languages

MDPI

Tutorial

Applying Mixed-Effects Models in Research on Second Language Acquisition: A Tutorial for Beginners

Marc Brysbaert

Department of Experimental Psychology, Ghent University, 9000 Gent, Belgium; marc.brysbaert@ugent.be

Abstract: Mixed-effects models have become indispensable tools for analyzing data in second language acquisition (SLA) research. This tutorial offers a step-by-step guide to conducting mixed-effects analyses for simple designs using the gamlj package in jamovi, a user-friendly, free statistical software. We begin by discussing the advantages of mixed-effects modeling over traditional methods, particularly for SLA data, and the rationale for focusing on simple designs. Subsequently, we introduce the gamlj package, highlighting its intuitive interface and error-prevention features. To illustrate the application of the package, we employ toy datasets that can be easily replicated and used with other statistical software. By providing a clear and accessible approach, this tutorial empowers SLA researchers to effectively analyze their data and draw meaningful conclusions.

Keywords: mixed-effects model; statistical analysis; second language research; jamovi

Academic Editor: Jeanine Treffers-Daller

Received: 26 September 2024
Revised: 19 January 2025
Accepted: 20 January 2025
Published: 23 January 2025

Citation: Brysbaert, M. (2025). Applying Mixed-Effects Models in Research on Second Language Acquisition: A Tutorial for Beginners. *Languages*, *10*(2), 20. https://doi.org/10.3390/languages10020020

Mixed-effects analysis is increasingly used to analyze data in language research (for good reason, as we will see below). Unfortunately, such analyses quickly become complicated, with the risk of drawing wrong conclusions.

This tutorial is an attempt to provide hands-on guidance on the use of mixed-effects analysis. Before doing so, it is important to remember that statistics are not meant to think for you. Two mistakes often made in statistical analysis are (1) performing the analysis without looking at the means and standard deviations of the conditions, and (2) throwing in every variable that seems sensible/interesting to see what holds up (i.e., is statistically significant).

The way statistics should be used is that you first look at the means and standard deviations of the data you obtained, see if the pattern makes sense and agrees with the expectations you had, and ONLY THEN look at whether the observed differences are statistically significant.[1] Similarly, variables should only be included in an analysis if you have good (theoretical) reasons to expect them to have an effect. Otherwise, variables can obscure or distort the effects you are interested in. This applies to mixed-effects analyses as well as any other statistical analysis. Similarly, control variables should be added to the model only if you have good evidence that they have effects that must be taken into account (Cinelli et al., 2024; Wysocki et al., 2022).

Statistics are most useful if the design is simple and directly related to the research question you want to answer (Cohen, 1992). Anything above a first-order interaction is a nightmare to interpret and, for typical effects studied by language researchers, requires hundreds of participants and observations to be replicable (Brysbaert, 2019). It is far better to include control variables in the selection of stimulus materials and in the study design than try to account for them by adding them to a statistical analysis.

Because it is good to aim for simple designs, the present discussion is limited to the analysis of such designs. There is little point in discussing complicated designs, as it is my

experience that such designs add more uncertainty to the literature than clarity. It is better to conduct a series of focused studies than a single, complex study that tries to cover all possible questions and take into account all possible criticisms.

The analyses will be illustrated with small toy datasets. This is to make the text more understandable and is not intended to give the impression that you can do interesting language research with such small numbers of participants and stimuli. In fact, it is now well known that many language research studies in the past were not sufficiently powerful. Given the typically investigated effect sizes, language research papers with fewer than 50–200 participants and 40 stimuli per condition rarely provide trustworthy results (Baker et al., 2021; Brysbaert, 2019, 2021; Brysbaert & Stevens, 2018; Kumle et al., 2021; Langenberg et al., 2023; Mahowald et al., 2016; Westfall et al., 2014). Keep in mind that a lot of data for a simple design makes your analysis and interpretation easier rather than more difficult: either the effect is clearly present, or it is too small to be of value. Statistics were not invented to draw far-reaching conclusions based on a few dozen data points.

Before going into the details of mixed-effects analysis, it is important to know why such an analysis is necessary. What is wrong with a simple *t*-test or ANOVA?[2]

1. Why Do We Need Mixed-Effects Analysis in Research on Second Language Acquisition?

In second language research, we want to draw conclusions that are true for all relevant participants and for all relevant stimuli. Suppose you have the hypothesis that it is easier to learn nouns in a second language (L2) than adjectives. To test the hypothesis, you teach a group of participants a sample of nouns and a sample of adjectives. Both samples are matched on important control variables, such as word length, word frequency, word valence, word concreteness, and so on.

Because of all the constraints (and time constraints for your participants), you limit your stimuli to a list of 15 L2 nouns and 15 matching L2 adjectives. Participants are given some time to study the list of new words and their translations in the dominant language (L1). After some distraction time, participants are asked to translate the L2 words into L1. The dependent variable is whether the translation is correct or not. There are 12 participants. Table 1 shows the outcome. This table can also be found on https://osf.io/f3hjb/, so you can repeat all the analyses that are described.

The traditional way to analyze the data in Table 1 would be to use a *t*-test for related samples or an ANOVA with a single repeated measure across participants. To do this, we take the total number of correct translations per participant for the 15 nouns and the 15 adjectives, as shown in Table 2. A *t*-test on the number of correct answers tells us that, contrary to the expectations, the adjectives were learned more often (M = 9.42, SD = 4.08) than the nouns (M = 8.17, SD = 4.00) and that the difference is significant ($t(11) = -3.19$, $p < 0.01$).

A different picture emerges if we take the number of correct translations per word as the dependent variable instead of the number of correct translations per participant, as shown in Figure 1. Now, we see that for both nouns and adjectives, there are some easy words that were learned by almost everyone and difficult words that were learned by almost no one. The noun W2 from Table 1 was translated by 10/12 participants, while noun W15 was translated by only 2/12 participants. The same is true for the adjectives. Given this variability, the difference between nouns and adjectives seems quite small. Our impression is confirmed when we perform a *t*-test for unrelated samples on the number of correct translations per item. This yields $t(28) = -0.88$, $p = 0.387$. We must use a test for unrelated samples, because each word belongs to either the nouns or the adjectives.

Table 1. A toy example of a study in which 12 participants learn the L1 translation of 15 L2 nouns and 15 matched L2 adjectives.

Word	Type	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
W1	Noun	1	0	0	1	1	1	1	0	0	0	1	1
W2	Noun	1	1	1	1	1	0	1	1	1	0	1	1
W3	Noun	0	1	0	0	1	1	1	0	0	0	1	1
W4	Noun	0	0	0	0	1	0	0	0	0	0	0	1
W5	Noun	1	1	0	1	1	0	1	1	1	0	1	0
W6	Noun	1	0	1	0	1	1	1	0	1	0	1	1
W7	Noun	0	1	0	0	0	0	0	1	1	0	0	1
W8	Noun	1	1	0	1	1	1	1	1	1	0	1	1
W9	Noun	1	0	0	0	1	1	1	0	0	0	1	0
W10	Noun	1	0	1	0	1	1	1	1	1	0	1	1
W11	Noun	0	0	0	0	1	0	0	0	0	0	1	0
W12	Noun	1	0	1	1	1	1	1	1	1	0	1	1
W13	Noun	1	0	0	0	1	1	1	0	0	0	1	1
W14	Noun	1	0	1	0	1	1	1	1	1	0	1	1
W15	Noun	0	0	0	0	1	0	0	0	0	0	1	0
W16	Adjective	1	1	1	0	1	1	1	1	1	1	1	1
W17	Adjective	1	1	0	1	1	1	1	1	1	0	1	1
W18	Adjective	1	1	0	1	1	1	1	1	1	0	1	1
W19	Adjective	1	0	0	0	1	1	1	1	0	0	1	1
W20	Adjective	0	0	0	0	1	0	0	0	0	0	1	1
W21	Adjective	1	1	1	1	1	1	1	1	1	0	1	1
W22	Adjective	1	0	1	0	1	1	1	0	1	0	1	1
W23	Adjective	1	1	0	1	1	1	1	1	1	1	1	1
W24	Adjective	1	0	0	0	1	1	1	1	1	0	1	1
W25	Adjective	0	0	0	0	1	0	1	0	0	0	0	1
W26	Adjective	1	0	0	0	1	1	1	1	1	0	1	1
W27	Adjective	0	1	0	0	1	0	1	0	1	0	1	1
W28	Adjective	1	1	0	1	1	1	1	1	1	1	1	1
W29	Adjective	1	0	0	0	1	1	0	0	0	0	1	0
W30	Adjective	0	0	0	0	1	0	1	0	0	0	0	0

Table 2. A summary table of Table 1 for an analysis across participants. The dependent variable is the number of correct translations.

Part	Noun	Adj
P1	10	11
P2	5	7
P3	5	3
P4	5	5
P5	14	15
P6	9	11
P7	11	13
P8	7	9
P9	8	10
P10	0	3
P11	13	13
P12	11	13

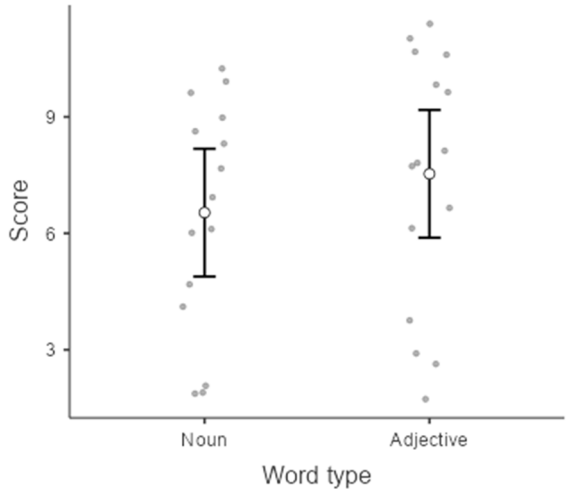

Figure 1. How well are the individual nouns and adjectives learned? Each dot represents one word. Also shown are the means and the 95% confidence intervals around the means. These show that the difference between nouns and adjectives is not significant in an analysis over items.

So, what can we conclude from our small study? We can say that if we present the same 15 nouns and 15 adjectives to another group of participants, we are likely to find the same difference (i.e., an advantage for the adjectives over the nouns). However, if we present two new samples of nouns and adjectives to (the same) participants, we have little guarantee that the adjectives will again outperform the nouns. For that, the differences within word classes are too great (as shown in Figure 1). So, on the basis of our analyses, we cannot conclude from our study that adjectives in general are easier to learn in L2 than nouns. We can only conclude that the sample of 15 adjectives we compiled was easier to learn than the sample of 15 nouns we compiled. Needless to say, this is not what we want to conclude.

Concluding that an effect is significant only if it is significant in both the analysis across participants (often called an F1 analysis) and in the analysis across words (an F2 analysis) is what language researchers did before the advent of mixed-effects methods. However, this approach has two drawbacks.

First, the analysis is cumbersome and not mathematically sound (Raaijmakers et al., 1999). Second, the F1 and F2 analyses often yield different descriptive statistics, especially when the design is not balanced or when values are missing. Missing values are also annoying because they can lead to a situation where all data from a participant or for a word must be discarded. This is not the case with mixed-effects modeling.

Once you know how to run a mixed-effects model, the analysis is actually easier than a combined F1 and F2 analysis, where you first have to calculate averages per participant or per stimulus.

2. A Mixed-Effects Analysis of Table 1

To perform a mixed-effects analysis, we need to turn the data from Table 1 into a long format. In such a format, there is one line per observation (see Wickham, 2014, for an introduction to data analysis). The first observation is from participant 1 (P1) about word 1 (W1), which was a noun and was correctly translated (1). The second observation is again from P1, about W2, which was a noun and was correctly translated, and so on. Figure 2 shows what the long notation looks like. In total, you have 360 data lines (12 participants $\times$ 30 words). The order of the lines is not important (for example, you can also have 30 words $\times$ 12 participants). What is important is that each line contains all the

information about the observation. In the example, that is the participant, the word, the type of word, and the score.

Participant	Word	Type	Score
P1	W1	Noun	1
P1	W2	Noun	1
P1	W3	Noun	0
P1	W4	Noun	0
P1	W5	Noun	1

Figure 2. The first five observations from Table 1 in a spreadsheet with long format. Each line is a different observation.

We can perform mixed-effects analysis in R, but as novice researchers, it is safer to perform the analysis in an environment that protects us from making mistakes. The default LME analysis in R works with dummy coding instead of sum coding, does not center the variables involved in an interaction, and codes random effects in a way which is difficult to understand for unexperienced users (see Brysbaert & Debeer, 2025, for more information). In addition, the user requires knowledge of other packages to obtain descriptive statistics and graphs. The gamlj package (Gallucci, 2022) in jamovi (The jamovi Project, 2022) has been developed to avoid all those issues and make LME as simple and robust as possible.

You can download jamovi for free at www.jamovi.org and find instructions for installing it on Youtube. Once jamovi is running, click on Modules and jamovi library, as shown in Figure 3. Select gamlj. It will appear on your jamovi dashboard as "Linear Models".

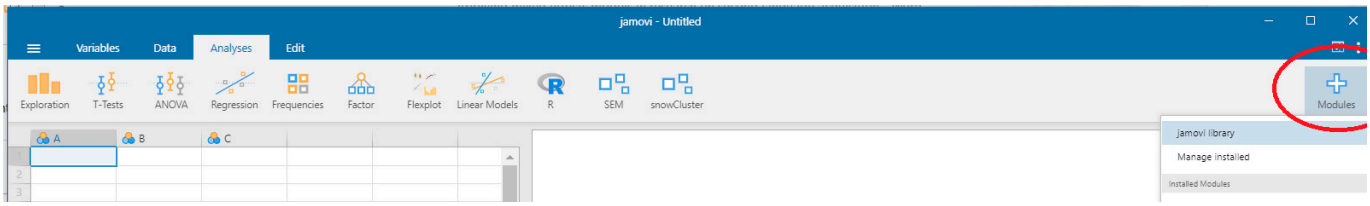

Figure 3. Adding gamlj to your jamovi dashboard. Click on modules, then jamovi library. Select gamlj, and you will see "Linear Models" added to your dashboard.

Now, open your file with the data from Figure 2 (jamovi reads Excel files) and select Mixed Model, as shown in Figure 4.

Then, indicate which column is the dependent variable, which column is the independent variable, and over which variables you want to generalize (the cluster variables). First, we say what the dependent variable is. That is the score (0 or 1). Next, we say which variables we manipulated. There is only one variable we manipulated and that is the type of words. It is a two-level, categorical variable (nouns and adjectives). It thus falls under Factors. If the variable had been continuous (e.g., frequency of words), we would have put it under Covariates. Finally, we list the cluster variables (often called the random factors). These are the participants and the words. The outcome is shown in Figure 5.

Next, we need to select the random effects, as shown in Figure 6. We almost always need to include the random intercepts, both for participants and for words. The participant intercept takes into account that participants differed in how much they learned; the word intercept takes into account that not all words were equally easy to learn. Thus, we include both Intercept | Participant and Intercept | Word.

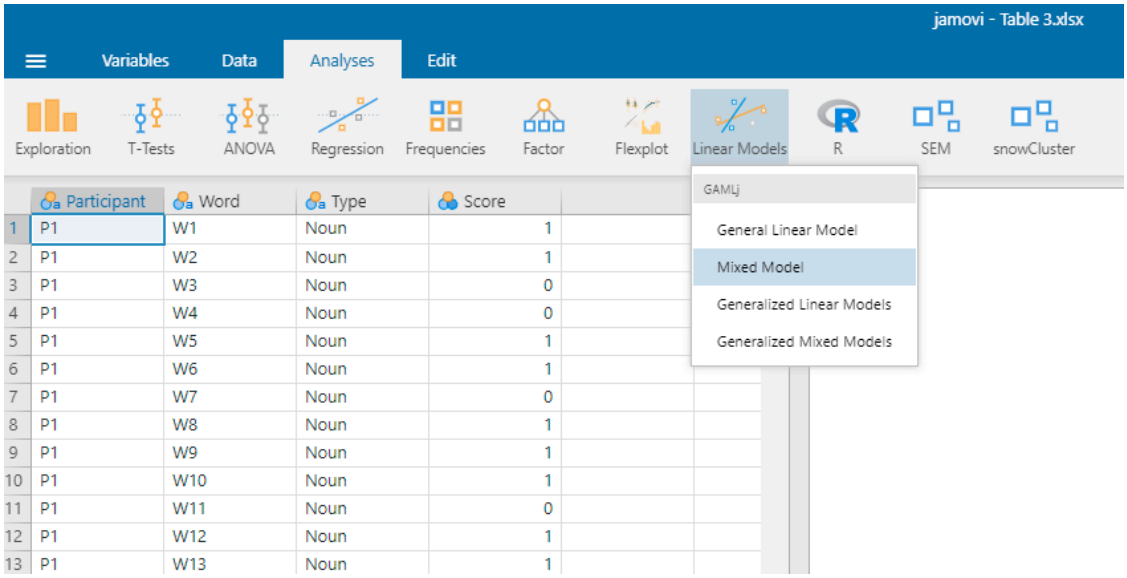

Figure 4. Select Mixed Model from gamlj.

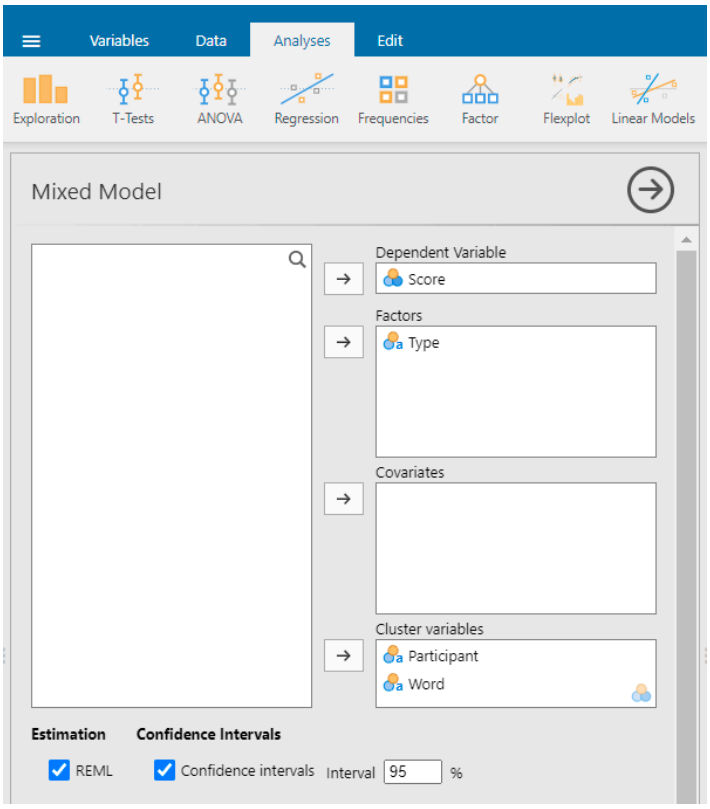

Figure 5. Tell gamlj which columns represent the dependent variable and the independent variable(s), and the variables over which you want to generalize (cluster variables).

There are two more random effects in our design: Type | Participant and Type | Word. These refer to the random slopes of the variable we are interested in. A detailed discussion of this part is given in Brysbaert and Debeer (2025). In short, random slopes must be included for repeated measures, but not for variables between groups. In most language studies, either the stimuli are a repeated measure (when two groups of participants are asked to respond to the same stimuli) or the participants are a repeated measure (when one group of participants sees different stimuli in the two conditions). Occasionally, both

the participants and the stimuli can be repeated measures (when one group of participants sees the same stimuli in different conditions).

Given that in the toy example, word type is a repeated measure across participants (each participant saw both nouns and adjectives), we must include Type | Participant in the random coefficients. Because word type is a between-word variable (a word was either a noun or an adjective), we should not include Type | Word. All in all, this yields the selection shown in Figure 7.

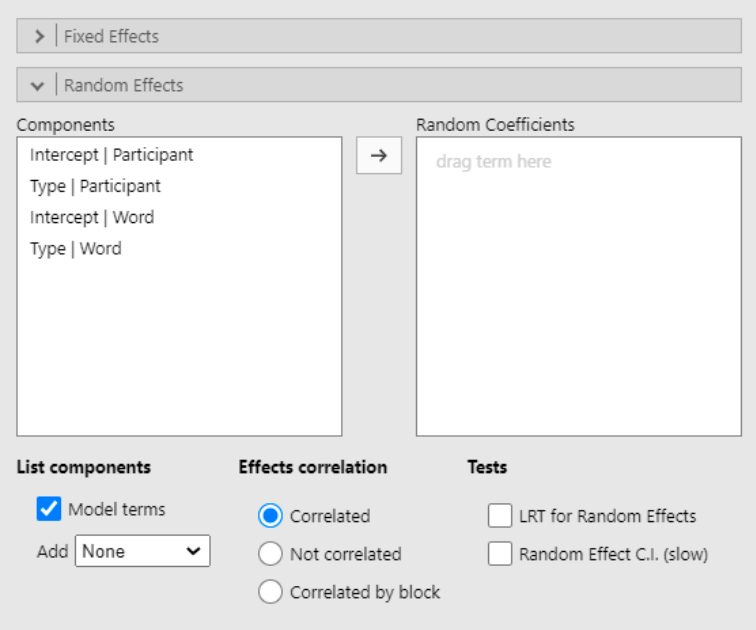

Figure 6. The panel for selecting the random variables.

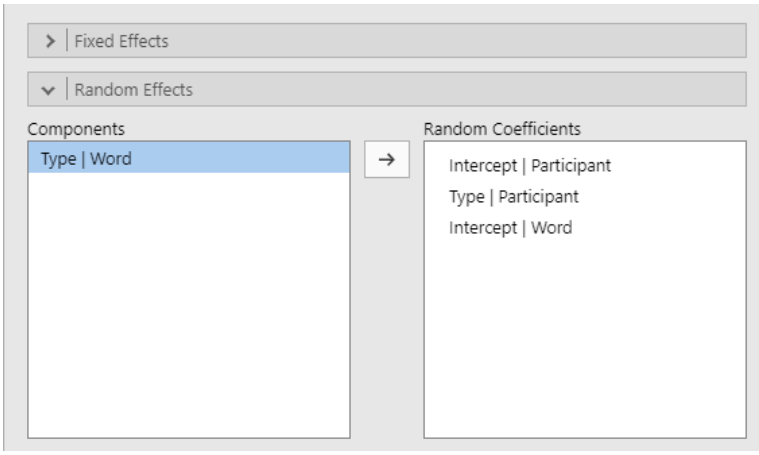

Figure 7. The selection of the random coefficients for the data shown in Figure 2.

Once the random coefficients are selected, jamovi automatically performs the analysis. The output is shown in Figure 8.

The most important part is that under Model Results. It tells us that the difference between nouns and adjectives is not significant ($t(28) = -0.879$, $p = 0.387$) if we want to generalize across both participants and words, as we might have expected from the F2 analysis.

Mixed Model

Model Info

Info	
Estimate	Linear mixed model fit by REML
Call	Score ~ 1 + Type+(1 + Type | Participant)+(1 | Word)
AIC	381.01320
BIC	414.16730
LogLikel.	−186.48231
R-squared Marginal	0.00689
R-squared Conditional	0.49302
Converged	yes
Optimizer	bobyqa

Note. (Almost) singular fit. Maybe random coefficients variances are too small or correlations among them too large.

Note. boundary (singular) fit: see ?isSingular

[3]

Model Results

Fixed Effect Omnibus tests

	F	Num df	Den df	p
Type	0.773	1	28.0	0.387

Note. Satterthwaite method for degrees of freedom

Fixed Effect Omnibus tests

	F	Num df	Den df	p
Type	0.773	1	28.0	0.387

Note. Satterthwaite method for degrees of freedom

Fixed Effects Parameter Estimates

				95% Confidence Interval				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	0.5861	0.0881	0.414	0.759	18.2	6.656	<0.001
Type1	Noun - Adjective	−0.0833	0.0948	−0.269	0.102	28.0	−0.879	0.387

Random Components

Groups	Name	SD	Variance	ICC
Word	(Intercept)	0.23803	0.0567	0.307
Participant	(Intercept)	0.25715	0.0661	0.341
	Type1	0.00551	3.04e−5	
Residual		0.35785	0.1281	

Note. Number of Obs: 360 , groups: Word 30, Participant 12

Random Parameters correlations

Groups	Param.1	Param.2	Corr.
Participant	(Intercept)	Type1	−1.00

Figure 8. Jamovi output for the analysis of Table 3.

The random components are also interesting, because they tell us that there is virtually no variability in the random slopes of type across participants (SD = 0.00551), meaning that the difference between adjectives and nouns was very similar for all participants. This is why we were warned about singular fit under the first table of the output. In such a case, it is better to omit Type | Participant from the random coefficients. If we do that, we see that everything remains the same, with no warnings this time, as shown in Figure 9.

Mixed Model

Model Info

Info	
Estimate	Linear mixed model fit by REML
Call	Score ~ 1 + Type+(1 | Participant)+(1 | Word)
AIC	377.03210
BIC	402.41410
LogLikel.	−186.49180
R-squared Marginal	0.00689
R-squared Conditional	0.49299
Converged	yes
Optimizer	bobyqa

[3]

Model Results

Fixed Effect Omnibus tests

	F	Num df	Den df	p
Type	0.774	1	28.0	0.387

Note. Satterthwaite method for degrees of freedom

Fixed Effects Parameter Estimates

				95% Confidence Interval				
Names	Effect	Estimate	SE	Lower	Upper	df	t	p
(Intercept)	(Intercept)	0.5861	0.0881	0.414	0.759	18.2	6.656	<0.001
Type1	Noun - Adjective	−0.0833	0.0947	−0.269	0.102	28.0	−0.880	0.387

Random Components

Groups	Name	SD	Variance	ICC
Word	(Intercept)	0.238	0.0567	0.307
Participant	(Intercept)	0.257	0.0661	0.341
Residual		0.358	0.1281	

Note. Number of Obs: 360 , groups: Word 30, Participant 12

Figure 9. The output when the random slope of word type over participants is omitted from the analysis.

Another nice aspect of gamlj is that it is easy to plot the results. Simply indicate you want a plot, as shown in Figure 10.

The plot in Figure 10 illustrates an important aspect of our data that we neglected: We have only 0 s and 1 s as dependent variable. This is not ideal for a typical mixed-effects model, which assumes a continuous (normally distributed) dependent variable. We use this with reaction times or average scores, but not with true/false observations.

A better analysis for binary dependent variables (consisting of 0 s and 1 s) is logistic analysis (Jaeger, 2008). For this, we select Generalized Mixed Models in gamlj. Everything else remains the same, as shown in Figure 11.

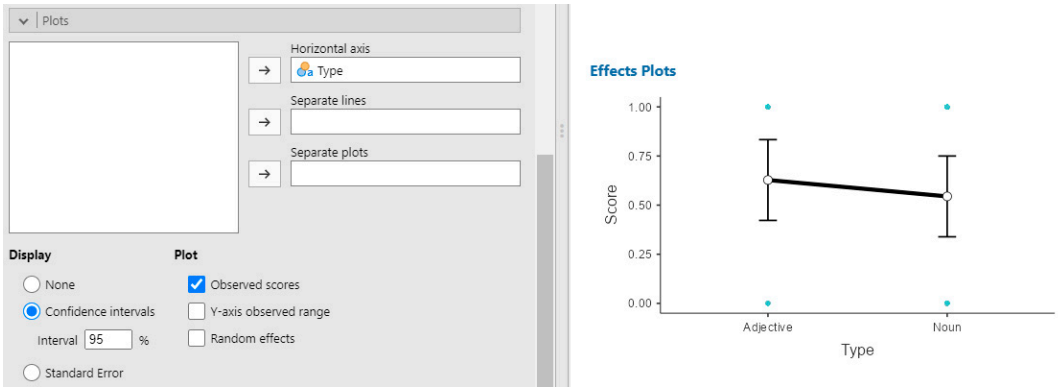

Figure 10. The commands to obtain a plot of the findings.

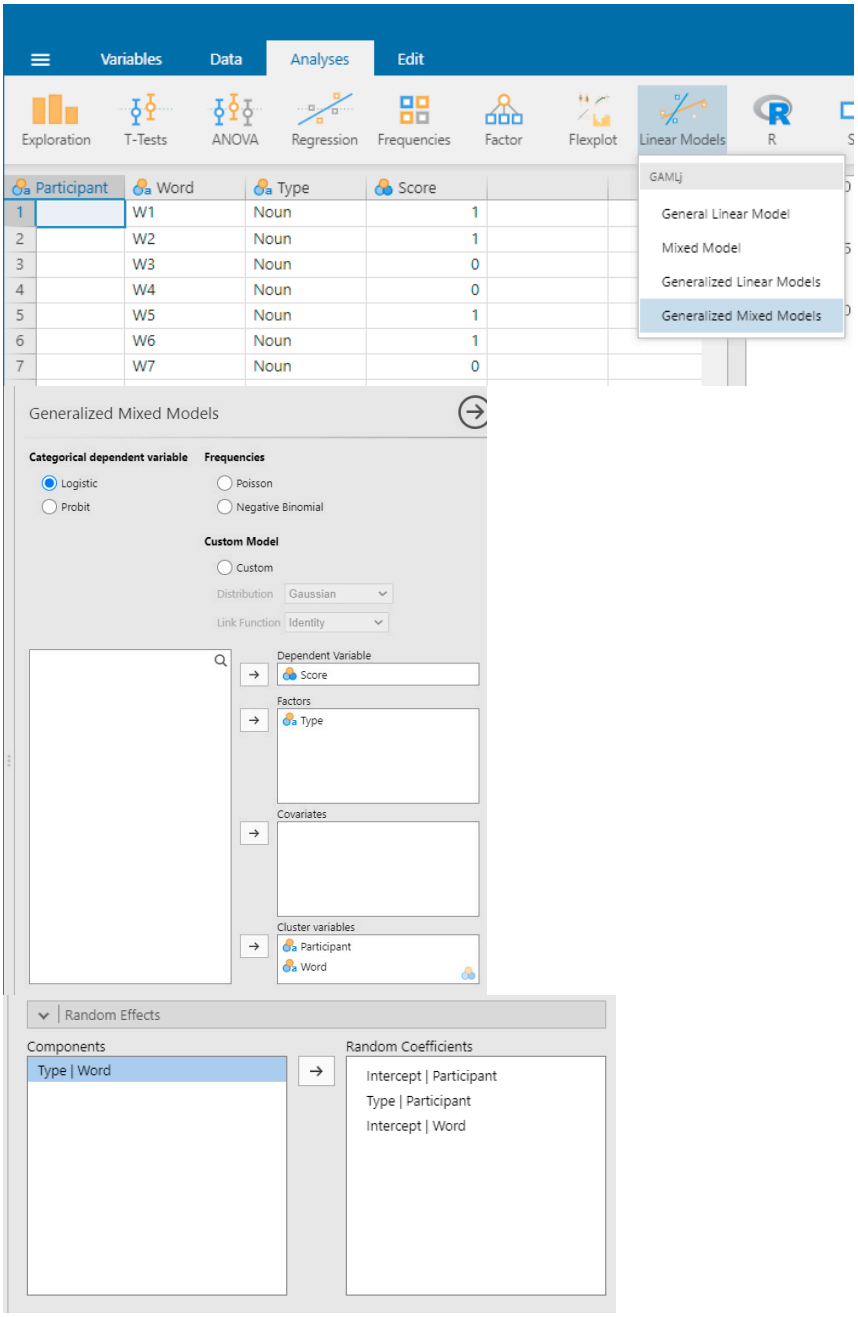

Figure 11. How to run a logistic mixed-effects model for the data from Figure 2 (which is better for yes/no responses).

In the output (Figure 12), we see that the *p*-value for word type is slightly lower, but still not significant (z = −1.10, $p = 0.270$), which it should be, given the variability in recognition rates for the individual words (both nouns and adjectives). We also see that there is now no problem with Type | Participant.

Generalized Mixed Models

Model Info

Info	Value	Comment
Model Type	Logistic	Model for binary y
Call	glm	Score ~ 1 + Type + (1 + Type | Participant) + (1 | Word)
Link function	Logit	Log of the odd of y=1 over y=0
Direction	P(y=1)/P(y=0)	P(Score = 1) / P(Score = 0)
Distribution	Binomial	Dichotomous event distribution of y
LogLikel.	−173.8521	Unconditional Log-Likelihood
-2*LogLikel.	347.7041	Unconditional absolute deviance
Deviance	230.9613	Conditional relative deviance
R-squared	0.0162	Marginal
R-squared	0.7098	Conditional
AIC	359.7000	Less is better
BIC	383.0208	Less is better
Residual DF	354.0000	
Chi-squared/DF	0.5414	Overdispersion indicator
Converged	yes	
Optimizer	bobyqa	

Note. boundary (singular) fit: see ?isSingular

[3]

Model Results

Fixed Effect Omnibus tests

	X^2	df	p
Type	1.22	1.00	0.270

Fixed Effects Parameter Estimates

					95% Exp(B) Confidence Interval			
Names	Effect	Estimate	SE	exp(B)	Lower	Upper	z	p
(Intercept)	(Intercept)	0.740	0.708	2.097	0.5235	8.40	1.05	0.296
Type1	Noun - Adjective	−0.856	0.775	0.425	0.0930	1.94	−1.10	0.270

Random Components

Groups	Name	SD	Variance	ICC
Word	(Intercept)	1.887	3.559	0.520
Participant	(Intercept)	2.060	4.243	0.563
	Type1	0.487	0.237	
Residuals		1.000	1.000	.

Note. Number of Obs: 360 , groups: Word 30, Participant 12

Random Parameters correlations

Groups	Param.1	Param.2	Corr.
Participant	(Intercept)	Type1	−1.00

Figure 12. The output of the logistic mixed-effects model.

3. An Example with an Interaction Between Two Independent Variables

We have already learned most of what there is to learn about mixed-effects analysis, as long as we stick to simple designs. Two more applications of mixed-effects analysis will be illustrated below: One with two independent variables and one with a continuous predictor.

Table 3 presents data from a toy experiment in which 12 participants wrote down as many words as they could after being given the first letter. Thus, when participants were given the letter "F," they had to produce as many words beginning with the letter "F" (foil, freeze, free, . . .) within one minute. Two variables were manipulated: the language in which the words were to be given (L1/L2) and whether the first letters were easy (P, S, D) or difficult (G, L, N) given their frequency as the first letter of an English word. The critical effect in which the researcher was interested was whether there would be an interaction between language (L1/L2) and condition (easy vs. difficult), the idea being that participants would have extra difficulty in the difficult L2 condition.

Table 3. The data of a toy experiment in which language and condition (easy/difficult) are manipulated. The dependent variable is the number of words produced in one minute.

	L1 Easy			L1 Difficult			L2 Easy			L2 Difficult		
Participant	**P**	**S**	**D**	**G**	**L**	**N**	**P**	**S**	**D**	**G**	**L**	**N**
P1	12	14	13	12	17	22	13	12	9	9	7	7
P2	13	13	17	10	13	14	15	12	7	9	6	7
P3	14	18	13	16	14	13	12	8	5	10	8	7
P4	16	15	17	14	13	8	11	7	10	4	0	4
P5	21	17	10	10	12	9	12	12	6	11	4	4
P6	16	17	18	19	13	16	20	10	13	10	10	3
P7	18	12	12	14	15	13	14	10	5	7	5	5
P8	13	18	17	16	16	10	15	13	12	7	9	5
P9	17	8	19	9	12	16	15	15	14	5	4	3
P10	15	19	16	11	10	9	18	13	17	8	8	6
P11	20	20	22	16	15	15	13	13	9	8	11	9
P12	18	19	16	11	12	10	12	13	11	14	6	8

If we want to make statements that generalize beyond the sample of participants and the specific letters used, we must perform a mixed-effects analysis with participants and letters as random (cluster) variables.

To perform a mixed-effects analysis, we must first convert the data from Table 3 into a long format. This produces a table, as shown in Figure 13, with 144 rows of data (12 participants × 12 observations).

Particip	Language	Cond	Letter	Score
P1	L1	Easy	P	12
P1	L1	Easy	S	14
P1	L1	Easy	D	13
P1	L1	Difficult	G	12
P1	L1	Difficult	L	17
P1	L1	Difficult	N	22
P1	L2	Easy	P	13
P1	L2	Easy	S	12

Figure 13. The lay-out of a dataset in long format for the data from Table 3.

Since the dependent variable is a continuous (normally distributed) variable, we can run Mixed Models in gamlj. We manipulated two factors (Language and Condition) and we have two cluster variables (Participant and Letter). Both variables are within participants. Language is within letters (the same letters were presented in L1 and L2), but Condition is a between variable across letters (different letters were used in the easy and difficult condition). Thus, we do not include Cond | Letter and Language:Cond | Letter in the random coefficients (see Brysbaert & Debeer, 2025, for further explanation).

Figure 14 shows the results of the analysis. It shows that the main effects of Language and Condition are significant, but the critical interaction is not ($t(6,18) = 1.26$, $p = 0.253$).

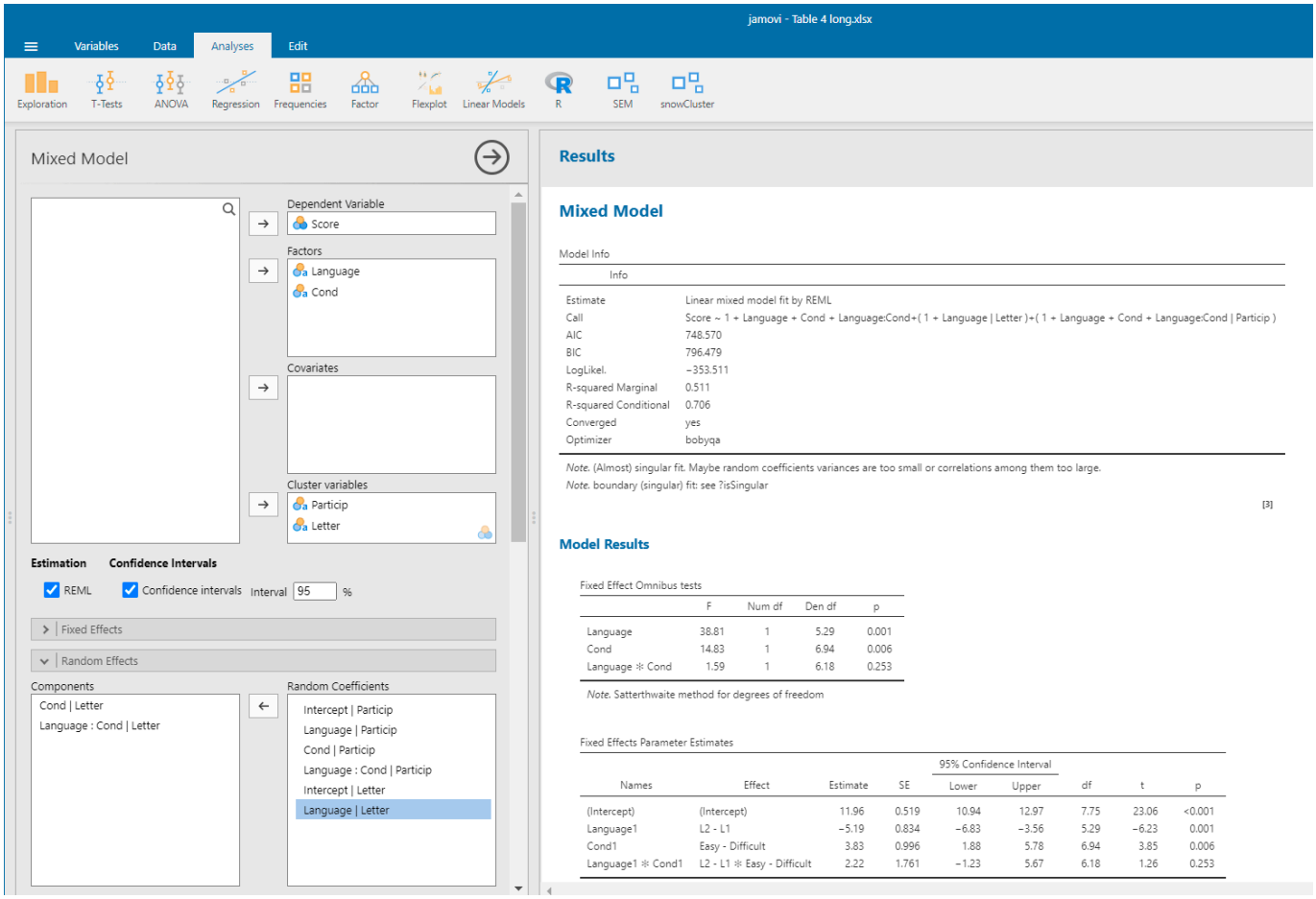

Figure 14. The outcome of the analysis of the data shown in Figure 12.

Figure 15 depicts a plot of the data. It shows that while there is a trend toward the expected interaction (the difference between difficult and easy is slightly larger in L2 than in L1), the evidence is not strong enough to be trustworthy. This will often be the conclusion for this type of experiment, since a small interaction between two variables (the effect is slightly larger in one language than in the other) requires at least four times as many participants as a main effect, so 200 at least (Brysbaert, 2019; Perugini et al., 2018).

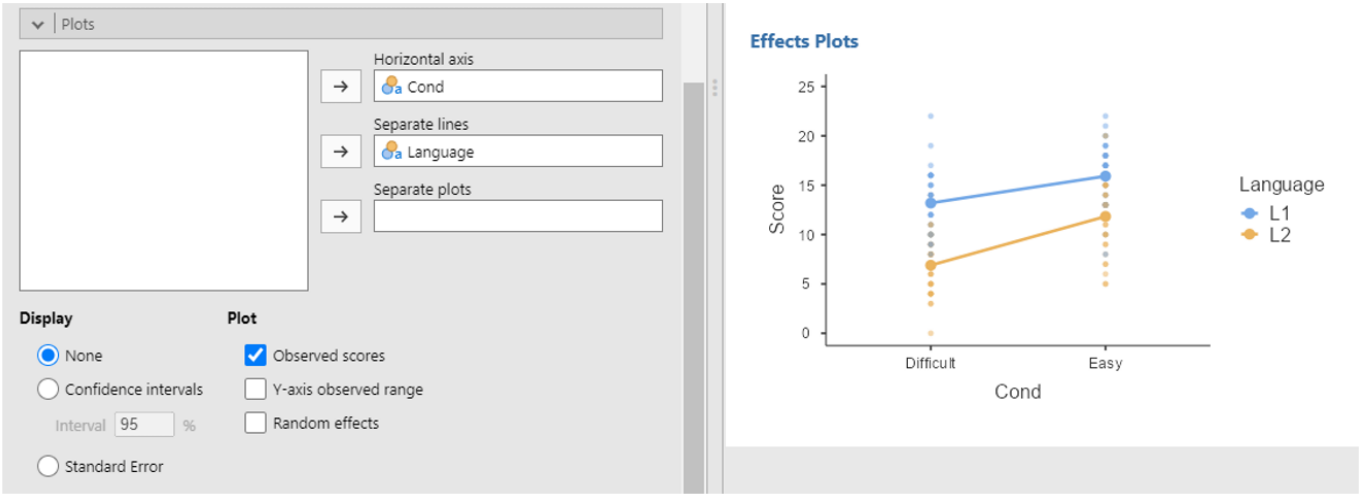

Figure 15. A plot summarizing the findings of the mixed-effects analysis of the dataset shown in Figure 12.

4. An Example with a Continuous Independent Variable

Mixed-effects analysis is also an interesting alternative (or actually extension) to linear regression. Suppose you want to investigate the concreteness effect in L2 word recognition. You do this by asking participants to name L2 words as quickly as possible.

For simplicity, we assume you had only 5 participants naming 20 L2 words that differed in concreteness as measured in a survey with a Likert scale of 1 (very abstract) to 5 (very concrete). Table 4 shows the results of your study.

Table 4. The data from a toy L2 word-naming study in which the effect of concreteness is investigated. The last column gives the mean response time (RT) of the five participants.

Stimulus	Concreteness	Part1	Part2	Part3	Part4	Part5	Mean
s1	5	372	365	393	398	321	370
s2	4.8	343	330	435	455	350	382
s3	4.6	342	409	467	354	367	388
s4	4.4	365	345	432	375	278	359
s5	4.2	441	348	503	288	404	397
s6	4	395	469	344	370	358	387
s7	3.8	327	451	341	400	416	387
s8	3.6	412	392	467	390	392	411
s9	3.4	337	272	513	377	324	365
s10	3.2	405	363	370	344	379	372
s11	3	385	376	485	362	450	412
s12	2.8	304	315	400	387	454	372
s13	2.6	412	439	454	372	445	424
s14	2.4	397	388	369	352	354	372
s15	2.2	402	318	442	409	416	398
s16	2	364	511	482	374	429	432
s17	1.8	411	370	383	399	461	405
s18	1.6	402	400	379	439	432	410
s19	1.4	457	431	399	313	389	398
s20	1.2	365	374	448	334	393	383

In a typical, multiple regression analysis, we take the average of the five participants (the last column of Table 4) and correlate that with word concreteness (the second column). We can easily do this in jamovi. We again use gamlj and select General Linear Model.

Figures 16 and 17 give the results and show that the effect of concreteness narrowly fails to be significant ($t(18) = -2.03$, $p = 0.058$).

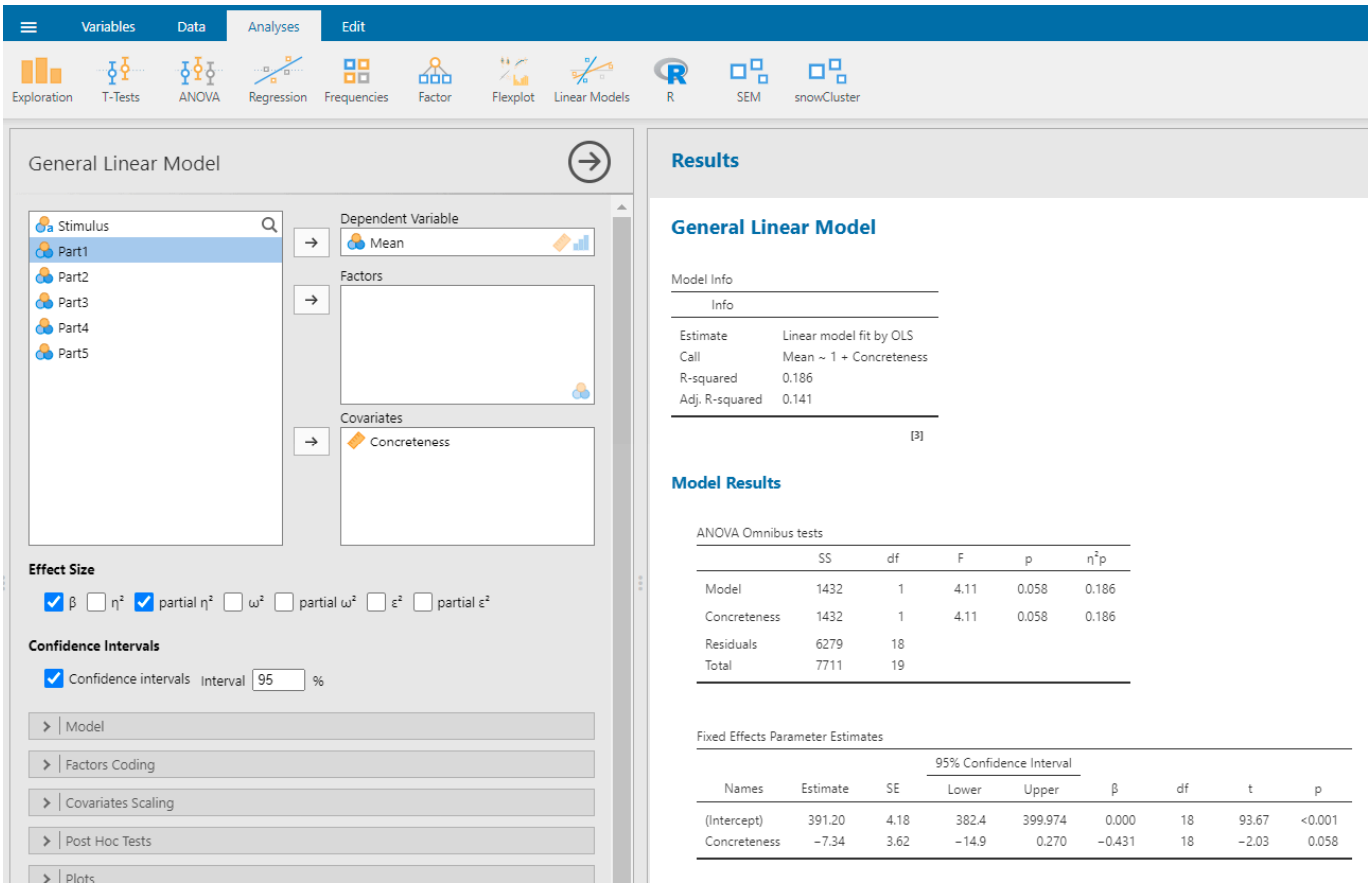

Figure 16. The outcome of a linear regression analysis for the mean data from Table 4.

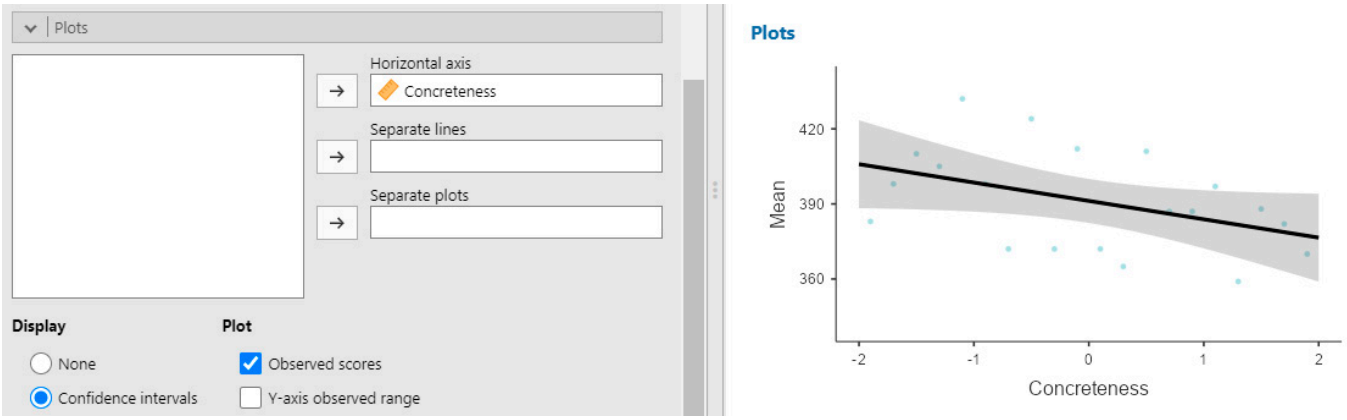

Figure 17. A plot of the effect of concreteness on mean word-naming latency.

A mainstream linear regression analysis is not bad (and indeed is used in many studies), but it does not actually allow you to generalize to a new group of participants. By averaging the effect across participants, you run the risk that the entire concreteness effect is caused by one participant (or a small percentage of participants). A better way to test the hypothesis is to see if the effect remains significant if participants are used as a random (cluster) variable. We can do this in gamlj with Mixed Models after converting Table 4 to long format (shown in Figure 18).

Participant	Stimulus	Concreteness	RT
P1	s1	5	372
P1	s2	4.8	343
P1	s3	4.6	342
P1	s4	4.4	365
P1	s5	4.2	441

Figure 18. The lay-out of the long format of Table 4, needed for a mixed-effects analysis (number of data lines = 5 × 20 = 100).

Figure 19 shows how we conduct the mixed-effects analysis. In the example, we only have one random variable, namely the participants, because each stimulus has a different concreteness value, so that we cannot disentangle the stimulus intercept from the concreteness effect (if try it in jamovi, you will see that there is no variance in the stimulus intercepts when you run the analysis with stimulus intercept as an additional cluster variable). Since all stimuli were seen by all participants, we need both random intercepts and random slopes for the participants.

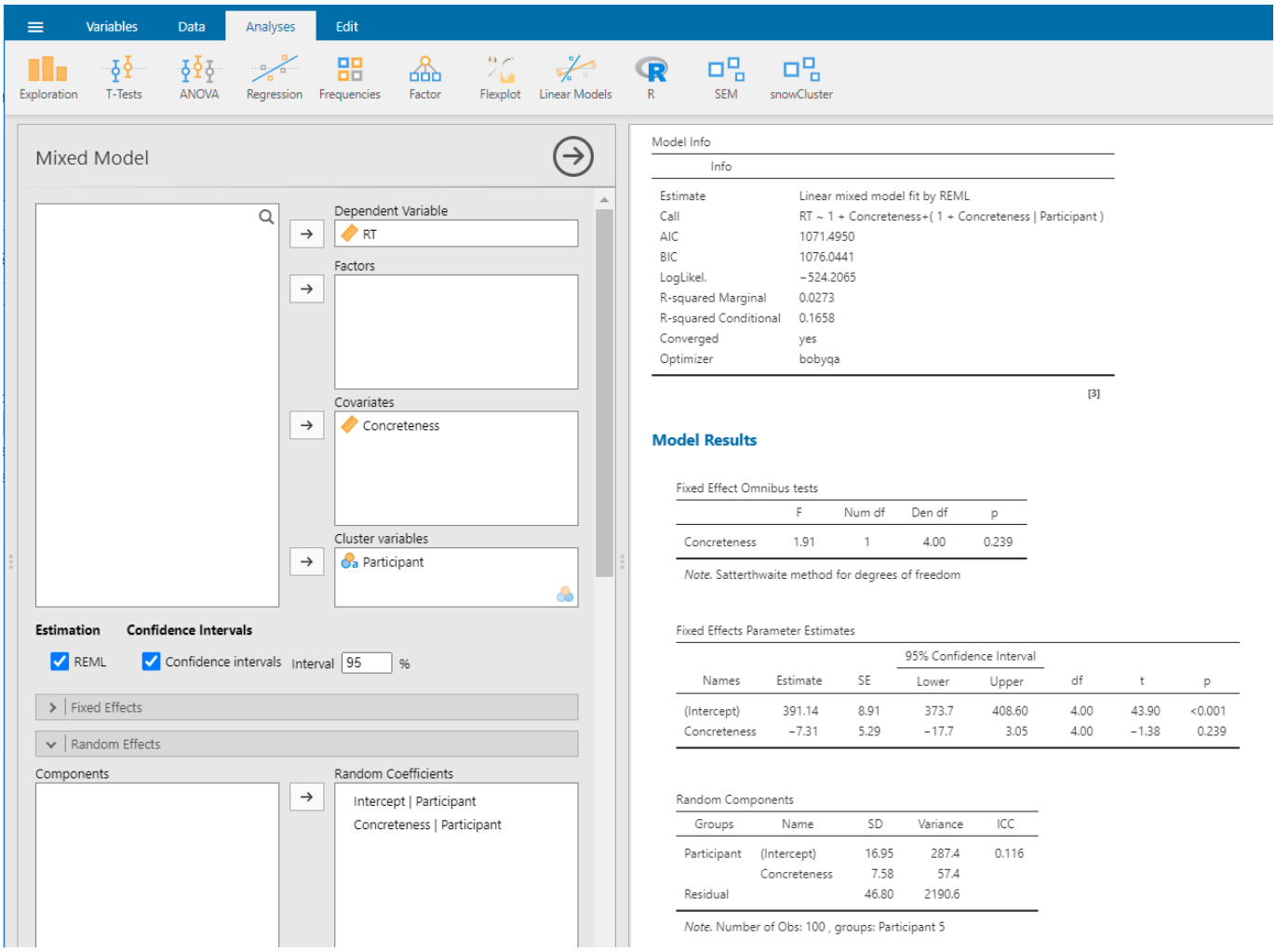

Figure 19. A mixed-effects analysis of the data from Table 4.

We see that the concreteness effect is now less certain ($t(4) = -1.8$, $p = 0.239$), because when we look at individual participants, only Participant 5 shows a clear concreteness effect. Participants 3 and 4 even show a small effect in the opposite direction. As such, the mixed-effects analysis protects us from drawing sweeping conclusions based on a pattern

present only in a small subset of the sample of participants. Figure 20 shows a plot of the results.

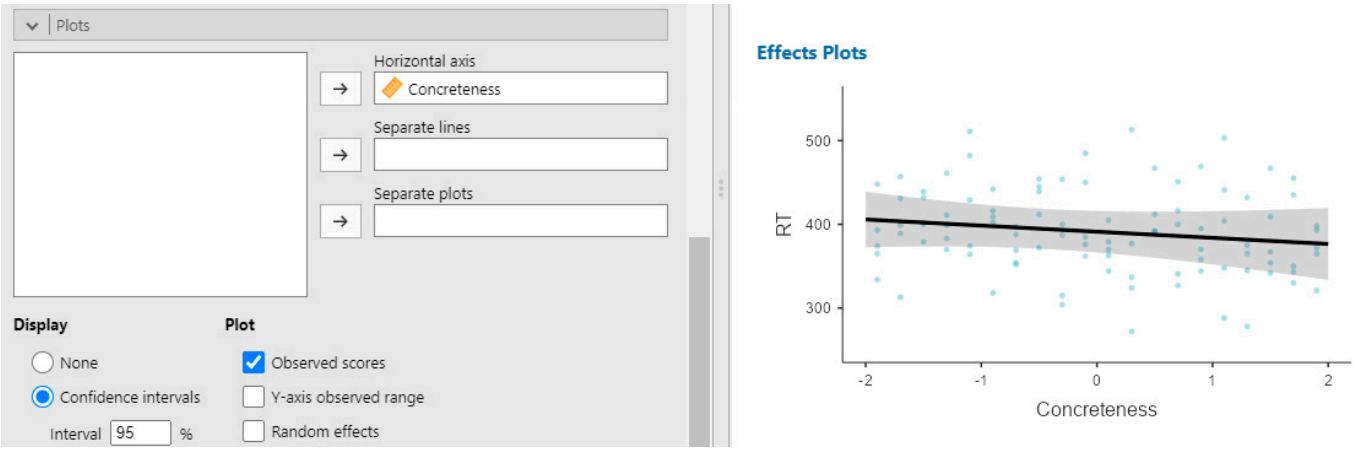

Figure 20. Plot of the findings of the mixed-effects analysis.

5. Discussion

This tutorial aims to make mixed-effects analysis available to novice researchers. It introduces a free statistical package (gamlj) that helps researchers avoid common mistakes (such as using dummy coding or forgetting to center variables that are part of an interaction). It uses sensible choices as default values and makes input and output as intuitive and simple as possible. Once the basics are mastered, researchers can explore advanced options (or see how to program the analysis in R).

Of course, a statistical package does not improve the data you start with. This remains the responsibility of the researcher. Since taking statistical power seriously, I have found that many of the problems we experience in our research are caused by datasets that are too small. Once you have a few 10 thousand observations, concerns such as outliers (Miller, 2023) and non-normal distributions (Burchill & Jaeger, 2024) become less of a problem. The effects of variables are obvious or are too small to be of practical importance (they can still be of theoretical importance if they relate to a critical prediction that would decide between two theories). Mixed-effects models are ideally suited for analyzing large databases. For such datasets, the long format is a blessing rather than a burden. A model works just as well (and even better) on a dataset of a million lines as it does on a dataset of 100 lines.

6. Further Reading

This tutorial is a hands-on version of Brysbaert and Debeer (2025). There, you find more information about the underlying mechanisms and the usual pitfalls in R (but not in gamlj). Other interesting articles are those by Brauer and Curtin (2018) and Brown (2020). Meteyard and Davies (2020) give a review of best practices in using mixed-effects models.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Notes

1 Notice that this is not data peeking, in which researchers look at the data and stop collecting data when the results are significant but continue when the findings fail to reach statistical significance. That is a bad research practice (John et al., 2012; Strube, 2006). What is meant here is that, once all data are collected, you look at the descriptive statistics before running statistical tests.

2 In my analyses, I often find it informative to also run simple *t*-tests and ANOVAs, in order to get a better feeling for the outcome of the mixed-effects analysis. If an effect is strong and reliable, you will find it in whatever analysis you carry out. Although

LME gives you the best analysis, looking at separate analyses across participants and stimuli often provides a useful additional understanding of what has been found.

References

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*, *26*(3), 295–314. [CrossRef] [PubMed]

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389–411. [CrossRef] [PubMed]

Brown, V. A. (2020). An introduction to linear mixed effects modeling in R. *PsyArxiv*. [CrossRef]

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), 16. [CrossRef]

Brysbaert, M. (2021). Power considerations in bilingualism research: Time to step up our game. *Bilingualism: Language and Cognition*, *24*(5), 813–818. [CrossRef]

Brysbaert, M., & Debeer, D. (2025). How to run linear mixed effects analysis for pairwise comparisons? A tutorial and a proposal for the calculation of standardized effect sizes. *Journal of Cognition*, *8*(1), 5. [CrossRef] [PubMed]

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, *1*(1), 9. [CrossRef] [PubMed]

Burchill, Z. J., & Jaeger, T. F. (2024). How reliable are standard reading time analyses? Hierarchical bootstrap reveals substantial power over-optimism and scale-dependent Type I error inflation. *Journal of Memory and Language*, *136*, 104494. [CrossRef]

Cinelli, C., Forney, A., & Pearl, J. (2024). A crash course in good and bad controls. *Sociological Methods & Research*, *53*(3), 1071–1104.

Cohen, J. (1992). Things I have learned (so far). In A. E. Kazdin (Ed.), *Methodological issues & strategies in clinical research* (pp. 315–333). American Psychological Association.

Gallucci, M. (2022). *gamlj/gamlj: GAMLj Suite for linear models* (Version 2.6.6). Available online: https://rdrr.io/github/gamlj/gamlj/ (accessed on 21 January 2025).

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. [CrossRef] [PubMed]

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. [CrossRef]

Kumle, L., Võ, M. L. H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*(6), 2528–2543. [CrossRef] [PubMed]

Langenberg, B., Janczyk, M., Koob, V., Kliegl, R., & Mayer, A. (2023). A tutorial on using the paired t test for power calculations in repeated measures ANOVA with interactions. *Behavior Research Methods*, *55*(5), 2467–2484. [CrossRef] [PubMed]

Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, *91*, 5–27. [CrossRef]

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092. [CrossRef]

Miller, J. (2023). Outlier exclusion procedures for reaction time analysis: The cures are generally worse than the disease. *Journal of Experimental Psychology: General*, *152*(11), 3189–3217. [CrossRef]

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, *31*(1), 20. [CrossRef]

Raaijmakers, J. G., Schrijnemakers, J. M., & 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. [CrossRef]

Strube, M. J. (2006). SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behavior Research Methods*, *38*, 24–27. [CrossRef]

The jamovi Project. (2022). *jamovi* (Version 2.3) [Computer software]. Available online: https://www.jamovi.org (accessed on 21 January 2025).

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. [CrossRef] [PubMed]

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*, 1–23. [CrossRef]
Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science*, *5*(2), 25152459221095823. [CrossRef]