552.30$, $p < .001$. Fourth, an adjusted non-zero variance for the random effect of participants is observed, $\tau_{0s}^2 = 13410.40$, $p < .001$. Fifth, an adjusted non-zero variance for the random effect of words is observed, $\tau_{0i}^2 = 11056.55$, $p < .001$. Sixth, we find that Model 6 fits the dataset significantly better than Model 5, χ^2 (1) = 6.35, $p < .025$.

Fixed effect for word animacy. The equation below adds the fourth, and final, predictor of word animacy to the CREM. Models are often termed "full" once all the predictors have been added.

$$\begin{aligned} Y_{si} &= \gamma_0 + \gamma_{1\ \mathrm{preexp}_s} + \gamma_{2\ \mathrm{freq}_i} \\ &\quad + \gamma_{3\ \mathrm{gender}_i} + \gamma_{4\ \mathrm{anim}_i} + u_{0s} + v_{0i} + \varepsilon_{si} \end{aligned} \tag{7}$$

where $\gamma_{4\ \mathrm{anim}_i}$ is the main effect of word animacy

Running the analysis for Model 7 we generate a new output (see Figure 8) with seven notable results (see also



Information Criteriaa

-2 Log Likelihood	256803.318
Akaike's Information Criterion (AIC)	256817.318
Hurvich and Tsai's Criterion (AICC)	256817.324
Bozdogan's Criterion (CAIC)	256879.236
Schwarz's Bayesian Criterion (BIC)	256872.236

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: Stimulus_RT.

Estimates of Fixed Effectsa

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	719.910203	17.427077	59.180	41.310	.000	685.040918	754.779487
P_Preexp_GMC	-71.941748	33.236271	48.971	-2.165	.035	-138.733527	-5.149970
W_Freq_LL_GMC	-.349361	.073615	396.386	-4.746	.000	-.494086	-.204637
W_Gender_GMC	27.719611	10.960967	397.221	2.529	.012	6.170853	49.268368

a. Dependent Variable: Stimulus_RT.

Estimates of Covariance Parametersa

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Residual		44552.29711	464.229452	95.970	.000	43651.65216	45471.52467
Intercept [subject = Part_ID]	Variance	13410.40462	2733.531446	4.906	.000	8993.647108	19996.22065
Intercept [subject = Word_ID]	Variance	11056.54988	852.025836	12.977	.000	9506.604247	12859.19684

a. Dependent Variable: Stimulus_RT.

Figure 7. SPSS output for Model 6.

Table 1, Model 7). First, an adjusted, yet still significant, regression coefficient for the intercept is observed, γ_0 = 719.90, p < .001. Second, an adjusted but still significant, regression coefficient main effect of word gender is observed, γ_3 = 27.69, p < .05. Third, a new and significant regression coefficient for the main effect of word animacy is observed, γ_4 = -41.60, p < .001. Fourth, an adjusted non-zero residual variance is observed, σ^2 = 44552.21, p < .001. Fifth, an adjusted non-zero variance for the random effect of participants is observed, τ_{0s}^2 = 13409.66, p < .001. Sixth, an adjusted non-zero variance for the random effect of words is observed, τ_{0i}^2 = 10625.14, p < .001. Seventh, we find that Model 7 fits the dataset significantly better than Model 6, χ^2 (1) = 14.64, p < .001.

5. Estimate the model's effect size

For this tutorial we will calculate an estimated local effect size since we are interested in level-1 variables. The estimated local effect size is calculated by determining the proportional reduction in variance using the equation below. Therefore, using the information from Table 1, we calculate the estimated local effect size to be .36 or 36%.

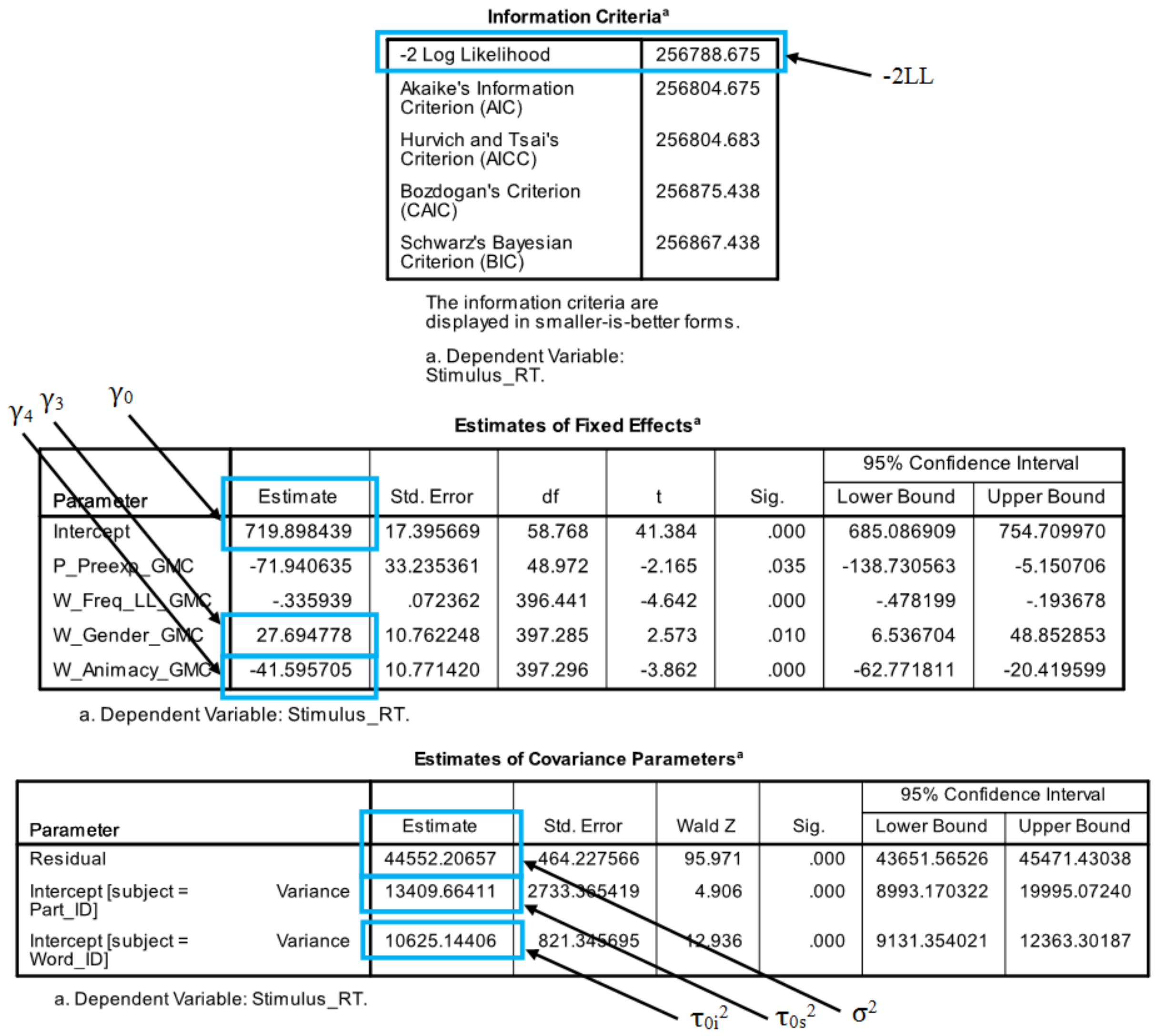
$$(\sigma^2_{\text{Model 1}} - \sigma^2_{\text{Model 7}})/\sigma^2_{\text{Model 1}} \tag{8}$$

This is the same percentage we obtained above when we calculated the proportion of variance explained by the random effects in our model using the ICC. At first glance, then, it appears that the predictors that we added to our model did not explain any of the variance accounted for. However, this is not the case. To determine the variance explained by the predictors over and above that explained by the random effects, we can compare the total variance of Model 3 to Model 7 using the equation below. We find that the predictors account for .036, or 4% of the total variance.

$$\frac{[\sigma^2_{\text{Model 3}} + \tau^2_{0s\text{ Model 3}} + \tau^2_{0i\text{ Model 3}}] - [\sigma^2_{\text{Model 7}} + \tau^2_{0s\text{ Model 7}} + \tau^2_{0i\text{ Model 7}}]}{\sigma^2_{\text{Model 3}} + \tau^2_{0s\text{ Model 3}} + \sigma^2_{0i\text{ Model 3}}}$$

(9)

Therefore, overall, we find a 36% change in the



Information Criteria[a]

-2 Log Likelihood	256788.675
Akaike's Information Criterion (AIC)	256804.675
Hurvich and Tsai's Criterion (AICC)	256804.683
Bozdogan's Criterion (CAIC)	256875.438
Schwarz's Bayesian Criterion (BIC)	256867.438

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: Stimulus_RT.

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	719.898439	17.395669	58.768	41.384	.000	685.086909	754.709970
P_Preexp_GMC	-71.940635	33.235361	48.972	-2.165	.035	-138.730563	-5.150706
W_Freq_LL_GMC	-.335939	.072362	396.441	-4.642	.000	-.478199	-.193678
W_Gender_GMC	27.694778	10.762248	397.285	2.573	.010	6.536704	48.852853
W_Animacy_GMC	-41.595705	10.771420	397.296	-3.862	.000	-62.771811	-20.419599

a. Dependent Variable: Stimulus_RT.

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		44552.20657	464.227566	95.971	.000	43651.56526	45471.43038
Intercept [subject = Part_ID]	Variance	13409.66411	2733.365419	4.906	.000	8993.170322	19995.07240
Intercept [subject = Word_ID]	Variance	10625.14406	821.345695	12.936	.000	9131.354021	12363.30187

a. Dependent Variable: Stimulus_RT.

Figure 8. SPSS output for Model 7.

proportion of total variance explained when we compare Model 1 to Model 7. Four percent of this variance is explained by the predictors we added after Model 3; the proportion of total variance explained by our random effects decreased for Model 7 in relation to Model 3.

Reporting the Results

Now that we have concluded our analyses, we need to summarize our findings. Below is one example of how this can be done (see Kärnä, Voeten, Poskiparta, & Salmivalli, 2010 and Konishi, Hymel, Zumbo, & Li, 2010 for additional examples on how to report MLM findings).

Analyses. We examined whether participant pre-exposure, word frequency, word gender, and word animacy predicted observed response times in a lexical decision task. A one-level CREM was used in order to encompass the random effects of both participants and words. All four predictor variables were grand mean centered.

Results. The results are organized in two sections. First, we present the CREM that tested the validity of labelling participants and words as random effects. Second, we present the CREM that tested the whether the predictor variables, in addition to the random effects, explain observed response times. An estimated effect size is also calculated to measure the amount of variance the full model explains.

Random effects alone model. We used the CREM below to test whether participants and words should be considered random effects.

$$Y_{si} = \gamma_0 + \mu_{0s} + \upsilon_{0i} + \varepsilon_{si}$$

This model states that observed response times (Y_{si}) can be explained by the general intercept (γ_0), the random effect of participants (u_{0s}, which allows response time to vary across participants), the random effect of words (v_{0i}, which allows response time to vary across words), and finally, by a certain amount of random error (ε_{si}).

The results of this model are summarized in Table 1, Model 3[4]. Both random effects were highly significant, indicating that observed response times differed across participants and words, which was expected; participants u_{0s}, Z = 4.91, $p < .001$, and words v_{0i}, Z = 13.05, $p < .001$,. Thus, these effects will be included in the predictor model as random effects.

Predictor model. We used the CREM below to test whether participant pre-exposure, word frequency, word gender, and word animacy helped predict observed response times. Predictors were entered one at a time to test their contribution to the model. All predictors were significant, therefore we only present the results of the final model.

$$Y_s i = \gamma_0 + \gamma_{1\ \mathrm{preexp}_s} + \gamma_{2\ \mathrm{freq}_i} + \gamma_{3\ \mathrm{gender}_i} + \gamma_{4\ \mathrm{anim}_i} + \mu_{0s} + v_{0i} + \varepsilon_{si}$$

The model states that, in addition to the general intercept (γ_0), the random effect of participants (u_{0s}), the random effect of words (v_{0i}), and the random error (ε_{si}), observed response times (Y_{si}) can be predicted by participant pre-exposure ($\gamma_{1(preexp)_s}$), word frequency ($\gamma_{2(freq)_i}$), word gender ($\gamma_{3(gender)_i}$), and word animacy ($\gamma_{4(anim)_i}$).

The results of this model are summarized in Table 1, Model 7. As mentioned above, all of the fixed effects were significant; participant pre-exposure, F(1,49[5]) = 4.69, p = .035, word frequency, F(1, 396) = 21.55, p < .001, word gender, F(1, 397) = 6.62, p = .010, word animacy, F(1, 397) = 14.91, p < .001. Additionally, both random effects remained highly significant; participants (u_{0s}), Z = 4.91, p < .001, and words (v_{0i}) Z = 12.94, p < .001. These findings indicate that all the parameters included in the model help explain observed response times. However, the magnitude of this relationship also needs to be tested.

In order to determine an estimated effect size, we calculated the proportion of variance explained by the predictor model using the formula below. The "empty" model contained only the general intercept and random error; no predictors or random effects were included.

$$\frac{\sigma^2_{\text{"empty" model}} - \sigma^2_{\text{predictor model}})}{\sigma^2_{\text{"empty" model}}}$$

Using the information from Table 1, we found that the estimated effect size for the predictor model was .036 or 36%. Therefore, it explains 36% of the variance for the observed response times.

Summary and Conclusion

This paper had two goals. The first was to provide an overview of MLM and CREM. The second was to provide a step-by-step tutorial on how to apply and report CREM analyses for psycholinguistic data that researchers familiar with SPSS could reference.

As was discussed early on in our paper, the analysis of hierarchical data has come a long way. There has been a clear transition away from ignoring the hierarchical structure of the data or ignoring the possibility of interactions among the hierarchical levels towards the use of MLM techniques. This is supported by the fact that several areas of research now use MLM where more traditional statistical techniques were used in the past. These include developmental research (Cheung, Goodman, Leckie, & Jenkins, 2011), educational research (Pustjens, Van de gaer, Van Damme, Onghena, & Van Landeghem, 2007), health research (Chen, Modin, Ji, & Hjern, 2011), personality research (West, Ryu, Kwok, & Cham, 2011), and romantic relationship research (Teachman, 2011), to name a few. The benefits of using MLM are numerous; it provides superior methods for dealing with problems that arise when applying more traditional statistical methods to hierarchical data. In addition, the limitations are virtually non-existent. The CREM, a type of MLM ideal for psycholinguistic data analysis, was introduced. The discussion surrounding the benefits of CREM over more traditional ANOVA based methods makes its value evident. We outlined the five basic steps required for performing a CREM analysis, along with the choices and theories behind each step.

To facilitate the use of CREM, we demonstrated the step-by-step process in SPSS in tutorial format. We provided a detailed explanation of the logic applied to each step of the analysis process. Important results were highlighted with supporting figures of the SPSS output. Furthermore, we presented an example of how to report your CREM findings in a research article.

In summary, we hope to have provided adequate evidence supporting the benefits of using CREM in psycholinguistics research, along with a clear applied statistical example through the tutorial to facilitate its implementation by researchers in the field.

[4] To preserve space we did not create a new table with the regression coefficient and variance estimates for the three models discussed in this section. Typically this table would be found in an article's results section.

[5] The denominator degrees of freedom are computed by SPSS using the Satherthwaite method; they do not correspond to the number of cases or items (Janssen, 2012). We have rounded them to the nearest whole integer.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412.

Baayen, R. H., Tweedie, F. J., & Schreuder, R. (2002). The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon. *Brain and Language*, *81*(1-3), 55-65.

Beretvas, S. N. (2011). Cross-classified and multiple membership models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 313-334). New York, NY: Routledge.

Bovaird, J. A. & Shaw, L. H. (2012). Multilevel structural equation modeling. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 501-518). New York, NY: The Guilford Press.

Chen, T.-J., Modin, B., Ji, C.-Y., & Hjern, A. (2011). Regional, socioeconomic and urban-rural disparities in child and adolescent obesity in China: A multilevel analysis. *Acta Paediatrica*, *100*(12), 1583-1589.

Cheung, C., Goodman, D., Leckie, G., & Jenkins, J. M. (2011). Understanding contextual effects on externalizing behaviors in children in out-of-home care: Influence of workers and foster families. *Children and Youth Services Review*, *33*(10), 2050-2060.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage Publications Inc.

Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research*, *32*(4), 385-410.

Heck, R. H. & Thomas, S. L. (2009). *An introduction to multilevel modeling techniques* (2nd ed.). New York, NY: Routledge.

Hoffman, L. & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, *39*(1), 101-117.

Janssen, D. P. (2012). Twice random, once mixed: Applying mixed models to simultaneously analyze random effects of language and participants. *Behavior Research Methods*, *44*(1), 232-247.

Kärnä, A., Voeten, M., Poskiparta, E., & Salmivalli, C. (2010). Vulnerable children in varying classroom contexts: Bystanders' behaviors moderate the effects of risk factors on victimization. *Merrill-Palmer Quarterly*, *56*(3), 261-282.

Konishi, C., Hymel, S., Zumbo, B. D., Li, Z. (2010). Do school bullying and student-teacher relationships matter for academic achievement? A multilevel analysis. *Canadian Journal of School Psychology*, *25*(1), 19-39.

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

Maas, C. J. M. & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 85-91.

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models: New simulation results. *Methodology*, *7*(3), 111-120.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*(1), 85-112.

Pustjens, H., Van de gaer, E., Van Damme, J., Onghena, P., & Van Landeghem, G. (2007). The short-term and the long-term effect of primary schools and classes on mathematics and language achievement scores. *British Educational Research Journal*, *33*(3), 419-440.

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

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

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

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

Snijerds, T. & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, England: Sage Publications Inc.

Teachman, J. (2011). Modeling repeatable events using discrete-time data: Predicting marital dissolution. *Journal of Marriage and Family*, *73*(3), 525-540.

West, S. G., Ryu, E., Kwok, O.-M., & Cham, H. (2011). Multilevel modeling: Current and future applications in personality research. *Journal of Personality*, *79*(1), 2-50.

Wike, E. L. & Church, J. D. (1976). Comments on Clark's "The language-as-fixed-effect fallacy". *Journal of Verbal Learning and Verbal Behavior*, *15*(3), 249-255.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology, 8*(1), 52-69.

Manuscript received 16 July 2012
Manuscript accepted 10 October 2012

Appendix follows.

Appendix: SPSS Syntax for Estimating Crossed Random Effects One-Level Models

```
*Empty model - Model #1.
MIXED Stimulus_RT WITH P_Preexp_GMC W_Freq_LL_GMC W_Gender_GMC W_Animacy_GMC
 /CRITERIA=MXITER(150)
 /FIXED=| SSTYPE(3)
 /METHOD=ML
 /PRINT=G  SOLUTION TESTCOV
 /EMMEANS=TABLES(OVERALL).

*Add random effect for participant ID - Model #2.
MIXED Stimulus_RT WITH P_Preexp_GMC W_Freq_LL_GMC W_Gender_GMC W_Animacy_GMC
 /CRITERIA=MXITER(150)
 /FIXED=| SSTYPE(3)
 /METHOD=ML
 /PRINT=G  SOLUTION TESTCOV
 /RANDOM=INTERCEPT | SUBJECT(Part_ID) COVTYPE(ID)
 /EMMEANS=TABLES(OVERALL).

*Add random effect for word ID - Model #3.
MIXED Stimulus_RT WITH P_Preexp_GMC W_Freq_LL_GMC W_Gender_GMC W_Animacy_GMC
 /CRITERIA=MXITER(150)
 /FIXED=| SSTYPE(3)
 /METHOD=ML
 /PRINT=G  SOLUTION TESTCOV
 /RANDOM=INTERCEPT | SUBJECT(Part_ID) COVTYPE(ID)
 /RANDOM=INTERCEPT | SUBJECT(Word_ID) COVTYPE(ID)
 /EMMEANS=TABLES(OVERALL).

*Add fixed effect for participant pre-exposure - Model #4.
MIXED Stimulus_RT WITH P_Preexp_GMC W_Freq_LL_GMC W_Gender_GMC W_Animacy_GMC
 /CRITERIA=MXITER(150)
 /FIXED=P_Preexp_GMC | SSTYPE(3)
 /METHOD=ML
 /PRINT=G  SOLUTION TESTCOV
 /RANDOM=INTERCEPT | SUBJECT(Part_ID) COVTYPE(ID)
 /RANDOM=INTERCEPT | SUBJECT(Word_ID) COVTYPE(ID)
 /EMMEANS=TABLES(OVERALL).

*Add fixed effect for word frequency -Model #5.
MIXED Stimulus_RT WITH P_Preexp_GMC W_Freq_LL_GMC W_Gender_GMC W_Animacy_GMC
 /CRITERIA=MXITER(150)
 /FIXED=P_Preexp_GMC W_Freq_LL_GMC | SSTYPE(3)
 /METHOD=ML
 /PRINT=G  SOLUTION TESTCOV
 /RANDOM=INTERCEPT | SUBJECT(Part_ID) COVTYPE(ID)
 /RANDOM=INTERCEPT | SUBJECT(Word_ID) COVTYPE(ID)
 /EMMEANS=TABLES(OVERALL).

*Add fixed effect for word gender - Model #6.
MIXED Stimulus_RT WITH P_Preexp_GMC W_Freq_LL_GMC W_Gender_GMC W_Animacy_GMC
 /CRITERIA=MXITER(150)
 /FIXED=P_Preexp_GMC W_Freq_LL_GMC W_Gender_GMC | SSTYPE(3)
 /METHOD=ML
 /PRINT=G  SOLUTION TESTCOV
 /RANDOM=INTERCEPT | SUBJECT(Part_ID) COVTYPE(ID)
 /RANDOM=INTERCEPT | SUBJECT(Word_ID) COVTYPE(ID)
 /EMMEANS=TABLES(OVERALL).
```

```
*Add fixed effect for word animacy - Model #7.
MIXED Stimulus_RT WITH P_Preexp_GMC W_Freq_LL_GMC W_Gender_GMC W_Animacy_GMC
 /CRITERIA=MXITER(150)
 /FIXED=P_Preexp_GMC W_Freq_LL_GMC W_Gender_GMC W_Animacy_GMC | SSTYPE(3)
 /METHOD=ML
 /PRINT=G  SOLUTION TESTCOV
 /RANDOM=INTERCEPT | SUBJECT(Part_ID) COVTYPE(ID)
 /RANDOM=INTERCEPT | SUBJECT(Word_ID) COVTYPE(ID)
 /EMMEANS=TABLES(OVERALL).
```