Unit 3 Lesson: Using Reaction Time and Mixed Models

Christer Johansson and Per Olav Folgerø

This lesson will introduce some concepts related to **empirical** studies and **statistical evaluation**. The focus is on evaluating a specified model with controlled **fixed factors** and several **control variables** in the context that we have *one continuous dependent variable*, such as reaction time.

It should be noted that it is important to clearly *state the expectations* before collecting data and that we assume a **null hypothesis** of no difference for our fixed factors. If we observe a **significant** difference for any of our fixed factors the difference can be explained in many ways, but a **first assumption** is that we can *take a gamble* and claim that there is a **real** difference, given that the probability of observing a difference by random chance is appropriately small. If there is a real difference this real difference may warrant an **explanation**, or at least an **interpretation**.

One general explanation for faster reaction times, associated with a factor, is that somehow the mental processing leading up to the decision to press a button, for our subjects, is easier. We may then talk about **facilitation**, or **priming**, i.e., that the processing was prepared, for example if there is evidence that information was presented before the decision that made the decision more fluent, and thus easier and faster. We would like that accuracy is unaffected, or at least equally good, so that the results are not simply due to a tradeoff between speed and accuracy.

Similarly, the experimental paradigm might predict **interference** when a decision is taken based on competing resources drawing on similar resources. For

C. Johansson (✉) · P. O. Folgerø
University of Bergen, Bergen, Norway
e-mail: Christer.Johansson@uib.no

P. O. Folgerø
e-mail: Per.Folgero@uib.no

T. Balinisteanu and K. Priest (eds.), *Neuroaesthetics*,
https://doi.org/10.1007/978-3-031-42323-9_9

example, if there is a choice between two items, the more conflict or the closer the items are, the harder it is to take a decision. However, people can have different preferences in such situations. One possible strategy for making close decisions could be that if two items are equal, it matters less which choice is made, both are good, or both are bad. This could be observed in a tendency to make a default choice, such as pressing the most frequent choice, or pressing the right hand side button, for right handed people.

It is not always possible to determine what the correct decision was or was supposed to be. In those cases, we may want supporting evidence, for example showing relevant correlations with other factors, for example features of the input, and showing that those correlations are not random. Argumentation is needed, and the formal results are typically not clearly associated with one and only one possible interpretation.

We are often interested in the intuitive fast decisions that people make given the available information, possibly in situations where the difficulty of the choice is varied. Ideally, we would like to exhaustively contrast the possibilities, but we should recognize that this is not always possible, for example because time is limited.

The **dependent variable** is typically a measurement that we want to explain by, or relate to, *fixed independent variables* that we can control and vary within the experiment. We are interested in the degree that we can influence the dependent variable by changing the value of our fixed variables in a principled way. We need the dependent variable to be a preferentially continuous variable on an interval scale, which means that we know that a change of one unit is worth the same wherever we are. One example is reaction time. One millisecond is the same time interval wherever we are on the scale, and wherever we are in the world. If the measurement also includes an absolute zero, meaning that 0 means absolutely nothing of whatever we measure we may also say something about proportions. The reaction times were twice as long, assume that there is such a thing as an absence of reaction time, simultaneous reaction to stimulus. However, it takes time from the presentation of the stimulus to the decision and the reaction. The real starting point is thus sometime after the stimulus has been presented, depending on how long it takes to perceive the stimulus. In real life, human reaction times cannot realistically be much smaller than 200ms, as it takes time to process that which we should react to, and it takes time to send the signal from our brain to the finger to activate the muscle that should press the button.

The **independent variable**s are the variables that we manipulate, or control, to affect the dependent variable. These are the **fixed effects** that we want to investigate. Fixed effects have a limited and exhaustive number of levels, and ideally, we should have included all those in our experiment. Example of a **condition** related to a fixed variable is if the **target stimulus** that we should react to is **primed** or not. Generally, we like conditions to be two levels, either primed or not primed, as it is easier to interpret the statistical model if this is the case. In the **experimental design**, we like to balance the conditions we are interested in, for example such that we have an equal number of primed and unprimed **events**. In this case, the unprimed is our **baseline**, and in the analysis, we could name the levels such that the **baseline** is included in the **intercept**, i.e., in the starting point of our regression equation.

The **control variables** are variables that we know, or suspect will affect the dependent variable, but these variables are not necessarily planned factors that we are interested in. Since they affect the dependent variable, without being a properly controlled independent variable (a fixed effect), we would like to regress out their effect. Examples of such variables are the **learning effect**, the effect of **hesitation**, the effect of **making an error** in the previous event, the effect of **exceptionally slow or fast responses** (positive and negative outliers), and the effect of having such an outlier in the previous response. Other examples are to control for the exact onset time of the stimuli we are interested in and to control the linear distance between say a **prime** and a **target**. A prime is an item that is presented before the target and is thought to affect the response to the target. The prime and the target could be placed at slightly different distances. With **continuous temporal stimuli**, like a speech signal, we can measure the time between the onset of the prime and the target. In a **discrete design**, it could be the number of items (typically written words) between them. Those effects could also be related to the conditions we are interested in. For example, it could be that the learning effect is larger for one of the conditions. This can be controlled by adding the **estimation of a slope** for each condition. A slope is the change in the response that depends on the condition compared to a baseline.

The aim of the model is to explain as much of the variance in the data as possible. Our tool for creating such models is **linear regression**, and **mixed effects models with random effects** in particular. In a linear regression, we try to estimate a baseline, the **intercept**, and estimate the change that our variables will have. For example, what is the effect of seeing a prime word compared to no prime word? The learning effect, and such control variables, can be **fitted regression lines** that depend on a numerical value to calculate the effect on the response variable, i.e., the dependent variable. Once the regression line is fitted it accounts for some variance, and thus the *estimates for the controlled fixed effects will become more precise*, as the effect of the control variables has been accounted for. In real experiments, there will be correlations between the variables that can be difficult to entangle, but the model makes a principled attempt at separating the sources of variance in the data.

There obviously need to be fairly large amounts of data, to estimate the effects of all the included variables. The restriction to a linear model helps. Linear equations have the property that adding the estimation of a factor (i.e., adding the estimation of a line) will result in a new linear equation, which can be solved if there is enough independent data.

Thus, to be able to estimate a line, for each subject, *we need at least two points.* It is a problem that subjects may have no responses which could potentially make it impossible to estimate lines. Therefore, we need to have more data points from each condition. A common **rule of thumb** is to have at least four data points in each of the conditions. For example, if we have a 2 by 2 design (say, two levels of prime and two levels of targets), we minimally need 4*2*2 (i.e., 16) **balanced** data points from each subject. Ideally, we aim for an equal number of data points in each condition, so to keep the balance in this example, increments will be in steps of four (2 by 2).

The data should ideally be balanced for other factors too, for example if we plan to use **gender** as an explanatory factor, we will need roughly the same number of each gender in our sample of subjects. Another common **rule of thumb** is that we need at least a thousand data points in total, assuming a small number of fixed effects. From this we can **estimate the number of subjects** needed in our study. If we increase the number of data points per subject, we will be able to better control the influence of individual variance, and if we do this by including more diverse test items, we will also better control the variance that is due to test items. In our example, the minimal number of subjects would thus be 64, as 1000 / 16 = 62.4, and if we balanced the genders we would need 32 of each gender. Here, we ignore the problems that arise from balancing the gender that the subjects might identify with. Some of that problem is handled by including a random factor for individual variance, rather than expanding the gender variable.

We need to consider *how much data we can collect from each subject*. In reaction time experiments, we can expect a subject to be concentrated for up to 20 minutes, and allowing 10 seconds for each test item (including presentation and reaction) would set a limit of 120 items per subject. Your experiment may have different demands. In the example, this would allow for 30 test items in a total of four conditions. Following the logic above, this would mean that minimally we need 10 subjects with 120 data points from each subject (1000 / 120 = 8.33, and the nearest higher even number is 10).

It is recommended to sample more than the minimum and to balance subject and item demands. If we go for 20 subjects and 16 test items in four conditions, we end up with 1280 data points in total, which is *close to the minimal demand*. We should also anticipate that not all subjects can be included. We may expect between 5 and 10 percent of the subjects to be excluded as outliers. That would mean adding 2 extra subjects to compensate. However, recall that if for some reason more outliers are detected we will get an underpowered study. If possible, we should add at least 25 percent of subjects, which will result in 20 + 5 subjects. Recall that it is *difficult to add more items*, as that would likely exceed the limit for fatigue in the subjects. If we discover that the planned number of subjects is too large to be realistic, we might consider repeating the study to sample more data from the same subjects. This might also allow us to test for the effect of repeating the study. In a longitudinal study, there is always a risk that the subjects will not return to the second session.

Ideally, we would like our study to generalize to the full population. However, this would mean that we would need to actively include a balanced sample from all the relevant sub-populations. A common choice is therefore to focus on a sub-population. This sub-population is typically a **convenience sample**. At a university, students between 18 and 32 might be a convenient sample. This limits generalization, but it also helps to control variance. University students are a pre-selected sample and may represent not only the young in the population, but also those with an interest in study and related to that better reading ability, and possibly better working memory capacity and reasoning skills. This may also result in better compliance with instructions. If the subjects in turn may help with recruiting new subjects, we may talk about a **snowball sample** approach. This can be actively used to test the effects of social networks, but often the recruiting strategy is just reported in the resulting article and is up to the external reviewers to evaluate if it is a reasonable choice. There is no such thing as an optimal research design, and limits on available resources are often an important part of the research. This will also allow other researchers to repeat the research under similar or other conditions. The proof of the pudding is in the tasting, and for experiments we are interested in how much variance is explained and in how easy it is to repeat the results. In that process, we will find out more details, and some of those details may be more important than the original study.

A last point, before moving on to another example. A **laboratory-controlled experiment** may often be unnatural, which in turn may affect the relevance of the study. For example, we may focus on reaction time for decision tasks, but such tasks might be only weakly related to the phenomena we study. Sometimes it is argued that **ecological validity** should be valued as well. That is, how well the task matches with what people are doing in the wild, outside of the laboratory. Control and ecological validity are often in opposition, but ecological validity is not only about bad control of variables, but also about allowing other sources of variance to estimate the true variance more closely. Ideally, different experiments may complement each other, or at least raise important and interesting questions.

Let us investigate a potential experiment, comparing faces in a hot-or-not contest, except that we will ask participants to pick the face that is either more attractive or separately the face that is more trustworthy. The faces are constructed from real faces morphed with a Jesus prototype. Each face can be labeled for the male or female substrate. Will people be able to detect which faces are male and which are female? Will the results be different between the attractiveness condition and the trustworthy condition. The dependent variable is the reaction time, i.e., the time it takes to make a choice.

Each face has slightly different facial features that can be measured in the pictures. We will look at features that indicate left/right symmetry in the face, the eyes, and the mouth. There is well-known research (the so-called Thatcher effect) that established that we typically judge eyes and mouth separately. We will also look at features such as how wide the eyes are apart, and how wide the mouth is. Furthermore, we will look at the proportion of the face in the vertical direction: how large is the forehead, the mid-section, and the chin. These measurements will be continuous measurements of proportions, which makes them scale-free, i.e., there is no measurement unit just proportions that can be compared regardless of how large the face is. All faces will be approximately the same size as they are presented on screen.

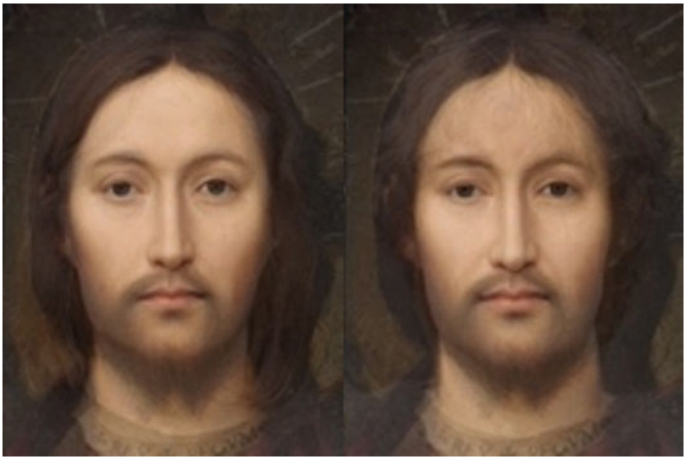


Fig. 1 Female and male competition: Attractiveness (female winner left, male winner right)

We will control the order of presentation. We expect the participants to make increasingly faster decisions as they get more familiar with the task. This learning effect can be handled by regression analysis, after we have made the order of presentation explicit in our data. The typical experimental program will give the order of presentation implicitly but will not create a variable for the order of presentation.

We will use the gender of the two pictures presented in pairs on the screen. The competition could be female–male (two directions), female–female, and male–male. This may interact with the gender of the selected picture, and the gender of the subject.

It might also interact with the gender preferences of the participants. This will be handled as a random effect that may explain more variance.

An example is given below (Figs. 1 and 2). Would you pick the same in each pair if it was a choice of attractiveness or a choice of trust? Can you identify the male and female substrates?

A Model of Face Proportions

The line K–L is the baseline for the horizontal dimension. This reference line is found by estimating the position of the zygomaticus bone, using information also from the ear lobes and the nose tip. The line G–H is the baseline in the vertical direction. This line is estimated from the highest point of the forehead to the lowest point of the chin, following the nose and through the philtrum. The point I is on the G_L and is the reference point for eye symmetry. The point J is similarly the reference point for mouth symmetry.

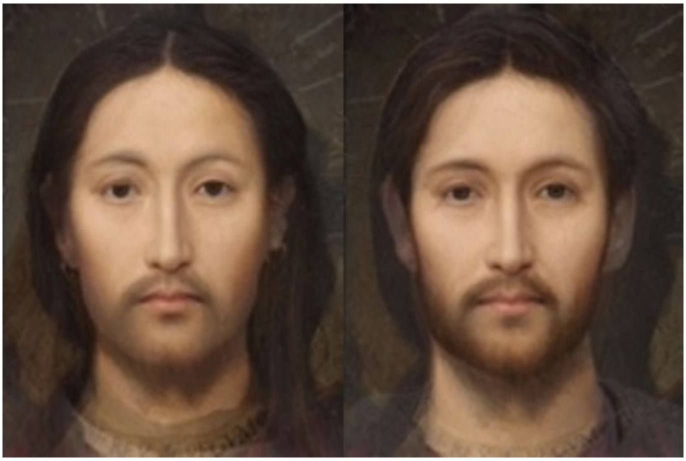


Fig. 2 Female and male competition: Trustworthiness (female winner left, male winner right)

The forehead is estimated by the area of triangle CGD compared to the larger triangle KGD.

The chin is estimated by the area of the triangle EHF similarly compared to KGD.

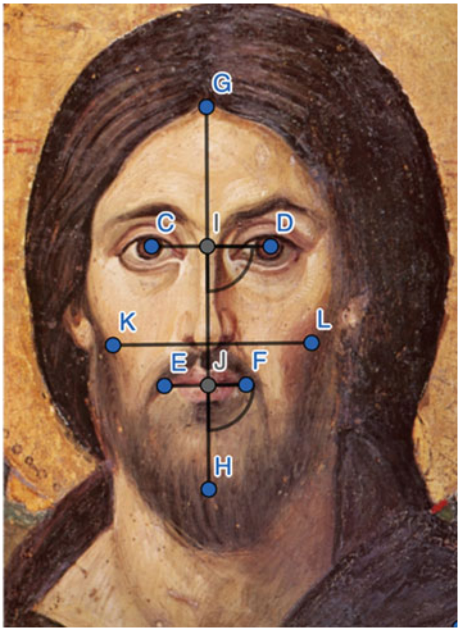
The mid-section is estimated by the area of the polygon KCDL compared to the larger polygon KGLH. Furthermore, the slant of the eyeline is estimated as the lower angle between CD and GH, and similarly for the slant of the mouth line, as marked in Fig. 3.

The face's right side is estimated by the area of the triangle GKH, and its left side is by GLH.

If GKH divided by GLH is larger than 1 the face's right side is larger, if smaller than one its left side is larger, and perfect symmetry is at a ratio of 1.

To visualize multi-factorial data of the same kind (e.g., proportions) we can use Correspondence Analysis (CA), cf. Glynn, 2014. This is just to mention the possibility. CA often gives a quick and intuitive overview of a dataset, as it projects a multitude of factors into a 2D plane. The points that are closer to the origin (0,0) are more as expected. The further from the origin the more distinct, and points that are close are more similar, but we should also value the angle of the line toward the origin. Points that are both close and have close angles are more the same. The axes of the CA graph are often possible to interpret, using the most extreme points along the x and y axes. In Fig. 4 we see an association from mouth width to mouth symmetry along the x axis, and from eye width (and proportion of mid-section, associated with eyes) to eye symmetry along the y axis. This has been detected by the algorithm from a limited set of investigated faces (marked in red, row points) and their anatomical proportions (marked in blue, anatomical column points). The clusters indicate prototypical faces (female (F#), male (MP#), human (H#)), Romanian Faces (X#), Men (M#), and Women(W#).

Fig. 3 Face proportions



Outlier Analysis

There are typically three types of outliers. The first is an analysis of the performance of the subjects. The subjects could solve the task differently. Given that there is a task that could be evaluated to be correct or not, it is possible to see if a subject has chosen more of the "incorrect" than other participants. One way to do this is to perform an association analysis. This is based on a cross-table analysis of the distribution of correct and incorrect answers for each subject. In each of the cells it is possible to calculate the **Pearson residual** (proportional to the contribution toward significance, **cf.** Cohen, 1980; Friendly, 1992) and the support in that cell, which is proportional to the number of observations in that cell. This type of analysis may objectively reveal if some subjects have solved the task in a different manner than other subjects. We do not know if their way of solving the task is better or worse, but we do know that it is likely to be different. For example, some subjects might always choose one of the decision buttons (for example, the left one). For some tasks it is not possible to know which decision is correct, but we may still see if the participants give similar answers and figure out if some of the participants have solved the task using some unintended strategy. We may also analyze the reaction times of each participant. Some people are faster than others, and this can be handled by having different **intercepts** (starting points) for each subject. However, it could also be that some subjects are faster because they solve the task in a different way, for example always pressing one of the buttons. In the analysis, we should pay attention to both the **trend** (e.g., the average time) and the

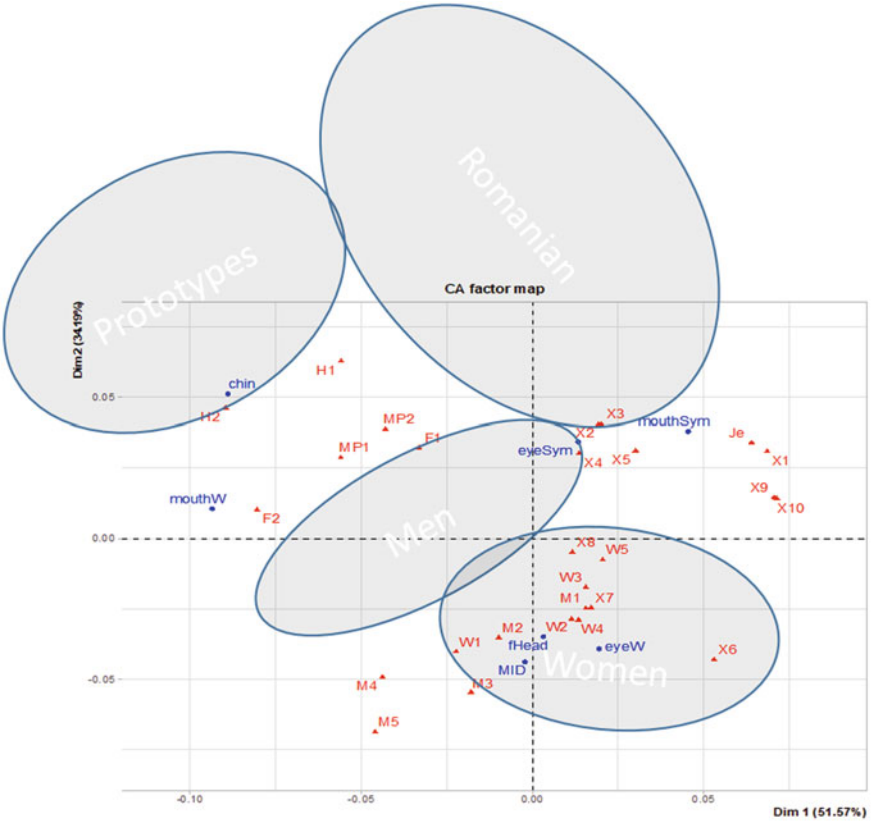


Fig. 4 CA map

variance (standard deviation). A significantly lower standard deviation may be an indicator, especially if this is combined with a low performance on the accuracy of the task. A significantly higher standard deviation might be an indicator that the subject has been unsure of the task or possibly been inattentive. It could also be that they have a stronger reaction, and this might be interesting.

In the item analysis, we are interested in the representativeness of our test items, i.e., the items that the participants are supposed to react to. Some items may be more interesting, or easier or harder to process, etc. This could be handled by different **intercepts** (starting points) for each item in the random effects analysis. It could also be that some items should be marked as outliers, and possibly removed from the data.

In the analysis of the data, we should look at those reactions that are outside of a **confidence interval** for the data. Those items that are slower than the trend plus two standard deviations are typically marked as outliers, as well as those items that are faster than the trend minus two standard deviations. We should also mark those items that are unrealistically fast. Items that are faster than 200ms are so fast that it is unrealistic that there was time enough to take a decision and press a button. In the analysis of data errors there are also some no responses (NR) that typically are marked with a reaction time of 0. No reactions are typically removed from the dataset, as no data was given by the participant. However, it is possible that there is a signal if the no responses are systematic in some way, for example, that they are more common for a participant or a test item. If this is the case, we may consider removing the participant or the test item from the dataset.

We would like to avoid removing data for many reasons. If we remove participants, it must be declared when we write up an article, as it affects how well the experiment may generalize. However, it is often the preferred choice, even if we lose many data points we may increase the quality of the data points. For items, we may consider estimating the contribution of the items in the analysis, for example as a random intercept. We may also mark the items in an item factor, say "extreme_item", as expected, too fast, or too slow, and try to regress out their effect on the trends For data we might see if it is better to mark the data points and regress out their effect. For example, we may have a factor "extreme" that tells if the data is as **expected** (i.e., within the confidence interval), too **fast**, or too **slow**. This has the benefit that we can keep more data, without letting the extreme values affect the trends disproportionately.

The Format of a Mixed Effects Model

A mixed effects model using linear regression is a model-based evaluation of an experiment. We can build up a model using the factors we suspect will have an effect, the correlations we think will play a role, and the sources of variance we know about. One philosophy is to start with a maximal model and reduce that model as necessary. Another philosophy is to start with a minimal model and add factors if they improve the model fit. Here we will suggest starting with a large model with interactions and looking at model fit. We may have to reduce the model if the model fails to converge (i.e., the model cannot be solved using the available data).

In R the format for the models can be stated as:

$$\begin{aligned}
&\textbf{dependent}\quad \textbf{factor1} * \textbf{factor2} * \textbf{factor3} * \textbf{factor4} + \\
&+ \textbf{control1} + \textbf{control2} + \textbf{control3} + \textbf{control4} + \textbf{control5} + \textbf{control6} + \textbf{control7} + \\
&+ \textbf{outlier} + \\
&+ (1|\textbf{Participant}) + (1|\textbf{Item})
\end{aligned}$$

The measurement variable (**dependent**) is explained by (~) the fixed factors 1…4, the control variables 1 … 5, the starting points of each Participant and Item. Testing for an interaction effect is noted by the use of a "*". An interaction between two items is when their independent values do not add up when they are together, at the same time. A common day example, if you buy eggs and bacon together you may pay less than if you buy them together because you have a rebate coupon when buying them together. A four way interaction between four factors as indicated above, relabeled for brevity as a, b, c, and d, will give rise to more than three interaction terms, in fact there will be four main effects (a, b, c, d) and 11 interactions (ab, ac, ad, bc, bd, cd, abc, abd, acd, bcd, abcd). This is obviously very costly to estimate. We can decide to only consider the main effects by using a different operator between the factors, i.e., a "+". It is often not possible to estimate the interaction between everything, because of lack of data and because of the correlation structure in the dataset. The inclusion of an interaction is useful if it explains data better, and models with and without an interaction can be compared using analysis of variance, as outlined by Baayen and Milin (2010).

The last two items in the formula above are called **random factors** that estimate the variance between the Participants and the Items. The Items are the controlled stimuli that the Participants will decide. An item is here thought of as an event, i.e., a test item presented in a specific context. It is also possible to structure the Items such that an item is presented in many different contexts and the formula will then be represented as (context | Item_name). Here we will prefer the notation with a different intercept for each presented stimulus (1|Item). The random effects are used to estimate the variance due to Participants and Items.

The fixed factors in the example experiment that we described earlier could be: **SelectedGender**, **CompetingGender**, **ParticipantGender**, and **Condition**.

Condition is whether the task is related to **trust** or **beauty**.

The control variables could be **PresentationOrder**, **FaceSymmetrySelected**, **MouthSymmetrySelected**, **MouthWidthSelected**, **EyeSymmetrySelected**, **EyeWidthSelected**, **MIDproportionSelected,** and **outlier**.

The presentation order is a measure of *learning*. The more items the participant has seen the faster the responses. This may be because the participant has learned something about the task and the items used in the testing, and/or has become more confident. Typical effects are about 5 ± 1ms faster per item seen, depending on the task.

The various measures of symmetry and relative width of eyes and mouth are anatomical scale-less proportions related to attractiveness. Attractiveness may lead to faster responses. However, one hypothesis is that the effects would be different for trust and beauty. Rather than having one very large model that investigates the interaction between all the control variables and the experimental condition, we might test trust and beauty separately in two similar models, to avoid very complicated interaction terms. The **outlier** factor marks if the data point is as **expected** inside the confidence interval, or if it is **faster** or **slower**.

Evaluating the Model

The model generated by the mixed effects analysis can be evaluated using an analysis of variance on the model. This will tell us which factors show significant differences and which correlation lines show significant trends. The call in

Table 1 The analysis of the model, in a table from analysis of variance

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)	
factor1	#	#	1	45.04	9.1539	0.0041	**
factor2	#	#	1	173.84	15.5468	0.0001	***
control1	#	#	1	1585.65	255.67	1E-16	***
factor1 × factor2	#	#	1	45.02	0.6693	0.4176	

Type III Analysis of Variance Table with Satterthwaite's method

R is simply **anova**(model). To get the full model, including the effects, we use a different call: **summary**(model).

As an example, consider a simpler model below.

$$> \text{anova(model)}$$

The sum squares can be useful and should be reported in a full table, as in Table 1. Here we focus on the information that is often stated in an article text. Here, **factor1** is significant (F(1, 45.04) = 9.15; p = 0.0041) and **factor2** is significant (F(1, 173.84) = 15.55; p = 0.0001). The regression line associated with **control1** is significant (F(1, 1585.65) = 255.67, p = 0.0000). The so-called *scientific notation* 1E-16 denotes that 1 is 16 decimals behind the decimal point, which is a very small number. There is no significant interaction between the two factors.

NumDF denotes the degrees of freedom between (which is a measure of the useful contrasts, here there are only two levels and thus one contrast). DenDF denotes the degrees of freedom within, which is a measure of the number of independent data points that was used—this is estimated mathematically, and it includes using correlations rather than the typical paired "*repeated measures*" structure. Satterthwaite's method is a reference to how this number has been estimated. Thus within degrees of freedom is a measure of how many independent data points were used to arrive at the estimate of significance. The F-value quantifies the deviance from expectations, and together with the between and within degrees of freedom it is possible to arrive at a p-value that we can use to take a decision for what constitutes a significant difference between the levels for the factor. The p-value is the probability of observing an F-value larger than the observed F-value, given the between and within degrees of freedom (i.e., the number of contrasts, and the size of the experiment).

If we want to see the resulting effects, we need to look at the summary of the model. Below, we will focus on the fixed effects (including the control variables). The correlation matrix and the random effects are typically not as "important" but should typically be included for reference in an appendix.

In Table 2, we start with the intercept. The intercept is the value associated with level1 of factor1 and factor2 and the starting point (0) of the regression line in control1. This value is here 1519.65. The values for the various factor levels must be calculated from the **offsets** in the table. Note that for the result of the

Table 2 Table of effects

	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	1519.65	442.12	46.8	3.437	0.0011	**
factor1Level2	184.88	66.85	44.9	2.766	0.0082	**
factor2Level2	−333.58	105.87	157.6	−3.151	0.0019	**
control1	−6.51	0.77	1576.1	−16.66	1E-16	***
factor1Level2 × factor2Level2	−76.45	93.45	45.02	−0.82	0.4176	

Table 3 Calculating the value for each combination of factor levels

factor1	factor2		
level1	level1	1519.65	
level2	level1	1519.65 **+ 184.88**	1704.53
level1	level2	1519.65 **− 333.58**	1186.07
level2	level2	1519.65 + 184.88 − 333.58 **− 76.45**	1294.50

interaction effect we first do the pure additive effects and then the effect due to interaction (see Table 3). For the regression line associated with control1 we get 6.51ms faster for each item, i.e., 1519.65 -6.51*control1 is the reaction time after control1 number of presentations. If we want to correct the dependent (reaction times) for this controlled "learning" effect we can cancel out the effect with a simple calculation (RT + 6.51*control1) as the corrected reaction time. This may be necessary for generating more accurate graphs.

Common graphs to illustrate the results include boxplots and interaction plots, and sometimes so-called lattice plots to investigate the interaction of more than two factors. Typically, these graphs are performed using the raw data, but if we have significant effects of control variables it might be worth considering correcting the dependent variable. This is easy to do for control variables that are main effect correlation, i.e., general regression lines that are not dependent on the combination of other factors.

Interpreting the Results

The formal analysis of our model will tell us which factors have a significant impact on the dependent variable. The word "significant" is used in the statistical sense and does not necessarily mean that this is an important difference. It just means that the impact is difficult to explain by random uncertainty in the data. However, variance and uncertainty can have causes that are not controlled in the experiment, and the causal structure might be different from the assumption in our model. There is always a chance r risk that other factors could explain the data better. We might have attributed a tentative causation where there is no causation, and what looks like a causal relation might just be a correlation.

When we have the results from the analysis it is therefore important that we interpret the results, and relate the results to our hypotheses, and other relevant research if such research exists. It is very often the case that we discover more detail and other predictions that can be tested when we explicitly argue for the interpretation of our findings. A model criticism might inform us of how well the model explains the data, and how much variance is still unexplained.

Summary

This lesson has introduced Mixed Effects models for evaluating experiments. First some vocabulary was introduced, and a brief introduction of some concerns when we plan an experiment. We need to know how many participants we need and how many items per participant we need. As it is likely that we will have outliers for participants, items, and data points, we would rather oversample. We introduced the use of control variables and some principles for outlier analysis, and finally how to interpret and report the result of the analysis.

Below is a short literature list. The presentation has assumed the availability of the **R Statistical Software**, the **lmerTest** package for *Linear Mixed Effects Models*, the **vcd** package for association graphs, and the package **FactoMiner** for *Correspondance Analysis*. It is possible to use other software to implement the formal analysis. R and all the mentioned packages are currently widely available for free download.

Literatures

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12–28. https://doi.org/10.21500/20112084.807

Cohen, A. (1980). On the graphical display of the significant components in a two-way contingency table. *Communications in Statistics—Theory and Methods, A9*, 1025–1041.

Friendly, M. (1992). Graphical methods for categorical data. *SAS User Group International Conference Proceedings, 17*, 190–200. http://www.math.yorku.ca/SCS/sugi/sugi17-paper.html

Glynn, D. (2014). Correspondence analysis: Exploring data and identifying patterns. In *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (D. Glynn & J. A. Robinson Ed., pp. 443–485). John Benjamins Publishing Company. http://digital.casalini.it/9789027270337

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An *R* package for multivariate analysis. *Journal of Statistical Software, 25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Meyer, D., Zeileis, A., & Hornik, K. (2006). The strucplot framework: Visualizing multi-way Contingency tables with vcd. *Journal of Statistical Software, 17*(3), 1–48. https://doi.org/10.18637/jss.v017.i03

Meyer, D., Zeileis, A., & Hornik, K. (2021). *vcd: Visualizing categorical data.* R package version 1.4-9.

R Core Team. (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Zeileis, A., Meyer, D., & Hornik, K. (2007). Residual-based shadings for visualizing (conditional) independence. *Journal of Computational and Graphical Statistics, 16*(3), 507–525.