

Psychological data analysis for graduate students

Rob Davies

2023-02-08

Table of contents

Preface	3
A change in approach	3
I MAKING THE MOST OF R	5
1 R knowledge	6
2 Data visualization	7
2.1 Aims	7
2.2 Why data visualization matters	8
2.3 Three kinds of honesty	8
2.4 Our approach: tidyverse	9
2.4.1 Grammar of graphics	10
2.4.2 Pipes	10
2.5 Key ideas	12
2.5.1 Goals	12
2.5.2 Psychological science of data visualization	13
2.6 A quick start	15
2.6.1 Sleepstudy data	15
2.6.2 Discovery and communication	16
2.6.3 Discovery and communication	18
2.6.4 Summary: Quick start lessons	24
2.7 A practical guide to visualization ideas	24
2.7.1 Set up for coding	25
2.7.2 Information about the Ricketts study and the datasets	26
2.7.3 Read the data into R	30
2.7.4 Process the data	32
2.7.5 Visualize the data: introduction	37
2.7.6 Examine the distributions of numeric variables	37
2.7.7 Comparing the distributions of numeric variables	41
2.7.8 Summary: Visualizing distributions	58
2.7.9 Examine the associations between numeric variables	59
2.7.10 Answering a scientific question: Visualize the effects of experimental conditions	72

2.7.11	Summary: Visualizing associations	81
2.8	Next steps for development	82
2.9	Helpful resources	82
2.9.1	Some helpful websites	82
2.9.2	Some helpful books	83

II MODELS **84**

3	Introduction to multilevel data	85
3.1	Motivations	85
3.1.1	A word about names	87
3.2	Challenges	87
3.3	The key idea to get us started	88
3.4	The approach we take	88
3.5	Targets	88
3.6	Study guide	89
3.7	The data we will work with: Brazilian school children	89
3.7.1	Locate and download the data file	90
3.7.2	Read in the data file using the <code>load()</code> function	90
3.7.3	Inspect the data	91
3.8	Tidy the data	91
3.8.1	Select the variables	92
3.8.2	Remove missing values	93
3.8.3	Getting R to treat a variable as an object of the type required using the <code>as...()</code> family of functions	94
3.9	Introduction to thinking about multilevel models	96
3.9.1	Main ideas – Phenomena and data sets in the social sciences often have a multilevel structure	96
3.9.2	Multilevel models – why they are more used and more useful than traditional methods	97
3.9.3	Practical applications: children sampled within classes	98
3.9.4	To understand the application of multilevel models: first, we ignore the multilevel structure in the data	99
3.9.5	A linear model ignores the multilevel structure in the data	102
3.9.6	Notation	104
3.9.7	Can we really ignore the multilevel structure?	105
3.9.8	Linear models for multilevel data – dealing with the hierarchical structure	108
3.9.9	Two-step or slopes-as-outcomes linear models as approximations to the Linear Mixed-effects or Multilevel modeling approach	108
3.9.10	Multilevel models	110
3.9.11	How should we think about the differences between the classes?	111
3.9.12	Fitting a multilevel model using the <code>lmer()</code> function	113

3.9.13	Reading the <i>lmer</i> results	115
3.9.14	Fixed and random effects	116
3.9.15	Is there a difference between linear model and linear mixed-effects model results?	117
3.10	Conclusions	117
3.10.1	Summary	118
3.10.2	Glossary: useful functions	118
3.11	Recommended reading	119
4	Introduction to linear mixed-effects models	120
4.1	Motivations: repeated measures designs and crossed random effects	120
4.2	The key idea to get us started	120
4.3	Targets	121
4.4	Study guide	121
4.5	The data we will work with: CP reading study	121
4.5.1	Our research question	122
4.5.2	The challenges of working with real (untidy) experimental data	122
4.5.3	Locate and download the data files	124
4.6	Tidy the data	125
4.6.1	Read in the data files by using the <code>read_csv</code> and <code>read_tsv</code> functions . .	126
4.6.2	Reshape the data from wide to long using the <code>gather()</code> function	128
4.6.3	Merging data from different data-sets using <code>_join()</code>	133
4.6.4	Select or transform the variables	136
4.6.5	Filter observations	137
4.6.6	Now we have some tidy data	140
4.6.7	We can output the data as a <code>.csv</code> file	140
4.6.8	Data tidying – conclusions	141
4.7	Repeated measures designs and crossed random effects	142
4.8	Working with mixed-effects models	143
4.8.1	Load the data if you need to	143
4.8.2	Linear model for multilevel data – ignoring the hierarchical structure . .	143
4.8.3	Can we ignore the hierarchical structure?	146
4.8.4	Multilevel – here, more appropriately known as – mixed-effects models .	148
4.8.5	Is there a difference between linear model and linear mixed-effects model results?	154
4.8.6	What we estimate when we estimate random effects	154
4.9	Variation between stimuli: the “language as fixed-effect fallacy”	158
4.9.1	Include the random effect of stimulus	160
4.10	Variances and covariances of random effects	164
4.10.1	But remember we excluded random effects covariance	164
4.11	Reporting the results of a mixed-effects model	164
4.12	Conclusions	165
4.12.1	Summary	166

4.12.2	Useful functions	167
4.13	R code and data file access for the class	167
4.14	References	168
4.14.1	Recommended reading	168
4.14.2	References list	168
5	Developing linear mixed-effects models	170
5.1	Motivations: to grow in sophistication	170
5.2	The key idea to get us started	170
5.3	Targets	171
5.4	Study guide	172
5.5	The data we will work with: ML word recognition study	173
5.5.1	Research hypotheses	175
5.5.2	Locate and download the data file	175
5.6	Tidy the data	175
5.6.1	Read-in the data file using <code>read_csv</code>	176
5.6.2	Examine the distribution of raw RT data using density plots	178
5.6.3	Filter observations	181
5.6.4	Select or transform the variables: the <code>log10</code> transformation of RT	184
5.6.5	Data tidying – conclusions	186
5.7	Repeated measures designs and crossed random effects	186
5.8	Working with mixed-effects models	187
5.8.1	Use facetting in <code>ggplot</code> to examine data by person	187
5.8.2	Approximations to Linear Mixed-effects models: complete pooling	192
5.8.3	Approximations to Linear Mixed-effects models: no pooling	194
5.9	The linear mixed-effects model	196
5.9.1	Fixed and random effects	196
5.9.2	Variance and covariance	197
5.9.3	Random effects of differences between stimuli	197
5.9.4	A model including random effects of differences between stimuli as well as participants	198
5.9.5	Fitting a mixed-effect model using <code>lmer()</code>	198
5.10	Mixed-effects models, partial pooling, and shrinkage or regularisation of estimates	201
5.10.1	Overfitting	201
5.10.2	Partial pooling: shrinkage or borrowing strength	201
5.11	Estimation methods – An intuitive account of estimation in mixed-effects models	204
5.11.1	Convergence problems	207
5.12	Fitting and evaluating Linear Mixed-effects models	208
5.12.1	Model comparison approach	208
5.12.2	Model comparison using information criteria, AIC and BIC	209
5.12.3	Model comparison using the Likelihood Ratio Test	211
5.13	Modeling steps recommendations	212
5.13.1	Maximum Likelihood and Restricted Maximum Likelihood	213

5.13.2	Comparing models of varying random effects but constant fixed effects	214
5.13.3	Evaluating random effects of subjects or items on slopes	220
5.13.4	Effects estimates and <i>significance</i> or p-values	222
5.14	Reporting results	224
5.14.1	Reporting comparisons of ML and REML models	224
5.14.2	Reporting the model: summary	226
5.15	Summary	226
5.15.1	Useful functions	227
5.16	R code and data file access for the class	227
5.17	References	228
5.17.1	Recommended reading	228
5.17.2	A <i>very</i> useful FAQ	228
5.17.3	References list	228
6	Introduction to Generalized Linear Mixed-effects Models	230
6.1	The key idea to get us started	230
6.2	Targets	230
6.3	Study guide	231
6.4	Motivations	231
6.4.1	Our focus is on the analysis of categorical outcome variables	231
6.4.2	Recognize the limitations of alternative methods for analyzing response accuracy	233
6.5	Understanding the Generalized part of the Generalized Linear Mixed-effects Models in practical terms	236
6.6	The data we will work with: the Ricketts word learning study	237
6.6.1	Study information	237
6.6.2	Locate and download the data file	240
6.7	Tidy the data	241
6.7.1	Read-in the data file using <code>read_csv</code>	241
6.7.2	Code categorical factors	242
6.8	Working with GLMMs in R	246
6.8.1	Specify a random intercepts model	247
6.8.2	Read the results	249
6.8.3	GLMMs and hypothesis tests	251
6.8.4	Presenting and visualizing the effects	251
6.9	Examining if we should include random effects	253
6.9.1	Examine the utility of random effects by comparing models with the same fixed effects but varying random effects	259
6.9.2	Bad signs	268
6.9.3	Comparison of models varying in random effects	268
6.9.4	Addressing convergence problems	270
6.10	Reporting model results	273

6.11	Summary	274
6.11.1	Useful functions	275
6.12	R code and data file access for the class	275
6.13	References	276
6.13.1	Recommended reading	276
6.13.2	A <i>very</i> useful FAQ	276
6.13.3	References list	276
6.14	Appendix: Example dataset variable information	278
III	WRITING ABOUT RESEARCH	281
7	Introduction: the why	282
7.1	The key ideas	282
7.2	Why: what is the motivation for the assignment?	283
7.2.1	The wider context: crisis and revolution	283
7.2.2	The specific context: what we need to look at, conceptually and practically	284
7.2.3	Multiverse analyses: multi- what?	285
7.2.4	Multiverse analyses	287
7.2.5	From the multiverse to kinds of reproducibility	293
7.2.6	The current state of the match between open science ideas and practices	295
7.3	This is why	300
7.3.1	Summary: this is why	301
8	What	303
8.1	PSYC401 Project – research report – what you are expected to do	303
8.1.1	What data can I analyse?	303
8.1.2	What structure should reports take?	304
8.1.3	What content should reports present?	304
8.1.4	What format?	306
9	How	307
9.1	The variety of things students do	307
9.2	Working with data associated with a published analysis	308
9.2.1	Locate, access and check the data	308
9.2.2	Plan the analysis you want to do	312
9.2.3	Summary: working with data associated with a published analysis	316
9.3	Working with data that are not associated with a published analysis	317
9.3.1	Looking for open data	317
9.3.2	Thinking about analyses of open data	319
9.4	Summary: how	320

IV END	322
10 Summary	323
References	324

Preface

A change in approach

We can, here, explain a development in the approach we take in teaching this course. Naturally, this development in approach will require a parallel development in your approach to learning.

We are going to focus on working in research in context (see Figure [Figure 0.1](#)).

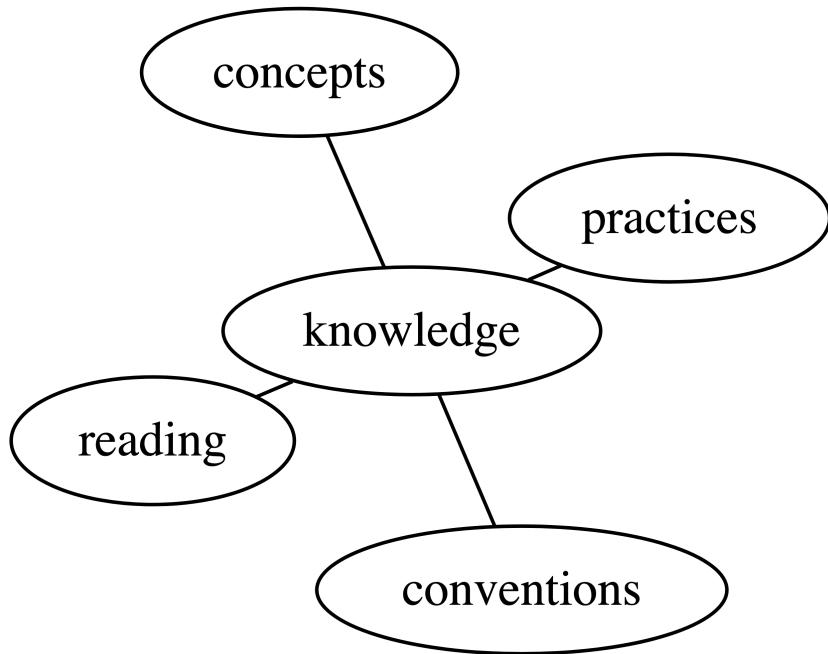


Figure 0.1: This is a simple graphviz graph.

You have been introduced to R. We know that some of you are new to R so we will practice the skills you are learning. We will consolidate, revise, and extend these skills.

We will encounter — some, for the first time — the linear model *also known as* regression analysis, multiple regression.

But the big change is this focus on the context. The reason is that *not* talking about the context has a dangerous impact on how you approach, do or think about data analysis.

In traditional methods teaching, the schedule of classes will progress through a series of tests, one test a week, from simpler to more complex tests (e.g., from t-test to multiple regression at the undergraduate level). Textbooks often mirror this structure, presenting one test per chapter. In this approach, the presentation is often brief about the context: the question the researchers are investigating; the methods they use to collect data, including the measurements; and the assumptions they make about how your reasoning can get you from the things you measure to the things you are trying to understand. In this approach, also, example data may be presented in a limited, partial, way.

The reasons for this are understandable: methods are complex, technical, subjects for learning, and teachers and students do not also have time, perhaps, to think about statistics and about theoretical or measurement assumptions. This is a mistake because it presents a misleading view of the challenge in learning methods: the challenge is *just* the (difficult enough) challenge of learning about statistical methods, or dealing with numbers. It is a mistake, also, because it implies that if you learn the method, and can match the textbook example – the variables, the state of the data – when it is your turn to do an analysis, all will be well.

Maybe. I think a more productive approach – this is the approach we will take – is to expose, and talk about some of the real challenges that anybody who handles data, or quantitative evidence, in professional life. These challenges include:

1. Thinking about the mapping from our concerns to the research questions, to the things we measure, to analysis we do, and then the conclusions we make.
2. Selecting or constructing valid measures that can be assumed to measure the things they are supposed to measure.
3. Taking samples of observations, and making conclusions about the population.
4. Making estimates and linking these estimates to an account that is explicit about causes.

Part I

MAKING THE MOST OF R

1 R knowledge

2 Data visualization

2.1 Aims

In writing this chapter, I have two aims.

1. The **first aim** for this chapter is to expose students to an outline summary of some key ideas and techniques for data visualization in psychological science.

There is an extensive experimental and theoretical literature concerning data visualization, what choices we can or should make, and how these choices have more or less impact, in different circumstances or for different audiences. Here, we can only give you a flavour of the on-going discussion. If you are interested, you can follow-up the references in the cited articles. But, using this chapter, I hope that you will gain a sense of the reasons *how or why* we may choose to do different things when we produce visualizations.

2. The **second aim** is to provide materials, and to show visualizations, to raise an awareness of what results come from making different choices. This is because we hope to encourage students to *make* choices based on reasons and it is hard to know what choices count without first seeing what the results might look like.

In my experience, knowing that there *are* choices is the first step. In proprietary software packages like Excel and SPSS there are plenty of choices but these are limited by the menu systems to certain combinations of elements. Here, in using R to produce visualizations, there is much more freedom, and much more capacity to control what a plot shows and how it looks, but knowing where to start has to begin with seeing examples of what some of the choices result in.

At the end of the chapter, I highlight some resources you can use in independent learning for further development, see Section 2.9.

So, we are aiming to (1.) start to build insight into the choices we make and (2.) provide resources to enable making those choices in data visualization.

2.2 Why data visualization matters

Data visualization is important. Building skills in visualization matters to you because, even if you do not go on to professional work in which you produce visualizations you will certainly be working in fields in which you need to work with, or read or evaluate, visualizations.

You have already been doing this: our cultural or visual environment is awash in visualizations, from weather maps to charts on the television news. It will empower you if you know a bit about how or why these visualizations are produced in the ways that they are produced. That is a complex development trajectory but we can get started here.

In the context of the research report exercise, see Section 7.2.3.1, I mention data visualization in relation to stages of the data analysis **pipeline** or **workflow**. But the reality is that, most of the time, visualization is useful and used at every stage of data analysis workflow.

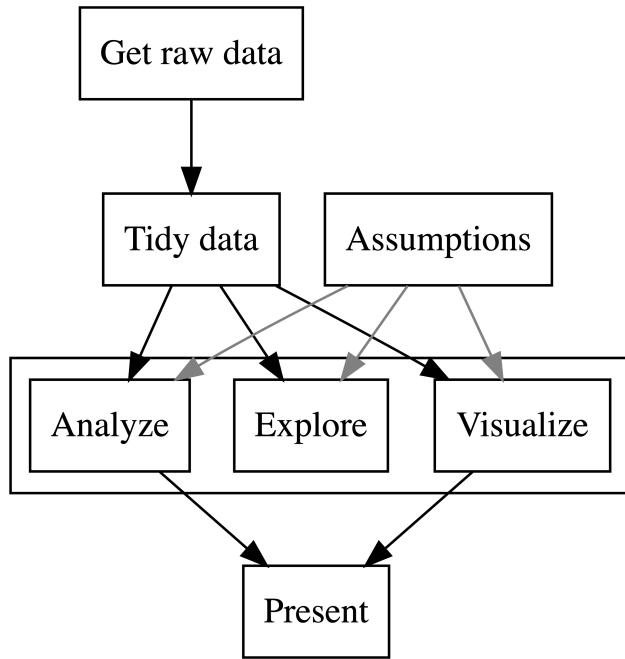


Figure 2.1: The data analysis pipeline or workflow

2.3 Three kinds of honesty

I write this chapter with three kinds of honesty in mind.

1. I will expose some of the process involved in thinking about and preparing for the production of plots.
 - I can assure you that when a professional data analysis worker produces plots in R they will be looking for information about what to do, and how to do it, online. I will provide links to the information I used, when I wrote this chapter, in order to figure out the coding to produce the plots.
 - I won't pretend that I got the plots "right first time" or that I know all the coding steps by memory. Neither is true for me and they would not be true for most professionals if they were to write a chapter like this. Looking things up online is something we all do so showing you where the information can be found will help you grow your skills.
2. I will show how we often prepare for the production of plots by processing the data that we must use to inform the plots.
 - We almost always have to process the data we collected or gathered together from our experimental work or our observations.
 - In this chapter, some of the coding steps I will outline are done in advance of producing a plot, to give the plotting code something to work with.
 - Knowing about these processing steps will ensure you have more flexibility or power in getting your plots ready.
3. I am going to expose *variation*, as often as I can, in observations.
 - We typically collect data about or from people, about their responses to things we may present (stimuli) or, given tasks, under different conditions, or concerning individual differences on an array of dimensions.
 - Sources of variation will be everywhere in our data, even though we often work with statistical analyses (like the t-test) that focus our attention on the average participant or the average response.
 - Modern analysis methods (like mixed-effects models) enable us to account for sources of variation systematically, so it is good to begin thinking about, say, how people vary in their response to different experimental conditions from early in your development.

2.4 Our approach: tidyverse

The approach we will take is to focus on step-by-step guides to coding. I will show plots and I will walk through the coding steps, explaining my reasons for the choices I make.

We will be working with plotting functions like `ggplot()` provided in libraries like `ggplot2` (Wickham, 2016) which is part of the `tidyverse` (Wickham, 2017) collection of libraries.

You can access information about the tidyverse collection [here](#).

2.4.1 Grammar of graphics

The `gg` in `ggplot` stands for the “Grammar of Graphics”, and the ideas motivating the development of the `ggplot2` library of functions are grounded in the ideas concerning the grammar of graphics, set out in the book of that name (Wilkinson, 2013).

What is helpful to us, here, is the insight that the code elements (and how they result in visual elements) can be identified as building blocks, or layers, that we can add and adjust piece by piece when we are producing a visualization.

A plot represents information and, critically, every time we write `ggplot` code we must specify somewhere the ways that our plot links data to something we see. In terms of `ggplot`, we specify *aesthetic mappings* using the `aes()` code to tell R what variables should be mapped e.g. to x-axis or y-axis location, to colour, or to group assignments. We then add elements to instruct R how to represent the aesthetic mappings as visual objects or attributes: geometric objects like a scatter of points `geom_point()` or a collection of bars `geom_bar()`; or visual features like colour, shape or size e.g. `aes(colour = group)`. We can add visual elements in a series of layers, as shall see in the practical demonstrations of plot construction. We can adjust how scaling works. And we can add annotation, labels, and other elements to guide and inform the attention of the audience.

You can read more about mastering the grammar [here](#).

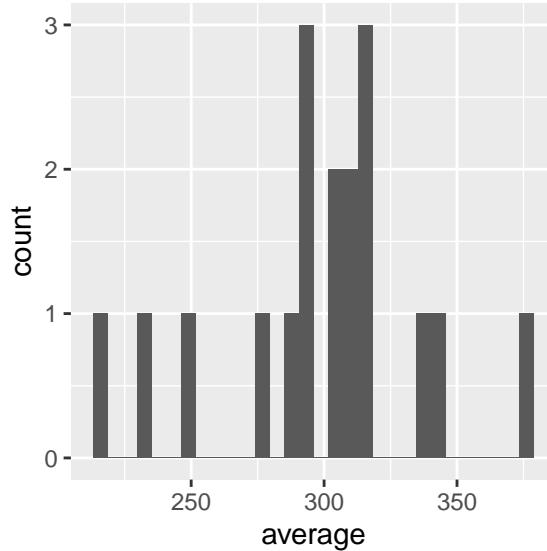
2.4.2 Pipes

We know that (some of) you want to see more use of pipes (represented as `%>%` or `|>`) in coding. There will be plenty of pipes in this chapter.

In using pipes in the code, I am structuring the code so that it works — and is presented — in a sequence of steps. There are different ways to write code but I find this way easier to work with and to read and I think you will too.

Let's take a small example:

```
1 sleepstudy %>%
2   group_by(Subject) %>%
3   summarise(average = mean(Reaction)) %>%
4   ggplot(aes(x = average)) +
5   geom_histogram()
```



Here, we work through a series of steps:

1. `sleepstudy %>%` we first tell R we want to work with the dataset called `sleepstudy` and the `%>%` pipe symbol at the end of the line tells R that we want it to pass that dataset on to the next step for what happens next.
2. `group_by(Subject) %>%` tells R that we want it to do something, here, group the rows of data according to the `Subject` (participant identity) coding variable, and pass the grouped data on to the next step for what happens following.
3. `summarise(average = mean(Reaction)) %>%` tells R to take the grouped variable and calculate a summary, the mean Reaction score, for each group of observations for each participant. The `%>%` pipe at the end of the line tells R to pass the summary dataset of mean Reaction scores on to the next process.
4. `ggplot(aes(x = average)) +` tells R that we want it to take these summary `average` Reaction scores and make a plot out of them.
5. `geom_histogram()` tells R that we want a histogram plot.

What you can see is that each line ending in a `%>` pipe passes something on to the next line. A following line takes the output of the process coded in the preceding line, and works with it.

Each step is executed in turn, in strict sequence. This means that if I delete line 3 `summarise(average = mean(Reaction)) %>%` then the following lines cannot work because the `ggplot()` function will be looking for a variable `average` that does not yet exist.

Warning

- You can see that in the data processing part of the code, successive steps in data processing end in a pipe `%>%`.
- In contrast, successive steps of the plotting code add `ggplot` elements line by line with each line (except the last) ending in a `+`.

Notice that none of the processing steps actually changes the dataset called `sleepstudy`. The results of the process exist and can be used only within the sequence of steps that I have coded. If you want to keep the results of processing steps, you need to assign an object name to hold them, and I show how to do this, in the following.

You can read a clear explanation of pipes [here](#).

Tip

You can use the code you see:

- Each chunk of code is highlighted in the chapter.
- If you hover a cursor over the highlighted code a little clipboard symbol appears in the top right of the code chunk.
- Click on the clipboard symbol to copy the code, paste it into your own R-Studio instance.
- Then *experiment*: try out things like removing or commenting out lines, or changing lines, to see what effect that has.
- Breaking things, or changing things, helps to show what each bit of code does.

2.5 Key ideas

Data visualization is not really about coding, as about thinking.

- What are our goals?
- Why do we make some choices instead of others?

2.5.1 Goals

A. Gelman & Unwin (2013) outline the goals we may contemplate when we produce or evaluate visual data displays. In general, they argue, we are doing one or both of two things.

1. Discovery
2. Communication

In practice, this may involve the following (I paraphrase them, here).

1. Discovery goals

- Getting a sense of what is in a dataset, checking assumptions, confirming expectations, and looking for distinct patterns.
- Making sense of the scale and complexity of the dataset.
- Exploring the data to reveal unexpected aspects. As we will see, using small multiples (grids of plots) can often help with this.

2. Communication goals

- We communicate about our data to ourselves and to others. The process of constructing and evaluating a plot is often one way we speak to ourselves about own data, developing an understanding of what we have got. Once we have done this for ourselves, we can better figure out how to do it to benefit the understanding of an audience.
- We often use a plot to tell a story: the story of our study, our data, or our insight and how we get to it.
- We can use visualizations to attract attention and stimulate interest. Often, in presenting data to an audience through a talk or a report we need to use effective visualizations to ensure we get attention and that we locate the attention of our audience in the right places.

2.5.2 Psychological science of data visualization

You will see a rich variety of data visualizations in media and in the research literature. You will know that some choices, in the production of those visualizations, appear to work better than others.

Some of the reasons why some choices work better will relate to what we can understand in terms of the psychological science of how visual data communication works. A useful recent review of relevant research is presented by Franconeri et al. (2021).

Franconeri et al. (2021) provide a reason for working on visualizations: they allow us humans to process an array of information at once, often faster than if we were reading about the information, bit by bit. Effective visualization, then, is about harnessing the power of the human visual system, or visual cognition, for quick, efficient, information processing. Critically for science, in addition, visualizations can be more effective for discovering or communicating the critical features of data than summary statistics, as we shall see.

In producing visualizations, we often work with a vocabulary or palette of objects or visual elements. Franconeri et al. (2021) discuss how visualizations rely on visual channels to transform numbers into images that we can process visually.

- Dot plots and scatterplots represent values as position.

- Bar graphs represent values as position (the heights of the tops of bars) but also as lengths.
- Angles are presented when we connect points to form a line, allowing us to encode the differences between points.
- Intensity can be presented through variation in luminance contrast or colour saturation.

These channels can be ordered by how precisely they have been found to communicate different numeric values to the viewer. Your audience may more accurately perceive the difference between two quantities if you communicate that difference through the difference in the location of two points than if you ask your audience to compare the angles of two lines or the intensity of two colour spots.

In constructing data visualizations, we often work with conventions, established through common practice in a research tradition. For example, if you are producing a scatterplot, then most of the time your audience will expect to see the outcome (or dependent variable) represented by the vertical height (on the y-axis) of points. And your audience will expect that higher points represent larger quantities of the y-axis variable.

In constructing visualizations, we need to be aware of the cognitive work that we require the audience to do. Comparisons are harder, requiring more processing and imposing more load on working memory. You can help your reader by guiding their attention, by grouping or ordering visual elements to identify the most important comparisons. We can vary colour and shape to group or distinguish visual elements. We can add annotation or elements like lines or arrows to guide attention.

Visualizations are presented in context, whether in presentations or in reports. This context should be provided, by you the producer, with the intention to support the communication of your key messages. A visual representation, a plot, will be presented with a title, maybe a title note, maybe with annotation in the plot, and maybe with accompanying text. You should use these textual elements to lead your audience, to help them make sense of what they are looking at.

The diversity of audiences means that we should habitually add alt text for data visualizations to help those who use screen readers by providing a summary description of what images show. This chapter has been written using Quarto and rendered to .html with alt text included along with all images. Please do let me know if you are using a screen reader and the alt text description is or is not so helpful.

You can read a helpful explanation of alt text [here](#).

If you use colour in images then we should use colour bind colour palettes.

You can read about using colour blind palettes [here](#) or [here](#).

In the following practical exercises, we work with many of the insights in our construction of visualizations.

2.6 A quick start

We can get started before we understand in depth the key ideas or the coding steps. This will help to show where we are going. We will work with the `sleepstudy` dataset.

I will model the process, to give you an example workflow:

- the data, where they come from — what we can find out;
- how we approach the data — what we *expect* to see;
- how we visualize the data — discovery, communication.

2.6.1 Sleepstudy data

When we work with R, we usually work with functions like `ggplot()` provided in libraries like `ggplot2` (Wickham, 2016). These libraries typically provide not only functions but also datasets that we can use for demonstration and learning.

The `lme4` library (Bates et al., 2015) provides the `sleepstudy` dataset and we will take a look at these data to offer a taste of what we can learn to do. Usually, information about the R libraries we use will be located on the [Comprehensive R Archive Network \(CRAN\)](#) web pages, and we can find the technical reference information for `lme4` in the CRAN reference manual for the library, where we see that the `sleepstudy` data are from a study reported by (Belenky et al., 2003). The manual says that the `sleepstudy` dataset comprises:

A data frame with 180 observations on the following 3 variables. [1.] Reaction – Average reaction time (ms) [2.] Days – Number of days of sleep deprivation [3.] Subject – Subject number on which the observation was made.

We can take a look at the first few rows of the dataset.

```
sleepstudy %>%  
  head(n = 4)
```

	Reaction	Days	Subject
1	249.5600	0	308
2	258.7047	1	308
3	250.8006	2	308
4	321.4398	3	308

What we are looking at are:

The average reaction time per day (in milliseconds) for subjects in a sleep deprivation study. Days 0-1 were adaptation and training (T1/T2), day 2 was baseline (B); sleep deprivation started after day 2.

The abstract for Belenky et al. (2003) tells us that participants were deprived of sleep and the impact of relative deprivation was tested using a cognitive vigilance task for which the reaction times of responses were recorded.

So, we can *expect to find*:

- A set of rows corresponding to multiple observations for each participant (**Subject**)
- A reaction time value for each participant (**Reaction**)
- Recorded on each **Day**

2.6.2 Discovery and communication

In data analysis work, we often begin with the objective to understand the structure or the nature of the data we are working with.

You can call this the *discovery* phase:

- what have we got?
- does it match our expectations?

If these are reaction time data (collected in an cognitive experiment) do they look like cognitive reaction time data *should* look? We would expect to see a skewed distribution of observed reaction times distributed around an average located somewhere in the range 200-700ms.

Figure 2.2 represents the distribution of reaction times in the `sleepstudy` dataset.

I provide notes on the code steps that result in the plot. Click on the **Notes** tab to see them. Later, I will discuss some of these elements.

2.6.2.1 Plot

```
sleepstudy %>%
  ggplot(aes(x = Reaction)) +
  geom_histogram(binwidth = 15) +
  geom_vline(xintercept = mean(sleepstudy$Reaction),
             colour = "red", linetype = 'dashed', size = 1.5) +
  annotate("text", x = 370, y = 20,
           colour = "red",
           label = "Average value shown in red") +
  theme_bw()
```

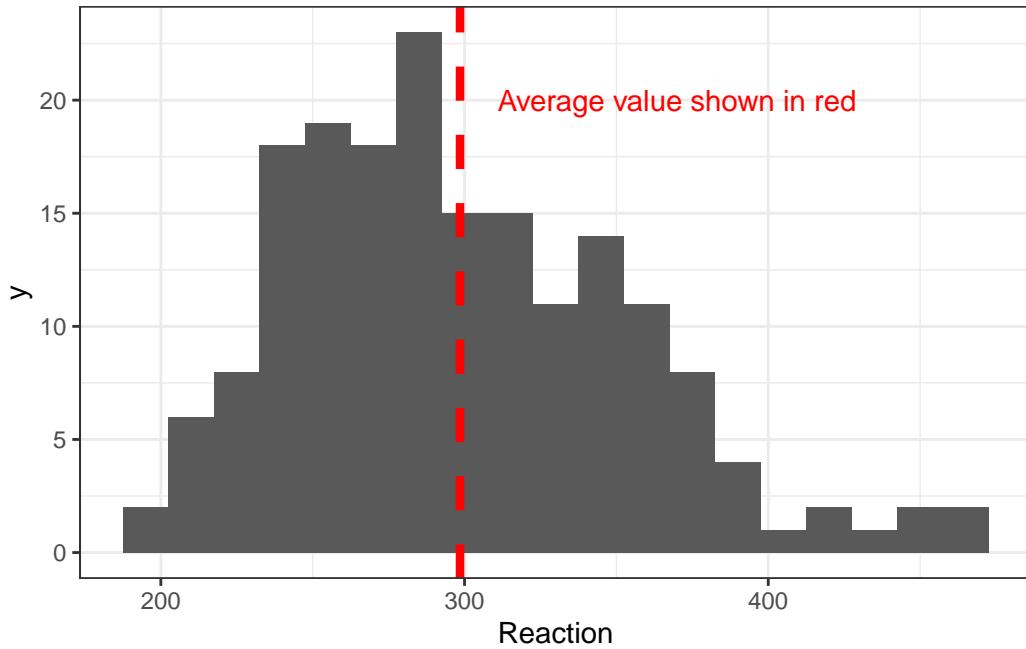


Figure 2.2: Figure showing a histogram of `sleepstudy` reaction time data

2.6.2.2 Notes

The plotting code pipes the data into the plotting code steps to produce the plot. You can see some elements that will be familiar to you and some new elements.

```
sleepstudy %>%
  ggplot(aes(x = Reaction)) +
  geom_histogram(binwidth = 15) +
  geom_vline(xintercept = mean(sleepstudy$Reaction),
             colour = "red", linetype = 'dashed', size = 1.5) +
  annotate("text", x = 370, y = 20,
           colour = "red",
           label = "Average value shown in red") +
  theme_bw()
```

Let's go through the code step-by-step:

1. `sleepstudy %>%` asks R to take the `sleepstudy` dataset and `%>%` pipe it to the next steps for processing.
2. `ggplot(aes(x = Reaction)) +` takes the `sleepstudy` data and asks R to use the `ggplot()` function to produce a plot.

3. `aes(x = Reaction)` tells R that in the plot we want it to map the `Reaction` variable values to locations on the x-axis: this is the aesthetic mapping.
4. `geom_histogram(binwidth = 15) +` tells R to produce a histogram then add a step.
5. `geom_vline(...)` + tells R we want to draw vertical line.
6. `xintercept = mean(sleepstudy$Reaction), ...` tells R to draw the vertical line at the mean value of the variable `Reaction` in the `sleepstudy` dataset.
7. `colour = "red", linetype = 'dashed', size = 1.5` tells R we want the vertical line to be red, dashed and 1.5 times the usual size.
8. `annotate("text", ...)` tells R we want to add a text note.
9. `x = 370, y = 20, ...` tells R we want the note added at the x,y coordinates given.
10. `colour = "red", ...`; and we want the text in red.
11. `label = "Average value shown in red") +` tells R we want the text note to say that this is where the average is.
12. `theme_bw()` lastly, we change the theme.

Figure 2.2 shows a distribution of reaction times, ranging from about 200ms to 500ms. The distribution has a peak around 300ms. The location of the mean is shown with a dashed red line. The distribution includes a long tail of longer times. This *is* pretty much what we would expect to see.

We may wish to communicate the information we gain through using this histogram, in a presentation or in a report.

2.6.3 Discovery and communication

Let us imagine that it is our study. (Here, we shall not concern ourselves too much — with apologies — with understanding what the original study authors actually did.)

If we are looking at the impact of sleep deprivation on cognitive performance, we might predict that reaction times got longer (responses slowed) as the study progressed. Is that what we see?

To examine the association between two variables, we often use scatterplots. Figure 2.3 is a scatterplot indicating the possible association between reaction time and days in the `sleepstudy` data. Points are ordered on x-axis from 0 to 9 days, on y-axis from 200 to 500 ms reaction time.

I provide notes on the code steps that result in the plot. Click on the Notes tab to see them. Later, I will discuss some of these elements.

2.6.3.1 Plot

```
sleepstudy %>%
  ggplot(aes(x = Days, y = Reaction)) +
  geom_point(size = 1.5, alpha = .5) +
  scale_x_continuous(breaks = c(0, 3, 6, 9)) +
  theme_bw()
```

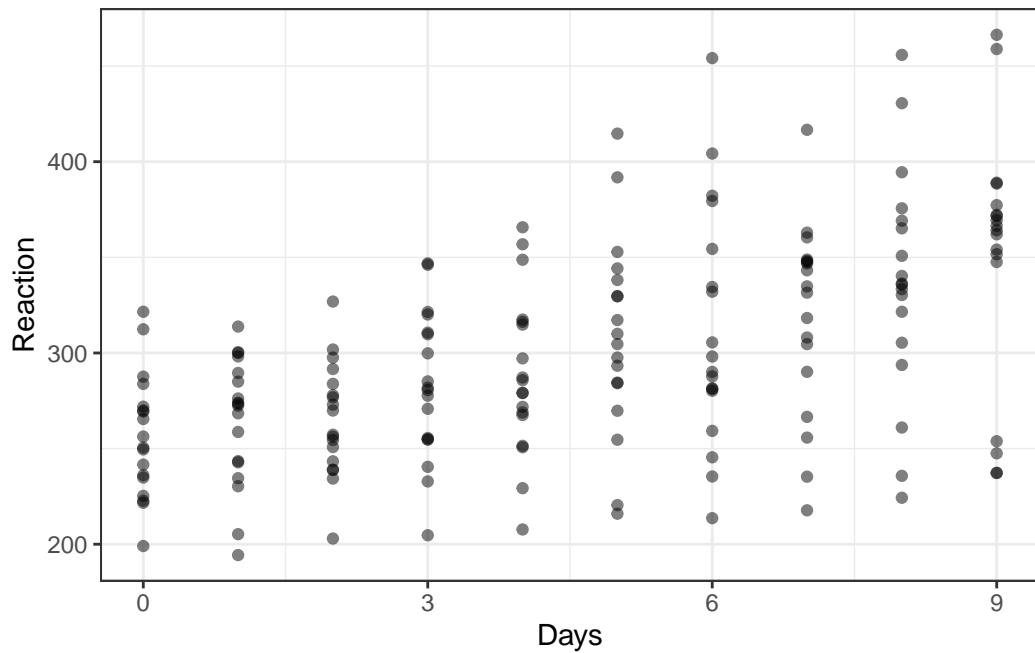


Figure 2.3: Figure showing a scatterplot of the relation between reaction time and days in the sleepstudy data

2.6.3.2 Notes

Notice the numbered steps in producing this plot.

```
1 sleepstudy %>%
  2   ggplot(aes(x = Days, y = Reaction)) +
  3   geom_point() +
  4   scale_x_continuous(breaks = c(0, 3, 6, 9)) +
  5   theme_bw()
```

1. Name the dataset: the dataset is called `sleepstudy` in the `lme4` library which makes it available therefore we use this name to specify it.
2. `sleepstudy %>%` uses the `%>%` pipe operator to pass this dataset to `ggplot()` to work with, in creating the plot. Because `ggplot()` now knows about the `sleepstudy` data, we can next specify what aesthetic mappings we need to use.
3. `ggplot(aes(x = Days, y = Reaction))` + tells R that we want to map `Days` information to x-axis position and `Reaction` (response time) information to y-axis position.
4. `geom_point()` + tells R that we want to locate points – creating a scatterplot – at the paired x-axis and y-axis coordinates.
5. `scale_x_continuous(breaks = c(0, 3, 6, 9))` + is new: we tell R that we want the x-axis tick labels – the numbers R shows as labels on the x-axis – at the values 0, 3, 6, 9 only.
6. `theme_bw()` requires R to make the plot background white and the foreground plot elements black.

You can find more information on `scale_` functions in the `ggplot2` reference information.

https://ggplot2.tidyverse.org/reference/scale_continuous.html

The plot suggests that reaction time increases with increasing number of days.

In producing this plot, we are both (1.) engaged in discovery and, potentially, (2.) able to do communication.

1. Discovery: is the relation between variables what we should expect, given our assumptions?
2. Communication: to ourselves and others, what relation do we observe, given our sample?

At this time, we have used and discussed scatterplots before, why we use them, how we write code to produce them, and how we read them.

With two additional steps we can significantly increase the power of the visualization. Figure 2.4 is a grid of scatterplots indicating the possible association between reaction time and days separately for each participant.

Again, I hide an explanation of the coding steps in the `Notes` tab: the interested reader can click on the tab to view the step-by-step guide to what is happening.

2.6.3.3 Plot

```
sleepstudy %>%
  group_by(Subject) %>%
  mutate(average = mean(Reaction)) %>%
  ungroup() %>%
```

```

mutate(Subject = fct_reorder(Subject, average)) %>%
ggplot(aes(x = Days, y = Reaction)) +
geom_point() +
geom_line() +
scale_x_continuous(breaks = c(0, 3, 6, 9)) +
facet_wrap(~ Subject) +
theme_bw()

```

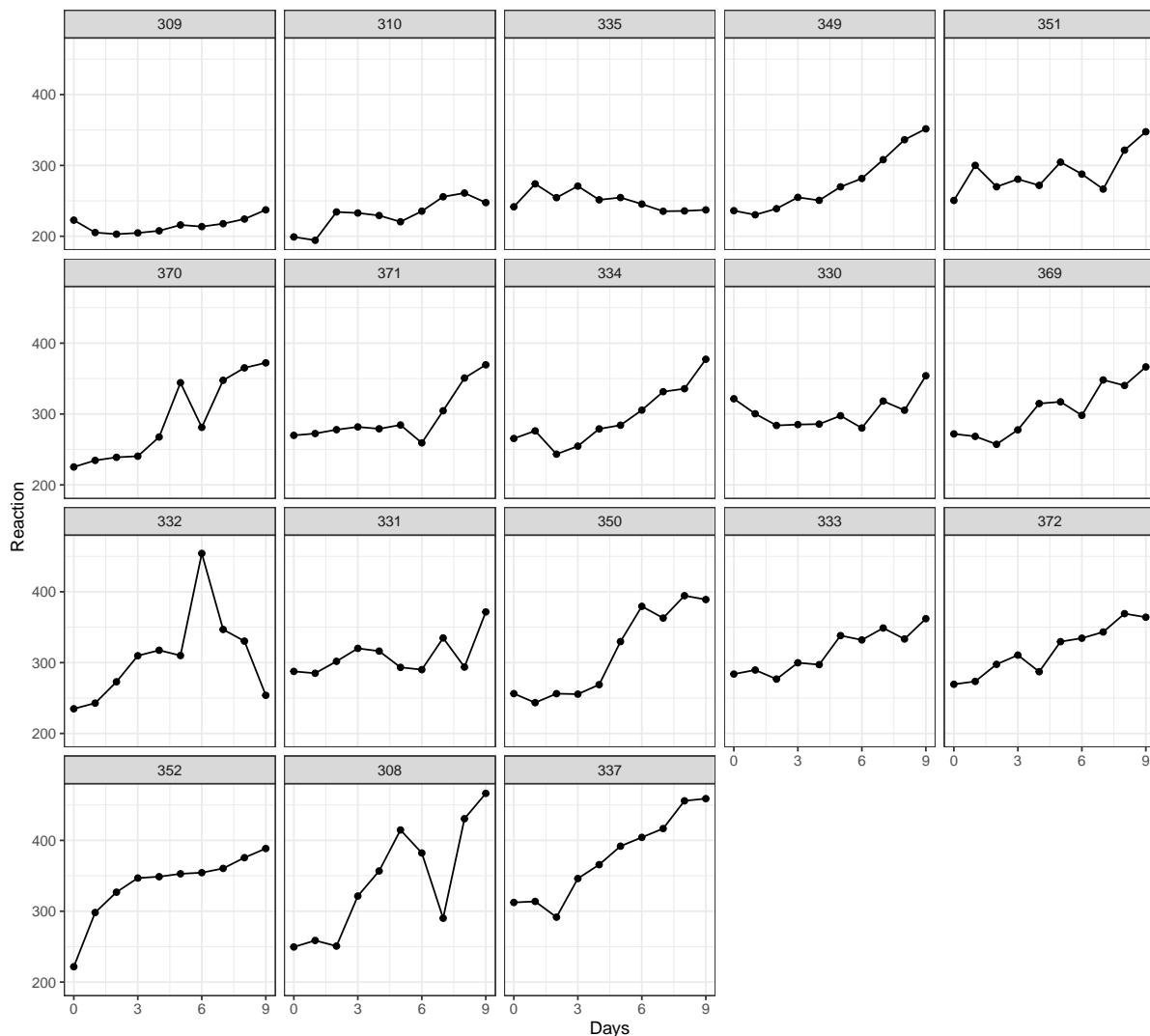


Figure 2.4: Figure showing a scatterplot of the relation between reaction time and days: here, we plot the data for each participant separately

2.6.3.4 Notes

Notice the numbered steps in producing this plot.

```
1 sleepstudy %>%
2   group_by(Subject) %>%
3   mutate(average = mean(Reaction)) %>%
4   ungroup() %>%
5   mutate(Subject = fct_reorder(Subject, average)) %>%
6   ggplot(aes(x = Days, y = Reaction)) +
7   geom_point() +
8   geom_line() +
9   scale_x_continuous(breaks = c(0, 3, 6, 9)) +
10  facet_wrap(~ Subject) +
11  theme_bw()
```

You can see that the block of code combines *data processing* and *data plotting* steps. Let's look at the data processing steps then the plotting steps in order.

First: why are we doing this? My aim is to produce a plot in which I show the association between `Days` and `Reaction` for each `Subject` individually. I suspect that the association between `Days` and `Reaction` may be stronger – so the trend will be steeper – for participants who are slower overall. I suspect this because, given experience, I know that slower, less accurate, participants tend to show larger effects.

So: in order to get a grid of plots, one plot for each `Subject`, in order of the average `Reaction` for each individual `Subject`, I need to first calculate the average `Reaction` then order the dataset rows by those averages. I do that in steps, using pipes to feed information from one step to the next step, as follows.

1. `sleepstudy %>%` tells R what data I want to use, and pipe it to the next step.
2. `group_by(Subject)` tells R I want it to work with data (rows) grouped by `Subject` identity code, `%>%` piping the grouped form of the data forward to the next step
3. `mutate(average = mean(Reaction))` uses `mutate()` to create a new variable `average` which I calculate as the `mean()` of `Reaction`, piping the data with this additional variable `%>%` forward to the next step.
4. `ungroup() %>%` tells R I want it to go back to working with the data in rows not grouped rows, and pipe the now ungrouped form of the data to the next step.
5. `mutate(Subject = fct_reorder(Subject, average))` tells R I want it to sort the rows of the whole `sleepstudy` dataset in order, moving groups of rows identified by `Subject` so that data for `Subject` codes associated with faster times are located near the top of the dataset.

These data, ordered by `Subject` by the average `Reaction` for each participant, are then `%>%` piped to `ggplot` to create a plot.

6. `ggplot(aes(x = Days, y = Reaction)) +` specifies the aesthetic mappings, as before.
7. `geom_point()` + asks R to locate points at the x-axis, y-axis coordinates, creating a scatterplot, as before.
8. `geom_line()` + is new: I want R to connect the points, showing the trend in the association between `Days` and `Reaction` for each person.
9. `scale_x_continuous(breaks = c(0, 3, 6, 9)) +` fixes the x-axis labels, as before.
10. `facet_wrap(~ Subject)` + is the big new step: I ask R to plot a separate scatterplot for the data for each individual `Subject`.

You can see more information about facetting here:

https://ggplot2.tidyverse.org/reference/facet_wrap.html

In short, with the `facet_wrap(~ .)` function, we are asking R to subset the data by a grouping variable, specified (`~ .`) by replacing the dot with the name of the variable.

Notice that I use `%>%` pipes to move the data processing forward, step by step. But I use `+` to add plot elements, layer by layer.

Figure Figure 2.4 is a grid or lattice of scatterplots *revealing* how the possible association between reaction time and days varies quite substantially between the participants in the `sleepstudy` data. Most plots indicate that reaction time increases with increasing number of days. However, different participants show this trend to differing extents.

What are the two additions I made to the conventional scatterplot code?

- I calculated the average reaction time per participant, and I ordered the data by those averages.
- I *facetted* the plots, breaking them out into separate scatterplots per participant.

Why would you do this? Variation between people or groups, in effects or in average outcomes, are often to be found in psychological data (Vasisht & Gelman, 2021). The variation between people that we see in these data — in the average response reaction time, and in how days affects times — would motivate the use of linear mixed-effects models to analyze the way that sleep patterns affect responses in the sleep study (Pinheiro & Bates, 2000).

💡 Tip

The data processing and plotting functions in the `tidyverse` collection of libraries enable us to discover and to communicate variation in behaviours that should strengthen our and others' scientific understanding.

2.6.4 Summary: Quick start lessons

What we have seen, so far, is that we can make dramatic changes to the appearance of visualizations (e.g., through faceting) and also that we can exert fine control over the details (e.g., adjusting scale labels). What we need to stop and consider are what we want to do (and why), in what order.

We have seen how we can feed a data process into a plot to first prepare then produce the plot in a sequence of steps. In processing the data, we can take some original data and extract or calculate information that we can use for our plotting e.g. calculating the mean of a distribution in order to then highlight where that mean is located.

We have also seen the use of plots, and the editing of their appearance, to represent information visually. We can verbalize the thought process behind the production of these plots through a series of questions.

1. Are we looking at the distribution of one variable (if yes: consider a histogram) or are we comparing the distributions of two or more variables (if yes: consider a scatterplot)?
2. Is there a salient feature of the plot we want to draw the attention of the audience to? We can add a visual element (like a line) and annotation text to guide the audience.
3. Are we interested in variation between sub-sets of the data? We can facet the plot to examine variation between sub-sets (facets) enabling the comparison of trends.

2.7 A practical guide to visualization ideas

In this guide, we illustrate some of the ideas about visualization we discussed at the start, working with practical coding examples. We will be working with real data from a published research project. We are going to focus the practical coding examples on the data collected for the analysis reported by Ricketts et al. (2021).

Advice

We will focus on working with the data from one of the tasks, in one of the studies reported by Ricketts et al. (2021).

- This means that you can consolidate your learning by applying the same code moves to data from the other task in the same study, or to data from the other study.
- In applying code to other data, you will need to be aware of differences in, say, the way that some things like the outcome response variable are coded.

You can then further extend your development by trying out the coding moves for yourself using the data collected by Rodríguez-Ferreiro et al. (2020).

- These data are from a quite distinct kind of investigation, on a different research

- topic than the topic we will be exploring through our working examples.
- However, some aspects of the data structure are similar.
 - Critically, the data are provided with comprehensive documentation.

2.7.1 Set up for coding

To do our practical work, we will need functions and data. We get these at the start of our workflow.

2.7.1.1 Get libraries

We are going to need the `lme4`, `patchwork`, `psych` and `tidyverse` libraries of functions and data.

```
library(ggeffects)
library(patchwork)
library(psych)
library(tidyverse)
```

2.7.1.2 Get the data

You can access the data we are going to use in two different ways.

2.7.1.2.1 Get the data from project repositories

The data associated with both (Ricketts et al., 2021) and (Rodríguez-Ferreiro et al., 2020) are freely available through project repositories on the Open Science Framework web pages.

You can get the data from the Ricketts et al. (2021) [paper](#) through the repository located [here](#).

You can get the data from the Rodríguez-Ferreiro et al. (2020) [paper](#) through the repository located [here](#).

These data are associated with full explanations of data collection methods, materials, data processing and data analysis code. You can review the papers and the repository material guides for further information.

In the following, I am going to abstract summary information about the Ricketts et al. (2021) study and data. I shall leave you to do the same for the Rodríguez-Ferreiro et al. (2020) study.

2.7.1.2.2 Get the data through a downloadable archive

Download the [data-visualization.zip](#) files folder and upload the files to RStudio Server.

The folder includes the Ricketts et al. (2021) data files:

- concurrent.orth_2020-08-11.csv
- concurrent.sem_2020-08-11.csv
- long.orth_2020-08-11.csv
- long.sem_2020-08-11.csv

The folder also includes the Rodríguez-Ferreiro et al. (2020) data files:

- PrimDir-111019_English.csv
- PrimInd-111019_English.csv

Warning

- These data files are collected together in a folder for download, for your convenience, but the *version of record* for the data for each study comprise the files located on the OSF repositories associated with the original articles.

2.7.2 Information about the Ricketts study and the datasets

Ricketts et al. (2021) conducted an investigation of word learning in school-aged children. They taught children 16 novel words in a study with a 2×2 factorial design. In this investigation, they tested whether word learning is helped by presenting targets for word learning with their spellings, and whether learning is helped by telling children that they would benefit from the presence of those spellings.

The presence of orthography (the word spelling) was manipulated within participants (orthography absent vs. orthography present): for all children, eight of the words were taught with orthography present and eight with orthography absent. Instructions (incidental vs. explicit) were manipulated between participants such that children in the explicit condition were alerted to the presence of orthography whereas children in the incidental condition were not.

A pre-test was conducted to establish participants' knowledge of the stimuli. Then, each child was seen for three 45-minute sessions to complete training (Sessions 1 and 2) and post-tests (Session 3). Ricketts et al. (2021) completed two studies: Study 1 and Study 2. All children, in both studies 1 and 2 completed the Session 3 post-tests.

In Study 1, longitudinal post-test data were collected because children were tested at two time points. Children were administered post-tests in Session 3, as noted: Time 1. Post-tests were

then re-administered approximately eight months later at Time 2 ($M = 241.58$ days from Session 3, $SD = 6.10$). In Study 2, the Study 1 sample was combined with an older sample of children. The additional Study 2 children were not tested at Time 2, and the analysis of Study 2 data did not incorporate test time as a factor.

The outcome data for both studies consisted of performance on post-tests.

The semantic post-test assessed knowledge for the meanings of newly trained words using a dynamic or sequential testing approach. I will not explain this approach in more detail, here, because the practical visualization exercises focus on the orthographic knowledge (spelling knowledge) post-test, explained next.

The orthographic post-test was included to ascertain the extent of orthographic knowledge after training. Children were asked to spell each word to dictation and spelling productions were transcribed for scoring. Responses were scored using a Levenshtein distance measure indexing the number of letter deletions, insertions and substitutions that distinguish between the target and child's response. The maximum score is 0, with higher scores indicating less accurate responses.

For the Study 1 analysis, the files are:

- `long.orth_2020-08-11.csv`
- `long.sem_2020-08-11.csv`

Where `long` indicates the longitudinal nature of the data-set.

For the Study 2 analysis, the files are:

- `concurrent.orth_2020-08-11.csv`
- `concurrent.sem_2020-08-11.csv`

Where `concurrent` indicates the inclusion of concurrent (younger and older) child participant samples.

Each column in each data-set corresponds to a variable and each row corresponds to an observation (i.e., the data are *tidy*). Because the design of the study involves the collection of repeated observations, the data can be understood to be in a *long* format.

Each child was asked to respond to 16 words and, for each of the 16 words, we collected post-test responses from multiple children. All words were presented to all children.

We explain what you will find when you inspect the .csv files, next.

2.7.2.1 Data – variables and value coding

The variables included in .csv files are listed, following, with information about value coding or calculation.

- **Participant** — Participant identity codes were used to anonymize participation. Children included in studies 1 and 2 – participants in the longitudinal data collection – were coded “EOF[number]”. Children included in Study 2 only (i.e., the older, additional, sample) were coded “ND[number]”.
- **Time** — Test time was coded 1 (time 1) or 2 (time 2). For the Study 1 longitudinal data, it can be seen that each participant identity code is associated with observations taken at test times 1 and 2.
- **Study** — Observations taken for children included in studies 1 and 2 – participants in the longitudinal data collection – were coded “Study1&2”. Children included in Study 2 only (i.e., the older, additional, sample) were coded “Study2”.
- **Instructions** — Variable coding for whether participants undertook training in the **explicit** or **incidental** conditions.
- **Version** — Experiment administration coding
- **Word** — Letter string values show the words presented as stimuli to children.
- **Consistency_H** — Calculated orthography-to-phonology consistency value for each word.
- **Orthography** — Variable coding for whether participants had seen a word in training in the orthography **absent** or **present** conditions.
- **Measure** — Variable coding for the post-test measure: **Sem_all** if the semantic post-test; **Orth_sp** if the orthographic post-test.
- **Score** — Variable coding for response category.

For the semantic (sequential or dynamic) post-test, responses were scored as corresponding to:

- 3 – correct response in the definition task
- 2 – correct response in the cued definition task
- 1 – correct response in the recognition task
- 0 – if the item wasn’t correctly defined or recognised

For the orthographic post-test, responses were scored as:

- 1 – correct, if the target spelling was produced in full
- 0 – incorrect

However, the analysis reported by Ricketts et al. (2021) focused on the more sensitive Levenshtein distance measure (see following).

- **WASI_mRS** — Raw score – Matrix Reasoning subtest of the Wechsler Abbreviated Scale of Intelligence

- **TOWREsweRS** — Raw score – Sight Word Efficiency (SWE) subtest of the Test of Word Reading Efficiency; number of words read correctly in 45 seconds
- **TOWREpdeRS** — Raw score – Phonemic Decoding Efficiency (PDE) subtest of the Test of Word Reading Efficiency; number of nonwords read correctly in 45 seconds
- **CC2regRS** — Raw score – Castles and Coltheart Test 2; number of regular words read correctly
- **CC2irregRS** — Raw score – Castles and Coltheart Test 2; number of irregular words read correctly
- **CC2nwRS** — Raw score – Castles and Coltheart Test 2; number of nonwords read correctly
- **WASIvRS** — Raw score – vocabulary knowledge indexed by the Vocabulary subtest of the WASI-II
- **BPVSRs** — Raw score – vocabulary knowledge indexed by the British Picture Vocabulary Scale – Third Edition
- **Spelling.transcription** — Transcription of the spelling response produced by children in the orthographic post-test
- **Levenshtein.Score** — Children were asked to spell each word to dictation and spelling productions were transcribed for scoring. Responses were scored using a Levenshtein distance measure indexing the number of letter deletions, insertions and substitutions that distinguish between the target and child's response. For example, the response 'epegram' for target 'epigram' attracts a Levenshtein score of 1 (one substitution). Thus, this score gives credit for partially correct responses, as well as entirely correct responses. The maximum score is 0, with higher scores indicating less accurate responses.

(Notice that, for the sake of brevity, I do not list the `z_` variables but these are explained in the study OSF repository materials.)

Warning

Levenshtein distance scores are higher *if* a child makes more errors in producing the letters in a spelling response.

- This means that if we want to see what factors help a child to learn a word, including its spelling, then we want to see that helpful factors are associated with *lower* Levenshtein scores.

To demonstrate some of the processes we can enact to process and visualize data, and some of the benefits of doing so, we are going to work with the `concurrent.orth_2020-08-11.csv` dataset. These are data corresponding to the Ricketts et al. (2021) Study 2. `concurrent` refers to the analysis (a concurrent comparison) of data from younger and older children.

2.7.3 Read the data into R

Assuming you have downloaded the data files, we first read the dataset into the R environment: `concurrent.orth_2020-08-11.csv`. We do the data read in a bit differently than you have seen it done before; we will come back to what is going on (in Section 2.7.4.1).

```
conc.orth <- read_csv("concurrent.orth_2020-08-11.csv",  
  
  col_types = cols(  
  
    Participant = col_factor(),  
    Time = col_factor(),  
    Study = col_factor(),  
    Instructions = col_factor(),  
    Version = col_factor(),  
    Word = col_factor(),  
    Orthography = col_factor(),  
    Measure = col_factor(),  
    Spelling.transcription = col_factor()  
  
)
```

We can inspect these data using `summary()`.

```
summary(conc.orth)  
  
Participant      Time       Study      Instructions  Version  
EOF001 : 16    1:1167    Study1&2:655   explicit     :592    a:543  
EOF002 : 16          Study2  :512    incidental:575   b:624  
EOF004 : 16  
EOF006 : 16  
EOF007 : 16  
EOF008 : 16  
(Other):1071  
Word      Consistency_H      Orthography      Measure  
Accolade  : 73    Min.    :0.9048    absent  :583    Orth_sp:1167  
Cataclysm : 73    1st Qu.:1.5043   present:584  
Contrition: 73    Median  :1.9142  
Debacle   : 73    Mean    :2.3253  
Dormancy  : 73    3rd Qu.:3.0436  
Epigram   : 73    Max.    :3.9681  
(Other)   :729
```

Score	WASIImRS	TOWREsweRS	TOWREpdeRS	CC2regRS
Min. :0.0000	Min. : 5	Min. :51.00	Min. :19.00	Min. :28.00
1st Qu.:0.0000	1st Qu.:13	1st Qu.:69.00	1st Qu.:35.00	1st Qu.:36.00
Median :0.0000	Median :17	Median :74.00	Median :41.00	Median :38.00
Mean :0.2913	Mean :16	Mean :74.23	Mean :41.59	Mean :36.91
3rd Qu.:1.0000	3rd Qu.:19	3rd Qu.:80.00	3rd Qu.:50.00	3rd Qu.:39.00
Max. :1.0000	Max. :25	Max. :93.00	Max. :59.00	Max. :40.00
CC2irregRS	CC2nwRS	WASIVRS	BPVSRs	
Min. :17.00	Min. :13.00	Min. :16.00	Min. :103.0	
1st Qu.:23.00	1st Qu.:29.00	1st Qu.:25.00	1st Qu.:119.0	
Median :25.00	Median :33.00	Median :29.00	Median :133.0	
Mean :25.24	Mean :32.01	Mean :29.12	Mean :130.9	
3rd Qu.:27.00	3rd Qu.:37.00	3rd Qu.:33.00	3rd Qu.:142.0	
Max. :35.00	Max. :40.00	Max. :39.00	Max. :158.0	
Spelling.transcription	Levenshtein.Score	zTOWREsweRS	zTOWREpdeRS	
Epigram : 57	Min. :0.000	Min. :-2.67807	Min. :-2.33900	
Platitude : 43	1st Qu.:0.000	1st Qu.:-0.60283	1st Qu.:-0.68243	
Contrition: 42	Median :1.000	Median :-0.02638	Median :-0.06122	
fracar : 39	Mean :1.374	Mean : 0.00000	Mean : 0.00000	
Nonentity : 39	3rd Qu.:2.000	3rd Qu.: 0.66537	3rd Qu.: 0.87061	
raconter : 35	Max. :7.000	Max. : 2.16415	Max. : 1.80243	
(Other) :912				
zCC2regRS	zCC2irregRS	zCC2nwRS	zWASIVRS	
Min. :-3.3636	Min. :-2.22727	Min. :-3.1053	Min. :-2.63031	
1st Qu.:-0.3435	1st Qu.:-0.60461	1st Qu.:-0.4920	1st Qu.:-0.82633	
Median : 0.4115	Median :-0.06373	Median : 0.1614	Median :-0.02456	
Mean : 0.0000	Mean : 0.00000	Mean : 0.0000	Mean : 0.00000	
3rd Qu.: 0.7890	3rd Qu.: 0.47716	3rd Qu.: 0.8147	3rd Qu.: 0.77721	
Max. : 1.1665	Max. : 2.64070	Max. : 1.3047	Max. : 1.97986	
zBPVSRs	mean_z_vocab	mean_z_read	zConsistency_H	
Min. :-1.9946	Min. :-2.06910	Min. :-2.39045	Min. :-1.4153	
1st Qu.:-0.8495	1st Qu.:-0.85941	1st Qu.:-0.43321	1st Qu.:-0.8181	
Median : 0.1525	Median :-0.01483	Median : 0.08829	Median :-0.4096	
Mean : 0.0000	Mean : 0.00000	Mean : 0.00000	Mean : 0.0000	
3rd Qu.: 0.7967	3rd Qu.: 0.72964	3rd Qu.: 0.68438	3rd Qu.: 0.7157	
Max. : 1.9418	Max. : 1.96083	Max. : 1.52690	Max. : 1.6368	

You should notice one key bit of information in the summary. Focus on the summary for what is in the Participant column. You can see that we have a number of participants in this

Participant	Time	Study	Instructions	Version	Word	Consistency_H	Orthography	Me
EOF001	1	Study1&2	explicit	a	Accolade	1.9142393	absent	Or
EOF001	1	Study1&2	explicit	a	Cataclysm	3.5060075	present	Or
EOF001	1	Study1&2	explicit	a	Contrition	1.7486898	absent	Or
EOF001	1	Study1&2	explicit	a	Debacle	2.9008386	present	Or
EOF001	1	Study1&2	explicit	a	Dormancy	1.6263089	absent	Or
EOF001	1	Study1&2	explicit	a	Epigram	1.3822337	present	Or
EOF001	1	Study1&2	explicit	a	Foible	2.7051987	present	Or
EOF001	1	Study1&2	explicit	a	Fracas	3.1443345	absent	Or
EOF001	1	Study1&2	explicit	a	Lassitude	0.9048202	present	Or
EOF001	1	Study1&2	explicit	a	Luminary	1.0985931	absent	Or
EOF001	1	Study1&2	explicit	a	Nonentity	3.9681391	absent	Or
EOF001	1	Study1&2	explicit	a	Platitude	0.9048202	present	Or
EOF001	1	Study1&2	explicit	a	Propensity	1.6861898	absent	Or
EOF001	1	Study1&2	explicit	a	Raconteur	3.8245334	absent	Or
EOF001	1	Study1&2	explicit	a	Syncopation	3.0436450	present	Or
EOF001	1	Study1&2	explicit	a	Veracity	2.8693837	present	Or

dataset, listed by `Participant` identity code in the `summary()` view e.g. `EOF001`. For each participant, we have 16 rows of data.

When we ask R for a `summary` of a nominal variable or *factor* it will show us the levels of each factor (i.e., each category or class of objects encoded by the categorical variable), and a count for the number of observations for each level.

Take a look at the rows of data for `EOF001`.

You can see that for `EOF001`, as for every participant, we have information on the conditions under which we observed their responses (`Instructions`, `Orthography`), as well as information about the stimuli that we asked participants to respond to (e.g., `Word`, `Consistency_H`), information about the responses or *outcomes* we recorded (`Measure`, `Score`, `Spelling.transcription`, `Levenshtein.Score`), and information about the participants themselves (e.g., `TOWREsweRS`, `TOWREpdeRS`).

2.7.4 Process the data

We almost always need to process data in order to render the information ready for discovery or communication data visualization.

2.7.4.1 Specify column data types

You will have seen that data processing began when we first read the data in for use. Let's go back and take a look at the code steps.

```
1 conc.orth <- read_csv("concurrent.orth_2020-08-11.csv",
2
3     col_types = cols(
4
5         Participant = col_factor(),
6         Time = col_factor(),
7         Study = col_factor(),
8         Instructions = col_factor(),
9         Version = col_factor(),
10        Word = col_factor(),
11        Orthography = col_factor(),
12        Measure = col_factor(),
13        Spelling.transcription = col_factor()
14
15    )
16)
```

The chunk of code is doing two things: first, we tell R what .csv file we want to read into the environment, and what we want to call the dataset; and then we tell R how we want to classify the data variable columns.

1. `conc.orth <- read_csv("concurrent.orth_2020-08-11.csv"` first reads the named .csv file, creating an object I will call `conc.orth`: a dataset or tibble we can now work with in R.
 - You have been using the `read.csv()` function to read in data files.
 - The `read_csv()` function is the more modern `tidyverse` form of the function you were introduced to.
 - Both versions work in similar ways but `read_csv()` is a bit more efficient, and it allows us to do what we do next.
2. `col_types = cols(...)` tells R how to interpret some of the columns in the .csv.
 - The `read_csv()` function is excellent at working out what types of data are held in each column but sometimes we have to tell it what to do.
 - Here, I am specifying with e.g. `Participant = col_factor()` that the `Participant` column should be treated as a categorical or nominal variable, a *factor*.

Using the `col_types = cols(...)` argument saves me from having to first read the data in then using code like the following to require, technically, *coerce* R into recognizing the nominal nature of variables like `Participant` with code like

```
conc.orth$Participant <- as.factor(conc.orth$Participant)
```

2.7.4.1.1 Exercise

I do not have to do step 2 of the read-in process, here. What happens if we use just `read_csv()`? Try it.

```
conc.orth <- read_csv("concurrent.orth_2020-08-11.csv")
```

2.7.4.1.2 Further information

You can read more about `read_csv()` [here](#)

You can read more about `col_types = cols()` [here](#)

2.7.4.2 Extract information from the dataset

The Ricketts et al. (2021) dataset `orth.conc` is a moderately sized and rich dataset with several observations, on multiple variables, for each of many participants. Sometimes, we want to extract information from a more complex dataset because we want to understand or present a part of it, or a relatively simple account of it. We look at an example of how you might do that now.

As you saw when you looked at the summary of the `orth.conc` dataset, we have multiple rows of data for each participant. Recall the design of the study. For each participant, we recorded their response to a stimulus word, in a test of word learning, for 16 words.

For each participant, we have a *separate* row for each response the participant made to each word. But you will have noticed that information about the participant is repeated. So, for participant `EOF001`, we have data about their performance e.g. on the BPVSRS vocabulary test (they scored 126). Notice that that score is repeated: the same value is copied for each row, for this participant, in the `BPVSRS` column. The reason the data are structured like this are not relevant here¹ but it does require us to do some data processing, as I explain next.

¹As you can see if you read the Ricketts et al. (2021) paper, and the associated guide to the data and analysis on the OSF repository, we analysed the word learning data using Generalized Linear Mixed-effects Models (GLMM). GLMMs are used when we are analyzing data with a *multilevel* structure. These structures are very common and can be identified whenever we have groups or clusters observations: here, we have multiple observations of the test response, for each participant and for each stimulus word. When we fit GLMMs, the functions we use to do the analysis require the data to be structured in this `tidy` fashion, with different

It is a very common task to want to present a summary of the attributes of your participants or stimuli when you are reporting data in a report of a psychological research project. We could get a summary of the participant attributes using the `psych` library `describe` function as follows.

```
conc.orth %>%
  select(WASImRS:BPVRS) %>%
  describe(ranges = FALSE, skew = FALSE)
```

	vars	n	mean	sd	se
WASImRS	1	1167	16.00	4.30	0.13
TOWREsweRS	2	1167	74.23	8.67	0.25
TOWREpdeRS	3	1167	41.59	9.66	0.28
CC2regRS	4	1167	36.91	2.65	0.08
CC2irregRS	5	1167	25.24	3.70	0.11
CC2nwRS	6	1167	32.01	6.12	0.18
WASIvRS	7	1167	29.12	4.99	0.15
BPVRS	8	1167	130.87	13.97	0.41

But you can see that part of the information in the summary does not appear to make sense at first glance. We do *not* have 1167 participants in this dataset, as Ricketts et al. (2021) report.

How do we extract the participant attribute variable data for each unique participant code for the participants in our dataset?

```
1 conc.orth.subjs <- conc.orth %>%
  2   group_by(Participant) %>%
  3     mutate(mean.score = mean(Levenshtein.Score)) %>%
  4       ungroup() %>%
  5         distinct(Participant, .keep_all = TRUE) %>%
  6           select(WASImRS:BPVRS, mean.score, Participant)
```

We create a new dataset `conc.orth.subjs` by taking `conc.orth` and piping it through a series of processing steps. As part of the process, we want to extract the data for each unique unique Participant identity code using `distinct()`. Along the way, we want to calculate the mean accuracy of response on the outcome measure (`Score`), that is, the average number of edits separating a child's spelling of a target word from the correct spelling.

This is how we do it.

rows for each response or outcome observation, and repeated information for each participant or stimulus (if present).

1. `conc.orth.subjs <- ...` tells R to create a new dataset `conc.orth.subjs`.
2. `conc.orth %>% ...` we do this by telling R to take `conc.orth` and pipe it through the following steps.
3. `group_by(Participant) %>%` first we group the data by `Participant` identity code.
4. `mutate(mean.score = mean(Score)) %>%` then we use `mutate()` to create the new variable `mean.score` by calculating the `mean()` of the `Score` variable values (i.e. the average score) for each participant. We then pipe to the next step.
5. `ungroup() %>%` we tell R to ungroup the data because we want to work with all rows for what comes next, and we then pipe to the next step.
6. `distinct(Participant, .keep_all = TRUE) %>%` requires R to extract from the full `orth.conc` dataset the set of (here, 16) data rows we have for each distinct (uniquely identified) `Participant`. We use the argument `.keep_all = TRUE` to tell R that we want to keep all columns. This requires the next step, so we tell R to pipe `%>%` the data.
7. `select(WASImRS:BPVRS, mean.score, Participant)` then tells R to select just the columns with information about participant attributes. (`WASImRS:BPVRS` tells R to select every column between `WASImRS` and `BPVRS` inclusive. `mean.score, Participant` tells R we also want those columns, specified by name, including the `mean.score` column of average response scores we calculated just earlier.

We can now get a sensible summary of the descriptive statistics for the participants in Study 2 of the Ricketts et al. (2021) investigation.

```
conc.orth.subjs %>%
  select(-Participant) %>%
  describe(ranges = FALSE, skew = FALSE)
```

	vars	n	mean	sd	se
WASImRS	1	73	16.00	4.33	0.51
TOWREsweRS	2	73	74.22	8.73	1.02
TOWREpdeRS	3	73	41.58	9.73	1.14
CC2regRS	4	73	36.90	2.67	0.31
CC2irregRS	5	73	25.23	3.72	0.44
CC2nwRS	6	73	32.00	6.17	0.72
WASIvRS	7	73	29.12	5.02	0.59
BPVRS	8	73	130.88	14.06	1.65
mean.score	9	73	1.38	0.62	0.07

Tip

This is exactly the kind of tabular summary of descriptive statistics we would expect to produce in a report, in a presentation of the participant characteristics for a study sample (in e.g., the Methods section).

Notice:

1. The table has not yet been formatted according to APA rules.
2. We would prefer to use real words for row name labels instead of dataset variable column labels, e.g., replace TOWREsweRS with: “TOWRE word reading score”.

2.7.4.2.1 Exercise

In these bits of demonstration code, we extract information relating just to participants. However, in this study, we recorded the responses participants made to 16 stimulus words, and we include in the dataset information about the word properties `Consistency_H`.

- Can you adapt the code you see here in order to calculate a mean score for each word, and then extract the word-level information for each distinct stimulus word identity?

2.7.4.2.2 Further information

You can read more about the `psych` library, which is often useful, [here](#). You can read more about the `distinct()` function [here](#).

2.7.5 Visualize the data: introduction

It has taken us a while but now we are ready to examine the data using visualizations. Remember, we are engaging in visualization to (1.) do discovery, to get a sense of our data, and maybe reveal unexpected aspects, and (2.) potentially to communicate to ourselves and others what we have observed or perhaps what insights we can gain.

We have been learning to use histograms, in other classes, so let's start there.

2.7.6 Examine the distributions of numeric variables

We can use histograms to visualize the distribution of observed values for a numeric variable. Let's start simple, and then explore how to elaborate the plotting code, in a series of edits, to polish the plot presentation.

```
1 ggplot(data = conc.orth.subjs, aes(x = WASImRS)) +  
2   geom_histogram()
```

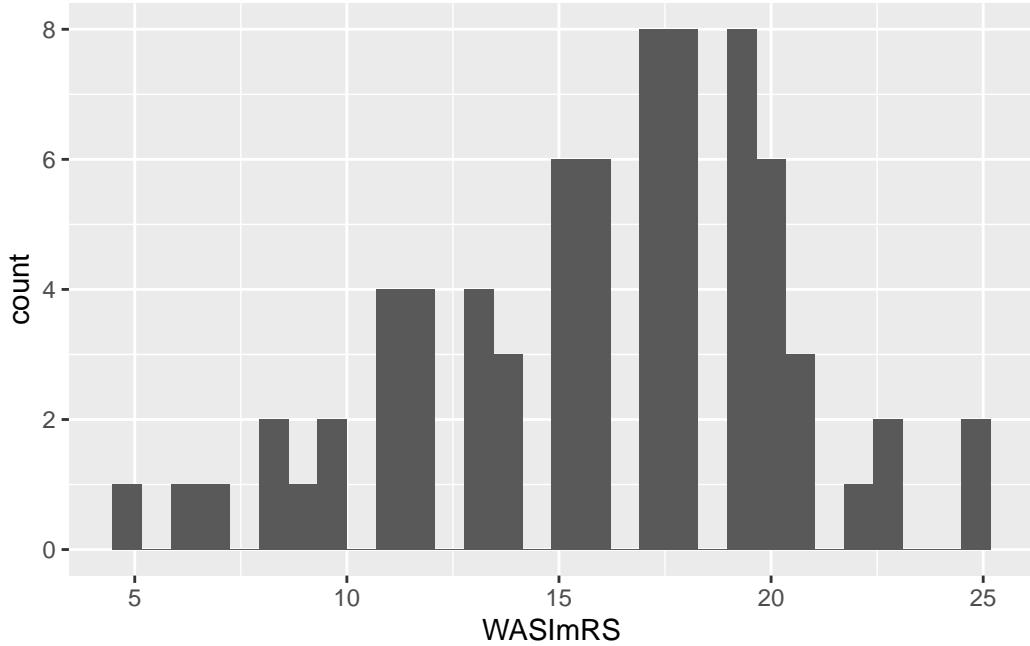


Figure 2.5: Distribution of WASImRS intelligence scores

This is how the code works.

1. `ggplot(data = conc.orth.subjs, ...)` tells R what function to use `ggplot()` and what data to work with `data = conc.orth.subjs`.
2. `aes(x = WASImRS)` tells R what aesthetic mapping to use: we want to map values on the WASImRS variable (small to large) to locations on the x-axis (left to right).
3. `geom_histogram()` tells R to construct a histogram, presenting a statistical summary of the distribution of intelligence scores.

With histograms, we are visualizing the distribution of a single continuous variable by dividing the variable values into bins (i.e. subsets) and counting the number of observations in each bin. Histograms display the counts with bars.

You can see more information about `geom_histogram` [here](#).

Figure 2.5 shows how intelligence (WASImRS) scores vary in the Ricketts Study 2 dataset. Scores peak around 17, with a long tail of lower scores towards 5, and a maximum around 25.

Advice

Where I use the word “peak” I am talking about the tallest bar in the plot (or, later the highest point in a density curve). At this point, we have the most observations of the

value under the bar. Here, we observed the score WASImRS = 17 for the most children in this sample.

A primary function of discovery visualization is to assess whether the distribution of scores on a variable is consistent with expectations, granted assumptions about a sample (e.g., that the children are typically developing). We would normally use research area knowledge to assess whether this distribution fits expectations for a sample of typically developing school-aged children in the UK. However, I shall leave that concern aside, here, so that we can focus on enriching the plot presentation, next.

There are two main problems with the plot:

1. The bars are “gappy” in the histogram, suggesting we have not grouped observed values in sufficiently wide subsets (bins). This is a problem because it weakens our ability to gain or communicate a visual sense of the distribution of scores.
2. The axis labeling uses the dataset variable name `WASImRS` but if we were to present the plot to others we could not expect them to know what that means.

We can fix both these problems, and polish the plot for presentation, through the following code steps.

```
1 ggplot(data = conc.orth.subjs, aes(x = WASImRS)) +  
2   geom_histogram(binwidth = 2) +  
3   labs(x = "Scores on the Wechsler Abbreviated Scale of Intelligence") +  
4   theme_bw()
```

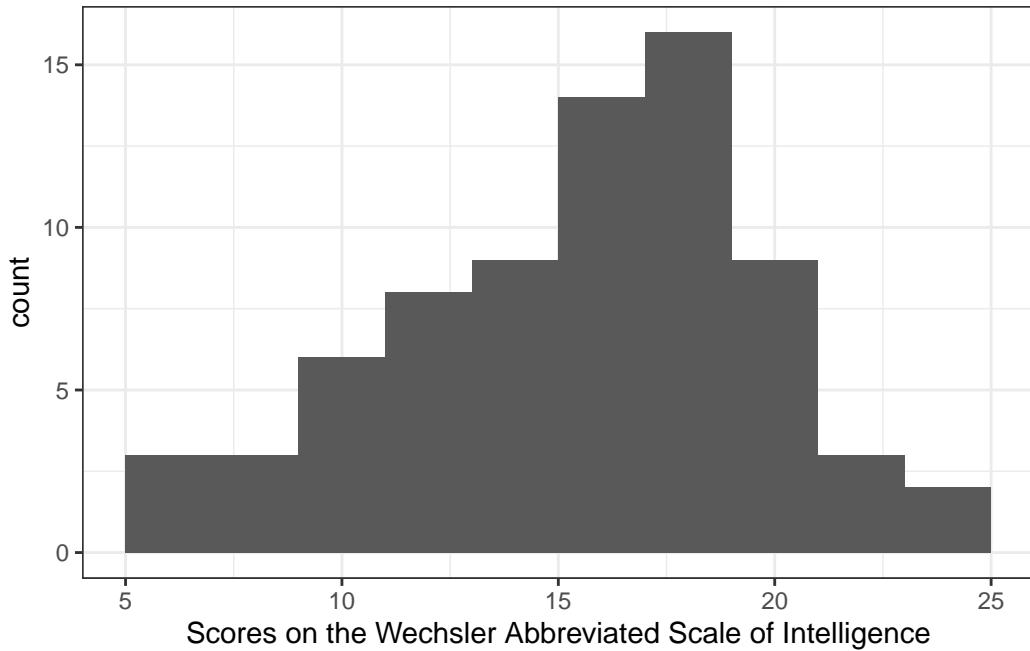


Figure 2.6: Distribution of WASImRS intelligence scores

Figure 2.6 shows the same data, and furnishes us with the same picture of the distribution of intelligence scores but it is a bit easier to read. We achieve this by making three edits.

1. `geom_histogram(binwidth = 2) +` we change the `binwidth`.
 - This is so that more different observed values of the data variable are included in bins (subsets corresponding to bars) so that the bars correspond to information about a wider range of values.
 - This makes the bars bigger, wider, and closes the gaps.
 - And this means we can focus the eyes of the audience for our plot on the visual impression we wish to communicate: the skewed distribution of intelligence scores.
2. `labs(x = "Scores on the Wechsler Abbreviated Scale of Intelligence") +` changes the label to something that should be understandable by people, in our audience, who do not have access to variable information (as we do) about the dataset.
3. `theme_bw()` we change the overall appearance of the plot by changing the theme.

2.7.6.1 Exercise

We could, if we wanted, add a line and annotation to indicate the mean value, as you saw in Figure 2.2.

- Can you add the necessary code to indicate the mean value of WASI scores, for this plot?

We can, of course, plot histograms to indicate the distributions of other variables.

- Can you apply the histogram code to plot histograms of other variables?

2.7.7 Comparing the distributions of numeric variables

We may wish to discover or communicate how values vary on dataset variables in two different ways. Sometimes, we need to examine how values vary on different variables. And sometimes, we need to examine how values vary on the same variable but in different groups of participants (or stimuli) or under different conditions. We look at this next. We begin by looking at how you might compare how values vary on different variables.

2.7.7.1 Compare how values vary on different variables

It can be useful to compare the distributions of different variables. Why?

Consider the Ricketts et al. (2021) investigation dataset. Like many developmental investigations (see also clinical investigations), we tested children and recorded their scores on a series of standardized measures, here, measures of ability on a range of dimensions. We did this, in part, to establish that the children in our sample are operating at about the level one might expect for typically developing children in cognitive ability dimensions of interest: dimensions like intelligence, reading ability or spelling ability. So, one of the aspects of the data we are considering is whether scores on these dimensions are higher or lower than typical threshold levels. But we also want to examine the distributions of scores because we want to find out:

- if participants are varied in ability (wide distribution) or if maybe they are all similar (narrow distribution) as would be the case if the ability measures are too easy (so all scores are at ceiling) or too hard (so all scores are at floor);
- if there are subgroups within the sample, maybe reflected by two or more peaks;
- if there are unusual scores, maybe reflected by small peaks at very low or very high scores.

We could look at each variable, one plot at a time. Instead, next, I will show you how to produce a set of histogram plots, and present them all as a single grid of plots.

Warning

I have to warn you that the way I write the code is not good practice. The code is written with repeats of the `ggplot()` block of code to produce each plot. This repetition is inefficient and leaves the coding vulnerable to errors because it is hard to spot a mistake in more code. What I *should* do is encapsulate the code as a function ([see here](#)). The

reason I do not, here, is because I want to focus our attention on just the plotting.

Figure 2.7 presents a grid of plots showing how scores vary for each ability test measure, for the children in the Ricketts et al. (2021) investigation dataset. We need to go through the code steps, next, and discuss what the plots show us (discovery and communication).

```
1 p.WASImRS <- ggplot(data = conc.orth.subjs, aes(x = WASImRS)) +
2   geom_histogram(binwidth = 2) +
3   labs(x = "WASI matrix") +
4   theme_bw()
5
6 p.TOWREsweRS <- ggplot(data = conc.orth.subjs, aes(x = TOWREsweRS)) +
7   geom_histogram(binwidth = 5) +
8   labs(x = "TOWRE words") +
9   theme_bw()
10
11 p.TOWREpdeRS <- ggplot(data = conc.orth.subjs, aes(x = TOWREpdeRS)) +
12   geom_histogram(binwidth = 5) +
13   labs(x = "TOWRE phonemic") +
14   theme_bw()
15
16 p.CC2regRS <- ggplot(data = conc.orth.subjs, aes(x = CC2regRS)) +
17   geom_histogram(binwidth = 2) +
18   labs(x = "CC regular words") +
19   theme_bw()
20
21 p.CC2irregRS <- ggplot(data = conc.orth.subjs, aes(x = CC2irregRS)) +
22   geom_histogram(binwidth = 2) +
23   labs(x = "CC irregular words") +
24   theme_bw()
25
26 p.CC2nwRS <- ggplot(data = conc.orth.subjs, aes(x = CC2nwRS)) +
27   geom_histogram(binwidth = 2) +
28   labs(x = "CC nonwords") +
29   theme_bw()
30
31 p.WASIVRS <- ggplot(data = conc.orth.subjs, aes(x = WASIVRS)) +
32   geom_histogram(binwidth = 2) +
33   labs(x = "WASI vocabulary") +
34   theme_bw()
35
```

```
36 p.BPVRS <- ggplot(data = conc.orth.subjs, aes(x = BPVRS)) +
37   geom_histogram(binwidth = 3) +
38   labs(x = "BPVS vocabulary") +
39   theme_bw()
40
41 p.mean.score <- ggplot(data = conc.orth.subjs, aes(x = mean.score)) +
42   geom_histogram(binwidth = .25) +
43   labs(x = "Mean orthographic test score") +
44   theme_bw()
45
46 p.mean.score + p.BPVRS + p.WASIvRS + p.WASImRS +
47   p.CC2nwRS + p.CC2irregRS + p.CC2regRS +
48   p.TOWREpdeRS + p.TOWREsweRS + plot_layout(ncol = 3)
```

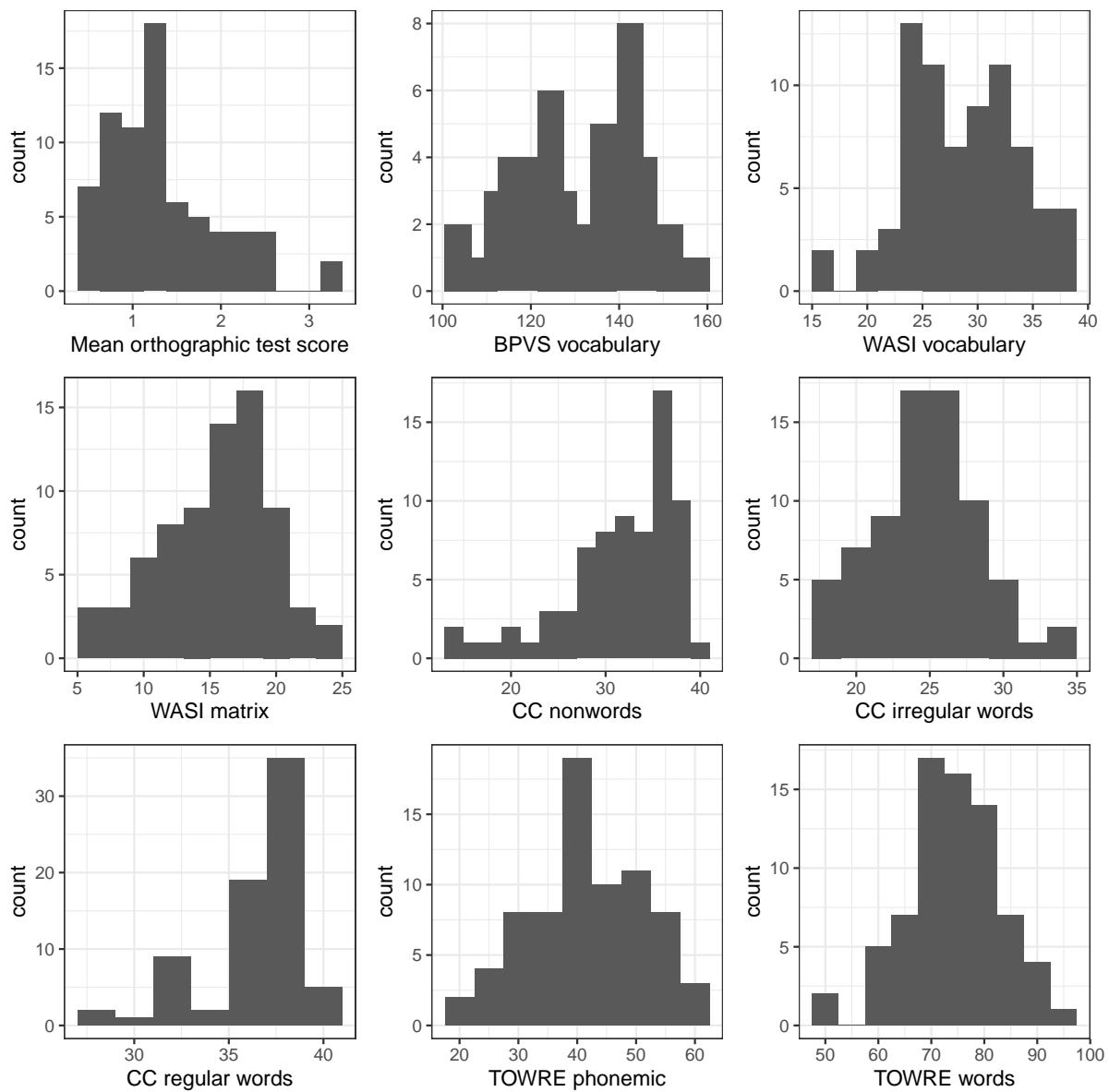


Figure 2.7: Distribution of childrens' scores on ability measures

This is how the code works, step by step:

1. `p.WASImRS <- ggplot(...)` first creates a plot object, which we call `p.WASImRS`.
2. `ggplot(data = conc.orth.subjs, aes(x = WASImRS)) +` tells R what data to use, and what aesthetic mapping to work with mapping the variable `WASImRS` here to the x-axis location.

3. `geom_histogram(binwidth = 2)` + tells R to sort the values of WASImRS scores into bins and create a histogram to show how many children in the sample present scores of different sizes.
4. `labs(x = "WASI matrix")` + changes the x-axis label to make it more informative.
5. `theme_bw()` changes the theme to make it a bit cleaner looking.

We do this bit of code separately for each variable. We change the plot object name, the `x =` variable specification, and the axis label text for each variable. We adjust the binwidth where it appears to be necessary.

We then use the following plot code to put all the plots together in a single grid.

```
p.mean.score + p.BPVRS + p.WASIVRS + p.WASIMRS +
p.CC2nwRS + p.CC2irregRS + p.CC2regRS +
p.TOWREpdeRS + p.TOWREsweRS + plot_layout(ncol = 3)
```

- In the code, we add a series of plots together e.g. `p.mean.score + p.BPVRS + p.WASIVRS ...`
- and then specify we want a grid of plots with a layout of three columns `plot_layout(ncol = 3)`.

This syntax requires the `library(patchwork)` and more information about this very useful library can be found [here](#).

What do the plots show us?

Figure 2.7 shows a grid of 9 histogram plots. Each plot presents the distribution of scores for the Ricketts et al. (2021) Study 2 participant sample on a separate ability measure, including scores on the BPVS vocabulary, WASI vocabulary, TOWRE words and TOWRE nonwords reading tests, as well as scores on the Castles and Coltheart regular words, irregular words and nonwords reading tests, and the mean Levenshtein distance (spelling score) outcome measure of performance for the experimental word learning post-test.

Take a look, you may notice the following features.

1. The mean orthographic test score suggests that many children produced spellings to the words they learned in the Ricketts et al. (2021) study that, on average, were correct (0 edits) or were one or two edits (e.g., a letter deletion or replacement) away from the target word spelling. The children were learning the words, and most of the time, they learned the spellings of the words effectively. However, one or two children tended to produce spellings that were 2-3 edits distant from the target spelling.
- We can see these features because we can see that the histogram peaks around 1 (at Levenshtein distance score = 1) but that there is a small bar of scores at around 3.

2. We can see that there are two peaks on the BPVS and WASI measures of vocabulary. What is going on there?
 - Is it the case that we have two sub-groups of children within the overall sample? For example, on the BPVS test, maybe one sub-group of children has a distribution of vocabulary scores with a peak around 120 (the peak shows where most children have scores) while another sub-group of children has a distribution of vocabulary scores with a peak around 140.
3. If we look at the CC nonwords and CC regular words tests of reading ability, we may notice that while most children present relatively high scores on these tests (CC nonwords peak around 35, CC regular words peak around 37) there is a skewed distribution. Many of the children's scores are piled up towards the maximum value in the data on the measures. But we can also see that, on both measures, there are long tails in the distributions because relatively small numbers of children have substantially lower scores.
 - Developmental samples are often highly varied (just like clinical samples). Are all the children in the sample at the same developmental stage, or are they all typically developing?

 Tip

Notice that in presenting a grid of plots like this, we offer a compact visual way to present the same summary information we might otherwise present using a table of descriptive statistics. In some ways, this grid of plots is more informative than the descriptive statistics because the mean and SD values do not tell you what you can see:

- the characteristics of the variation in values, like the presence of two peaks;
- or the presence of unusually high or low scores (for this sample).

Grids of plots like this can be helpful to inspect the distributions of variables in a concise approach. They are not really too useful for *comparing* the distributions because they require your eyes to move between plots, repeatedly, to do the comparison.

Here is a more compact way to code the grid of histograms using the `library(ggridges)` function `geom_density_ridges()`. I do not discuss it in detail because I want to focus your attention on core `tidyverse` functions (I show you more information in the **Notes** tab).

Notice that if you produce all the plots so that they are in line in the same column with a shared x-axis it becomes *much easier* to compare the distributions of scores. You lose some of the fine detail, discussed in relation to Figure 2.7, but this style allows you to gain an impression, quickly, of how distributions of scores compare between measures. For example, we can see that within the Castles and Coltheart (CC) measures of reading ability, children do better on regular words than on nonwords, and on nonwords better than on irregular words.

2.7.7.2 Plot

```
1 library(ggridges)
2 conc.orth.subjs %>%
3   pivot_longer(names_to = "task", values_to = "score", cols = WASIMRS$mean.score) %>%
4   ggplot(aes(y = task, x = score)) +
5   geom_density_ridges(stat = "binline", bins = 20, scale = 0.95, draw_baseline = FALSE) +
6   theme_ridges()
```



Figure 2.8: Distribution of childrens' scores on ability measures

2.7.7.3 Notes

1. `library(gggridges)` get the library we need.
2. `conc.orth.subjs %>%` pipe the dataset for processing.
3. `pivot_longer(names_to = "task", values_to = "score", cols = WASImRS:mean.score) %>%` pivot the data so all test scores are in the same column, “scores” with coding for “task” name, and pipe to the next step for plotting.
4. `ggplot(aes(y = task, x = score)) +` create a plot for the scores on each task.
5. `geom_density_ridges(stat = "binline", bins = 20, scale = 0.95, draw_baseline = FALSE) +` show the plots as histograms.
6. `theme_ridges()` change the theme to the specific theme suitable for showing a grid of ridges.

You can find more information on `gggridges` [here](#).

2.7.7.4 Compare between groups how values vary on different variables

We will often want to compare the distributions of variable values between groups or between conditions. This need may appear when, for example, we are conducting a between-groups manipulation of some condition and we want to check that the groups are approximately matched on dimensions that are potentially linked to outcomes (i.e., on potential *confounds*). The need may appear when, alternatively, we have recruited or selected participant (or stimulus) samples and we want to check that the sample sub-groups are approximately matched or detectably different on one or more dimensions of interest or of concern.

As a demonstration of the visualization work we can do in such contexts, let’s pick up on an observation we made earlier, that there are two peaks on the BPVS and WASI measures of vocabulary. I asked: Is it the case that we have two sub-groups of children within the overall sample? Actually, we know the answer to that question because Ricketts et al. (2021) state that they recruited one set of children for their Study 1 and then, for Study 2:

Thirty-three children from an additional three socially mixed schools in the South-East of England were added to the Study 1 sample (total N = 74). These additional children were older ($M_{age} = 12.57$, SD = 0.29, 17 female)

Do the younger (Study 1) children differ in any way from the older (additional) children?

We can check this through data visualization. Our aim is to present the distributions of variables side-by-side or *superimposed* to ensure easy comparison. We can do this in different ways, so I will demonstrate one approach with an outline explanation of the actions, and offer suggestions for further approaches.

I am going to process the data before I do the plotting. I will re-use the code I used before (see Section 2.7.4.2) with one additional change. I will add a line to create a group coding

variable. This addition shows you how to do an action that is *very often* useful in the data processing part of your workflow.

2.7.7.4.1 Data processing

You have seen that the Ricketts et al. (2021) report states that an additional group of children was recruited for the investigation's second study. How do we know who they are? If you recall the summary view of the complete dataset, there is one variable we can use to code group identity.

```
summary(conc.orth$Study)
```

Study1&2	Study2
655	512

This summary tells us that we have 512 observations concerning the additional group of children recruited for Study 2, and 655 observations for the (younger) children whose data were analyzed for both Study 1 and Study 2 (i.e., coded as Study1&2 in the Study variable column). We can use this information to create a coding variable. (If we had age data, we could use that instead but we do not.) This is how we do that.

```
conc.orth.subjs <- conc.orth %>%
  group_by(Participant) %>%
  mutate(mean.score = mean(Levenshtein.Score)) %>%
  ungroup() %>%
  distinct(Participant, .keep_all = TRUE) %>%
  mutate(age.group = fct_recode(Study,
    "young" = "Study1&2",
    "old" = "Study2"
  )) %>%
  select(WASImRS:BPVSRS, mean.score, Participant, age.group)
```

The code block is mostly the same as the code I used in Section Section 2.7.4.2 to extract the data for each participant, with two changes:

1. First, `mutate(age.group = fct_recode(...))` tells R that I want to create a new variable `age.group` through the process of recoding, with `fct_recode(...)` the variable I specify next, in the way that I specify.
2. `fct_recode(Study, ...)` tells R I want to recode the variable `Study`.
3. `"young" = "Study1&2", "old" = "Study2"` specifies what I want recoded.

- I am telling R to look in the `Study` column and (a.) whenever it finds the value `Study1&2` replace it with `young` whereas (b.) whenever it finds the value `Study2` replace it with `old`.
 - Notice that the syntax in recoding is `fct_recode`: “new name” = “old name”.
 - Having done that, I tell R to pipe the data, including the recoded variable, to the next step.
4. `select(WASImsRS:BPVSRs, mean.score, Participant, age.group)` where I add the new recoded variable to the selection of variables I want to include in the new dataset `conc.orth.subjs`.

💡 Tip

Notice that R handles categorical or nominal variables like `Study` (or, in other data, variables e.g. gender, education or ethnicity) as *factors*.

- Within a classification scheme like education, we may have different classes or categories or groups e.g. “further, higher, school”. We can code these different classes with numbers (e.g. `school = 1`) or with words “further, higher, school”. Whatever we use, the different classes or groups are referred to as *levels* and each level has a name.
- In factor recoding, we are *changing level names* while keeping the underlying data the same.

The `tidyverse` collection includes the `forcats` library of functions for working with categorical variables (`forcats = factors`). These functions are often very useful and you can read more about them [here](#).

Changing factors level coding by hand is, for many, a common task, and the `fct_recode()` function makes it easy. You can find the technical information on the function, with further examples, [here](#).

2.7.7.4.2 Group comparison visualization

There are different ways to examine the distributions of variables *so that* we can compare the distributions of the same variable between groups.

Figure 2.9 presents some alternatives as a grid of 4 different kinds of plots designed to enable the same comparison. Each plot presents the distribution of scores for the Ricketts et al. (2021) Study 2 participant sample on the BPVS vocabulary measure so that we can compare the distribution of vocabulary scores between age groups.

The plots differ in method using:

- a. faceted histograms showing the distribution of vocabulary scores, separately for each group, in side-by-side histograms for comparison;
- b. boxplots, showing the distribution of scores for each group, indicated by the y-axis locations of the edges of the boxes (25% and 75% quartiles) and the middle lines (medians);
- c. superimposed histograms, where the histograms for the separate groups are laid on top of each other but given different colours to allow comparison; and
- d. superimposed density plots where the densities for the separate groups are laid on top of each other but given different colours to allow comparison.

 Tip

There is one thing you should notice about all these plots.

- It looks like the BPVS vocabulary scores have their peak – most children show this value – at around 120 for the **young** group and at around 140 for the **old** group.
- We return to this shortly.

I am going to hide the coding and the explanation of the coding behind the **Notes** tab. Click on the tab to get a step-by-step explanation. Of these alternatives, I focus on one which I explain in more depth, following: d. Superimposed density plots.

2.7.7.5 Plot

2.7.7.6 Notes

```
p.facet.hist <- ggplot(data = conc.orth.subjs, aes(x = BPVSRS)) +
  geom_histogram(binwidth = 5) +
  labs(x = "BPVS vocabulary score", title = "a. Faceted histograms") +
  facet_wrap(~ age.group) +
  theme_bw()

p.colour.boxplot <- ggplot(data = conc.orth.subjs, aes(y = BPVSRS, colour = age.group)) +
  geom_boxplot() +
  labs(x = "BPVS vocabulary score", title = "b. Boxplots") +
  theme_bw()

p.colour.hist <- ggplot(data = conc.orth.subjs, aes(x = BPVSRS, colour = age.group, fill =
```

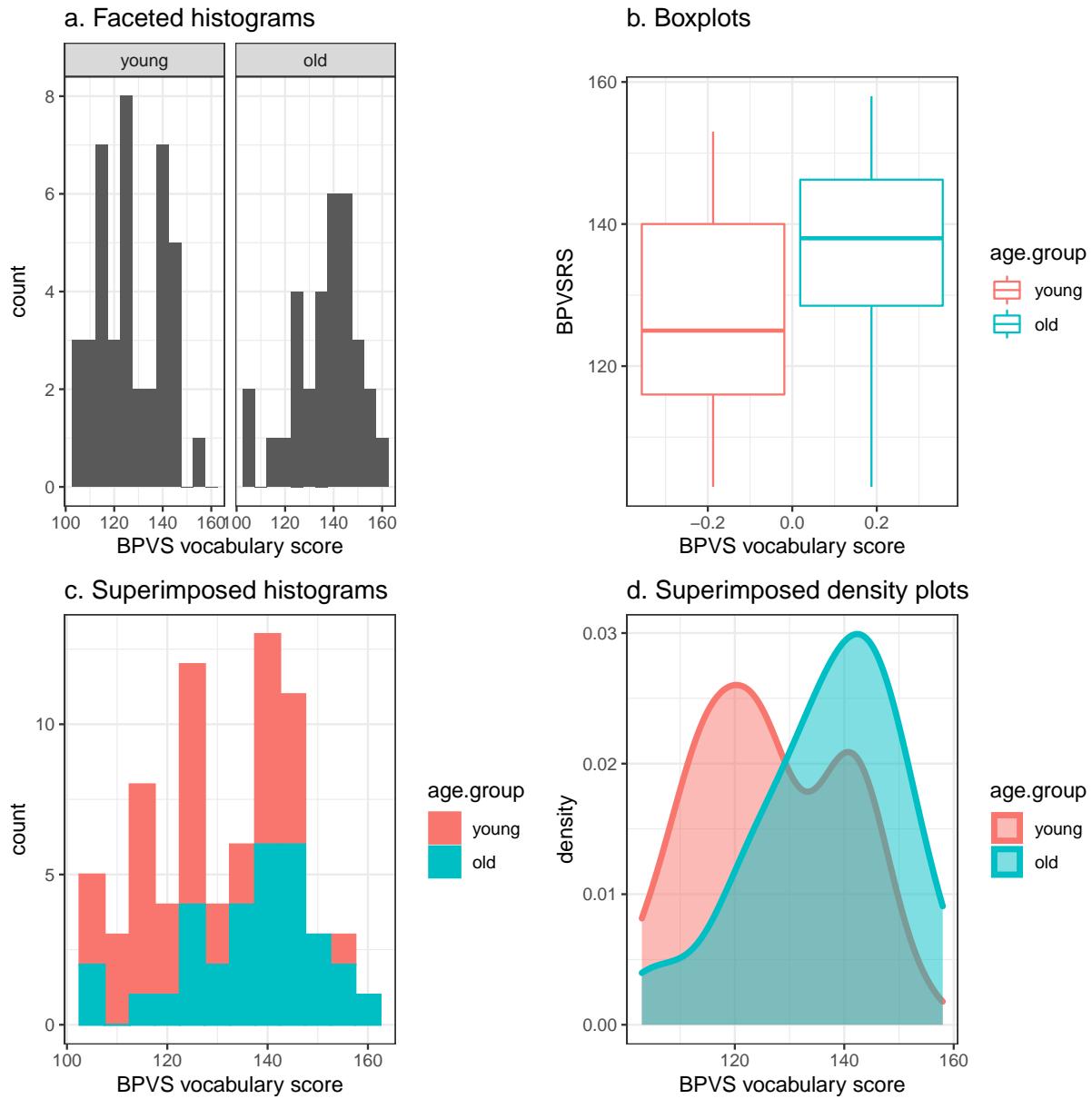


Figure 2.9: Distribution of childrens' scores on the BPVS vocabulary measure: distributions are compared between the younger and older age groups

```

p.colour.density <- ggplot(data = conc.orth.subjs, aes(x = BPVRS, colour = age.group, fill = age.group) +
  geom_density(alpha = .5, size = 1.5) +
  labs(x = "BPVS vocabulary score", title = "d. Superimposed density plots") +
  theme_bw()

p.facet.hist + p.colour.boxplot + p.colour.hist + p.colour.density

```

1. In plot “a. Faceted histograms”, we use the code to construct a histogram but the difference is we use:
 - `facet_wrap(~ age.group)` to tell R to split the data by `age.group` then present the histograms indicating vocabulary score distributions *separately* for each group.
2. In plot “b. Boxplots”, we use the `geom_boxplot()` code to construct a boxplot to summarize the distributions of vocabulary scores – as you have seen previously – but the difference is we use:
 - `aes(y = BPVRS, colour = age.group)` to tell R to assign different colours to different levels of `age.group` to help distinguish the data from each group.
3. In plot “c. Superimposed histograms”, we use the code to construct a histogram but the difference is we use:
 - `aes(x = BPVRS, colour = age.group, fill = age.group)` to tell R to assign different colours to different levels of `age.group` to help distinguish the data from each group.
 - Notice that the `fill` gives the colour inside the bars and `colour` gives the colour of the outline edges of the bars.
4. In plot “d. Superimposed density plots”, we use the code `geom_density(...)` to construct what is called a density plot.
 - A density plot presents a smoothed histogram to show the distribution of variable values.
 - We add arguments in `geom_density(alpha = .5, size = 1.5)` to adjust the thickness of the line (`size = 1.5`) drawn to show the shape of the distribution and adjust the transparency of the colour fill inside the line `alpha = .5`.
 - We use `aes(x = BPVRS, colour = age.group, fill = age.group)` to tell R to assign different colours to different levels of `age.group` to help distinguish the data from each group.
 - Notice that the `fill` gives the colour inside the density plots and `colour` gives the colour of the outline edges of the densities.

Density plots can be helpful when we wish to compare distributions. This is because we can *superimpose* distribution plots on top of each other, enabling us or our audience to directly compare the distributions: *directly* because the distributions are shown on the same scale, in the same image.

We can (roughly) understand a density plot as working like a smoothed version of the histogram. Imagine how the heights of the bars in the histogram represent how many observations we have of the values in a particular bin. If we draw a smooth curving line through the tops of the bars then we are representing the chances that an observation in our sample has a value (the value under the curve) at any specific location on the x-axis. You can see that in Figure 2.10.

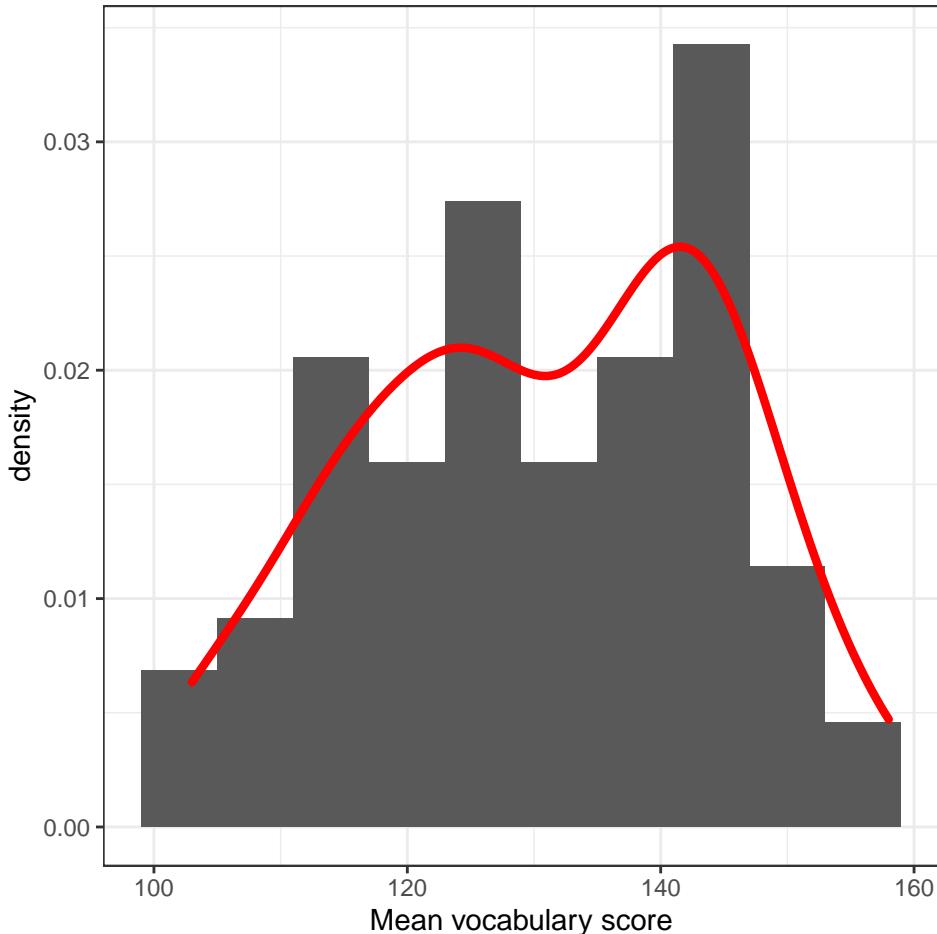


Figure 2.10: Distribution of childrens' scores on the BPVS vocabulary measure. The figure shows the histogram versus density plot representation of the same data distribution

You can find the `ggplot2` reference information on the `geom_density()` function, with further examples, [here](#). You can find technical information on density functions [here](#) and [here](#).

We can develop the density plot to enrich the information we can discover or communicate through the plot. Figure 2.11 shows the distribution of scores on both the BPVS and WASI vocabulary knowledge measures.

```
1 p.BPVRS.density <- ggplot(data = conc.orth.subjs, aes(x = BPVRS, colour = age.group, fill = age.group) +  
2   geom_density(alpha = .5, size = 1.5) +  
3   geom_rug(alpha = .5) +  
4   geom_vline(xintercept = 120, linetype = "dashed") +  
5   geom_vline(xintercept = 140, linetype = "dotted") +  
6   labs(x = "BPVRS vocabulary score") +  
7   theme_bw()  
8  
9 p.WASIVRS.density <- ggplot(data = conc.orth.subjs, aes(x = WASIVRS, colour = age.group, fill = age.group) +  
10  geom_density(alpha = .5, size = 1.5) +  
11  geom_rug(alpha = .5) +  
12  labs(x = "WASI vocabulary score") +  
13  theme_bw()  
14  
15 p.BPVRS.density + p.WASIVRS.density + plot_layout(guides = 'collect')
```

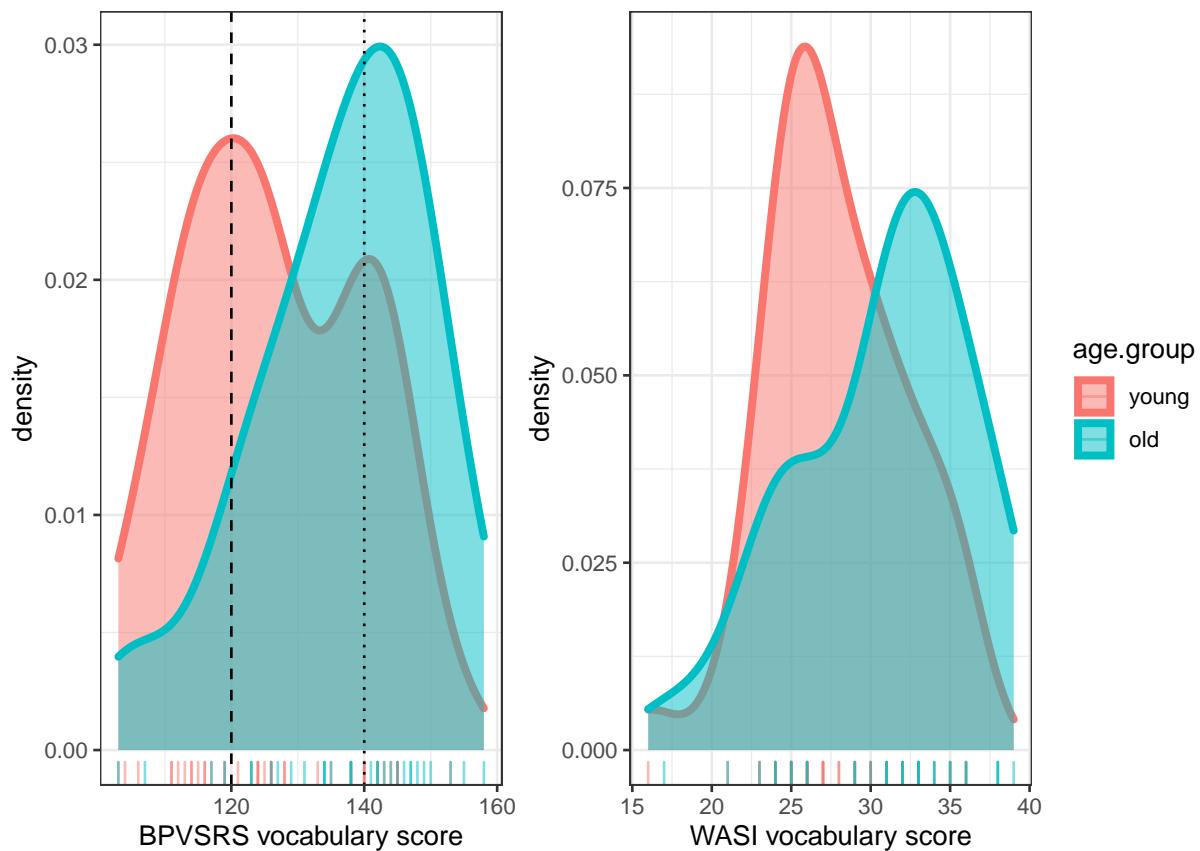


Figure 2.11: Distribution of childrens' scores on the BPVS and WASI vocabulary measures.

Here is what the code does:

1. `p.BPVRS.density <- ggplot(...)` creates a plot object called `p.BPVRS.density`.
2. `data = conc.orth.subjs, ...` says we use the `conc.orth.subjs` dataset to do this.
3. `aes(x = BPVRS, colour = age.group, fill = age.group)) +` says we want to map BPVRS scores to x-axis location, and `age.group` level coding (`young`, `old`) to both `colour` and `fill`.
4. `geom_density(alpha = .5, size = 1.5) +` draws a density plot; note that we said earlier what we want for `colour` and `fill` but here we also say that:
 - `alpha = .5` we want the fill to be transparent;
 - `size = 1.5` we want the density curve line to be thicker than usual.
5. `geom_rug(alpha = .5) +` adds a one-dimensional plot, a series of tick marks, to show where we have observations of BPVRS scores for specific children. We ask R to make the tick marks semi-transparent.

6. `geom_vline(xintercept = 120, linetype = "dashed")` + draws a vertical dashed line where BPVRS = 120.
7. `geom_vline(xintercept = 140, linetype = "dotted")` + draws a vertical dotted line where BPVRS = 140.
8. `labs(x = "BPVS vocabulary score")` + makes the x-axis label something understandable to someone who does not know about the study.
9. `theme_bw()` changes the theme.

2.7.7.6.1 Critical evaluation: discovery and communication

As we work with visualization, we should aim to develop skills in reading plots, so:

- What do we see?

When we look at Figure 2.11, we can see that the younger and older children in the Ricketts et al. (2021) sample have broadly overlapping distributions of vocabulary scores. However, as we have noticed previously, the peak of the distribution is a bit lower for the younger children compared to the older children. This appears to be the case whether we are looking at the BPVS or at the WASI measures of vocabulary, suggesting that the observation does not depend on the particular vocabulary test. Is this observation unexpected? Probably not, as we should hope to see vocabulary knowledge increase as children get older. Is this observation a problem for our analysis? You need to read the paper to find out what we decided.

2.7.7.6.2 Exercise

In the demonstration examples, I focused on comparing age groups on vocabulary, what about the other measures?

I used superimposed density plots: are other plotting styles more effective, for you? Try using boxplots or superimposed or faceted histograms instead.

2.7.8 Summary: Visualizing distributions

So far, we have looked at how and why we may examine the distributions of numeric variables. We have used histograms to visualize the distribution of variable values. We have explored the construction of grids of plots to enable the quick examination or concise communication of information about the distributions of multiple variables at the same time. And we have used histograms, boxplots and density plots to examine how the distributions of variables may differ between groups.

The comparison of the distributions of variable values in different groups (or, similarly, between different conditions) may be the kind of work we would need to do, in data visualization, as part of an analysis ending in, for example, a t-test comparison of mean values.

While boxplots, density plots and histograms are typically used to examine how the values of a numeric variable vary, scatterplots are typically used when we wish to examine, to make sense of or communicate potential associations or relations between two (or more) numeric variables. We turn to scatterplots, next.

2.7.9 Examine the associations between numeric variables

Many of us start learning about scatterplots in high school math classes. Using the modern tools made available to us through the `ggplot2` library (as part of `tidyverse`), we can produce effective, nice-looking, scatterplots for a range of discovery or communication scenarios.

We continue working with the Ricketts et al. (2021) dataset. In the context of the Ricketts et al. (2021) investigation, there is interest in how children vary in the reading, spelling and vocabulary abilities that may influence the capacity of children to learn new words. So, in this context, we can begin to progress our development in visualization skills by usefully considering the potential association between participant attributes in the Study 2 sample.

Later on, we will look at more advanced plots that help us to communicate the impact of the experimental manipulations implemented by Ricketts et al. (2021), and also to discover the ways that these impacts may vary between children.

2.7.9.1 Getting started: Scatterplot basics

We can begin by asking a simple research question we can guess the answer to:

- Do vocabulary knowledge scores on two alternative measures, the BPVS and the WASI, relate to each other?

If two measurement instruments or tests are intended to measure individual differences in the same psychological attribute, here, vocabulary knowledge, then we would reasonably expect that scores on one test should covary with scores on the second test.

```
1 ggplot(data = conc.orth.subjs, aes(x = WASIvRS, y = BPVSR)) +
2   geom_point() +
3   labs(x = "WASI vocabulary score",
4        y = "BPVSR vocabulary score",
5        title = "Are WASI and BPVSR vocabulary scores associated?") +
6   theme_bw()
```

Are WASI and BPVS vocabulary scores associated?

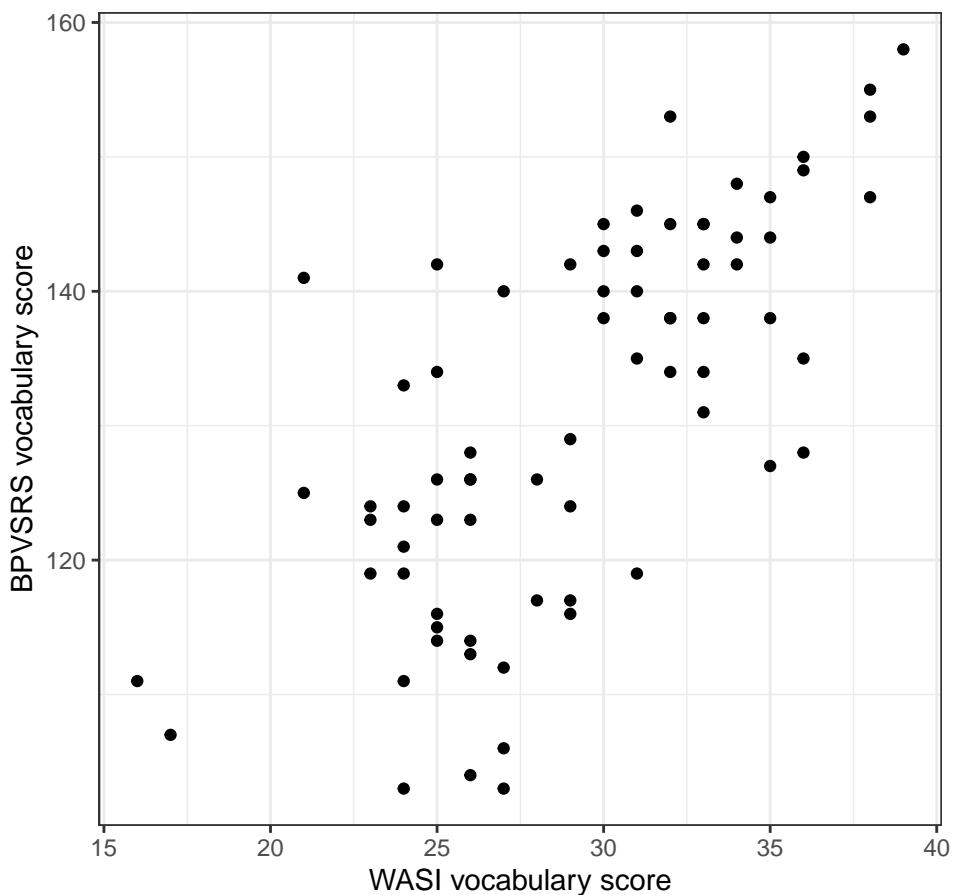


Figure 2.12: Scatterplot indicating the potential association of childrens' scores on the BPVS and WASI vocabulary measures.

What does the plot show us?

As a reminder of how scatterplots work, we can recall that they present integrated information. Each point, for the Ricketts et al. (2021) data, represents information about *both* the BPVS and the WASI score for each child.

- The vertical height of a point tells us the BPVS score recorded for a child: higher points represent higher scores.
- The left-to-right horizontal position of the same point tells us the WASI score for the same child: points located more on the right represent higher scores.

Figure 2.12 is a scatterplot comparing variation in childrens' scores on the BPVS and WASI vocabulary measures: variation in BPVS scores are shown on the y-axis and variation in WASI

scores are shown on the x-axis. Critically, the scientific insight the plot gives us is this: higher WASI scores are associated with higher BPVS scores.

How does the code work? We have seen scatterplots before but, to ensure we are comfortable with the coding, we can go through them step by step.

1. `ggplot(data = conc.orth.subjs...) +` tells R we want to produce a plot using `ggplot()` with the `conc.orth.subjs` dataset.
2. `aes(x = WASIvRS, y = BPVSRs)` tells R that, in the plot, WASIvRS values are mapped to x-axis (horizontal) position and BPVSRs values are mapped to y-axis (vertical) position.
3. `geom_point()` + constructs a scatterplot, using these data and these position mappings.
4. `labs(x = "WASI vocabulary score", ...)` fixes the x-axis label.
5. `y = "BPVSRs vocabulary score", ...)` fixes the y-axis label.
6. `title = "Are WASI and BPVS vocabulary scores associated?"` + fixes the title.
7. `theme_bw()` changes the theme.

2.7.9.2 Building complexity: adding information step by step

For this pair of variables in this dataset, the potential association in the variation of scores is quite obvious. However, sometimes it is helpful to guide the audience by imposing a *smoother*. There are different ways to do this, for different objectives and in different contexts. Here, we look at two different approaches. In addition, as we go, we examine how to adjust the appearance of the plot to address different potential discovery or communication needs.

We begin by adding what is called a LOESS smoother.

```
1 ggplot(data = conc.orth.subjs, aes(x = WASIvRS, y = BPVSRs)) +
2   geom_point() +
3   geom_smooth() +
4   labs(x = "WASI vocabulary score",
5        y = "BPVSRs vocabulary score",
6        title = "Are WASI and BPVS vocabulary scores associated?") +
7   theme_bw()
```

Are WASI and BPVS vocabulary scores associated?

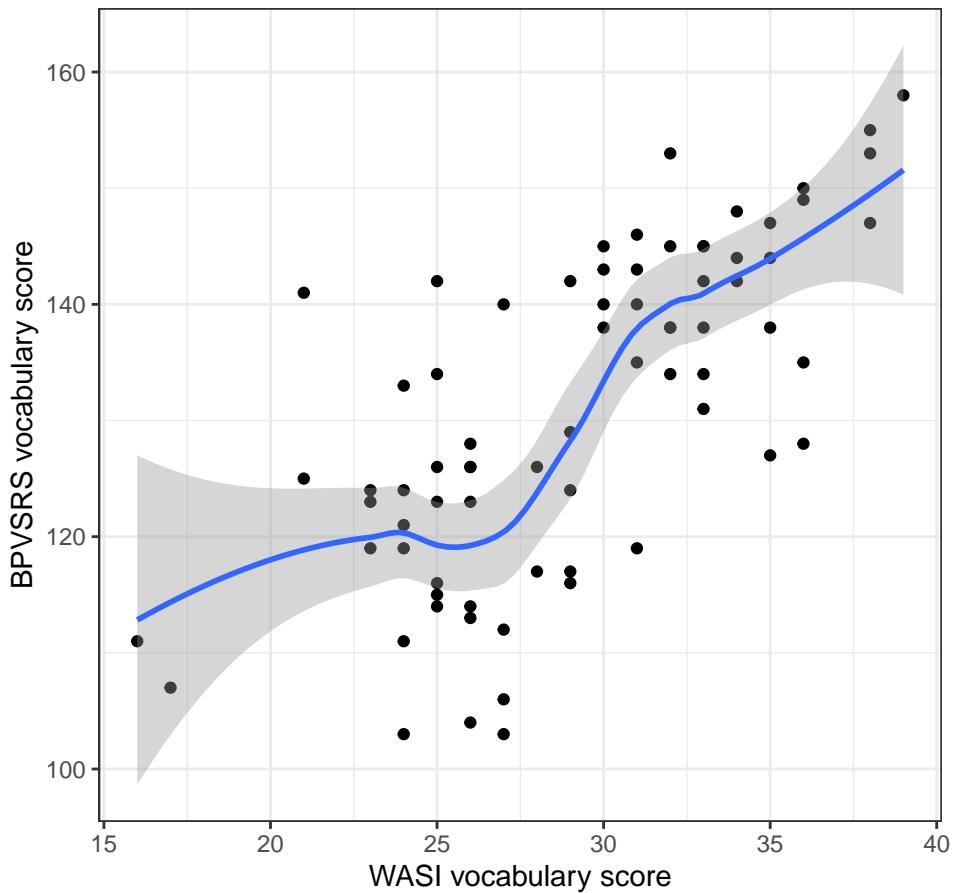


Figure 2.13: Scatterplot indicating the potential association of childrens' scores on the BPVS and WASI vocabulary measures.

The only coding difference between this plot Figure 2.13 and the previous plot Figure 2.12 appears at line 3:

- `geom_smooth()`

The addition of this bit of code results in the addition of the curving line you see in Figure 2.13. The blue line is curving, and visually suggests that the relation between BPVS and WASI scores is different – sometimes more sometimes less steep – for different values of WASI vocabulary score.

This line is generated by the `geom_smooth()` code, by default, in an approach in which the dataset is effectively split into sub-sets, dividing the data up into sub-sets from the lowest to the highest WASI scores, and the predicted association between the y-axis variable (here,

BPVS score) and the x-axis variable (here, WASI score) is calculated bit by bit, in a series of regression analyses, working in order through sub-sets of the data. This calculation of what is called the LOESS (locally estimated scatterplot smoothing) trend is done by `ggplot` for us. And this approach to visualizing the trend in a potential association between variables is often a helpful way to discover curved or non-linear relations.

You can find technical information on `geom_smooth()` [here](#) and an explanation of LOESS [here](#).

For us, this default visualization is not helpful for two reasons:

1. We have not yet learned about linear models, so learning about LOESS comes a bit early in our development.
2. It is hard to look at Figure 2.13 and identify a convincing curvilinear relation between the two variables. A lot of the curve for low WASI scores appears to be linked to the presence of a small number of data points.

At this stage, it is more helpful to adjust the addition of the smoother. We can do that by adding an argument to the `geom_smooth()` function code.

```
1 ggplot(data = conc.orth.subjs, aes(x = WASIvRS, y = BPVSR)) +
2   geom_point() +
3   geom_smooth(method = 'lm') +
4   labs(x = "WASI vocabulary score",
5        y = "BPVSR vocabulary score",
6        title = "Are WASI and BPVSR vocabulary scores associated?") +
7   theme_bw()
```

Are WASI and BPVS vocabulary scores associated?

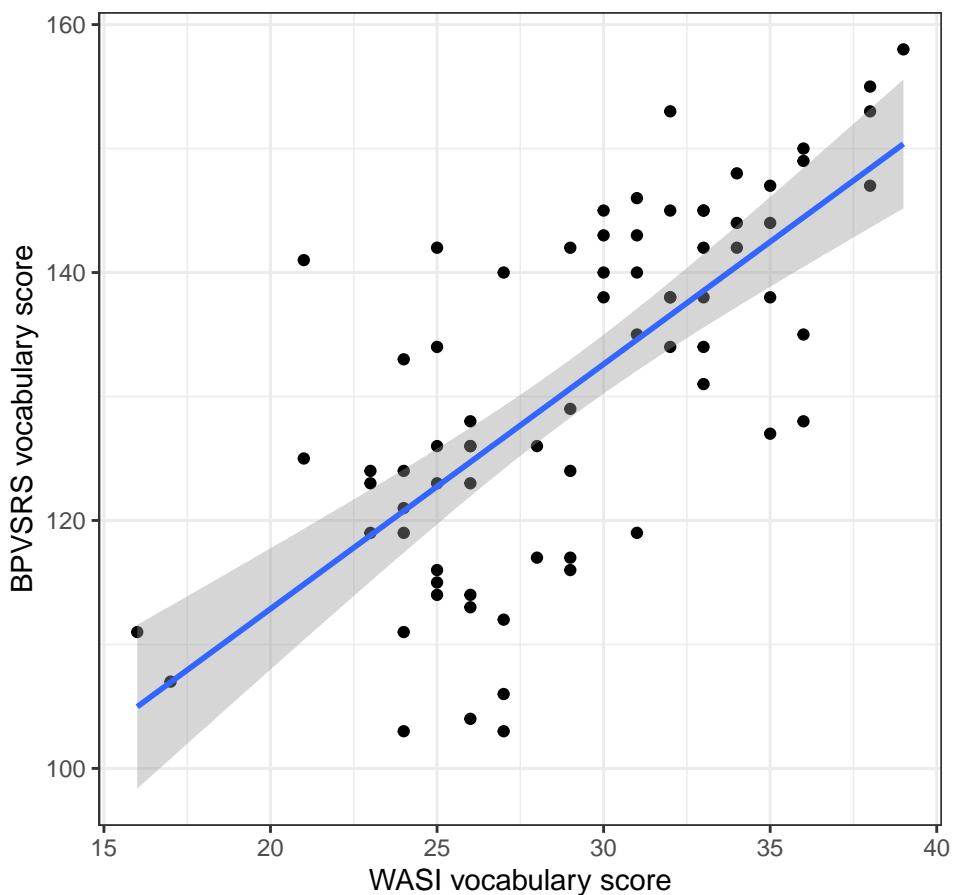


Figure 2.14: Scatterplot indicating the potential association of childrens' scores on the BPVS and WASI vocabulary measures.

Notice the difference between Figure 2.13 and Figure 2.14:

- `geom_smooth(method = 'lm')` tells R to draw a trend line, a smoother, using the `lm` method.

The `lm` method requires R to estimate the association between the two variables, here, BPVS and WASI, assuming a linear model. Of course, we are going to learn about linear models but, in short, right now, what we need to know is that we assume a “straight line” relationship between the variables. This assumption requires that for any interval of WASI scores – e.g., whether we are talking about WASI scores between 20-25 or about WASI scores between 30-35 – the relation between BPVS and WASI scores has the same shape: the direction and steepness of the slope of the line is the same.

2.7.9.3 Exercise

Advice

Developing skill in working with data visualizations is not just about developing coding skills, it is also about developing skills in reading, and critically evaluating, the information the plots we produce show us.

Stop and take a good look at the scatterplot in Figure 2.14. Use the visual representation of data to critically evaluate the potential association between the BPVS and WASI variables. What can you see?

You can train your critical evaluation by asking yourself questions like the following:

1. How does variation in the x-axis variable relate to variation in values of the y-axis variable?
 - We can see, here, that higher WASI scores are associated with higher BPVS scores.
2. How strong is the relation?
 - The strength of the relation can be indicated by the steepness of the trend indicated by the smoother, here, the blue line.
 - If you track the position of the line, you can see, for example, that going from a WASI score of 20 to a WASI score of 40 is associated with going from a BPVS score of a little over 110 to a BPVS score of about a 150.
 - That seems like a big difference.
3. How well does the trend we are looking at capture the data in our sample?
 - Here, we are concerned with how close the points are to the trend line.
 - If the trend line represents a set of predictions about how the BPVS scores vary (in height) given variation in WASI scores, we can see that in places the prediction is not very good.
 - Take a look at the points located at WASI 25. We can see that there are points indicating that different children have the same WASI score of 25 but BPVS scores ranging from about 115 to 140.

2.7.9.4 Polish the appearance of a plot for presentation

Figure 2.14 presents a satisfactory looking plot but it is worth checking what edits we can make to the appearance of the plot, to indicate some of the ways that you can exercise choice in determining what a plot looks like. This will be helpful to you when you are constructing plots for presentation and report and you want to ensure the plots are as effective as possible.

```

1 ggplot(data = conc.orth.subjs, aes(x = WASIvRS, y = BPVSRs)) +
2   geom_point(alpha = .5, size = 2) +
3   geom_smooth(method = 'lm', colour = "red", size = 1.5) +
4   labs(x = "WASI vocabulary score",
5        y = "BPVSRs vocabulary score",
6        title = "Are WASI and BPVSRs vocabulary scores associated?") +
7   xlim(0, 40) + ylim(0, 160) +
8   theme_bw()

```

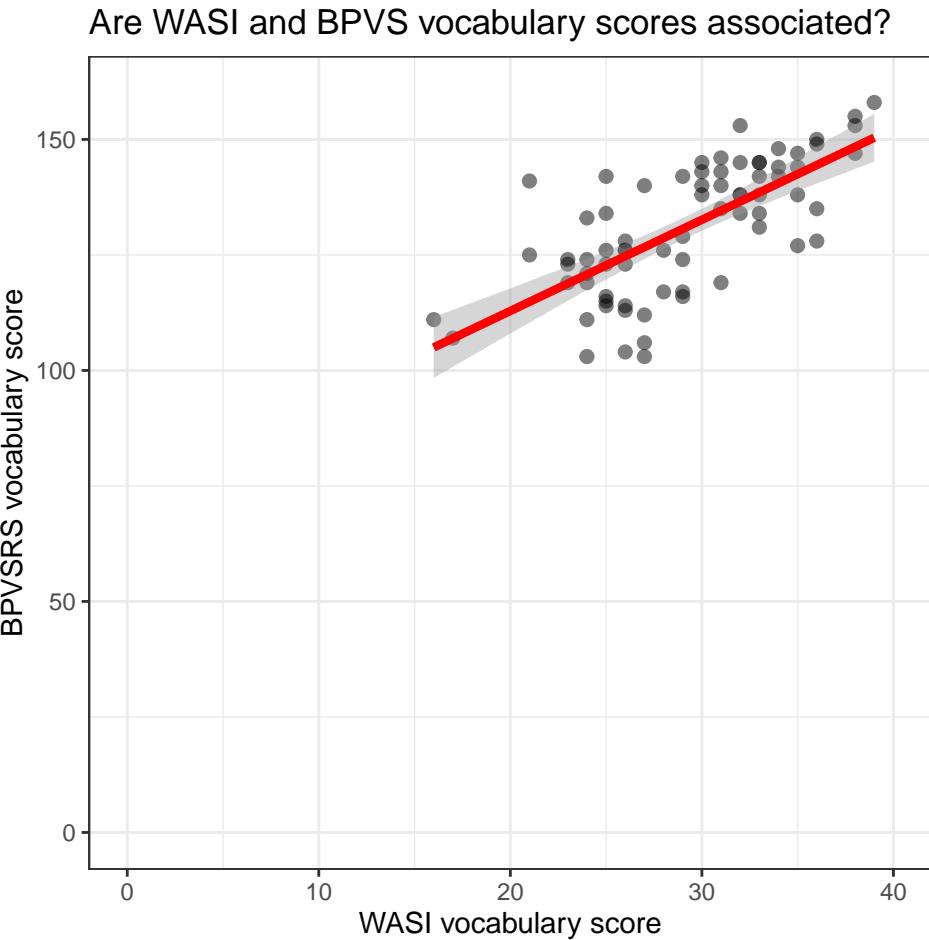


Figure 2.15: Scatterplot indicating the potential association of childrens' scores on the BPVS and WASI vocabulary measures.

If you inspect the code, you can see that I have made three changes:

1. `geom_point(alpha = .5, size = 2)` changes the `size` of the points and their transparency.

- parency (using `alpha`).
2. `geom_smooth(method = 'lm', colour = "red", size = 1.5)` change the colour of the smoother line, and the thickness (`size`) of the line.
 3. `xlim(0, 40) + ylim(0, 160)` changes the axis limits.

The last step — changing the axis limits — reveals how the sample data can be understood in the context of possible scores on these ability measures. Children *could* get BPVS scores of 0 or WASI scores of 0. By showing the start of the axes we get a more realistic sense of how our sample compares to the possible ranges of scores we could see in the wider population of children. This perhaps offers a more honest or realistic visualization of the potential association between BPVS and WASI vocabulary scores.

2.7.9.5 Examining associations among multiple variables

As we have seen previously, we can construct a series of plots and present them all at once in a grid or lattice. Figure 2.16 presents just such a grid: of scatterplots, indicating a series of potential associations.

Let's suppose that we are primarily interested in what factors influence the extent to which children in the Ricketts et al. (2021) word learning experiment are able to correctly spell the target words they were given to learn. As explained earlier, in Section 2.7.2, Ricketts et al. (2021) examined the spellings produced by participant children in response to target words, counting how many string edits (i.e., letter deletions etc.) separated the spelling each child produced from the target spelling they should have produced.

We can calculate the mean spelling accuracy score for each child, over all the target words we observed their response to. We can identify mean spelling score as the *outcome variable*. We can then examine whether the outcome spelling scores are or are not influenced by participant attributes like vocabulary knowledge.

Figure 2.16 presents a grid of scatterplots indicating the potential association between mean spelling score and each of the variables we have in the `conc.orth` dataset, including the Castles and Coltheart (CC) and TOWRE measures of word or nonword reading skill, WASI and BPVS measures of vocabulary knowledge, and the WASI matrix measure of intelligence, as well as (our newly coded) age group factor.

I hide an explanation of the coding behind the Notes tab, because we have seen how to produce grids of plots, but you can take a look if you want to learn how the plot is produced.

2.7.9.6 Plot

2.7.9.7 Notes

The code to produce the figure is set out as follows.

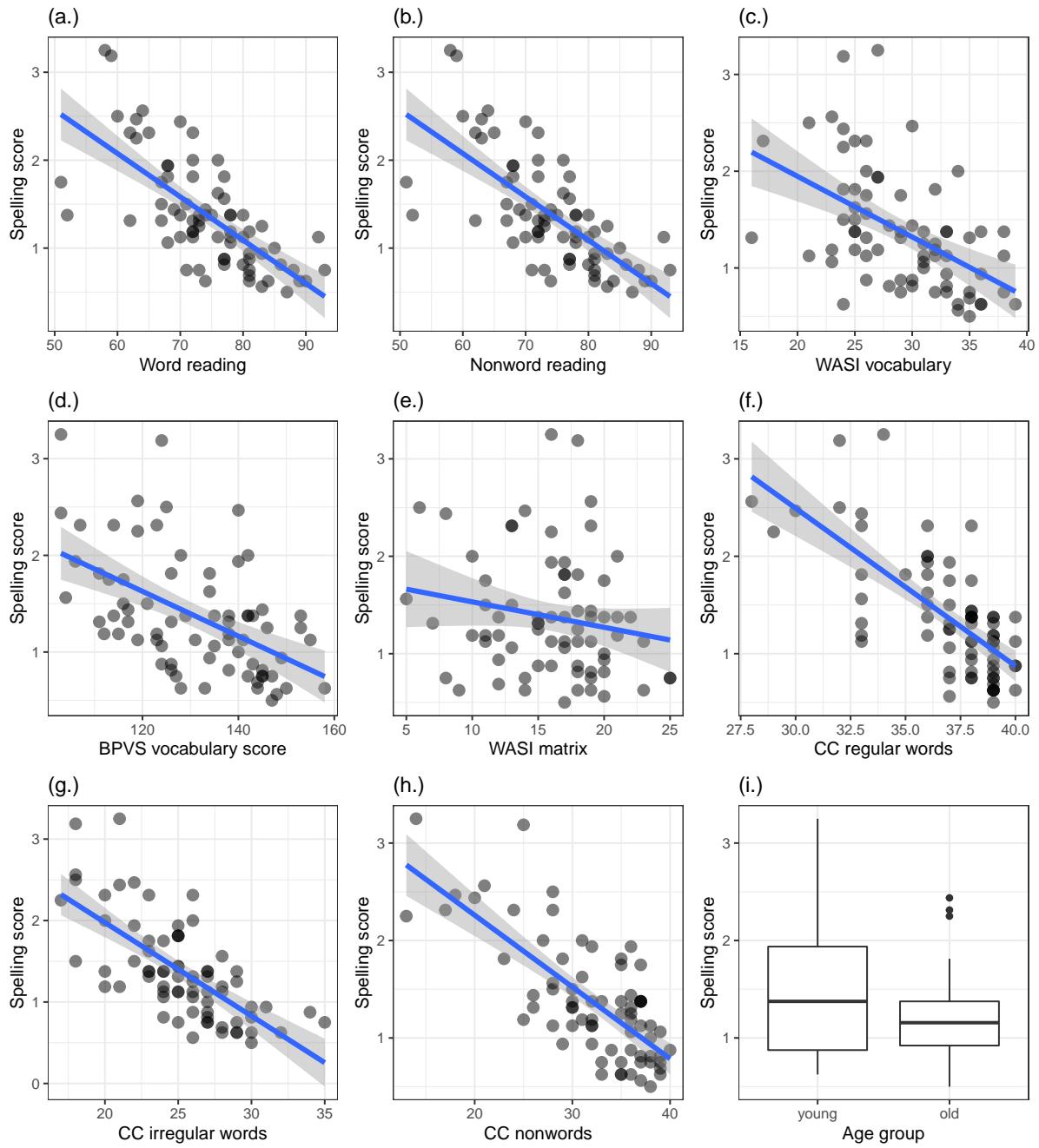


Figure 2.16: Grid of scatterplots showing the potential association between mean spelling score, for each child, and variation in the Castles and Coltheart (CC) and TOWRE measures of word or nonword reading skill, WASI and BPVS measures of vocabulary knowledge, the WASI matrix measure of intelligence, and age group factor

```

p.wordsvsmean.score <- ggplot(data = conc.orth.subjs,
                               aes(x = TOWREsweRS,
                                   y = mean.score)) +
  geom_point(alpha = .5, size = 3) +
  geom_smooth(method = 'lm', size = 1.5) +
  labs(x = "Word reading",
       y = "Spelling score",
       title = "(a.)") +
  theme_bw()

p.nonwordsvsmean.score <- ggplot(data = conc.orth.subjs,
                                   aes(x = TOWREsweRS,
                                       y = mean.score)) +
  geom_point(alpha = .5, size = 3) +
  geom_smooth(method = 'lm', size = 1.5) +
  labs(x = "Nonword reading",
       y = "Spelling score",
       title = "(b.)") +
  theme_bw()

p.WASIVRSvsmean.score <- ggplot(data = conc.orth.subjs,
                                   aes(x = WASIVRS,
                                       y = mean.score)) +
  geom_point(alpha = .5, size = 3) +
  geom_smooth(method = 'lm', size = 1.5) +
  labs(x = "WASI vocabulary",
       y = "Spelling score",
       title = "(c.)") +
  theme_bw()

p.BPVRSVsmean.score <- ggplot(data = conc.orth.subjs,
                                 aes(x = BPVRS,
                                     y = mean.score)) +
  geom_point(alpha = .5, size = 3) +
  geom_smooth(method = 'lm', size = 1.5) +
  labs(x = "BPVS vocabulary score",
       y = "Spelling score",
       title = "(d.)") +
  theme_bw()

p.WASImRSvsmean.score <- ggplot(data = conc.orth.subjs,

```

```

aes(x = WASImRS,
    y = mean.score)) +
geom_point(alpha = .5, size = 3) +
geom_smooth(method = 'lm', size = 1.5) +
labs(x = "WASI matrix",
    y = "Spelling score",
    title = "(e.)") +
theme_bw()

p.CC2regRSvsmean.score <- ggplot(data = conc.orth.subjs,
                                    aes(x = CC2regRS,
                                        y = mean.score)) +
geom_point(alpha = .5, size = 3) +
geom_smooth(method = 'lm', size = 1.5) +
labs(x = "CC regular words",
    y = "Spelling score",
    title = "(f.)") +
theme_bw()

p.CC2irregRSvsmean.score <- ggplot(data = conc.orth.subjs,
                                       aes(x = CC2irregRS,
                                           y = mean.score)) +
geom_point(alpha = .5, size = 3) +
geom_smooth(method = 'lm', size = 1.5) +
labs(x = "CC irregular words",
    y = "Spelling score",
    title = "(g.)") +
theme_bw()

p.CC2nwRSvsmean.score <- ggplot(data = conc.orth.subjs,
                                    aes(x = CC2nwRS,
                                        y = mean.score)) +
geom_point(alpha = .5, size = 3) +
geom_smooth(method = 'lm', size = 1.5) +
labs(x = "CC nonwords",
    y = "Spelling score",
    title = "(h.)") +
theme_bw()

p.age.groupvsmean.score <- ggplot(data = conc.orth.subjs,
                                    aes(x = age.group,

```

```

y = mean.score)) +
geom_boxplot() +
labs(x = "Age group",
y = "Spelling score",
title = "(i.)") +
theme_bw()

p.wordsVsmean.score + p.nonwordsVsmean.score + p.WASIvRSvsmean.score +
p.BPVSRVsmean.score + p.WASImRSvsmean.score + p.CC2regRSvsmean.score +
p.CC2irregRSvsmean.score + p.CC2nwRSvsmean.score + p.age.groupVsmean.score

```

1. To produce the grid of plots, we first create a series of plot objects using code like that shown in the chunk.

```

p.wordsVsmean.score <- ggplot(data = conc.orth.subjs,
                               aes(x = TOWREsweRS,
                                   y = mean.score)) +
  geom_point(alpha = .5, size = 3) +
  geom_smooth(method = 'lm', size = 1.5) +
  labs(x = "Word reading",
       y = "Spelling score",
       title = "(a.)") +
  theme_bw()

```

- `p.wordsVsmean.score <- ggplot(...)` creates the plot.
 - `data = conc.orth.subjs` tells R what data to work with.
 - `aes(x = TOWREsweRS, y = mean.score)` specifies the aesthetic data mappings.
 - `geom_point(alpha = .5, size = 3)` tells R to produce a scatterplot, specifying the size and transparency of the points.
 - `geom_smooth(method = 'lm', size = 1.5)` tells R to add a smoother, specifying the method and the thickness of the line.
 - `labs(x = "Word reading", y = "Spelling score", title = "(a.)")` fixes the labels.
 - `theme_bw()` adjusts the theme.
2. We then put the plots together, using the `patchwork` syntax where we list the plot objects by name, separating each name by a `+`.

```

p.BPVSRVsmean.score + p.WASImRSvsmean.score + p.CC2regRSvsmean.score +
p.CC2irregRSvsmean.score + p.CC2nwRSvsmean.score + p.age.groupVsmean.score

```

Figure 2.16 allows us to visually represent the potential association between an outcome measure, the average spelling score, and a series of other variables that may or may not have an influence on that outcome. Using a grid in this fashion allows us to compare the extent to which different variables appear to have an influence on the outcome. We can see, for example, that measures of variation in word reading skill appear to have stronger association (the trend lines are more steeply slowed) than measures of vocabulary knowledge or intelligence, or age group.

Using grids of plots like this allow us to compactly communicate these potential associations in a single figure.

 Warning

Levenshtein distance scores are higher *if* a child makes more errors in producing the letters in a spelling response.

- This means that if we want to see what factors help a child to learn a word, including its spelling, then we want to see that helpful factors are associated with *lower* Levenshtein scores.

2.7.10 Answering a scientific question: Visualize the effects of experimental conditions

As explained in Section 2.7.2, in the Ricketts et al. (2021) study, we taught children taught 16 novel words in a study with a 2×2 factorial design. The presence of orthography (orthography absent vs. orthography present) was manipulated within participants: for all children, eight of the words were taught with orthography (the word spelling) present and eight with orthography absent. Instructions (incidental vs. explicit) were manipulated between participants such that children in the explicit condition were alerted to the presence of orthography whereas children in the incidental condition were not. The Ricketts et al. (2021) investigation was primarily concerned with the effects on word learning of presenting words for learning with or without showing the words with their spellings, with or without instructing students explicitly that they would be helped by the presence of the spellings.

We can analyze the effects of orthography and instruction using a linear model.

```
model <- lm(Levenshtein.Score ~ Instructions*Orthography, data = conc.orth)
```

The model code estimates variation in spelling score (values of the `Levenshtein.Score`) variable, given variation in the levels of the `Instructions` and `Orthography` factors, and their interaction.

Table 2.1: Model summary

term	estimate	std.error	statistic	p.value
(Intercept)	1.584	0.072	21.857	0.000
Instructionsincidental	-0.041	0.103	-0.396	0.692
Orthographypresent	-0.409	0.103	-3.987	0.000
Instructionsincidental:Orthographypresent	0.060	0.146	0.409	0.683

This model is a *limited* approximation of the analysis we would need to do with these data to estimate the effects of orthography and instruction; see Ricketts et al. (2021) for more information on what analysis is required (in our view). However, it is good enough as a basis for exploring the kind of data visualization work — in terms of both discovery and communication — that you can do when you are working with data from an experimental study.

We can get a summary of the model results which presents the estimated effect of each experimental factor. These estimates represent the predicted change in spelling score, given variation in Orthography (present, absent) or Instruction (explicit, incidental), and given the possibility that the effect of the presence of orthography is different for different levels of instruction.

Notice that some of the p-values are incorrectly shown as 0.000. This is a result of using functions to automatically take a model summary and generate a table. I am going to leave this error with a warning because our focus is on visualization, next.

Very often, when we complete a statistical analysis of outcome data, in which we estimate or test the effects on outcomes of variation in some variables or of variation in experimental conditions, then we present a table summary of the analysis results. However, these estimates are typically difficult to interpret (it gets easier with practice) and talk about. Take a look at the summary table. We are often to focus on whether effects are significant or not significant. But, really, what we should consider is *how much* the outcome changes given the different experimental conditions.

How do we get that information from the analysis results? We can communicate results — to ourselves or to an audience — by constructing plots from the model information. The `ggeffects` library extends `ggplot2` to enable us to do this quite efficiently.

When we write code to fit a linear model like:

```
model <- lm(Levenshtein.Score ~ Instructions*Orthography, data = conc.orth)
```

We record the results as an object called `model` because we specify `model <- lm(...)`. We can take these results and ask R to create a plot showing predicted change in outcome (spelling) given our model. We can then present the effects of the variables, as shown in Figure 2.17.

```

1 dat <- ggpredict(model, terms = c("Instructions", "Orthography"))
2 plot(dat, facet = TRUE) + ylim(0, 3)

```

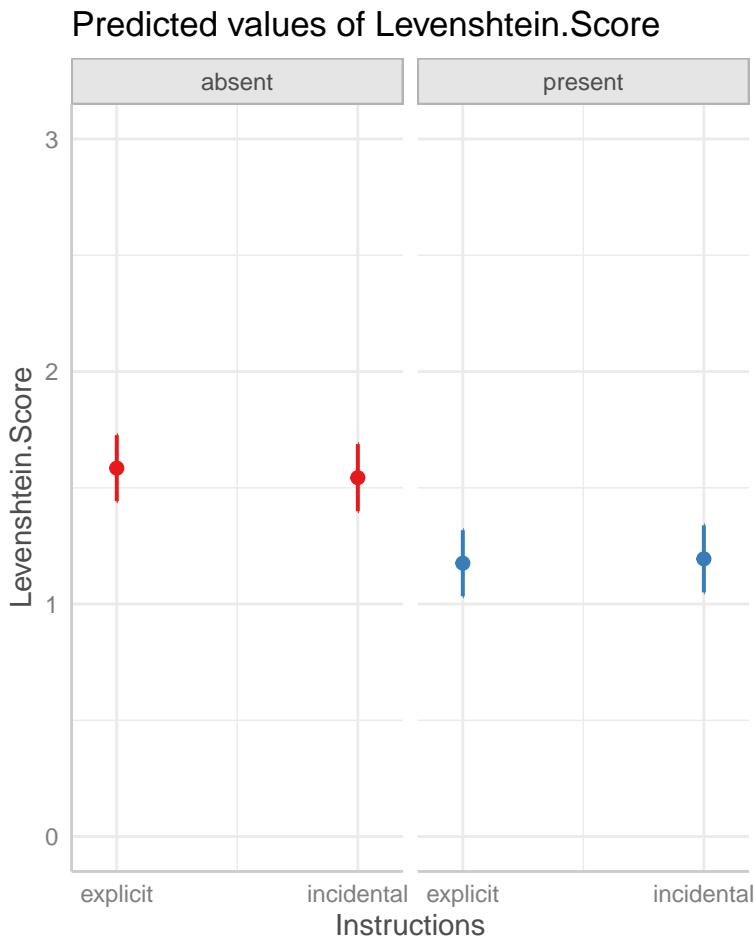


Figure 2.17: Dot and whisker plots showing the predicted effect on outcome spelling (Levenshtein) score, given different experimental conditions: Orthography (present, absent) x Instruction (explicit, incidental).

The code works as follows:

1. `dat <- ggpredict(model, terms = c("Instructions", "Orthography"))` tells R to calculate predicted outcomes, given our `model` information, for the factors `"Instructions"`, `"Orthography"`.
2. `plot(dat, facet = TRUE)` plot the effects, given the predictions, showing the effect of different instruction conditions in different plot facets (the left and right panels).

3. `ylim(0, 3)` fix the y-axis to show a more honest indication of the effect on outcomes, given the potential range of spelling scores can start at 0.

In Figure 2.17, the dots represent the linear model estimates of outcome spelling, predicted under different conditions. The plots indicate that spelling scores are predicted to be lower when orthography is present. There appears to be little or no effect associated with different kinds of instruction.

The vertical lines (often termed “whiskers”) indicate the 95% confidence interval about these estimates. Confidence intervals (CIs) are often mis-interpreted so I will give the quick definition outlined by Hoekstra et al. (2014) here:

A CI is a numerical interval constructed around the estimate of a parameter [i.e. the model estimate of the effect]. Such an interval does not, however, directly indicate a property of the parameter; instead, it indicates a property of the procedure, as is typical for a frequentist technique. Specifically, we may find that a particular procedure, when used repeatedly across a series of hypothetical data sets (i.e., the sample space), yields intervals that contain the true parameter value in 95 % of the cases.

In short, the interval shows us the range of values within which we can expect to capture the effects of interest, in the long run, if we were to run our experiment over and over again.

Given our data and our model, these intervals indicate where the outcome might be expected to vary, given different conditions, and that is quite useful information. If you look at Figure 2.17, you can see that the presence of orthography (present versus absent) appears to shift outcome spelling, on average, by about a quarter of a letter edit: from over 1.5 to about 1.25. This is about one quarter of the difference, on average, between getting a target spelling correct and getting it wrong by one letter (e.g., the response ‘epegram’ for the target ‘epigram’). This is a relatively small effect but we may consider how such small effects add up, over a child’s development, cumulatively, in making the difference between wrong or nearly right spellings to correct spellings.

In the Ricketts et al. (2021) paper, we conducted *Bayesian* analyses which allow us to plot the estimated effects of experimental conditions along with what are called *credible* intervals indicating our uncertainty about the estimates. In a Bayesian analysis, we can indicate the probable or plausible effect of conditions, or range of plausible effects, given our data and our model. (This intuitive sense of the probable location of effects is, sometimes, what researchers and students mis-interpret confidence intervals as showing; Hoekstra et al. (2014).) Accounting for our uncertainty is a productive approach to considering how much we learn from the evidence we collect in experiments.

But this gets ahead of where we are now in our development of skills and understanding. There is another way to discover how uncertain we may be about the results of our analysis. This

is an approach we have already experienced: plotting trends or estimates together with the observed data points. We present an example in Figure 2.18.

```
1 plot(dat, add.data = TRUE)
```

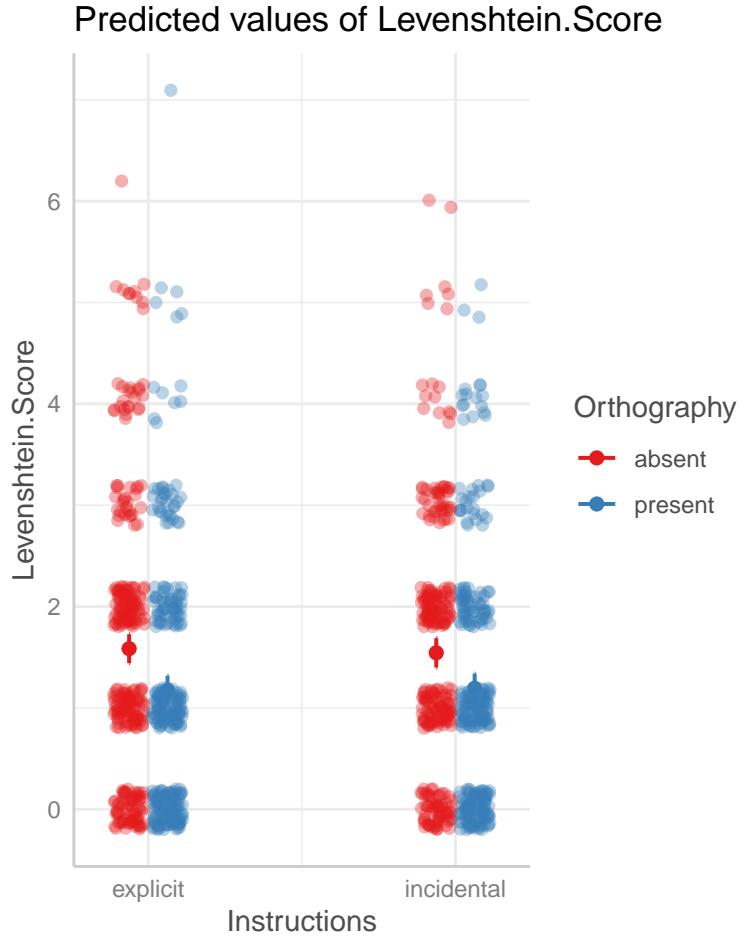


Figure 2.18: Dot and whisker plots showing the predicted effect on outcome spelling (Levenshtein) score, given different experimental conditions: Orthography (present, absent) x Instruction (explicit, incidental). The estimates are shown as dot-whisker points. In addition, the plot shows as points the spelling score observed for each child for each response recorded in the `conc.orth` dataset.

Figure 2.18 reveals the usefulness of plotting model estimates of effects alongside the raw observed outcomes. We can make two critical observations.

1. We can see that the observed scores clearly cluster around outcome spelling values of 0, 1, 2, 3, 4, and 5.

- This is not a surprise because Ricketts et al. (2021) scored each response in their test of spelling knowledge by counting the number of letter edits (letter deletions, additions etc.) separating a spelling response from a target response.
 - But the plot does suggest that the linear model is missing something about the outcome data because there is no recognition in the model or the results of this bunching or clustering around whole number values of the outcome variable. (This is why Ricketts et al. (2021) use a different analysis approach.)
2. We can also see that it is actually quite difficult to distinguish the effects of the experimental condition differences on the observed spelling responses. There is a lot of variation in the responses.

How can we make sense of this variation?

Another approach we can take to experimental data is to examine visually how the effects of experimental conditions vary between individual participants. Usually, in teaching, learning and doing foundation or introductory statistical analyses we think about the *average impact* on outcomes of the experimental conditions or some set of predictor variables. It often makes sense, also, or instead, to consider the ways that the impact on outcomes vary between individuals.

Here, it might be worthwhile to look at the effect of the conditions for each child. We can do that in different ways. In the following, we will look at a couple of approaches that are often useful. We will focus on the effect of variation in the Orthography condition (present, absent)

To begin our work, we first calculate the average outcome (`Levenshtein.Score`) spelling score for each child in each of the experimental conditions (`Orthography`, present versus absent):

We do this in a series of steps.

```
1 score.by.subj <- conc.orth %>%
2   group_by(Participant, Orthography) %>%
3     summarise(mean.score = mean(Levenshtein.Score))
```

1. `score.by.subj <- conc.orth %>%` create a new dataset `score.by.subj` by taking the original data `conc.orth` and piping it through a series of processing steps, to follow.
2. `group_by(Participant, Orthography) %>%` first group the rows of the original dataset and piped the grouped data to the next bit. We group the data by participant identity code and by Orthography condition
3. `summarise(mean.score = mean(Levenshtein.Score))` then calculate the mean `Levenshtein.Score` for each participant, for their responses in the Orthography present and in the Orthography absent conditions.

Participant	Orthography	mean.score
EOF001	absent	1.750
EOF001	present	0.875
EOF002	absent	1.375
EOF002	present	2.125
EOF004	absent	1.625
EOF004	present	1.000
EOF006	absent	0.750
EOF006	present	0.500
EOF007	absent	1.500
EOF007	present	0.625

This first step produces a summary version of the original dataset, with two mean outcome spelling scores for each child, for their responses in the Orthography present and in the Orthography absent conditions. This arranges the summary mean scores in rows, with two rows per child: one for the absent, one for the present condition. You can see what we get in the extract from the dataset, shown next.

In the second step, we also calculate the difference between spelling scores in the different Orthography conditions. We do this because Ricketts et al. (2021) were interested in whether spelling responses were different in the different conditions.

```

1 score.by.subj.diff <- score.by.subj %>%
2   pivot_wider(names_from = Orthography, values_from = mean.score) %>%
3   mutate(difference.score = absent - present) %>%
4   pivot_longer(cols = c(absent, present),
5                 names_to = 'Orthography',
6                 values_to = 'mean.score')
```

1. `score.by.subj.diff <- score.by.subj %>%` creates a new version of the summary dataset from the dataset we just produced.
2. `pivot_wider(names_from = Orthography, values_from = mean.score) %>%` re-arranges the dataset so that the `absent`, `present` mean scores are side-by-side, in different columns, for each child.
3. `mutate(difference.score = absent - present) %>%` calculates the difference between the `absent`, `present` mean scores, creating a new variable, `difference.score`.
4. `pivot_longer(cols = c(absent, present) ...)` re-arranges the data back again so that the dataset is in tidy format, with one column of mean spelling scores, with two rows for each participant for the `absent`, `present` mean scores.

This code arranges the summary mean scores in rows, with two rows per child: one for the absent, one for the present condition — plus a difference score.

Participant	difference.score	Orthography	mean.score
EOF001	0.875	absent	1.750
EOF001	0.875	present	0.875
EOF002	-0.750	absent	1.375
EOF002	-0.750	present	2.125
EOF004	0.625	absent	1.625
EOF004	0.625	present	1.000
EOF006	0.250	absent	0.750
EOF006	0.250	present	0.500
EOF007	0.875	absent	1.500
EOF007	0.875	present	0.625

Now we can use these data to consider how the impact of the experimental condition (Orthography: present versus absent) varies between individual participants. We do this by showing the mean outcome spelling score, separately for each participant, in each condition.

Figure 2.19 shows dot plots indicating the different outcome spelling (Levenshtein) scores, for each participant, in the different experimental conditions: Orthography (present, absent). Plots are ordered, from top left to bottom right, by the difference between mean spelling scores in the absent versus present conditions. The plots indicate that some children show higher spelling scores in the present than in the absent condition (top left plots), some children show little difference between conditions (middle rows), while some children show higher spelling scores in the absent than in the present condition (bottom rows).

```

1 ggplot(data = score.by.subj.diff,
2         aes(x = Orthography, y = mean.score,
3               colour = Orthography)) +
4   geom_point() +
5   facet_wrap(~ reorder(Participant, difference.score)) +
6   theme(axis.text.x = element_blank())

```

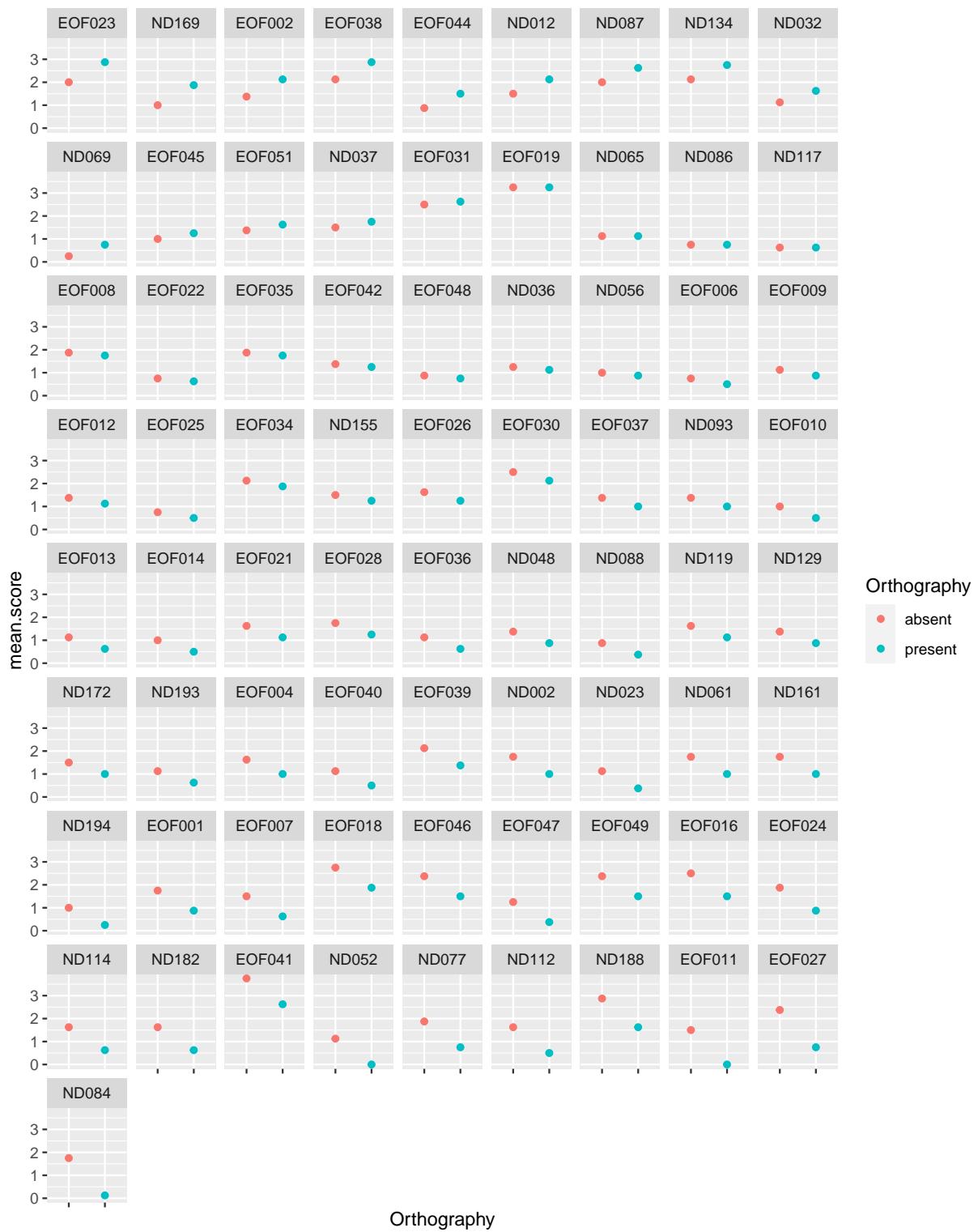


Figure 2.19: Dot plots showing the different outcome spelling (Levenshtein) scores, for each participant, in the different experimental conditions: Orthography (present, absent). Plots are ordered, from top left to bottom right, by the difference between mean spelling scores in the absent versus present conditions.

Once we have done the data processing in preparation, the code to produce the plot is fairly compact.

1. `ggplot(data = score.by.subj.diff ...` tells R to produce a plot, using `ggplot()` and the newly created `score.by.subj.diff` dataset.
2. `aes(x = Orthography, y = mean.score,...` specifies the aesthetic mappings: we tell R to locate `mean.score` on the y-axis and `Orthography` condition on the x-axis/
3. `aes(...colour = Orthography)) +` specifies a further aesthetic mapping: we tell R to map different `Orthography` conditions to different colours.
4. `geom_point()` + tells R to take the data and produce a scatterplot, given our mapping specifications.
5. `facet_wrap(...)` + tells to split the dataset into sub-sets (facets).
6. `facet_wrap(~ reorder(Participant, difference.score))` tells R that we want the sub-sets to be organized by Participant, and we want the facets to be ordered by the `difference.score` calculated for each participant.
7. `theme(axis.text.x = element_blank())` removes the x-axis labels because it is too crowded with the axis labels left in, and the information is already present in the colour guide legend shown on the right of the plot.

2.7.11 Summary: Visualizing associations

Visualizing associations between variables encompasses a wide range of the things we have to do, in terms of both discovery and communication, when we work with data from psychological experiments.

The conventional method to visualize how the distribution of values in one variable covaries with the distribution of values in another variable is through using a scatterplot. However, the construction of a scatterplot can be elaborated in various ways to enrich the information we present or communicate to our audiences, or to ourselves.

- We can add elements like smoothers to indicate trends.
- We can add annotation, as with the histograms, to highlight specific thresholds.
- We can facet the plots to indicate how trends may vary between sub-sets of the data.

In the final phases of our practical work, we started by presenting model-based predictions of the effects of experimental manipulations. However, you will have noticed that presenting plots of effects is not where we stop when we engage with a dataset. Further plotting indicates quite marked variation between participants in the effects of the conditions. This kind of insight is something we can and should seek to reveal through our visualization work.

2.8 Next steps for development

To take your development further, take a look at the resources listed in Section 2.9.

In my experience, the most productive way to learn about visualization and about coding the production of plots, is by doing. And this work is most interesting if you have a dataset you care about: for your research report, or for your dissertation study.

As you have the alternate datasets described in Section 2.7.1.2.1, you can start with the data from the other task or the other study in Ricketts et al. (2021). Ricketts et al. (2021) recorded children’s responses in two different outcome tasks, the orthographic spelling task we have looked at, and a semantic or meaning-based task. It would be a fairly short step to adapt the code you see in the example code chunks to work with the semantic datasets.

Alternatively, you can look at the data reported by Rodríguez-Ferreiro et al. (2020). Rodríguez-Ferreiro et al. (2020) present both measures of individual differences (on schizotypal traits) and experimental manipulations (of semantic priming) so you can do similar things with those data as we have explored here.

2.9 Helpful resources

2.9.1 Some helpful websites

- We typically use the `ggplot` library (part of the `tidyverse`) to produce plots. Clear technical information, with useful examples you can copy and run, can be found in the reference webpages:

<https://ggplot2.tidyverse.org/reference/index.html>

- A source of inspiration can be found here:

<https://r-graph-gallery.com>

If you are trying to work out how to do things by searching for information online, you often find yourself at tutorial webpages. You will develop a sense of quality and usefulness with experience. Most often, what you are looking for is a tutorial that provides some explanation, and example code you can adapt for your own purposes. Here are some examples.

- Cedric Scherer on producing raincloud plots:

<https://www.cedricscherer.com/2021/06/06/visualizing-distributions-with-raincloud-plots-and-how-to-create-them-with-ggplot2/>

- Winston Chang on colours and colour blind palettes:

[http://www.cookbook-r.com/Graphs/Colors_\(ggplot2\)/](http://www.cookbook-r.com/Graphs/Colors_(ggplot2)/)

- Thomas Lin Pedersen (and others) on putting together plots into a single presentation using the `patchwork` library functions:

<https://patchwork.data-imaginist.com/articles/patchwork.html>

2.9.2 Some helpful books

- The book “R for Data Science” (Wickham & Grolemund, 2016) will guide you through the data analysis workflow, including data visualization, and the latest version can be accessed in an online free version here:

<https://r4ds.hadley.nz>

- The “ggplot2: Elegant Graphics for Data Analysis” book (Wickham, 2016) corresponding to the `ggplot` library was written by Hadley Wickham in its first edition, it is now in its third edition (as a work in progress, co-authored by Wickham, Danielle Navarro and Thomas Lin Pedersen) and this latest version can be accessed in an online free version here:

<https://ggplot2-book.org/index.html>

- The “R graphics cookbook” (Chang, 2013), and the latest version can be accessed in an online free version here:

<https://r-graphics.org>

- The book “Fundamentals of Data Visualization” (Wilke, n.d.) is about different aspects of visualization, and can be accessed in an online free version here:

<https://clauswilke.com/dataviz/>

Part II

MODELS

3 Introduction to multilevel data

3.1 Motivations

In this chapter, we shall start to develop skills in using a method or approach that is essential in modern data analysis: *multilevel modeling*. We are going to invest four weeks in working on this approach. This investment is designed to give you a specific, important, advantage in your work as a psychologist, or as someone who produces or consumes psychological research.

 Note

Multilevel models: *why* do we need to do this? - Four weeks is a lot of time to spend on one method.

It is now clear that someone who works in psychological research *has* to know about multilevel or hierarchically structured data, and has to know how to apply multilevel models (or mixed-effects models). Growth in the popularity of these kinds of analysis has been very very rapid, as can be seen in Figure 3.1. It is now, effectively, the standard or default method for professional data analysis in most areas of psychological and other social or clinical sciences (or it soon will be). There are good reasons for this (Baayen et al., 2008a).

We continue to teach ANOVA and multiple regression (linear models) in our courses because the research literature is full of the results of analyses done using these methods and because many psychologists continue to use these methods in their research. However, there is increasingly wide-spread recognition that these classical methods have serious problems when applied to data with hierarchical structure. Because most psychological data (not all) will have hierarchical structure, this makes learning about multilevel or mixed-effects methods a key learning objective.

But, because it is relatively new, many professional psychologists struggle to understand why or how to use these methods effectively. This means that students who acquire the skill graduate with a clear *employability* advantage. It also means that we have to take seriously the challenge of learning about these methods. This is why we will spend a bit of time on them. In my experience, in over a decade of teaching multilevel models, in working with both students and professionals, we shall need to develop understanding and skills gradually. We will work patiently, so that we can secure understanding by building our learning through a series of practical examples, increasing the scope of our practical skills, and developing the sophistication of our understanding, step-by-step, as we go.

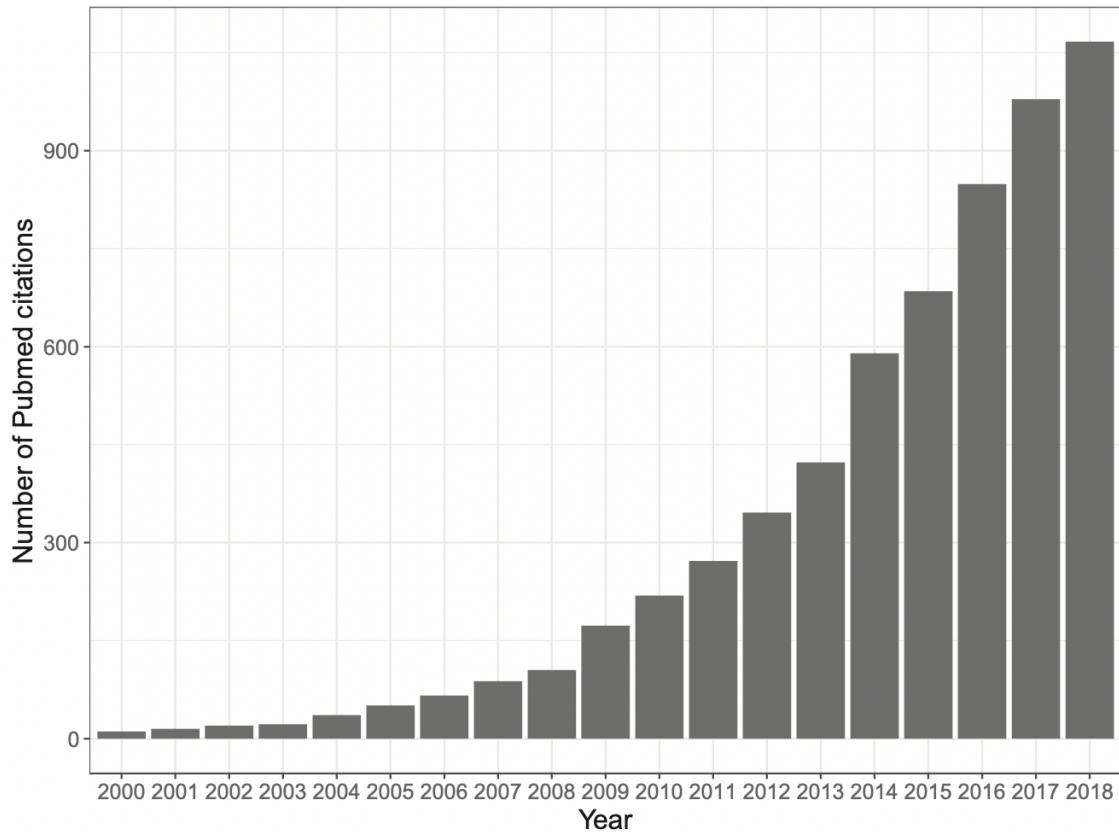


Figure 3.1: Number of Pubmed citations for ‘Linear Mixed Models’ by year. Generated using the tool available at <http://dan.corlan.net/medline-trend.html>, entering “Linear Mixed Models” as the phrase search term and using data from 2000 to 2018, from Meteyard and Davies (2020) – used without permission

3.1.1 A word about names

Multilevel models are also known as hierarchical models or linear mixed-effects models or random effects models. People use these terms interchangeably. They also use the abbreviations LMMs or LMEs. Sorry about that: humans make methods, and the names we use for things do vary.

I will only use the terms *multilevel* or *linear mixed-effects* models.

⚠ Warning

- In this chapter, we emphasize the multilevel perspective but, to anticipate future development, we will come to think in terms of *mixed-effects models*.

3.2 Challenges

The key challenges for learning should be explained at the start so that we know what we shall have to do to overcome them.

1. Even though most psychological data has some sort of multilevel or hierarchical structure, we are not used to recognizing it. This is because the structure has often been hidden or ignored in the education and practice of traditional research methods in Psychology.

In time, you will come to see **multilevel structure everywhere** (Kreft & Leeuw, 1998). But first you will need to get some practice so that you can become familiar with the idea and learn to recognize what it looks like when data have a multilevel or hierarchical structure. This is why we will examine multilevel structured data across a range of different kinds of experiments or surveys, over a series of weeks.

We will learn to identify and understand hierarchical structure in psychological data by just looking at datasets, by producing visualizations, by doing analyses, and by trying to explain to ourselves and each other what we think we see.

2. The ideas that support an understanding of why and how we use multilevel models can be intimidating when we first encounter them. The mathematics behind how the models work *is* both profound and sophisticated. But the good news is that we can practice the application of the analysis method while talking and thinking about the critical ideas using just words or plots.

We cannot or should not avoid engaging with the ideas – because we have to be able to explain what we are doing – but there are many routes to an effective understanding. For those who want to develop a more mathematically-based perspective, I will provide references to important texts in the literature on multilevel models (see Section 3.11).

3.3 The key idea to get us started

! Important

Multilevel models are a *general* form of linear model.

Another way of saying this is: *linear models are a special form of multilevel models*.

This is because linear models assume that observations are independent. We often cannot make this assumption, as we shall see. More generally, then, we do not assume that observations are independent and so we use multilevel models.

3.4 The approach we take

We are *not* going to take a mathematical approach to learning about multilevel models. We do not have to. The approach we are going to take is:

1. **verbal** – We will talk about the main ideas in words. Sometimes, we will present formulas but that is just to save having to use *too many* words.
2. **visual** – We will show ourselves and each other what multilevel structure in data looks like, and what that structure means for our analyses of behaviour.
3. **practical** – We will use R to complete analyses, so we will learn about coding models in practice. Fortunately, through coding we can get a clear idea of what we want the models to do.

3.5 Targets

Our learning objectives include the development of key concepts and skills.

1. **concepts** – how data can have multilevel structures and what this requires in models
2. **skills** – where skills comprise the capacity to:
 - i. use visualization to examine observations within groups
 - ii. run linear models over all data and within each class
 - iii. use the `lmer()` function to fit models of multilevel data

We are just getting started. Our plan will be to build depth and breadth in understanding as we progress over the next few weeks.

3.6 Study guide

I have provided a collection of materials you can use. Here, I explain what they are and how I suggest you use them.

1. Video recordings of lectures

1.1 I have recorded a lecture in three parts. The lectures should be accessible by anyone who has the link.

- [Part 1](#) – about 16 minutes
- [Part 2](#) – about 13 minutes
- [Part 3](#) – about 13 minutes

1.2 I suggest you watch the recordings then read the rest of this chapter. The lectures provide a summary of the main points.

2. Chapter: 01-multilevel

2.1 I have written a chapter discussing the main ideas and setting out the practical steps you can follow to start to develop the skills required to analyse multilevel structured data. 2.2 The practical elements include data tidying, visualization and analysis steps. 2.3 You can read the chapter, run the code, and do the exercises.

- Read in the example data `BAFACALO_DATASET.RData` and identify how the data are structured at multiple levels.
- Use visualizations to explore the impact of the structure.
- Run analyses using linear models (revision) and linear mixed-effects models (extension) code.
- Review the recommended readings (Section [3.11](#)).

3. Practical workbook materials

3.1 In the following sections, I describe the practical steps, and associated resources, you can use for your learning.

3.7 The data we will work with: Brazilian school children

This week, we will be working with data taken from a study on education outcomes in Brazilian children, reported by Golino & Gomes (2014). First, we will progress through the steps required to download and prepare the files for analysis in R.

The `BAFACALO_DATASET.RData` data were collected and shared online by Golino & Gomes (2014). Information about the background motivating the study, the methods of data collection, along with the dataset itself, can be found [here](#).

Golino & Gomes (2014) collected school end-of-year subject grades for a sample of 292 children recruited from multiple classes in a school in Brazil. Here, each subject is a theme or course that children studied in school, e.g., Physics or English language, and for which each child was awarded a grade at the end of the school year. So, we have information on different children, and information about their subject grades. Children were taught in different classes but the classes appear to be units of the school organization (the information is not quite clear) not subject or course groupings. Thus, we also have information on which classes children were in when data about them were collected.

The data were compiled to an .RData format file.

.RData is R's own file format so the code you use to load and access the data for analysis is a bit simpler than you are used to. (You will have used `read.csv()` or `read_csv()` previously.)

In my own practice, I prefer to keep files in formats like .csv which can be opened and read in other software applications (like Excel), so using an .RData is an exception in these materials.

3.7.1 Locate and download the data file

You can access the data from the link associated with the Golino & Gomes (2014) article [here](#).

Or you can download the [data-01-multilevel.zip](#) files folder for this chapter.

The file is located in a .zip folder called `data-01-multilevel`. The data file is collected together with the .R scripts:

- `01-multilevel-workbook.R` the workbook you will need to do the practical exercises.
- `01-multilevel-workbook-answers.R` with answers to questions and code for exercises.

3.7.2 Read in the data file using the `load()` function

You can read the `BAFACALO_DATASET.RData` file into the R workspace or environment using the following code.

```
load("BAFACALO_DATASET.RData")
```

3.7.3 Inspect the data

The dataset consists of rows and columns. Take a look. If you have successfully loaded the dataset into the R environment, then you should be able to view it.

You could look at the dataset by using the `View()` function.

```
View(BAFACALO_DATASET)
```

Or you can use the `head()` function to see the top few rows of the dataset.

```
head(BAFACALO_DATASET)
```

You can check for yourself that each row holds data about a child, including their participant identity code, as well as information about their parents, household, gender, age, school class, and their grades on end-of-year subject attainment (e.g., how well they did in English language).

There are many more variables in the dataset than we need for our exercises, and a summary would fill pages. You can see for yourself if you inspect the dataset using

```
summary(BAFACALO_DATASET)
```

3.8 Tidy the data

When you inspect the file, you will see that it includes a large number of variables, but we only really care about those we will use in our exercises:

- `participant_id` gives the participant identity code for each child;
- `class_number` gives the class identity code for the school class for each child;
- and values in `portuguese`, `english`, `math`, and `physics` columns give the score for each child in subject class attainment measures.

You can see that I am not explaining the variables in great depth. For our aims, we do not need more detailed information but please do read the data article if you wish to find out more.

We need to tidy the data before we can get to the analysis parts of this chapter. We are going to:

1. Select the variables we want to work with;
2. Filter out missing values, if any are present;

3. And make sure R knows how we want each variable to be identified; what type the variable should have.

We shall need the `tidyverse` library of functions.

```
library(tidyverse)
```

3.8.1 Select the variables

We can start by selecting the variables we want, which include those named here, ignoring all the rest. We take the `BAFACALO_DATASET`. We then use the `select()` function to select the variables we want.

```
brazil <- BAFACALO_DATASET %>%  
  select(  
    class_number, participant_id,  
    portuguese, english, math, physics  
  )
```



Tip

Notice the code is written to create the new selected dataset and, at the same time, gave it a more usable (shorter) name, with `brazil <- BAFACALO_DATASET ...`

Inspect the data to see what you have.

```
summary(brazil)
```

```
class_number participant_id      portuguese      english       math  
M11      : 21   Min.    : 1.00     60      : 10     100      : 17     60      : 13  
M15      : 20   1st Qu.: 73.75    76      : 9      79      : 11     69      : 9  
M14      : 19   Median   :146.50    65      : 8      96      : 8      70      : 9  
M36      : 19   Mean     :146.50    73      : 8      81      : 7      62      : 8  
M18      : 18   3rd Qu.:219.25    74      : 7      89      : 7      71      : 7  
(Other):133  Max.    :292.00    (Other):188  (Other):180  (Other):184  
NA's     : 62                           NA's    : 62    NA's    : 62    NA's    : 62  
  physics  
  60      : 16  
  60.1    : 5  
  0       : 4  
  61.8    : 4
```

```
66      : 4  
(Other):197  
NA's   : 62
```

3.8.2 Remove missing values

If you look at the results of the selection, you can see that there are missing values, written as e.g. NA's 62 at the bottom of each summary of each variable (if a variable column includes missing values).



Tip

Remember that in R NA means “not available” i.e. missing.

We will need to get rid of the missing values. It is simpler to do this at the start rather than wait for an error message, later, when some arithmetic function tells us it cannot give us a result because there are NAs present.

We can get rid of the missing values using `na.omit()`

```
brazil <- na.omit(brazil)
```

If you then look at a summary of the data again then you will see that the NAs are gone.

```
summary(brazil)
```

```
class_number participant_id    portuguese      english       math  
M11      : 21  Min.    : 3.0    60      : 10    100     : 17    60      : 13  
M15      : 20  1st Qu.: 77.5   76      : 9     79      : 11    69      : 9  
M14      : 19  Median  :144.5   65      : 8     96      : 8     70      : 9  
M36      : 19  Mean    :146.6   73      : 8     81      : 7     62      : 8  
M18      : 18  3rd Qu.:222.8   74      : 7     89      : 7     71      : 7  
M21      : 17  Max.    :291.0   82      : 7     86      : 6     66      : 6  
(Other):116                               (Other):181   (Other):174   (Other):178  
physics  
60      : 16  
60.1    : 5  
0       : 4  
61.8    : 4  
66      : 4  
74.2    : 4  
(Other):193
```

3.8.3 Getting R to treat a variable as an object of the type required using the `as...()` family of functions

But if you look closely at the output from `summary(brazil)` you will see that the `portuguese` and `english` variables are summarized in the way that R summarizes factors.

When you ask R to summarize factors, R gives you a count of the number of observations associated with each factor level, that is, each category in each variable. Here, it is treating a grade score like 100 in `english` as a category (like you might treat *dog* as a category of pets), and you can see that the count shows you that 17 children were recorded as having scored 100 in their class. We do not want numeric variables like subject grades (e.g. children's grades in English) treated as categorical variables, factors.

You can also see that R gives you a numeric summary of the recorded values in the `participant_id` variable. This makes no sense because the identity code numbers are (presumably) assigned at random so identity numbers provide no useful numeric information for us. We do not want this either.

We want R to treat the educational attainment scores as numbers. We can do this using the `as.numeric()` function. We want R to treat the class and participant identity numbers as factors (categorical variables). We can do this using the `as.factor()` function.

We could do this one variable at a time.

```
brazil$portuguese <- as.numeric(brazil$portuguese)
brazil$english <- as.numeric(brazil$english)
brazil$math <- as.numeric(brazil$math)
brazil$physics <- as.numeric(brazil$physics)

brazil$class_number <- as.factor(brazil$class_number)
brazil$participant_id <- as.factor(brazil$participant_id)
```

But it is simpler and more efficient in `tidyverse` style. (You can see a discussion [here](#) that helped me to figure this out.)

```
brazil <- brazil %>%
  mutate(across(c(portuguese, english, math, physics), as.integer),
        across(c(class_number, participant_id), as.factor))
```

If you now look at the summary of the data, you can see that R will give you mean etc. for the subject class score variables e.g. `english`, showing that it is now treating them as numeric variables. In comparison, R gives you counts of the numbers of observations for each level (category) of categorical or nominal variables like `participant_id`.

```
summary(brazil)
```

```
  class_number participant_id   portuguese      english       math
M11     : 21      3        : 1    Min.   : 1.00   Min.   : 1.00   Min.   : 1.00
M15     : 20      4        : 1    1st Qu.:26.25  1st Qu.:29.00  1st Qu.:39.00
M14     : 19      5        : 1    Median  :42.00   Median  :55.00   Median  :58.00
M36     : 19      6        : 1    Mean    :43.55   Mean    :50.25   Mean    :58.17
M18     : 18      7        : 1    3rd Qu.:58.75  3rd Qu.:73.00  3rd Qu.:79.75
M21     : 17      9        : 1    Max.    :83.00   Max.    :93.00   Max.    :104.00
(Other):116   (Other):224
               physics
Min.   : 1.00
1st Qu.:34.00
Median :71.50
Mean   :71.57
3rd Qu.:107.75
Max.   :148.00
```

3.8.3.1 Coercion: a quick lesson

R treats things like variables as vectors. A vector can be understood to be a set or list of elements: things like numbers or words.

R gives each vector a **type** (factor, numeric etc.) which helps to inform different functions how to handle that vector. Usually, R assigns type correctly but sometimes it does not. We can use what is called *coercion* to force R to assign the correct type to a variable.

It is more efficient to do this at the start of an analysis workflow.

Normally, I would use `read_csv()` from `tidyverse` and assign type to variable using `col_types()` specification. (See [here](#) for more information and an example.) But it is useful to learn what you need to do if you need to change the way a variable is treated after you have got the data into the R environment.

3.8.3.2 Exercise – experiment with coercion

In R, there are a family of functions that work together. You can test whether a variable (a vector, technically) is or is not a certain type using the `is.[something]` function. For example:

```
is.factor(brazil$english)
is.numeric(brazil$english)
is.character(brazil$english)
```

And you can coerce variables so that they are treated as having certain types.

```
brazil$english <- as.factor(brazil$english)
brazil$english <- as.numeric(brazil$english)
brazil$english <- as.character(brazil$english)
```

Now try it out.

1. Test out type for different variables using `is...()` for some of the variables.
2. Test out coercion – and its results – using `as...()` for some of the variables.
3. Look at the results using `summary()`.

3.9 Introduction to thinking about multilevel models

3.9.1 Main ideas – Phenomena and data sets in the social sciences often have a multilevel structure

We (Psychologists) often adopt *Repeated Measures* or *Clustered* designs in our studies, and these designs yield data that have a multilevel structure. Examples of research which results in data with a multilevel structure include:

- Studies where we test the same people multiple times, maybe in developmental or longitudinal investigations;
- Intervention, learning or treatment studies where we need to make pre- and post-treatment comparisons;
- Studies where we present multiple stimuli and everyone sees the same stimuli;
- Studies that involve multi-stage sampling e.g. selecting a sample of classes or schools then testing a sample of children within each classes or within each school.

The key insight to keep in mind when considering the analysis of such data is that observations are clustered and are *not independent*: they are correlated. What is correlated with what?

Imagine testing a number of subjects by giving them all the same test. In that test, you might present them all with the same stimuli over a series of trials, so that everybody sees the same set of stimuli. Do you think an observed response recorded for any one individual will be uncorrelated i.e. independent of that person's other responses?

People are different and one usually finds that a slow or inaccurate subject is slow or inaccurate for most of their responses. That means that if you have information about one of their

responses you can predict, in part, what the time or accuracy of one of their other responses would be. That capacity to predict one response from another is what we mean when we talk about a lack of independence.

Alternatively, imagine going to test children in a school. Will the children in one class be more like each other than they are like children in other classes? In other words, is there an effect of class – maybe due to the approach of the teacher, the effect of the class environment etc. – so that outcomes for children in a class are correlated with each other?

! Important

We are talking, here, about a really quite general property of data collected in certain, very widely used, designs in Psychology: the *clustering or hierarchical ordering* of data.

This week, we will see that we must deal with the dependence of observations within a class, where the observed responses were made about the different pupils tested in a class, for a number of different classes.

3.9.2 Multilevel models – why they are more used and more useful than traditional methods

The utility of **multilevel models** to analyze **multilevel data** or hierarchically structured data is well established in education. In educational research, we often need to think about effects of interventions or correlations in the context of observing children in classes, schools or districts, perhaps over time or at different time points. This means that many of the critical textbooks present examples that are based on educational research data (Goldstein, 1995; Kreft & Leeuw, 1998; Raudenbush & Bryk, 2002; Snijders & Bosker, 2004).

Multilevel models are growing in popularity in Psychology as well as in Education because they can be used to account for systematic and random sources of variance in observed outcomes when data are hierarchically structured. A hierarchical structure is present in data when a researcher: tests participants who belong to different groups like classes, clinics or schools; presents a sample of stimuli to each member of a sample of participants; or makes repeated observations for each participant over a series of test occasions.

In these circumstances, the application of traditional analytic methods has typically required the researcher to aggregate their data (e.g., averaging the responses made by a participant to different stimuli) or to ignore the hierarchical structure in their data, (e.g., analyzing the responses made by some pupils while ignoring the fact that the pupils were tested in different classes). But the application of traditional analysis approaches (e.g., regression, ANOVA) to multilevel structured data extracts scientific costs (Baayen et al., 2008a; Barr et al., 2013a).

Ignoring structure by ignoring or averaging over sources of variability like differences between classes, participants, or stimulus items can mean that analyses are less sensitive because they

fail to fully account for error variance. Where differences between classes, participants or stimuli include variation in the impact of experimental variables, e.g., individual differences in response to an experimental manipulation, the application of traditional methods can be associated with an increased risk of false positives in discovery. Yet these costs need no longer be suffered because the capacity to perform multilevel modeling is now readily accessible.

3.9.3 Practical applications: children sampled within classes

If a researcher tests participants belonging to different groups, e.g., records the educational attainment of different children recruited from different classes in a school, the test scores for the participants are observations that occupy the lowest level of a hierarchy (see Figure 3.2). In multilevel modeling, those observations are understood to be nested within higher-level sampling units, here, the classes. We can say that the children are sampled from the population of children. And the classes are sampled from the population of classes. Critically, we recognize that the children's test scores are **nested within the classes**. This multi-stage sampling has important consequences, as we shall see.

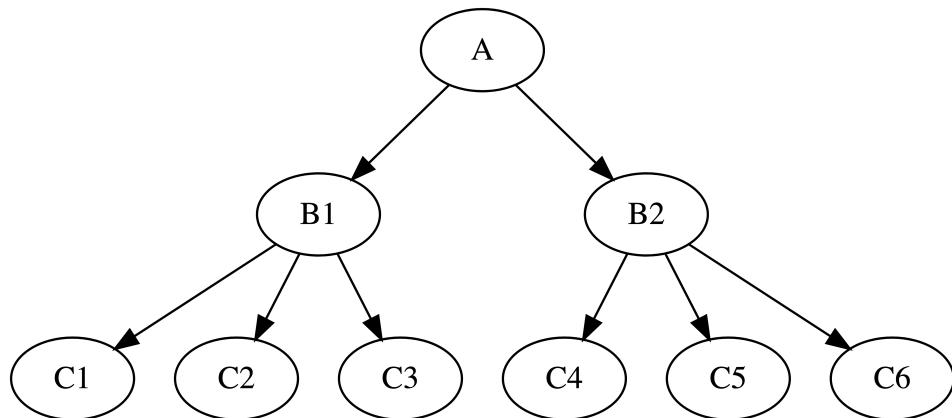


Figure 3.2: Multilevel or hierarchically structured data

We are going to be working with the school data collected by Golino & Gomes (2014) in Brazil, so let's take another look at that dataset. Figure ?@fig-golino2014 presents the first 25 rows of the selected-variables `brazil` dataset. (I have arranged the rows by class number ID.) You can see that there are multiple rows of data for each `class_number`, one row for each child, with multiple children (you can see different `participant_id` numbers). This presentation of the dataset illustrates in practical terms – what you can actually see when you look at your data – what multilevel structured data can look like.

```

brazil %>%
  arrange(desc(class_number)) %>%
  head(n = 25)

  class_number participant_id portuguese english math physics
1          M36             11       80     55   60     81
2          M36             30       48     55   43     57
3          M36             93       83     75   55     80
4          M36             95       80     66   55     66
5          M36            111      80     57   99    129
6          M36            117      27     76   83     89
7          M36            153       2     66   89    141
8          M36            163      68     88   74    137
9          M36            176      80     75   60    127
10         M36            216      75     79   64    124
11         M36            223      81     56   58     78
12         M36            234      81     75   49     74
13         M36            242      81     67   45     77
14         M36            254      76     51   33    103
15         M36            260      81     59   39     82
16         M36            271      64     69   82    138
17         M36            278      80     57   56    107
18         M36            280      83     64   56     84
19         M36            286      80     91   64     98
20         M35             20      42     61   97     92
21         M35            118      33     59   33     80
22         M35            140      28     40   33     58
23         M35            179      38     69   56    145
24         M35            247      65     51   64     97
25         M33             15      76     70   73     85

```

3.9.4 To understand the application of multilevel models: first, we ignore the multilevel structure in the data

Recognizing that the children's scores data are observed within classes means that, when we examine the factors that influence variance in observed outcomes, we need to take into account the fact that the children can be grouped by (or under) the class they were in when their grades were recorded. We can develop an understanding of what this means by moving through a series of steps.

To illustrate the understanding we need to develop, we analyze the end-of-year school subject grades for the sample of 292 children studied by Golino & Gomes (2014). With these data, we

can examine whether differences between children in their Portuguese language grades predicts differences in their English language grades. (We do not have a theoretical reason to make this prediction though it does not seem unreasonable.) Let's make this our research question.

i Note

- Research question: Does Portuguese grade predict English grade?

Now, the first step we take in our development of understanding will be to *first ignore* the importance of the potential differences between classes.

We can begin our analysis in order to address the research question by plotting the possible association between Portuguese and English grades. We shall create a scatterplot to do this, and we create the plot by running the following code and *ignoring group (class) membership*. In the following chunk of code, I pipe the `brazil` data using `%>%` to `ggplot()` and then create the plot step-by-step, with each `ggplot()` step separated by a `+` except for the last step.

```
brazil %>%
  ggplot(aes(x = portuguese, y = english)) +
  geom_point(colour = "black", size = 3, alpha = .5) +
  geom_smooth(method = "lm", size = 2, se = FALSE, colour = "red") +
  xlab("Portuguese") + ylab("English") +
  theme_bw()
```

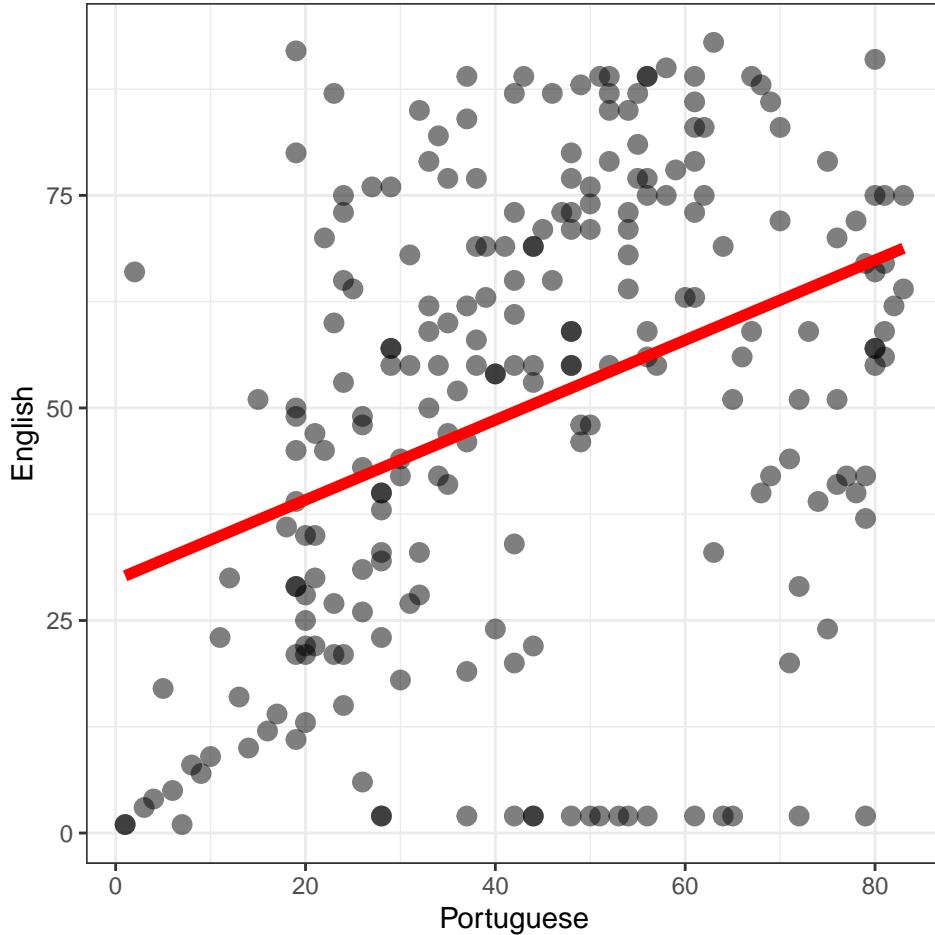


Figure 3.3: Portuguese compared to English grades: each point represents the scores for one child

Remember that each child ID is associated with a pair of grades: their grade in English and their grade in Portuguese. Each point in Figure 3.3 represents the paired grade data for one child.

Because I say that we are interested in predicting English grades then, by convention, we map English grades to the height of the points (i.e. English grade differences are shown as differences on the y-axis). Because we are using Portuguese grades to do the predicting, by convention, we map Portuguese grades to the horizontal position of the points (i.e. Portuguese grade differences are shown as differences on the x-axis).

I have added a line using `geom_smooth()` in red to indicate the trend of the potential association between variation in Portuguese grades and variation in English grades.

Figure 3.3 suggests that children with higher grades in Portuguese tend, on average, to also have higher grades in English.

3.9.4.1 Exercise – edit plots

Notice that when we write the code to produce the plot, we add arguments to the `geom_point()` and `geom_smooth()` function calls to adjust the appearance of the points and the smoother line. Notice, also, that we adjust the labels for the x-axis and y-axis and, finally, that we determine the overall appearance of the plot using the `theme_bw()` function call.

Do the following exercises to practice your `ggplot()` skills

1. Change the x and y variables to `math` and `physics`
2. Change the theme from `theme_bw()` to something different
3. Change the appearance of the points, try different colours, shapes or sizes.

Further information to help you try out coding options can be found [here, on scatterplots](#), and [here, on themes](#).

3.9.5 A linear model ignores the multilevel structure in the data

Our plot indicates the relationship between English and Portuguese language grades, in Figure 3.3, ignoring the fact that children in the sample belonged to different classes when they were tested. The plot shows us only information about the children and grades, with each point representing the observed English and Portuguese grades for each i child.

Can we predict variation in English grades given information about child Portuguese grades? Are English grades related to Portuguese grades? We can estimate the relationship between English and Portuguese grades using a linear model in which the English grades variable is the **outcome** variable (or the dependent variable) and the Portuguese grades variable is the **predictor** or independent variable:

$$y_i = \beta_0 + \beta_1 X_i + e_i$$

- where y_i represents the English grade for each i child;
- β_0 represents the intercept, the outcome value obtained if values of the explanatory variable are zero;
- β_1 represents the effect of variation in X_i the Portuguese grade for each child, with the effect of that variation estimated as the the rate of change in English grade for unit change in Portuguese grade;
- e_i represents differences, for each child, between the observed English grade and the English grade predicted by the relationship with Portuguese grades.

As you have seen, the code for running a linear model corresponds transparently to the statistical model given in the formula.

```
summary(lm(english ~ portuguese, data = brazil))
```

Call:

```
lm(formula = english ~ portuguese, data = brazil)
```

Residuals:

Min	1Q	Median	3Q	Max
-64.909	-17.573	2.782	20.042	53.292

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	29.77780	3.81426	7.807	2.11e-13 ***							
portuguese	0.47001	0.07897	5.952	9.91e-09 ***							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'..'	0.1	' '	1

Residual standard error: 25.02 on 228 degrees of freedom

Multiple R-squared: 0.1345, Adjusted R-squared: 0.1307

F-statistic: 35.43 on 1 and 228 DF, p-value: 9.906e-09

The linear model yields the estimate that for unit increase in Portuguese grade there is an associated increase of about .47 in English grade, on average (for the model, $F(1, 228) = 35, p < .001$; $adj.R^2 = .13$). So, we have a preliminary answer to our research question.

Note

- Research question: Does Portuguese grade predict English grade?
- Result: Our analysis shows that children who score one grade higher in their Portuguese class (e.g., 61 compared to 60) tended to score .47 of a grade higher in their English class.

We shall see that we will need to revise this estimate when we do an analysis that *does* take school class differences into account.

Tip

Notice: here, I have dispensed with the creation of a model object by name. The model is estimated anyway, and I have embedded the `lm()` function call within a `summary()`

function so that I am asking R to do two things:

1. `lm()` fit a linear model
2. `summary()` print out a summary of the fitted model object

I do this to give an example of the way in which code can be varied. Also, to show you how one function can be embedded inside another.

3.9.5.1 Exercise – fit a different linear model

Do the following exercises to practice your `lm()` skills

1. Change the outcome and predictor variables to `math` and `physics`
2. What do the results tell you about the relationship between maths and physics ability?

3.9.6 Notation

In the following discussion, I will present some model formulas as equations. I am not doing that because the discussion is going to consider the modeling in terms of the underlying mathematics. I am doing it with the aim of clarifying how quantities – observed, estimated or predicted – add up, in terms of the linear and then the linear mixed-effects model.

1. I am going to refer to the dependent or outcome variable (e.g., school grade) as y ,
2. and to explanatory or experimental or independent variables (e.g., language skill) as X .
3. Models shall be fitted to estimate the coefficients, written β , of the effects of the explanatory variables on the outcome variable
4. To distinguish the different coefficients of the different effects, I am going to number the coefficients so ...
 - β_0 is the coefficient of the intercept;
 - β_1 will be the coefficient of the effect of a first explanatory variable: in the Brazilian schools example, the coefficient of the effect of variation in Portuguese language skill on variation in English language scores.

To make it clear that each observation can be understood as part of a complex multilevel or crossed random effects structure, I am going to use indices as subscripts for variables.

1. I will, here, index individual participants (children) using i ;
2. I will index index classes in which children were tested using j .

Thus, in the Brazilian schools example, we shall see that we are concerned with observations about children's school grades, where children are sampled as individuals *nested* in samples of classes. We will examine how an outcome variable (English language grade) is related to a predictor variable (Portuguese language grade) such that:

- y_{ij} is the outcome English language grade recorded for each child i in each class j ;
- while X_{ij} represents the explanatory variable, see following, the Portuguese language grade.

X_{ij} is subscripted ij because values of the variable depend upon child identity, and children are identified as child i in class j to represent the multilevel structure of the data

3.9.7 Can we really ignore the multilevel structure?

The linear model ignores the higher-level structure, the distinction between classes: **does this matter?**

We can see the answer to that question if we inspect Figure 3.4. We create the plot using the following chunk of code; we discuss that later, first reflecting on what the plot shows us.

```
brazil %>%
  ggplot(aes(x = portuguese, y = english)) +
  geom_point(colour = "darkgrey") +
  geom_smooth(method = "lm", se = FALSE, colour = "black") +
  facet_wrap(~ class_number) +
  xlab("Portuguese") + ylab("English") +
  theme_bw() +
  scale_x_continuous(breaks=c(25,50,75)) + scale_y_continuous(breaks=c(0,50,100))
```

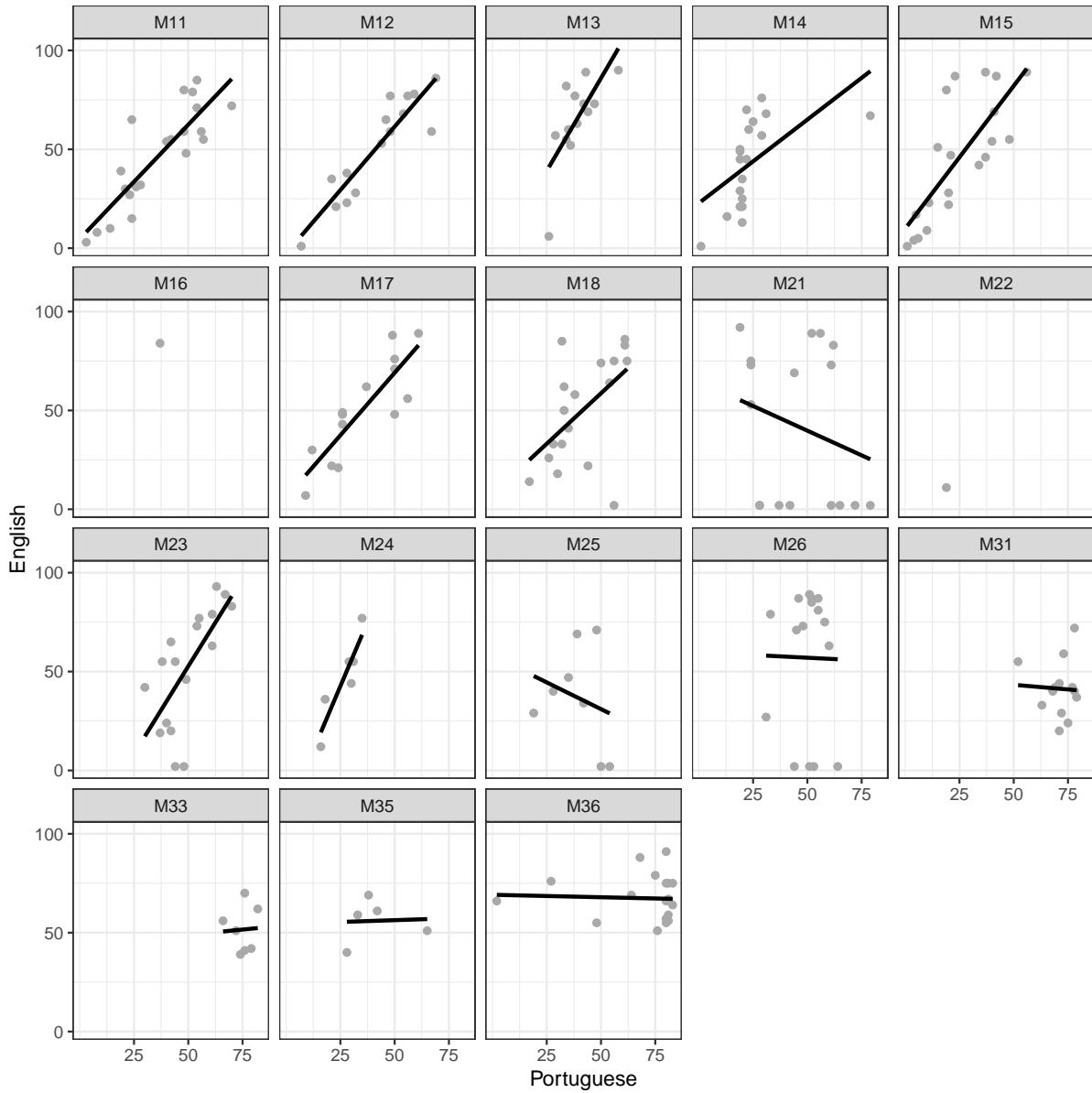


Figure 3.4: Plot of child grades, comparing English with Portuguese grades, shown separately for each school class

Figure 3.4 presents a grid of scatterplots, with a different scatterplot to show the relationship between children's Portuguese and English grades for the children in each different class. We can see that the relationship between Portuguese and English grades is (roughly) similar across classes: in general, children with higher Portuguese grades also tend to have higher English grades. However, Figure 3.4 makes it obvious that there are important differences between

classes.

We can see that the slope of the best fit line (shown in black) varies between classes. And we can see that the intercept (where the line meets the y-axis) also varies between classes. Further, we can see that in some classes, there is no relationship or a negative relationship between Portuguese and English grades.

Critically, we can see that classes differ in how much data we have for each. For some classes, we have many observations (e.g., M11) and for other classes we have few or one observation (e.g., M22, with one child). We know, in advance, that variation in sample size will be associated with variation in the uncertainty we have about the estimated relationship between Portuguese and English grades. You will remember that the Central Limit Theorem allows us to calculate the standard error of an estimate like the mean, given the sample size, and that the standard error is an index of our uncertainty of the estimate.

 Tip

Facetting – notice that, in the plotting code, the key expression is `facet_wrap(~ class_number)` This means:

1. The `class_number` variable is a factor: we ask R to check the summary for the dataframe, and that factor codes what class a child is in.
2. The `facet_wrap(...)` function then asks R to produce separate plots for each *facet* for the data – the word facet means face or aspect.
3. We use the formula `facet_wrap(~ ...)` to ask R to split the data up by using the classification information in the named variable, here `class_number`.

The production of a grid or lattice of plots is a useful method for comparing patterns between data sub-sets or groups. You can see more information about `facet_wrap()` [here](#)

 Tip

Adjusting scales – notice that in these plots I modified the axes to show x-axis and y-axis ticks at specific locations using scale functions. The tick is the notch or short line where we show the numbers on the axes.

1. `scale_x_continuous(breaks=c(25,50,75))` means: set the x-axis ticks at 25, 50 and 75, defined using a vector `c()` of values
2. `scale_y_continuous(breaks=c(0,50,100))` means: set the y-axis ticks at 0, 50, 100, defined using a vector of values.

It can be useful to adjust the ticks on axes, where other plot production factors cause crowding of labels.

You can see more information on scales [here](#).

3.9.7.1 Exercise – edit plots

Do the following exercises to practice your `facet_wrap()` skills

1. Change the x and y variables to `math` and `physics`.
2. Experiment with showing the differences between classes in a different way: instead of using `facet_wrap()` in `aes()` add `colour = class_number`: what happens?

3.9.8 Linear models for multilevel data – dealing with the hierarchical structure

Figure 3.4 shows that there is variation between classes in both the average English grade, shown as differences in the y-axis intercept, and the ‘effect’ of Portuguese language skill, shown as differences in the slope of the best fit line for the predicted relationship between Portuguese and English grades.

We put ‘effect’ in quotes to signal the fact that we do not wish to assert a causal relation. The interpretation of the results of the linear model assumes those results are valid provided that the observations are independent, among other things. We can see that we cannot make the assumption of independence because individuals in classes with high average English grades are more likely to have higher English grades.

We can represent in our analysis the information we have about hierarchical structure in the data (child within class) by allowing the regression coefficients to vary between groups. We therefore modify the subscripting to take into account the fact that we must distinguish which child i and which class j we are examining, adapting our model to:

$$y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$

- where y_{ij} represents the outcome measure, the English grade for each i child in each j class;
- β_{0j} represents the average grade, different in different classes;
- $\beta_{1j}X_{ij}$ represents the variation in X the Portuguese grade for each i child in j class, with the effect of that variation estimated as the coefficient β_{1j} , different in different classes;
- and e_{ij} represents differences between observed and predicted English grades for each i child in j class.

3.9.9 Two-step or slopes-as-outcomes linear models as approximations to the Linear Mixed-effects or Multilevel modeling approach

In practice, we could capture the variation between classes by performing a **two-step analysis**.

- First, we estimate the coefficient of the ‘Portuguese’ effect for each class separately. We do multiple (per-class) analyses. In each of these analyses, we estimate the coefficient looking only at the data for one class.
- Second, we can take those per-class coefficients as the outcome variable in a ‘slopes-as-outcomes analysis’ to examine if the per-class estimates of the experimental effect are reliably different from zero or, more usefully, if the per-class estimates vary in relation to some explanatory variable like teacher skill.

The problem with the approach is apparent in Figure 3.5: the figure shows the estimated intercept and coefficient of the slope of the ‘Portuguese’ effect for each class, when we have analyzed the data for each class in a separate linear model.

The estimate for each class is shown as a black dot. The standard error of the estimate is shown as a black vertical line, shown above and below a point.

You can say that where there is a longer line there we have more uncertainty about the location of the estimate. Notice that **the standard errors vary** a lot between classes. In some classes, the standard error is small (the black line is short) so we can maybe have more certainty over the estimated intercept or slope for those classes. In other classes, the standard error is large (the black line is long) so we can maybe have less certainty over the estimated intercept or slope for those classes.

! Important

The **key idea** here is that standard errors vary widely between classes but the two-step modeling approach, while it can take into account the between-class differences in estimates, cannot account for the variation in the standard errors about those estimates.

! Tip

Notice that the code I used fits a separate model for each class, and then plots the per-class estimates of intercepts and slopes of the `english ~ portuguese` relationship.

3.9.9.1 Discussion: slopes-as-outcomes analyses as a common approach

The ‘slopes-as-outcomes’ approach is quite common and can be found in a number of papers in the psychological and educational research literatures. An influential example can be found in the report by Balota et al. (2004) of their analysis of psycholinguistic effects in older and younger adults. Balota et al. (2004) wanted to examine if or how effects of variables like word frequency, on reading response latencies, were different in different age groups. To do this, they first estimated the effect of the (word-level) psycholinguistic variables in separate linear model (multiple regressions) for each adult. They then took the estimated coefficients as the dependent variable (slopes-as-outcomes) for a second analysis in which they tested the effect of

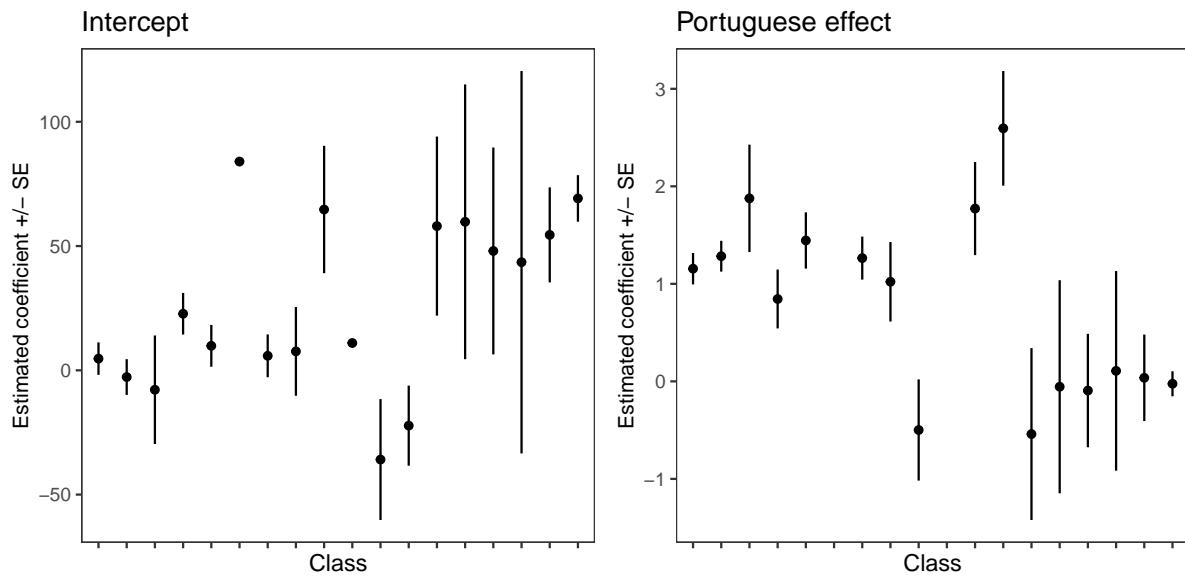


Figure 3.5: Plot showing the estimated intercept and coefficient of the slope of the Portuguese effect for each class analysed separately

age group on variation in the estimated coefficients (see Lorch & Myers (1990) for a discussion of the approach).

You may be asked to do this. I would advise you against it because there is a better way, as we see next.

3.9.10 Multilevel models

Multilevel models incorporate estimates of the intercept and the effect of independent (experimental or correlational) variables plus estimates of the random variation between classes in intercepts and slopes. Multilevel models are also known as **mixed-effects models** because they involve the estimation of **fixed effects** (effects due to independent variables) and **random effects** (effects due to random differences between groups).

We model the intercept (varying between classes) as:

$$\beta_{0j} = \gamma_0 + U_{0j}$$

- where β_{0j} values equal γ_0 the overall average intercept,
- plus U_{0j} the differences, for each class, between that average intercept and the intercept for that class.

We model the coefficient of the (Portuguese language) skill effect as:

$$\beta_{1j} = \gamma_1 + U_{1j}$$

- where β_{1j} values equal γ_1 the average slope,
- plus U_{1j} the differences, for each class, between that average slope and the slope for that class.

These models can then be combined:

$$y_{ij} = \gamma_0 + \gamma_1 X_{ij} + U_{0j} + U_{1j} X_{ij} + e_{ij}$$

such that the English grade observed for each child y_{ij} can be predicted given:

- the average grade overall γ_0 plus
- the average relationship between English and Portuguese language skills γ_1 plus
- adjustments to capture
 1. the difference between the average grade overall and the average grade for the child's class U_{0j} ,
 2. as well as the difference between the average slope of the Portuguese effect and the slope of that effect for their class $U_{1j} X_{ij}$,
 3. plus any residual differences between the observed and the model predicted English grade e_{ij} .

3.9.11 How should we think about the differences between the classes?

If you have had some experience analyzing psychological data, then there is a potential approach to thinking about the effect of the differences between classes that would seem natural. This approach has, in fact, been proposed (see e.g. Lorch & Myers, 1990) and applied in the literature. I would not recommend using it but thinking about it helps to develop our understanding of what we are doing with multilevel models

We could seek to estimate (1.) the relationship of interest – here, the association between English and Portuguese grades, while (2.) also estimating the relationship between (outcome) English grades and the impact made by what class a child is in.

In this approach, we would construct a model in which we have English grades as the outcome variable, Portuguese grades as one predictor variable, and then add a variable to code for class identity. The ‘effect’ of the class variable would then capture the differences in intercepts between classes. You could add a further variable to code for differences between classes in the slope of the relationship between English and Portuguese grades. This would allow for the fact that the relationship is positive or negative, stronger or weaker, in different classes.

If we added that further variable to code for differences between classes in the English-Portuguese relationship, then we would be estimating those differences as *interactions* between (1.) the predictive ‘effect’ of Portuguese grades on English grades and (2.) the class effect. An interaction effect is what we have when the effect of one variable (the predictive ‘effect’ of Portuguese grades) is different for different levels of the other variable (the predictive ‘effect’ of Portuguese grades is different for different classes).

Thinking about this approach helps us to think about what we are doing when we are working with multilevel structured data. But the problem with the approach is that it does not allow us to generalize beyond the sample we have. The estimates we would have, given a model in which we code directly for class, would tell us only about the classes in our sample.

Most of the time, we would prefer to do analyses whose results could be generalized: to other children, to other classes, etc.. For this reason, it makes more sense to suppose that $U_{0j} + U_{1j}$, the class-level deviations, are unexplained differences drawn at random from a population of potential class differences.

What does this mean? Think back to your understanding of the linear model. You have learnt that when we fit a linear model like $y_i = \beta_0 + \beta_1 X_i + e_i$ we include a term e_i to represent the differences, for each child, between the *observed* (outcome) English grade and the English grade *predicted* by the relationship with Portuguese grades. Those differences between observed and predicted outcomes are called **residuals**. We assume that the direction and size of any one residual, for any one child, is randomly determined because we typically have no idea why there might be a big difference between the predicted and observed grade for one child but a small residual difference for another.

Now, we can imagine that there will be many classes in many schools, and we can surely expect that there will be differences between the classes. These differences will result from plenty of factors we do not measure or cannot explain. Indeed, we have seen that there are differences between the average (outcome) English grade i.e. the *predicted* intercept and the observed intercept for each class, and we have seen that there are differences between the average slope of the English-Portuguese grades relationship i.e. the *predicted* slope and the slope for each class. These differences would, in effect, be *random differences*. And we can see how the variation in these differences are, for us, just a kind of random error variance, which we can see as **class-level residuals**.

We suppose, technically, that the differences between classes, controlling for the effect of the explanatory variable, are exchangeable (it does not matter which class is which), with classes varying at random. Multilevel models incorporate estimation of the explanatory variables effects $\gamma_0 + \gamma_1$ accounting for group-level error variance $U_{0j} + U_{1j}$.

The difference between the two-step approach and the multilevel modelling approach is this:

- In the two-step approach, seen earlier, we estimate – separately, for each class – the intercept and the slope of the Portuguese effect, using *just the data for that class* and ignoring the data for other classes.

- In the multilevel model, in contrast, we use all the observations, estimating the average intercept and the average slope of the Portuguese effect *plus the variance* due to the difference for each class (1.) between the average intercept and the class intercept or (2.) between the average slope and the class slope.
- We can understand these differences between the average (intercept or class) and the class differences as class-level residuals $U_{0j} + U_{1j}$ in addition to the child-level residuals e_{ij} .

This leads us to a conclusion that represents a critical way to understand multilevel models.

! Important

The multilevel model is a linear model but with **multiple random effects** terms to take into account the hierarchical structure in the data.

3.9.12 Fitting a multilevel model using the `lmer()` function

We can fit a multilevel model of the `english ~ portuguese` relationship, taking into account the fact that the pupils tested in the study were recruited from different classes, using the convenient and powerful `lmer()` function.

Happily, the code we use to define a model in `lmer()` is much like the code we use to define a model in `lm()` with *one important difference* which I shall explain. Let's try it out.

To use the `lmer()` function you need to make the `lme4` library available.

```
library(lme4)
```

The model you are going to code will correspond to the statistical model that we have been discussing:

$$y_{ij} = \gamma_0 + \gamma_1 X_{ij} + U_{0j} + U_{1j} X_{ij} + e_{ij}$$

And the code is written as follows.

```
1 porto.lmer1 <- lmer(english ~ portuguese +
2
3             (portuguese + 1|class_number),
4
5             data = brazil)
```

You can see that the `lmer()` function call is closely similar to the `lm()` function call, with one critical exception, as I explain next.

First, we have a chunk of code mostly similar to what we do when we do a regression analysis.

1. `porto.lmer1 <- lmer(...)` creates a *linear mixed-effects model* object using the `lmer()` function.
2. `english ~ portuguese` is a formula expressing the model in which we estimate the fixed effect on the outcome or dependent variable `english` (English grades) predicted `~` by the independent or predictor variable `portuguese` (Portuguese grades).
3. If there were more terms in the model, the terms would be added in series separated as variable names separated by `+ sum` symbols.
4. `...(..., data = brazil)` specifies the dataset in which you can find the variables named in the model fitting code.
5. `summary(porto.lmer1)` gets a summary of the fitted model object.

Second, we have a bit that is specific to multilevel or mixed-effects models.

6. Critically, we add `(...|class_number)` to tell R about the random effects corresponding to random differences between sample groups (classes) coded by the `class_number` variable.
7. `(...1 |class_number)` says that we want to estimate random differences between sample groups (classes) in intercepts coded `1`.
8. `(portuguese... |class_number)` adds random differences between sample groups (classes) in slopes of the `portuguese` effect coded by using the `portuguese` variable name.

If you run the model code as written – see the .R workbook file for an example of the code – and it works then the code will be shown in the console window in R-Studio. To show the model results, you need to get a summary of the model, using the model name.

```
summary(porto.lmer1)
```

3.9.12.1 Exercise – fitting linear mixed-effects models

Mixed-effects model code is hard to get used to *at first*. A bit of practice helps to show you which bits of code are important, and which bits you will change for your own analyses

1. Change the outcome (from `english`) and the predictor (from `portuguese`): this is about changing the fixed effect part of the model.
2. Vary the random effects part of the model.
 - Change it from `(portuguese + 1 | class_number)` to `(1 | class_number)` what you are doing is asking R to ignore the differences in the slope of the effect of Portuguese grades.

- Change it from `(portuguese + 1 | class_number)` to `(portuguese + 0 | class_number)` what you are doing is asking R to ignore the differences in the intercept.

Try out these variations and *look carefully* at the different results.

3.9.13 Reading the *lmer* results

Now let's take a look at the results.

```
summary(porto.lmer1)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: english ~ portuguese + (portuguese + 1 | class_number)
Data: brazil

REML criterion at convergence: 2104.3

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-2.81321 -0.59584  0.04359  0.60018  2.23722

Random effects:
 Groups           Name        Variance Std.Dev. Corr
 class_number (Intercept) 341.4803 18.479
               portuguese   0.3295  0.574   -0.98
 Residual          493.1009 22.206
Number of obs: 230, groups: class_number, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept) 25.2837    6.2669   4.034
portuguese   0.6590    0.1729   3.811

Correlation of Fixed Effects:
  (Intr)
portuguese -0.943
```

Notice that the output has a number of elements.

1. First, we see information about the function used to fit the model, and the model object created by the `lmer()` function call.

2. Then, we see the model formula `english ~ portuguese + (portuguese + 1 | class_number)`.
3. Then, we see ‘REML criterion at convergence: about the model fitting process, which we can usually ignore.
4. Then, we see information about the distribution of the model residuals.
5. Then, we see information about the error terms estimated (technically, predicted) by the model.
 - Residuals, just like a linear model plus error terms specific to multilevel or mixed-effects models, group-level residuals;
 - differences between the average intercept and, here, the intercept (average ‘english: score) per class;
 - and differences between the average slope capturing the *english ~ portuguese* relationship and the slope (of the effect) per class.
6. Then, just as for linear models, we see estimates of the coefficients (of the slopes) of the fixed effects, the intercept and the slope of the *english ~ portuguese* relationship.

Note that we see coefficient estimates like in a linear model summary **but no p-values**. We will come back to p-values later but note that their absence is not a bug. Note also that we do not see an R^2 estimate. We will come back to that too.

3.9.13.1 Exercise – How do we report mixed-effects models results?

There is no convention, yet, on how to report the results of these models. Lotte Meteyard and I argue for a set of conventions that will help researchers to understand each others’ results better.

- In this exercise, go read the bit where we advise Psychologists how to write about the results, in our paper (Meteyard & Davies, 2020).

3.9.14 Fixed and random effects

You will have noticed the reference to fixed and random effects in the discussion in this chapter, and the use of fixed and random effects as titles for sections of the model output. The terms *fixed effects* and *random effects* are not used consistently in the statistical literature (Gelman & Hill, 2007). I am going to use these terms because they are helpful, at first, and because they are widely used, not least in the results of our analyses in R.

It is common in the psychological literature to refer to the effects on outcomes of experimental manipulations (e.g., the effect on outcomes of differences between experimental conditions) or to the effects of correlated variables of theoretical or explanatory interest (e.g., the effect of differences in Portuguese language skill) as *fixed effects*. Typically, we are aiming to get

estimates of the coefficients of these effects. And, much like we would do when we use linear models, we expect that these coefficients represent an estimate of the effects of these variables, on average, across the population. (Hence, some analysts prefer to talk about these effects as population average effects).

In comparison, when we are thinking about the effects on outcomes of what we understand to be the random differences between sampled children (e.g., the child-level residuals) or the random differences between sampled classes (e.g., the class-level residuals) then we refer to these effects as *random effects*. As we have seen, we usually estimate random effects as variances and can estimate them as random error variances. While we may care to estimate how differences in, say, Portuguese language score, is associated with English grade, we typically do not care about the impact of the specific difference between any two classes in English grade.

However, if you take your education in this area further, you will find that the way that fixed and random effects are talked about in the statistical literature can be, at best, inconsistent. And, ultimately, you might ask yourself if there are principled distinctions between fixed and random effects. We can leave these problems aside, here, because they do not influence how we shall learn and practice using multilevel models in the early to medium term in our development of skills and understanding.

3.9.15 Is there a difference between linear model and linear mixed-effects model results?

Recall that the linear model yields the estimate that for unit increase in Portuguese grade there is an associated .5 increase in English grade, on average ($F(1, 228) = 35, p < .001; adj.R^2 = .13$). We can see that the estimate of the effect **for the mixed-effects model** is $\beta = .659$ which is somewhat different from the effect estimate we got from the linear model.

i Note

- Research question: Does Portuguese grade predict English grade?
- Result: Our analysis shows that children who score one grade higher in their Portuguese class (e.g., 61 compared to 60) tended to score about .66 of a grade higher in their English class.

3.10 Conclusions

Psychological studies frequently result in hierarchically structured data. The structure can be understood in terms of the grouping of observations, as when there are multiple observations per group, participant or stimulus. The existence of such structure must be taken into account in analyses. Multilevel or mixed-effects models can be specified by the researcher to include

random effects parameters that capture unexplained differences between participants or other sampling units in the intercepts or the slopes of explanatory variables. Where a sample of participants is asked to respond to a sample of stimuli, structure relating both to participants and to stimuli can be incorporated.

Many researchers will be aware of concerns over the non-replication of published effects in psychological science (Pashler & Wagenmakers, 2012). As Gelman (2014) discusses, non-replication of results may arise if effects vary between contexts groups while traditional analytic methods assume that effects are constant. Psychological researchers can expect average outcomes and the effects of independent variables to vary among sampling units, whether their investigations involve multiple observations per child or stimulus, or children sampled within classes, clinics or schools. However, traditional analytic methods often require us to ignore this variation by averaging, say, responses over multiple trials for each child, to collapse an inherently multilevel data structure into a single level of observations that can be analysed using regression or ANOVA. By using multilevel models, researchers will, instead, be able to properly estimate the effects of theoretical interest, and better understand how those effects vary.

3.10.1 Summary

We outlined the features of a *multilevel* structure dataset. We then discussed visualizing and modeling the relationship between outcome and predictor variables:

1. When you ignore the multilevel structure;
2. Why ignoring the structure is a bad idea;
3. How slopes-as-outcomes analyses are an approximate method to take the structure into account;
4. How multilevel modeling is a better method to estimate the relationship between outcome and predictor variables

3.10.2 Glossary: useful functions

We used functions to read-in the libraries of functions we needed.

- `library(tidyverse)`
- `library(lme4)`

We used some new functions, or focused on functions we have seen before but not discussed.

- We used `load()` to load an `.RData` file into R workspace.
- For visualization, we used `facet_wrap()` to show plots of the relationship between English and Portuguese scores in each class separately.
- We used `lmer()` to fit a multilevel model.

We used the `summary()` function to get model results for both linear models and for the multilevel or liner mixed-effects model.

3.11 Recommended reading

Snijders & Bosker (2004) present a helpful overview of multilevel modelling. Readers familiar with the book will see that I rely on it to construct the formal presentation of the models.

Baayen et al. (2008b; see also Barr et al., 2013b; Judd et al., 2012) discuss mixed-effects models with crossed random effects. This is the topic we shall discuss next.

I wrote a tutorial article on mixed-effects models with Lotte Meteyard (Meteyard & Davies, 2020a). We discuss how important the approach now is for psychological science, what researchers worry about when they use it, and what they should do and report when they use the method.

4 Introduction to linear mixed-effects models

4.1 Motivations: repeated measures designs and crossed random effects

In the **Introduction to multilevel data**, we looked at a multilevel structured dataset in which there are observations about children's grades, and it is evident that those children can be grouped by or under classes. As we discussed, this kind of data structure will come from studies with a very common design in which the researcher recorded observations about a sample of children who are members of a sample of classes. In working with these kind of data, it is common to say that the observations of children's grades are *nested* within classes in a hierarchy.

Many Psychologists conduct studies where observations are properly understood to be structured in groups of some form but where, nevertheless, it is inappropriate to think of the observations as being nested (Baayen et al., 2008). We are talking, here, about **repeated-measures designs** where the experimenter presents a sample of multiple stimuli for response to each participant in a sample of for multiple participants. This is another *very* common experimental design in psychological science. Studies with this kind of design will produce data with a structure that, also, requires the use of mixed-effects models but, as we shall see, the way we think about the structure will be a bit more complicated. We could say that observations of the responses made by participants to each stimulus can be grouped by participant: each person will tend to respond in similar ways to different stimuli. Or, we could say that observations of responses can be grouped by stimulus because each stimulus will tend to evoke similar kinds of responses in different people. Or, we could say that both forms of grouping should be taken into account at the same time.

We shall take the third position and this chapter will concern why, and how we will adapt our thinking and practice.

4.2 The key idea to get us started

Linear mixed-effects models and multilevel models are basically the same.

This week, we again look at data with multilevel structure. But we are looking at data where participants were asked to respond to a set of stimuli (words) so that our observations consist of recordings made of the response made by each child to each stimulus. We use the same procedure we did for multilevel data but with one significant change which we shall identify and explain.

4.3 Targets

Our learning objectives again include the development of both understanding and practical skills.

skills Practice how to tidy experimental data for mixed-effects analysis

concepts Begin to develop an understanding of crossed random effects of subjects and stimuli

skills and concepts Practice fitting linear mixed-effects models incorporating random effects of subjects and stimuli

4.4 Study guide

1. Read in the example CP study data
2. Identify how the data are structured by both participant and stimulus differences
3. Use visualizations to explore the impact of the structure
4. Run analyses using linear mixed-effects models involving multiple random effects
5. Review readings provided

4.5 The data we will work with: CP reading study

This week, we will be working with the **CP reading study** dataset. CP tested 62 children (aged 116-151 months) on reading aloud in English. In the experimental reading task, she presented 160 words as stimuli. The same 160 words were presented to all children. The words were presented one at a time on a computer screen. Each time a word was shown, the children had to read the word out loud and their response was recorded. Thus, the CP reading study dataset comprised observations about the responses made by 62 children to 160 words.

In addition to the reading task, CP administered tests of reading skill (TOWRE sight word and phonemic tests, Torgesen et al., 1999), reading experience (CART, Stainthorp, 1997), the Spoonerisms sub-test of the Phonological Awareness test Battery (Frederickson et al., 1997),

and an orthographic choice test measure of orthographic knowledge (based on Olson et al., 1985). She also recorded the gender and the handedness of the children.

Ultimately, the CP dataset were incorporated in an analysis of the impact of age on reading skills over the life-span, reported by Davies, Arnell, Birchenough, Grimmond and Houlson (2017). You can find more details on the data and the methods in that paper.

The CP study resembles many studies in psychological science. The critical features of the study are that we have an outcome measure – the reading response – observed multiple times (for each stimulus) for each participant. We have 160 responses recorded for each participant, one response for each stimulus word. And we have 62 responses recorded for each word, one response for each participant. The presence of these features is the reason why we need to use mixed-effects models in our analysis. These features are common across a range of study designs so the lessons we learn will apply frequently in psychological research. This is the reason why it is important we teach and learn how to use mixed-effects models.

4.5.1 Our research question

We are going to use these data to examine the answers to the following question:

RQ.1. What word properties influence responses to words in a test of reading aloud?

We can look at the answers to this question while also taking into account the impacts of random differences – between sampled participants or between sampled words – using mixed-effects models. But, first, we are going to look at how we get the data ready for analysis.

4.5.2 The challenges of working with real (untidy) experimental data

Ordinarily, textbooks and guides to data analysis give you the data ready for analysis but this situation will never be true for your professional practice (at least, not at first). Instead of pretending that data arrive ready for analysis, we are going to look at the process of **data tidying**, step-by-step. This will help you to get ready for the same process when you have to develop and use it in your own research.

We are going to spend a bit of time looking at the data tidying process. This process involves identifying and resolving a series of challenges, in order. Looking at the tidying process will give you a concrete sense of the structure in the data. You should also take this opportunity to reflect on the nature of the process itself – what we have to do and why, in what order and why – so that you can develop a sense of the process you might need to build when the time comes for you to prepare your own data for analysis.

The time that we spend looking at data tidying is an investment in learning that will save you time later, in your professional work. If, however, you want to skip it, go to section @ref(crossed-random).

4.5.2.1 The data we need to use for analysis are not all in the same file

In analyzing psychological data, the first step is usually to collect the data together. In psychological research, the data may exist, at first, in separate files. For the CP study, we have *separate files* for each of the pieces of information we need to use in our analyses:

Participant attributes – information about participants' age, gender, identifier code, and abilities on various measures.

Stimulus attributes – information about stimulus words, e.g., the word, its item number, its value on each variable in a set of psycholinguistic properties (like word length, frequency).

Behaviour – behavioural observations e.g. reaction time or accuracy of responses made by each participant to each stimulus word.

Often, we need all these kinds of information for our analyses but different pieces of information are produced in separate ways and come to us in separate files. For example, we may collect experimental response data using software like PsychoPy, E-Prime, Qualtrics or DMDX. We may collect information about participant characteristics using standardized measures, or by asking participants to complete a set of questions on their age, gender, and so on.

4.5.2.2 The data we need to use are untidy

Often, the files we get are untidy: not in a useful or *tidy* format. For example, if you open the file `CP_study_word_naming_rt_180211.dat` (a .dat or tab delimited file) in Excel, you will see a spreadsheet that looks like Figure @ref(fig:CPrt).

Typical of the output from data collection software, we can see a data table with:

1. in the top row, column header labels `item_name`, `AislingoC`, `AllanaD` ...;
2. in the first (leftmost) column, row labels `item_name`, `act`, `ask`, `both` ...;
3. for each row, we see values equal to the reaction time (RT) observed for the response made to each stimulus (listed in the row labels);
4. for each column, we see values equal to the RTs observed for each person (listed in the column labels);
5. and at each intersection of row and column (for each cell), we see the RT observed for a response made by a participant to a stimulus.

Data laid out like this are sometimes said to be in *wide* format. You can see that the data are *wide* because at least one variable – here, reading reaction time – is held not in one column but spread out over several columns, side-by-side. Thus, the dataset is wide with fewer rows and many columns.

We want the data in what is called the *tidy* format.

	A	B	C	D	E	F	G	H								
1	item_name	Aisling	oC	Alex	B	Allana	D	Amy	R	Andy	D	Anna	F	Aoife	H	C
2	act		594.8		586	-999		693		597		627		649		
3	ask		481.5		864	1163		694.4		616		631		538		
4	both		457.5		670	1114.3		980		1019		796.1		545.2		
5	box		546		748.6	975		678		589		604		574		
6	broad		580		1473.5	-999		789		684		-999		815.6		
7	bronze		546		861.2	-999		845		731.9		803.4		487.1		
8	calf		552		766	-999		927.8		670		618		751.4		
9	can		508		726	1148.2		852		649		617		473.6		
10	care		572		909.8	-999		948		766		544.2		558.5		
11	carve		664.3		702	-999		934.4		629.9		883.5		892.3		
12	chance		757.1		751.3	-999		982.1		833.8		667		527.6		

Figure 4.1: CP study RTs .dat file

4.5.2.3 How tidy data are tidy

There are three inter-related rules which make data *tidy* (Grolmund & Wickham, 2019):

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

You can read more about tidy data here:

<http://r4ds.had.co.nz/tidy-data.html>

For our purposes, the reason we want the data in *tidy* format is that it is required for the functions we are going to use for mixed-effects modelling. However, in general, *tidy* format is maximally flexible, and convenient, for use with different R functions.

4.5.3 Locate and download the data files

Go to the 402 Moodle folder for week 18, and download the .zip (compressed) folder labeled **PSYC402-01-multilevel-resources**

Or, download the same folder by clicking on the link:

<https://modules.lancaster.ac.uk/mod/resource/view.php?id=1795341>

In this folder, we have got four files that we will need to import or read in to R:

- CP study word naming rt 180211.dat
- CP study word naming acc 180211.dat
- words.items.5 120714 150916.csv
- all.subjects 110614-050316-290518.csv

The `words.items` file holds information about the 160 stimulus words presented in the experimental reading (word naming) task. The `all.subjects` file holds information about the 62 participants who volunteered to take part in the experiment. The `.csv` files are *comma separated values* files. The `.dat` files are *tab delimited* files holding behavioural data: the latency or reaction time `rt` (in milliseconds) and the accuracy `acc` of response made by each participant to each stimulus.

4.6 Tidy the data

To answer our research question, we will need to combine the behavioural data with information about the participants (age, gender ...) and about the words (word, frequency ...) We will need to ensure that the data-set we construct will be in *tidy* format. We will need to *select* variables (columns) to get just those required for our later analyses. And we will need to *filter* cases (rows), excluding errors or outliers.

We shall need to do this work in a series of processing steps:

1. Import the data or read the data into R, see Section @ref(import)
2. Restructure the data, see Section @ref(restructure)
3. Select or transform variables, see Section @ref(transform)
4. Filter observations, see Section @ref(filter)

We will use `tidyverse` library functions from the begining, starting with the import stage.

```
library(tidyverse)
```

(Every step can also be done in alternative processing steps with the same result using *base R* code.)

4.6.1 Read in the data files by using the `read_csv` and `read_tsv` functions

I am going to assume you have downloaded the data files, that they are all in the same folder, and that you know where they are on your computer or server. We need to use different versions of the `read_` function to read all four files into R.

```
behaviour.rt <- read_tsv("CP study word naming rt 180211.dat", na = "-999")
behaviour.acc <- read_tsv("CP study word naming acc 180211.dat", na = "-999")
subjects <- read_csv("all.subjects 110614-050316-290518.csv", na = "-999")
words <- read_csv("words.items.5 120714 150916.csv", na = "-999")
```

These different versions respect the different ways in which the `.dat` and `.csv` file formats work. We need `read_tsv()` when data files consist of tab separated values. We need `read_csv()` when data files consist of comma separated values.

You can read more about the **tidyverse** `readr` library of helpful functions here:

<https://readr.tidyverse.org/>

It is *very* common to get experimental data in all sorts of different formats. Learning to use **tidyverse** functions will make it easier to cope with this when you do research.

4.6.1.1 Code tip

Notice, here, that we use the `read_` function to read in the data, entering two arguments inside the brackets after the function name. For example, we write the code as:

```
behaviour.rt <- read_tsv("CP study word naming rt 180211.dat", na = "-999")
```

Take a look at what this line of code includes, element by element.

1. We write `behaviour.rt <- read_tsv(...)` to create an object in the R environment, which we call `behaviour.rt` – the object with this name is the dataset we read into R using `read_tsv(...)`.
2. When we write the function `read_tsv(...)` we include two arguments inside it.
3. `read_tsv("CP study word naming rt 180211.dat", ...)` first, the name of the file, given in quotes "" and then a comma.
4. `read_tsv(..., na = "-999")` second, we tell R that there are some missing values `na` which are coded with the value `"-999"`.

4.6.1.2 A quick lesson about missing value codes

In R, a missing value is said to be “not available”: NA.

In the datasets – typically, the spreadsheets – we create in our research, we will have values missing for different reasons. Take another look at the data spreadsheet you saw earlier, Figure @ref(fig:CPrt-2).

	A	B	C	D	E	F	G	H	C
1	item_name	AislingC	AlexB	AllanaD	AmyR	AndyD	AnnaF	AoifeH	C
2	act	594.8	586	-999	693	597	627	649	
3	ask	481.5	864	1163	694.4	616	631	538	
4	both	457.5	670	1114.3	980	1019	796.1	545.2	
5	box	546	748.6	975	678	589	604	574	
6	broad	580	1473.5	-999	789	684	-999	815.6	
7	bronze	546	861.2	-999	845	731.9	803.4	487.1	
8	calf	552	766	-999	927.8	670	618	751.4	
9	can	508	726	1148.2	852	649	617	473.6	
10	care	572	909.8	-999	948	766	544.2	558.5	
11	carve	664.3	702	-999	934.4	629.9	883.5	892.3	
12	chance	757.1	751.3	-999	982.1	833.8	667	527.6	

Figure 4.2: CP study RTs .dat file

You should be able to see that the spreadsheet holds information, as explained, about the RTs of the responses made by each child to each stimulus word. Each of the cells in the spreadsheet (i.e. the box where a column intersects with a row) includes a number value. Most of the values are positive numbers like 751.3: the reaction time of a response, recorded in milliseconds. The values have to be positive because they represent the length of time between the moment the stimulus word is presented on the test computer screen and the moment the child’s spoken word response has begun to be registered by the computer microphone and sound recording software.

Some of the cells hold the value -999, however. Obviously, we cannot have negative RT. The value represents the fact that we have no data. Take a look at Figure @ref(fig:CPrt-2): we have a -999 where we should have a RT for the response made by participant AllanaD to the word broad. This -999 is there because, for some reason, we did not record an RT or a response for that combination of participant and stimulus.

We can choose any value we like, as researchers, to code for missing data like this. Some researchers choose not to code for the absence of a response recording or leave the cell in

a spreadsheet blank or empty where data are missing. This is **bad practice** though it is common.

There are a number of reasons why it is bad practice to just leave a cell empty when it is empty because no observation is to be recorded.

1. Data may be missing for different reasons: maybe a child did not make any response to a stimulus (often called a “null response”); or maybe a child made a response but there was a microphone or other technical fault; or maybe a child made a response but it was an error and (here) the corresponding performance measure (RT) cannot be counted.
2. If you do not code for missingness in the data then the software you use will do it for you, but you may not know how it does so, or where.
3. If you have missing data, you ought to be able to identify where the data are missing.

I use -999 to code for missing values because you should never see a value like that in real reading RT data. You can use whatever value you like but you should make sure you *do* code for missing data somehow.

4.6.2 Reshape the data from wide to long using the `gather()` function

We are going to need to restructure these data from a wide format to a longer format. We need to restructure both behavioural data-sets, accuracy and RT. We do this using the `pivot_longer()` function.

```
rt.long <- behaviour.rt %>%  
  pivot_longer(2:62, names_to = "subjectID", values_to = "RT")  
  
acc.long <- behaviour.acc %>%  
  pivot_longer(2:62, names_to = "subjectID", values_to = "accuracy")
```

Doing data-set construction programmatically, using R functions, is generally more reliable, and faster, than doing it by hand. Researchers used to have to do this sort of thing by hand, using copying and pasting, in Excel or SPSS. Doing the process by hand takes many hours or days. And you *always* make errors.

4.6.2.1 Code tip

Here, we use a function you may not have seen before: `pivot_longer()`.

```
rt.long <- behaviour.rt %>%  
  pivot_longer(2:62, names_to = "subjectID", values_to = "RT")
```

The name of the function comes from the fact that we are starting with data in wide format e.g. `behaviour.rt` where we have what should be a single variable of observations (RTs) arranged in a wide series of multiple columns, side-by-side (one column for each participant). But we want to take those wide data and *lengthen* the dataset, increasing the number of rows and decreasing the number of columns.

Let's look at this line of code bit by bit. It includes a powerful function that accomplishes a lot of tasks, so it is worth explaining this function in some detail.

1. `rt.long <- behaviour.rt %>%`

- At the start, I tell R that I am going to create a new longer dataset (more rows, fewer columns) that I shall call `rt.long`.
- I will create this longer dataset from `<-` the original wide dataset `behaviour.rt`.
- and I will create the new longer dataset by taking the original wide dataset and piping it `%>%` to the pivot function coded on the next line:

2. `pivot_longer(2:62, names_to = "subjectID", values_to = "RT")`

- On this next line, I tell R how to do the pivoting by using three arguments.

a. `pivot_longer(2:62...)`

- First, I tell R that I want to re-arrange all the columns that can be found in the dataset from the second column to the sixty-second column.
- In a spreadsheet, we have a number of columns.
- Columns can be identified by their position in the spreadsheet.
- The position of a column in a spreadsheet can be identified by number, from the leftmost column (column number 1) to the rightmost column (here, column number 62) in our dataset.
- So this argument tells R exactly which columns I want to pivot.

b. `pivot_longer(..., names_to = "subjectID", ...)`

- Second, I tell R that I want it to take the column labels and put them into a new column, called `subjectID`.
- In the wide dataset `behaviour.rt`, each column holds a list of numbers (RTs) but begins with a word in the topmost cell, the name code for a participant, in the column label position.
- We want to keep the information about which participant produces which response when we pivot the wide data to a longer structure.
- We do this by asking R to take the column labels (the participant names) and listing them in a new column, called `subjectID` which now holds the names as participant ID codes.

c. `pivot_longer(...values_to = "RT")`

- Third, we tell R that all the RT values should be put in a single column.
- We can understand that this new column RT will hold RT observations in a vertical stack, one cell for each response by a person to a word, with rows ordered by `subjectID`.

There are 61 columns of data listed by participant though 62 children were tested because we lost one child's data through an administrative error. As a result, in the wide data sets there are 62 columns, with the first column holding `item_name` data.

You can find more information about pivoting data here:

<https://tidyverse.org/articles/pivot.html>

And you can find more information specifically about the `pivot_longer()` operation here:

<https://tidyverse.org/articles/pivot.html>

4.6.2.1.1 Why we restructure the data

As I noted, one problem with the wide format is that the data are structured so that the column names are not names of variables. In our example wide format dataset `behaviour.rt`, the columns are headed by a participant identity code or name but a participant code is not the name of a variable, it is a value of the variable I call `subjectID`. In the design of the CP reading study, we want to take into account the impact of differences between participants on response RT (so, we need to identify which participant makes which response). But we do not see the responses made by a participant as a predictor variable.

A second problem is that, in a wide format file like `behaviour.rt`, information about the responses made to each stimulus word is all on the same row (that seems good) but in different columns. Each person responded to all the words. But the response made to a word e.g. `act` made by one participant is in a different column (e.g., 594.8ms, for `AislingoC`) from the response made to the same word by a different participant (e.g., 586ms, for `AlexB`). This means that information about the responses made to each stimulus word are spread out as values across multiple columns.

You can see this for yourself if you inspect the source data using `head()`.

```
head(behaviour.rt)

# A tibble: 6 x 62
  item_name Aisli~1 AlexB AllanaD AmyR AndyD AnnaF AoifeH Chloe~2 ChloeF ChloeS
  <chr>      <dbl> <dbl>    <dbl> <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl>
1 act        595.  586     NA     693   597   627   649    1081  642   623.
2 ask        482.  864     1163   694.  616   631   538    799.  603   526
3 both       458.  670     1114.  980   1019  796.  545.    NA    581   568.
4 box         546   749.    975   678   589   604   574    658   689.  492
5 broad      580   1474.   NA    789   684   NA    816.   NA    NA    798
```

```

6 bronze      546   861.     NA   845   732.   803.   487.   1701   871   574
# ... with 51 more variables: CianR <dbl>, ConorF <dbl>, DavidL <dbl>,
# DillonF <dbl>, DJHerlihy <dbl>, EamonD <dbl>, EimearK <dbl>, EllenH <dbl>,
# EoinL <dbl>, GrainneH <dbl>, JackBr <dbl>, JackK <dbl>, JackS <dbl>,
# JamesoC <dbl>, JenniferoS <dbl>, KateF <dbl>, KayleighMc <dbl>, KenW <dbl>,
# KevinL <dbl>, KieranF <dbl>, KillianB <dbl>, KirstyC <dbl>, LeeJ <dbl>,
# MarkC <dbl>, MatthewC <dbl>, MeganOB <dbl>, MichaelaoD <dbl>,
# NataliaR <dbl>, NiallG <dbl>, NiallGavin <dbl>, NiallW <dbl>, ...

head(behaviour.acc)

# A tibble: 6 x 62
  item_name Aisli~1 AlexB AllanaD AmyR AndyD AnnaF AoifeH Chloe~2 ChloeF ChloeS
  <chr>       <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 have        1     1     1     1     1     1     1     1     1     1     1
2 cheer       1     1     0     1     1     1     1     1     1     1     1
3 ask         1     1     1     1     1     1     1     1     1     1     1
4 care        1     1     0     1     1     1     1     1     1     1     1
5 with        1     1     1     1     1     1     1     1     1     1     1
6 false       1     1     0     1     1     1     1     1     1     1     1
# ... with 51 more variables: CianR <dbl>, ConorF <dbl>, DavidL <dbl>,
# DillonF <dbl>, DJHerlihy <dbl>, EamonD <dbl>, EimearK <dbl>, EllenH <dbl>,
# EoinL <dbl>, GrainneH <dbl>, JackBr <dbl>, JackK <dbl>, JackS <dbl>,
# JamesoC <dbl>, JenniferoS <dbl>, KateF <dbl>, KayleighMc <dbl>, KenW <dbl>,
# KevinL <dbl>, KieranF <dbl>, KillianB <dbl>, KirstyC <dbl>, LeeJ <dbl>,
# MarkC <dbl>, MatthewC <dbl>, MeganOB <dbl>, MichaelaoD <dbl>,
# NataliaR <dbl>, NiallG <dbl>, NiallGavin <dbl>, NiallW <dbl>, ...

```

This structure is a problem for visualization and for analysis because the functions we will use require us to specify *single* columns for an outcome variable like reaction time.

We are looking at the process of tidying data because untidiness is very common. Learning how to deal with it will save you a lot of time and grief later.

You should check for yourself how `subjectID` and `RT` or `accuracy` scores get transposed from the old structure to the new structure.

```

head(rt.long)

# A tibble: 6 x 3
  item_name subjectID    RT
  <chr>      <chr>     <dbl>
1 act        AislingoC  595.
2 act        AlexB      586
3 act        AllanaD    NA

```

```
4 act      AmyR      693
5 act      AndyD     597
6 act      AnnaF     627
```

```
head(acc.long)
```

```
# A tibble: 6 x 3
  item_name subjectID accuracy
  <chr>      <chr>       <dbl>
1 have      AislingoC     1
2 have      AlexB        1
3 have      AllanaD      1
4 have      AmyR         1
5 have      AndyD        1
6 have      AnnaF        1
```

If you compare the `rt.long` or `acc.long` data with what you see in when the data are in the original wide format then you can see how – in going from wide – we have re-arranged the data to a longer and narrower set of columns, one column listing each word, one column for `subjectID` and one column for `RT` or `accuracy`. What a check will show you is that we have multiple rows for responses to each item so that the item is repeated multiple times in different rows.

These data are *now tidy*.

- Each column has information about one variable
- And each row has information about one observation, here, the response made by a participant to a word

But these data are *incomplete*. Next we shall combine behavioural observations with data about stimulus words and about participants.

4.6.2.1.2 The tidyverse evolves

Over the years, different ways of reshaping data have evolved. This reflects how important and common the task is. An older way to do the same operation uses the function `gather()`.

You can read more about `gather()` here:

<http://r4ds.had.co.nz/tidy-data.html#spreading-and-gathering>

In `tidyverse` the functions designed to enable you to restructure data have evolved through a series of different forms. This change is one of the real benefits of using open software like R. In my experience, the newer functions can be useful for *really* untidy data. I expect things will continue to evolve and improve over time.

4.6.3 Merging data from different data-sets using `_join()`

To answer our research question, we next need to combine the **RT** with the **accuracy** data, and then the combined behavioural data with **participant** information and **stimulus** information. This is because, as we have seen, information about behavioural responses, about participant attributes or stimulus word properties, are located in separate files.

Many researchers have completed this kind of operation by hand. This involves copying and pasting bits of data in a spreadsheet. It can take hours or days. I know because I have done it, and I have seen others do it. **Please don't.** There are better ways to spend your time. And you will make mistakes that you will not then be able to identify.

We can combine the datasets, in the way that we need, using the `tidyverse full_join()` function. This gets the job done quickly, and accurately.

First, we join RT and accuracy data together.

```
long <- rt.long %>%
    full_join(acc.long)
```

Then, we join subject and item information to the behavioural data.

```
long.subjects <- long %>%
    full_join(subjects, by = "subjectID")

long.all <- long.subjects %>%
    full_join(words, by = "item_name")
```

Notice, we can let R figure out how to join the pieces of data together. If we were doing this by hand then we would need to check *very carefully* the correspondences between observations in different datasets.

4.6.3.1 Code tip

Here, in a series of steps, we take one dataset and join it (merge it) with the second dataset. Let's look at an example.

```
long <- rt.long %>%
    full_join(acc.long)
```

The code work as follows.

1. `long <- rt.long %>%`

- We create a new dataset we call `long`.
- We do this by taking one original dataset `rt.long` and `%>%` piping it to the operation defined in the second step.

2. `full_join(acc.long)`

- In this second step, we use the function `full_join()` to add observations from a second original dataset `acc.long` to those already from `rt.long`

The addition of observations from one database joining to those from another happens through a matching process.

- R looks at the datasets being merged.
- It identifies if the two datasets have columns in common. Here, the datasets have `subjectID` and `item_name` in common).
- R can use these common columns to identify rows of data. Here, each row of data will be identified by both `subjectID` and `item_name` i.e. as data about the response made by a participant to a word.
- R will then do a series of identity checks, comparing one dataset with the other and, row by row, looking for matching values in the common columns.
- If there is a match then R joins the corresponding rows of data together.
- If there isn't a match then it creates `NAs` where there are missing entries in one row for one dataset which cannot be matched to a row from the joining dataset.

Note that in one example, the example of code I discuss here, I did not specify identifying columns in common, allowing the function to do the work. In the other code chunks I did: `long.all <- long.subjects %>% full_join(words, by = "item_name")` using the `by = ...` argument.

4.6.3.2 Relational data

In the `tidyverse` family of `dplyr` functions, when you work with multiple datasets (tables of data), we call the datasets **relational** data.

<http://r4ds.had.co.nz/relational-data.html#relational-data>

There are three families of verbs designed to work with relational data:

- Mutating joins, which add new variables to one data frame from matching observations in another.
- Filtering joins, which filter observations from one data frame based on whether or not they match an observation in the other table.

- Set operations, which treat observations as if they were set elements.

We can connect datasets – relate them – according to shared variables like `subjectID`, `item_name` (for our data). In `tidyverse`, the variables that connect pairs of tables are called keys where, and this is what counts, *key(-s) are variable(-s) that uniquely identify an observation*.

For the experimental reading data, we have observations about each response made by a participant (one of 61 subjects) to an item (one of 160 words). For these data, we can match up a pair of RT and accuracy observations for each (unique) `subjectID-item_name` combination.

If you reflect, we could not combine the RT and accuracy data correctly:

1. If we did not have both identifying variables for both datasets, both required to uniquely identify each observation.
2. If there were mismatches in values of the identifying variable.

Sometimes, I have done this operation and it has gone wrong because a `subjectID` has been spelled one way in one dataset e.g. `hugh` and another way in the other dataset e.g. `HughH`. This means I am careful about spelling identifiers and I always check my work after merger operations, calculating dataset lengths to ensure the number of rows in the new dataset matches my expectations.

4.6.3.3 _join functions

We used the `full_join()` function.

There are three kinds of joins.

- A left join keeps all observations in x.
- A right join keeps all observations in y.
- A full join keeps all observations in x and y.

I used `full_join()` because I wanted to retain all observations from both datasets, whether there was a match (as assumed) or not, in the identifying variables, between observations in each dataset.

4.6.3.4 Exercise

Break the join You could examine how the `full_join()` works by experimenting with stopping it from working

As I discuss, you need to have matches in values on key (common) variables. If the `subjectID` is different on different datasets, you will lose data that would otherwise be merged to form the merged or composite dataset. So, check what happens if you deliberately mis-spell one of the `subjectID` values in one of the original source wide behavioural data files.

To be safe, you might want to do this exercise with copies of the source files kept in a folder you create for this purpose. If it goes wrong, you can always re-access the source files and read them in again.

You can check what happens before and after you break the match by counting the number of rows in the dataset that results from the merger. We can count the number of rows in a dataset with:

```
length(long.all$RT)
```

```
[1] 9762
```

This bit of code takes the length of the vector (i.e. variable column `RT` in dataset `long.all`), thus counting the number of rows in the dataset.

4.6.4 Select or transform the variables

OK, now we have all the data about everything all in one big, long and wide, dataset. But we do not actually need all this stuff. We next need to do two things. First, we need to get rid of variables we will not use: we do that by using `select()`. Then, we need to remove errors and outlying short RT observations: we do that by using `filter()` in Section @ref(filter).

We are going to select just the variables we need using the `select()` function.

```
long.all.select <- long.all %>%
  select(item_name, subjectID, RT, accuracy,
         Lg.UK.CDcount, brookesIMG, AoA_Kup_lem,
         Ortho_N, regularity, Length, BG_Mean,
         Voice, Nasal, Fricative, Liquid_SV,
         Bilabials, Labiodentals, Alveolars,
         Palatals, Velars, Glottals, age.months,
         TOWREW_skill, TOWRENW_skill, spoonerisms, CART_score)
```

Notice that these variables do not have *reader-friendly* names. Naming things well is important, as Jenny Bryan teaches:

<https://speakerdeck.com/jennybc/how-to-name-files>

I would say that this true for variables as much as for files. The names we have in the CP study data were fine for internal use within my research group but we should be careful to ensure that variables have names that make sense to others and to our future selves. We can adjust variable names using the `rename()` function but I will leave that as an exercise for you to do.

4.6.4.1 Exercise

Select different variables You could analyze the CP study data for a research report. What if you wanted to analyze a different set of variables, could you select different variables?

4.6.5 Filter observations

We now have a tidy dataset `long.all.select` with 26 columns and 9762 rows.

The dataset includes missing values, designated `NA` for *not available* (to you). Here, every error (coded 0, in `accuracy`) corresponds to an `NA` in the `RT` column.

The dataset also includes outlier data. In this context, $RT < 200$ are probably response errors or equipment failures. We will want to analyse `accuracy` later, so we shall need to be careful about getting rid of NAs.

At this point, I am going to exclude two sets of observations only.

- observations corresponding to correct response reaction times that are too short: $RT < 200$.
- plus observations corresponding to the word *false* which (because of stupid Excel auto-formatting) dropped item attribute data.

We can do this using the `filter()` function, setting conditions on rows, as arguments.

```
# step 1
long.all.select.filter <- long.all.select %>%
  filter(item_name != 'FALSE')

# step 2
long.all.select.filter <- long.all.select.filter %>%
  filter(RT >= 200)
```

4.6.5.1 Code tip

Here, I am using the function `filter()` to ...

- Create a new dataset `long.all.select.filter <- ...` by
- Using functions to work on the data named immediately to the right of the assignment arrow: `long.all.select`
 - An observation is included in the new dataset if it matches the condition specified as an argument in the `filter()` function call, thus:
 1. `filter(item_name != `FALSE`)` means: include in the new dataset `long.all.select.filter` all observations from the old dataset `long.all.select` that are not != (! not = equal to) the value `FALSE` in the variable `item_name`
 2. then recreate the `long.all.select.filter` as a version of itself (with no name change) by including in the new version only those observations where RT was greater than or equal to 200ms using `RT >= 200`

4.6.5.2 The difference between = and ==

You need to be careful to distinguish these signs.

- `=` assigns a value, so `x = 2` means "x equals 2"
- `==` tests a match so `x == 2` means: "is x equal to 2?"

4.6.5.3 Using multiple arguments in filtering

You can supply multiple arguments to `filter()` and this may be helpful if (1.) you want to filter observations according to a match on condition-A **and** condition-B (logical “and” is coded with `&`) or (2.) you want to filter observations according to a match on condition-A or condition-B (logical “or” is coded `|`).

You can read more about using multiple arguments to filter observations here:

<https://dplyr.tidyverse.org/reference/filter.html>

4.6.5.4 Exercise

Vary the filter conditions in different ways

1. Change the threshold for including RTs from `RT >= 200` to something else
2. Can you assess what impact the change has? Note that you can count the number of observations (rows) in a dataset using e.g. `length(data.set.name$variable.name)`

Filtering or re-coding observations is an important element of the research workflow in psychological science. How we do or do not remove observations from original data may have an impact on our results (as explored by, e.g., Steegen et al., 2014). It is important, therefore, that we learn how to do this reproducibly using R scripts that we can share with our research reports.

You can read further information about filtering here:

<https://r4ds.had.co.nz/transform.html?q=filter#filter-rows-with-filter>

4.6.5.5 Remove missing values

We will be working with the `long.all.select.filter.csv` dataset collated from the experimental, subject ability scores, and item property data collected for the CP word naming study.

For convenience, I am going to remove missing values before we go any further, using the `na.omit()` function.

```
long.all.noNAs <- na.omit(long.all.select.filter)
```

4.6.5.6 Code tip

The `na.omit()` function is powerful. In using this function, I am asking R to create a new dataset `long.all.noNAs` from the old dataset `long.all.select.filter` in a process in which the new dataset will have *no* rows in which there is a missing value `NA` in *any* column. You need to be reasonably sure, when you use this function, where your `NAs` may be because, otherwise, you may end the process with a new filtered dataset that has many fewer rows in it than you expected.

4.6.6 Now we have some tidy data

```
head(long.all.noNAs, n = 10)

# A tibble: 10 x 26
  item_n~1 subje~2      RT accur~3 Lg.UK~4 brook~5 AoA_K~6 Ortho_N regul~7 Length
  <chr>    <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 act      Aislin~  595.     1   4.03     4   6.42     5   1     3
2 act      AlexB   586      1   4.03     4   6.42     5   1     3
3 act      AmyR    693      1   4.03     4   6.42     5   1     3
4 act      AndyD   597      1   4.03     4   6.42     5   1     3
5 act      AnnaF   627      1   4.03     4   6.42     5   1     3
6 act      AoifeH  649      1   4.03     4   6.42     5   1     3
7 act      ChloeB~ 1081     1   4.03     4   6.42     5   1     3
8 act      ChloeF  642      1   4.03     4   6.42     5   1     3
9 act      ChloeS  623.     1   4.03     4   6.42     5   1     3
10 act     CianR   701      1   4.03     4   6.42     5   1     3
# ... with 16 more variables: BG_Mean <dbl>, Voice <dbl>, Nasal <dbl>,
#   Fricative <dbl>, Liquid_SV <dbl>, Bilabials <dbl>, Labiodentals <dbl>,
#   Alveolars <dbl>, Palatals <dbl>, Velars <dbl>, Glottals <dbl>,
#   age.months <dbl>, TOWREW_skill <dbl>, TOWRENW_skill <dbl>,
#   spoonerisms <dbl>, CART_score <dbl>, and abbreviated variable names
#   1: item_name, 2: subjectID, 3: accuracy, 4: Lg.UK.CDcount, 5: brookesIMG,
#   6: AoA_Kup_lem, 7: regularity
```

If we inspect the `long.all.noNAs` data-set, we can see that we have now got a tidy data-set with all the data we need for our analyses:

- One observation per row, corresponding to data about a response made by a participant to a stimulus in an experimental trial
- One variable per column
- We have information about the speed and accuracy of responses
- And we have information about the children and about the words.

We have removed the missing values and we have filtered outliers.

4.6.7 We can output the data as a .csv file

Having produced the tidy dataset, we may wish to share it, or save ourselves the trouble of going through the process again. We can do this by creating a .csv file.

```
write_csv(long.all.noNAs, "long.all.noNAs.csv")
```

This function will create a `.csv` file from the dataset you name `long.all.noNAs` which R will put in your working directory.

4.6.8 Data tidying – conclusions

Most research work involving quantitative evidence requires a *big* chunk of data tidying or other processing before you get to the statistics. Most of the time, this is work *you* will have to do. The lessons you can learn about the process will generalize to many future research scenarios.

4.7 Repeated measures designs and crossed random effects

Our focus this week is on analyzing data that come from studies with **repeated-measures designs** where the experimenter presents multiple stimuli for response to each participant. In our working example, the **CP reading study**, CP asked all participants in her study to read a selection of words. All participants read the same selection of words, and every person read every word. For each participant, we have multiple observations and these (within-participant) observations will not be independent of each other. One participant will tend to be slower or less accurate compared to another participant, on average. Likewise, one participant's responses will reveal a stronger (or weaker) impact of the effect of an experimental variable than another participant. These between-participant differences will tend to be apparent for each set of observations we have for each participant, across the sample of participants.

You could say that the lowest trial-level observations can be grouped with respect to participants, that observations are nested within participant. But the data can also be grouped by stimuli. Remember that in the CP study, all participants read the same selection of words, and every person read every word. This means that for each stimulus word, there are multiple observations because all participants responded to each word, and these (within-item) observations will not be independent of each other. One word may prove to be more challenging compared to another, eliciting slower or less accurate responses, on average. Likewise, participants' responses to a word will reveal a stronger (or weaker) impact of the effect of an experimental variable than the responses to another word. Again, these between-stimulus differences will tend to be apparent for each set of observations we have for each stimulus word, across the sample of words.

Under these circumstances, are observations about the responses made by different participants nested under words or are observations about the responses to different words nested under participants? We do not have to make a decision.

Given this common **repeated-measures** design, we can analyze the outcome variable in relation to:

fixed effects the impact of independent variables like participant reading skill or word frequency

random effects the impact of random or unexplained differences between participants and also between stimuli

In this situation, we can say that the random effects are crossed (Baayen et al., 2008). When multilevel models require the specification of crossed random effects, they tend to be called **mixed-effects models**.

4.8 Working with mixed-effects models

To illustrate the approach, we examine observations from the CP study. We begin, as we did previously, by ignoring differences due to grouping variables (like participant or stimulus). We pretend that all observations are independent. In this fantasy situation, we address our research question.

RQ.1. What word properties influence responses to words in a test of reading aloud?

4.8.1 Load the data if you need to

If you did not go through the process of tidying the CP study data from the component source data files, then you can import the pre-tidied data here.

```
long.all.noNAs <- read_csv("long.all.noNAs.csv",
  col_types = cols(
    subjectID = col_factor(),
    item_name = col_factor()
  )
)
```

4.8.1.1 Code tip

Notice that I am using `read_csv()` with an additional argument `col_types = cols(...)` through which I control how `read_csv()` processes specific column variables in the data. Here, I am requesting that `read_csv()` treats `subjectID` and `item_name` as factors.

This is a very useful capacity, and a more efficient way to work than, say, first reading in the data and then using *coercion* to ensure that variabels are assigned appropriate types. You can read more about it here.

<https://readr.tidyverse.org/articles/readr.html>

4.8.2 Linear model for multilevel data – ignoring the hierarchical structure

We begin by asking if reading reaction time (RT) varies in association with word frequency. A scatterplot shows that response latencies decrease with increasing word frequency (Figure @ref(fig:pfreqc5)).

```

long.all.noNAs %>%
  ggplot(aes(x = Lg.UK.CDcount, y = RT)) +
  geom_point(alpha = .2) +
  geom_smooth(method = "lm", se = FALSE, size = 1.5, colour="red") +
  theme_bw() +
  xlab("Word frequency: log context distinctiveness (CD) count")

```

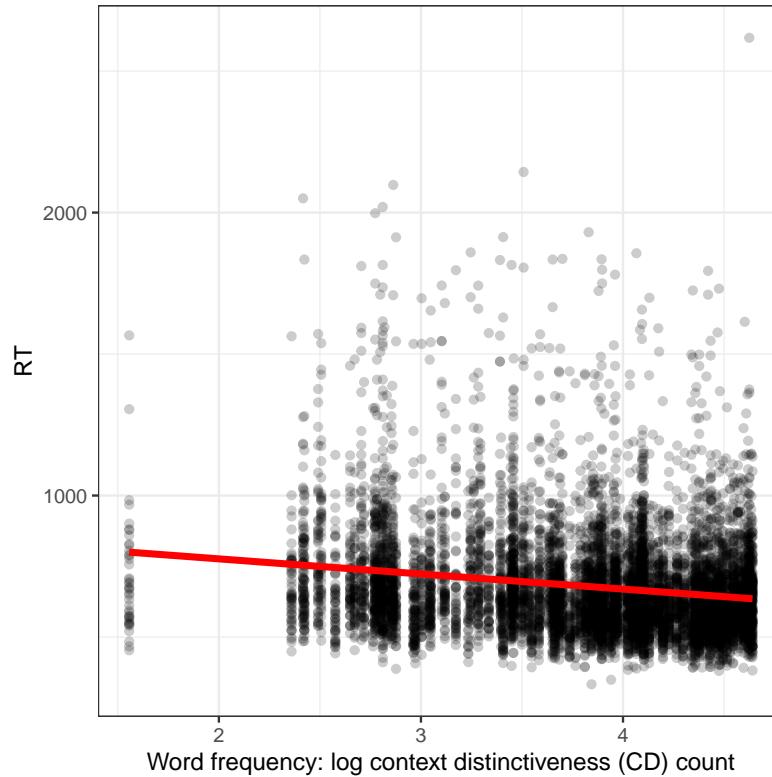


Figure 4.3: Reading reaction time compared to word frequency, all data

In the plot, we see that the best fit line drawn with `geom_smooth()` trends downward for higher values of word frequency. This means that Figure @ref(fig:pfreqc5) suggests that RT decreases with increasing word frequency. (I know there is a weird looking line of points around 0 but we can ignore that here.)

We can estimate the relationship between RT and word frequency using a linear model in which we ignore the possibility that there may be differences (between subjects, or between items) in the intercept or (between subjects) in the slope of the frequency effect:

$$Y_{ij} = \beta_0 + \beta_1 X_j + e_{ij} \quad (4.1)$$

- where Y_{ij} is the value of the observed outcome variable, the RT of the response made by the i participant to the j word;
- $\beta_1 X_j$ refers to the fixed effect of the explanatory variable (here, word frequency), where the frequency value X_j is different for different words j , and β_1 is the estimated coefficient of the effect due to the relationship between response RT and word frequency;
- e_{ij} is the residual error term, representing the differences between observed Y_{ij} and predicted values (given the model).

The linear model can be fit in R using the `lm()` function, as we have done previously.

```
lm.all.1 <- lm(RT ~ Lg.UK.CDcount,
                 data = long.all.noNAs)

summary(lm.all.1)
```

Call:
`lm(formula = RT ~ Lg.UK.CDcount, data = long.all.noNAs)`

Residuals:

Min	1Q	Median	3Q	Max
-346.62	-116.03	-38.37	62.05	1981.58

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	882.983	11.901	74.19	<2e-16 ***
Lg.UK.CDcount	-53.375	3.067	-17.40	<2e-16 ***

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1 '	'	'	1

Residual standard error: 185.9 on 9083 degrees of freedom
Multiple R-squared: 0.03227, Adjusted R-squared: 0.03216
F-statistic: 302.8 on 1 and 9083 DF, p-value: < 2.2e-16

We can see that, in this first analysis, the estimated effect of word frequency is $\beta = -53.375$ (here, word frequency information is in the `Lg.UK.CDcount` variable). This means that, in the linear model, RT decreases by about 54 milliseconds for each unit increase in log word frequency. (In our analyses, in common with many in the reading literature, we transformed the frequency estimate to the Log base 10 of the frequency of occurrence estimated for each word.) The model does not explain much variance, as $R^2 = .03$ but, no doubt due to the large sample, the regression model is overall significant $F(1, 9083) = 302.8, p < .001$.

4.8.2.1 Exercise

Vary the linear model using different outcomes or predictors

The CP study dataset is rich with possibility. It would be useful to experiment with it.

1. Change the predictor from frequency to something else: what do you see when you visualize the relationship between variables using scatterplots?
2. Specify linear models with different predictors: do the relationships you see in plots match the coefficients you see in the model estimates?

4.8.3 Can we ignore the hierarchical structure?

In this linear model, the observations are assumed to be independent but the assumption of independence is questionable given the expectation that participants will differ, with one participant's responses perhaps slower or less accurate than another, perhaps more or less affected by word frequency than another. We can examine that variation by estimating the intercept and the slope of the frequency effect separately using the data for each participant alone.

We can start by examining the frequency effect for each child in a grid of plots, with each plot representing the $RT \sim frequency$ relationship for the data for a child (Figure @ref(fig:freqperchildtrellis)).

We discussed how the plotting code functions in the previous chapter.

```
long.all.noNAs %>%
  ggplot(aes(x = Lg.UK.CDcount, y = RT)) +
  geom_point(alpha = .2) +
  geom_smooth(method = "lm", se = FALSE, size = 1.25, colour = "red") +
  theme_bw() +
  xlab("Word frequency (log10 UK SUBTLEX CD count)") +
  facet_wrap(~ subjectID)
```

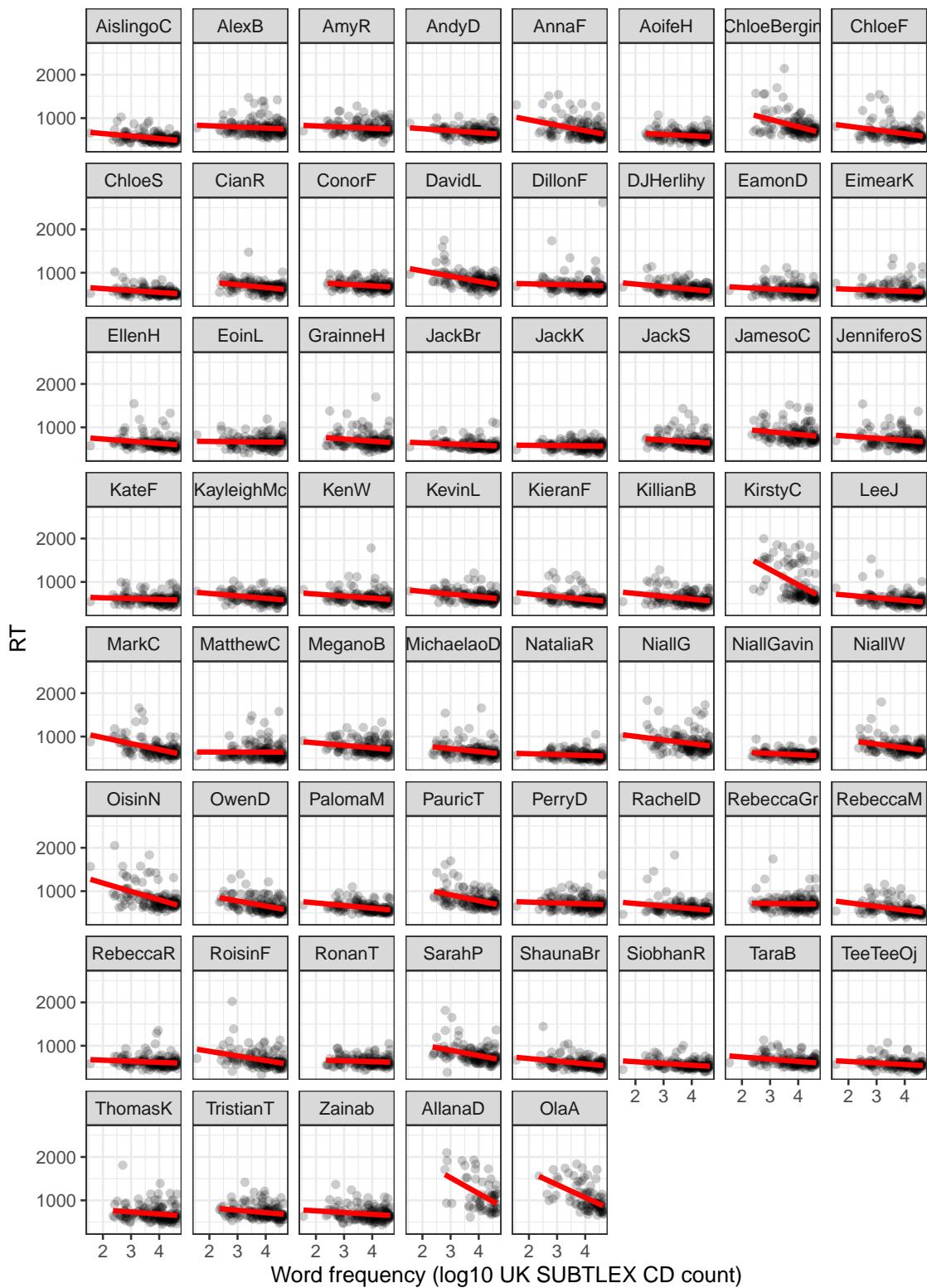


Figure 4.4: RT vs. word frequency, considered separately for data for each child

Figure @ref(fig:freqperchildtrellis) shows how, on average, more frequent words are associated with shorter reaction time, faster responses. The plot further shows, however, that the effect of frequency varies considerably between children. Some children show little or no effect; the best fit line is practically level. Other children show a marked effect, with a steep fit line indicating a strong frequency effect.

We can get more insight into the differences between children, however, if we plot the estimated intercept and frequency effect coefficients for each child directly. This allows more insight because it focuses the eye on the differences between children in the estimates.

Figure @ref(fig:freqperchildlm) presents a plot showing the estimates of the intercept and the coefficient of the effect of word frequency on reading RT, calculated separately for each child. The estimate for each child is shown as a black dot. The standard error of the estimate is shown as a black vertical line, shown above and below a point. You can say that where there is a longer line there we have more uncertainty about the location of the estimate.

Figure @ref(fig:freqperchildlm) presents the estimates of intercept and the frequency coefficient, calculated for each child, ordered by the size of the estimate. Drawn this way, we can see how the estimates of both the intercept and the slope of the frequency effect vary substantially between children. We can see also how the standard errors vary greatly between children.

Notice that if there is an average intercept for everyone in the sample or, better, an intercept we could estimate for everyone in the population, then the different intercepts we have estimated for each child would be distributed around that population-level average. Some children will have slower (here, larger) intercepts and other children will have faster (shorter) intercepts. (Here, the intercept can be taken to be the average RT when all other effects in the model are set to zero. RT varies for this sample around somewhere like $\beta_0 = 883ms$ so a slower larger intercept might be e.g. $\beta_0 = 1000ms$.)

In the same way, if there is an average slope for the frequency effect, an effect of frequency on reading RT, averaged across everyone in the population, then, again, the different slopes we have estimated for each child would be distributed around that population-level effect. Some children will have larger (here, more negative) frequency effects and other children will have smaller (less negative) frequency effects. (Here, the frequency effect is associated with a negative coefficient e.g. $\beta_1 = -53$ so a larger frequency effect will be a bigger negative number e.g. $\beta_1 = -100$.)

4.8.4 Multilevel – here, more appropriately known as – mixed-effects models

In a mixed-effects model, we account for this variation: the differences between participants in intercepts and slopes. We do this by modeling the intercept as two terms:

$$\beta_{0i} = \gamma_0 + U_{0i} \quad (4.2)$$

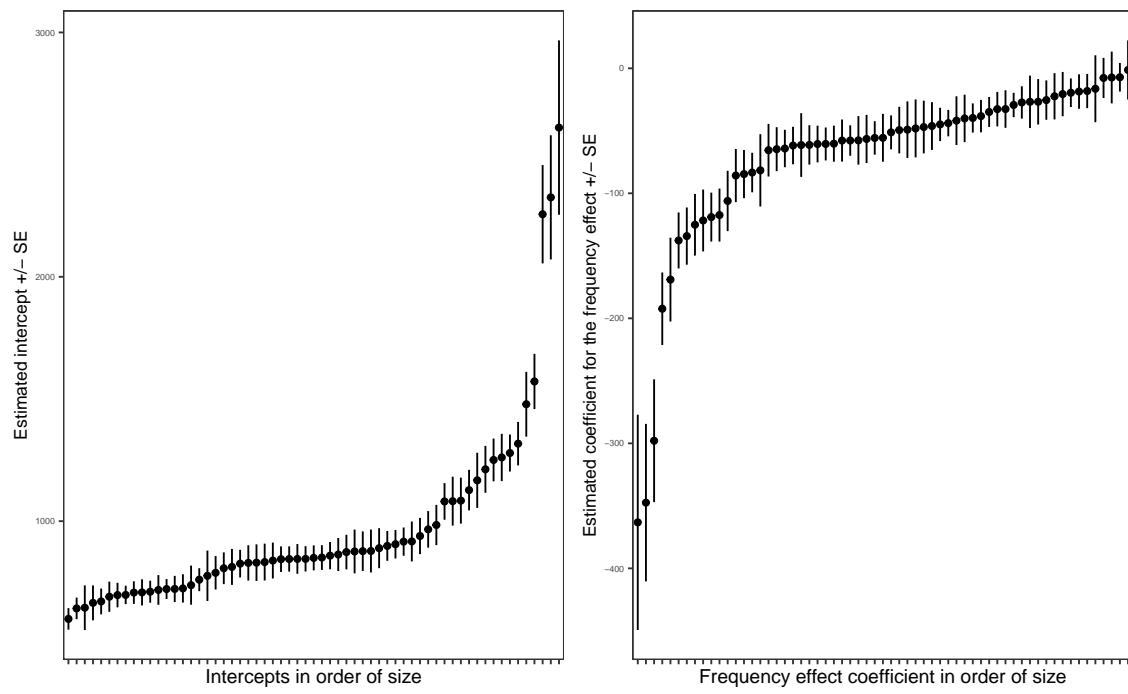


Figure 4.5: Estimated intercepts and frequency effect slopes (with SEs) calculated for each child analysed separately, with point estimates presented in order of size

- where γ_0 is the average intercept and U_{0i} is the difference for each i child between their intercept and the average intercept.

We model the frequency effect as two terms:

$$\beta_{1i} = \gamma_1 + U_{1i} \quad (4.3)$$

- where γ_1 is the average slope and U_{1i} represents the difference for each i child between the slope of their frequency effect and the average slope.

We can then incorporate in a single model the **fixed effects** due to the average intercept and the average frequency effect, as well as the **random effects**, error variance due to unexplained differences between participants in intercepts and frequency effects:

$$Y_{ij} = \gamma_0 + \gamma_1 X_j + U_{0i} + U_{1i} X_j + e_{ij} \quad (4.4)$$

- where the outcome Y_{ij} is related to ...
- the average intercept γ_0 and differences between i children in the intercept U_{0i} ;
- the average effect of the explanatory variable frequency $\gamma_1 X_j$ and differences between i participants in the slope $U_{1i} X_j$;
- in addition to residual error variance e_{ij} .

4.8.4.1 What are we doing with these random effects terms?

Note that in sections @ref(BLUPS) and @ref(variance-covariance), we look at what *exactly* is captured in these random effects terms U_{0i}, U_{1i} . Let's first look at the practicalities of analysis then come back to deepen our understanding a bit more.

Right now, it is important to understand that in our analysis we do not care about the differences between *specific* children. We care that there are differences. And we care how widely spread are the differences between child A and the average intercept (or slope), or between child B and the average intercept (or slope), or between child C ... (you get the idea). Therefore, in our analysis, we estimate the spread of the differences as a *variance term*. We can see this when we look at the results of the mixed-effects model we specify, next.

4.8.4.2 Fitting a mixed-effect model using the lmer() function

We can fit a mixed-effects model of the $RT \sim frequency$ relationship, taking into account the random differences between participants. I first go through the model fitting code bit by bit. (I then go through the output, the results.)

```

lmer.all.1 <- lmer(RT ~ Lg.UK.CDcount + (Lg.UK.CDcount + 1 || subjectID) ,
                     data = long.all.noNAs)

summary(lmer.all.1)

```

You have seen the `lmer()` function code before but *practice makes perfect* so we shall go through the code step by step, as we did previously. This time, notice what is different versus what stays the same.

First, we have a chunk of code mostly similar to what we do when we do a regression analysis.

1. `lmer.all.1 <- lmer(...)` creates a *linear mixed-effects model* object using the `lmer()` function.
2. `RT ~ Lg.UK.CDcount` is a formula expressing the model in which we estimate the fixed effect on the outcome or dependent variable `RT` (reaction time, in milliseconds) predicted \sim by the independent or predictor variable `Lg.UK.CDcount` (word frequency).
3. `...(..., data = long.all.noNAs)` specifies the dataset in which you can find the variables named in the model fitting code.
4. `summary(lmer.all.1)` gets a summary of the fitted model object, showing you the results.

Second, we have the bit that is specific to multilevel or mixed-effects models.

- We add `(... || subjectID)` to tell R about the random effects corresponding to random differences between sample groups (here, observations grouped by child) that are coded by the `subjectID` variable.
- `(...1 || subjectID)` says that we want to estimate random differences between sample groups (observations by child) in intercepts, where the intercept is coded by 1.
- `(Lg.UK.CDcount... || subjectID)` adds random differences between sample groups (observations by child) in slopes of the frequency effect coded using the `Lg.UK.CDcount` variable name.

4.8.4.2.1 What does `||` mean?

I want you to notice something that looks like nothing much: `||`. We are going to need to defer until later a (necessary) discussion of exactly why we need the two double lines. In short, the use of `||` asks R to fit a model in which we estimate random effects associated with

- variance due to differences in intercepts
- variance due to differences in slopes
- but *not* covariance between the two sets of differences

I do this because otherwise the model I specify will not converge. We shall need to discuss these things: **convergence**, and failures to converge; as well as **random effects specification** and simplification. We will discuss random effects covariance in Section @ref(variance-covariance). For now, the most important lesson is learnt by seeing how the analysis approach we saw last week can be extended to examining the effects of experimental variables in data from repeated measures design studies.

4.8.4.3 Reading the `lmer()` results

The `lmer()` model code we discussed in Section @ref(lmer-first) gives us the following output.

```
Linear mixed model fit by REML ['lmerMod']
Formula: RT ~ Lg.UK.CDcount + ((1 | subjectID) + (0 + Lg.UK.CDcount |
    subjectID))
Data: long.all.noNAs

REML criterion at convergence: 117805.3

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-3.7839 -0.5568 -0.1659  0.3040 12.4850

Random effects:
 Groups      Name        Variance Std.Dev. 
subjectID   (Intercept) 87575    295.93  
subjectID.1 Lg.UK.CDcount 2657     51.55  
Residual       23734    154.06  
Number of obs: 9085, groups: subjectID, 61

Fixed effects:
            Estimate Std. Error t value
(Intercept)  950.913   39.216  24.248
Lg.UK.CDcount -67.980    7.092  -9.586

Correlation of Fixed Effects:
            (Intr)
Lg.UK.CDcnt -0.093
```

We discussed the major elements of the results output last week. We expand on that discussion, a little, here.

The output from the model summary first gives us information about the model.

- First, we see information about the function used to fit the model, and the model object created by the `lmer()` function call
- Then, we see the model formula `RT ~ Lg.UK.CDcount + (Lg.UK.CDcount + 1|subjectID)`
- Then, we see `REML criterion at convergence` about the model fitting process, which we can usually ignore
- Then, we see information about the distribution of the model residuals.

We then see information listed under Random effects.

This is where you can see information about the error variance terms estimated by the model.

The information is listed in four columns: 1. Groups; 2. Name; 3. Variance; and 4. Std.Dev. You will recall that we have talked about how observations can be grouped by participant (because we have multiple response observations for each person in the study) just as previously we talked about how observations could be grouped by class (because we saw that children were nested under class). That is what we mean when we refer to **Groups**, we are identifying the grouping variables that give hierarchical structure to the data. The **Name** lists whether the estimate we are looking at corresponds to, here, random differences between participants in intercepts (listed as `(Intercept)`), or in slopes (listed as `Lg.UK.CDcount`). As we discuss later, in Section @ref(variance-covariance), mixed-effects models estimate the spread in random differences. We are not interested in the specific differences in intercept or slope between specific individuals. What we want is to be able to take into account the variance associated with those differences.

Thus, we see in the **Random Effects** section, the variances associated with:

- `subjectID Intercept` 87575, differences between participants in the intercepts;
- `subjectID.1 Lg.UK.CDcount` 2657, differences between participants in the slopes of the frequency effect;
- Alongside `Residual` 23734, residuals where, just like a linear model, we have variance associated with differences between model estimates and observed RT, here, at the trial level.

We do not usually discuss the specific variance estimates in research reports. However, the relative size of the variances does provide useful information (see also Meteyard & Davies, 2020), as we shall see when we discuss the different estimates we get when we include a random effect due to differences between items (Section @ref(random-effect-items)).

Lastly, we see estimates of the coefficients (of the slopes) of the fixed effects.

In this model, we see estimates of the fixed effects of the intercept and the slope of the `RT ~ Lg.UK.CDcount` model. We discuss these estimates next.

4.8.5 Is there a difference between linear model and linear mixed-effects model results?

Recall that the linear model yields the estimate for the frequency effect on reading RT such that RT decreases by about 53 ms for unit increase in log word frequency ($\beta = -53.375$). Now, when we have taken random differences between participants into account, we see that the estimate of the effect **for the mixed-effects model** is $\beta = -67.980$. This is a noteworthy difference in our estimate for the effect. As we saw last, we see, again, that taking into account random differences has an impact on results.

Which coefficient estimate should you trust? Well, it is obvious that the linear model and the linear mixed-effects model estimate are relatively similar. However, it is also obvious that the linear model makes an assumption – the *assumption of independence of observations* – that does not make sense theoretically (we can readily expect that reading responses will be similar within a child) and does not make sense empirically (responses clearly differ between children, Figure @ref(fig:freqperchildlm)). Thus, I think we have good grounds for supposing that the linear mixed-effects model estimate for the frequency effect is likely to be closer to the true underlying population effect (whatever that might be).

That being said, it is important to remember, in this discussion, that whatever estimate we can produce is the estimate we can produce *given* the sample of words we used, the measurement of reading RT we were able to make, and the estimate of word frequency we were able to collect. How far our estimate actually generalizes to the wider population is not something we can settle in the context of a single study.

Further, we have not finished in our consideration of the random effects that the account should include. We need to do more work by thinking about the differences between stimuli (Section @ref(fixed-fallacy)).

4.8.5.1 Why aren't there p-values?

We will come back to this but note that if $t > 2$ we can suppose that an effect is significant at the .05 significance level.

4.8.6 What we estimate when we estimate random effects

We have said that we can incorporate, in a mixed-effects model, **fixed effects** (e.g., the average frequency effect) and **random effects**, error variance due to unexplained differences between participants in intercepts and in frequency effects:

$$Y_{ij} = \gamma_0 + \gamma_1 X_j + U_{0i} + U_{1i} X_j + e_{ij} \quad (4.5)$$

So we distinguish:

- the average intercept γ_0 and differences between i children in the intercept U_{0i} ;
- the average effect of the explanatory variable frequency $\gamma_1 X_j$ and differences between i participants in the slope $U_{1i} X_j$.

When we think about the differences between participants (or between the units of any grouping variable), in intercepts or in slopes, we should understand that for the mixed-effects model, the differences are:

- random;
- should be normally distributed;
- and are distributed around the population or average fixed effects.

We should understand that the mixed-effects model sees the differences between participants **relative to the fixed effect intercept or slope**, that is, relative to the population level or average effects. We can illustrate this by plotting, in Figure @ref(fig:BLUPS-prediction), the differences as estimated (technically, predicted) by the mixed-effects model that we discussed in sections @ref(lmer-first) and @ref(lmer-results).

What you can see in Figure @ref(fig:BLUPS-prediction) are distributions. The centers of the distributions are on zero (shown by a red line). For each distribution (a. and b.), that is where the model estimate of the intercept or the slope of the frequency effect is located. Spread around that central point, you see the adjustments the model makes to account for differences between participants.

You can see how in Figure @ref(fig:BLUPS-prediction) (a.), some children have intercepts that are smaller than the population-level or average intercept – so their adjustments are negative (to decrease their intercepts). In comparison, some children have intercepts that are larger than the population-level or average intercept – so their adjustments are positive (to increase their intercepts). Strikingly, you can see that a few children have intercepts that are as much as 1000ms larger than the population-level or average intercept.

You can see also how in Figure @ref(fig:BLUPS-prediction) (b.), some children have frequency effects (coefficients) that are smaller than the population-level or average frequency effect – so their adjustments are positive (to decrease their frequency effect, by making it *less* negative). (Remember the estimated β coefficient for the frequency effect is negative because higher word frequency is associated with smaller RT.) In comparison, some children have frequency effects that are larger than the population-level or average frequency effect – so their adjustments are negative (to increase their frequency effect, by making it *more* negative). Strikingly, you can see that a few children have frequency effects that are as much as 200ms larger (see plot (b.) around $x = -200$) than the population-level or average effect.

When a mixed-effects model is fitted to a dataset, its set of estimated parameters includes the coefficients for the fixed effects as well as the standard deviations for the random effects (Baayen, 2008). The individual values of the adjustments made to intercepts and slopes are

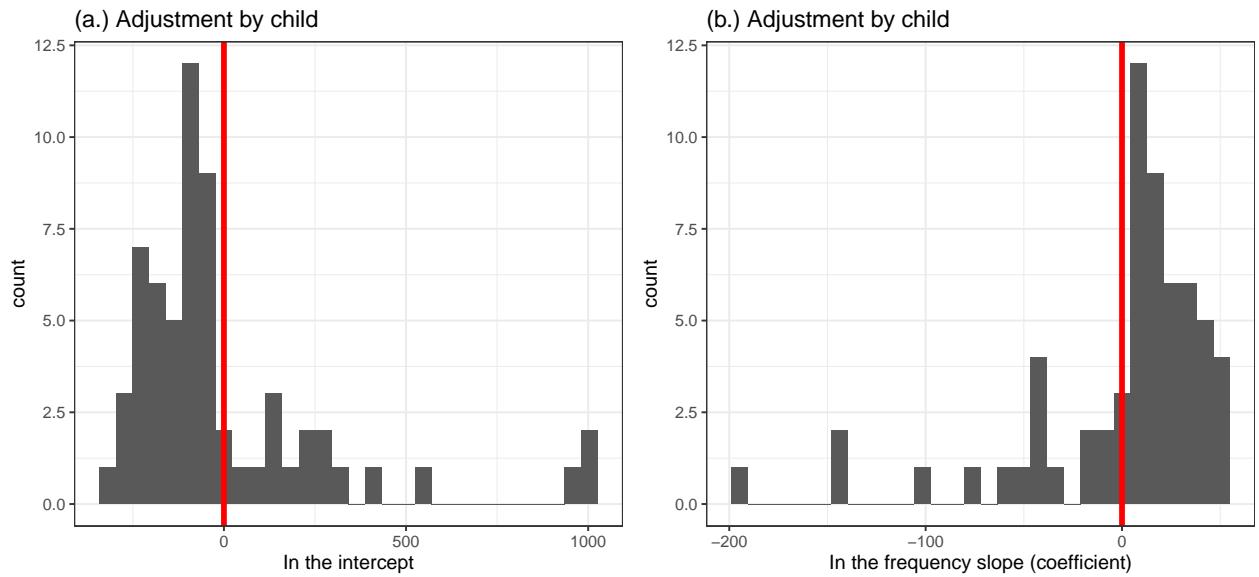


Figure 4.6: Plot showing histograms indicating the distribution of participant adjustments to account for between-child differences in intercept or slope (known as Best Linear Unbiased Predictions)

calculated once the random effects have been estimated. If you read the literature on mixed-effects models, you will see that the adjustments are called Best Linear Unbiased Predictors (or BLUPs).

4.8.6.1 Exercise

Mixed-effects modeling is hard to get used to *at first*. A bit more practice helps to show you how the different parts of the model work. We again focus on the random effects.

- In the model we have seen so far, we specify `(Lg.UK.CDcount + 1 || subjectID)`
- We can change this part – and only this part – to see what happens to the results, do this:
 1. `lmer(RT ~ Lg.UK.CDcount + (1|subjectID) ...)` gives us a *random intercepts* model accounting for just random differences between participants in the intercept
 2. `lmer(RT ~ Lg.UK.CDcount + (Lg.UK.CDcount + 0|subjectID) ...)` gives us a *random slope* model accounting for just random differences between participants in the slope of the frequency effect

3. `lmer(RT ~ Lg.UK.CDcount + (Lg.UK.CDcount + 1|subjectID) ...)` gives us a *random intercepts and slopes* model accounting for both random differences between participants in the intercept and in the slope, as well as covariance in these differences.

Try out these variations and *look carefully* at the different results. Look, especially, at what happens to the **Random effects** part of the summary.

We can *visualize* the differences between the models in a plot showing the different predictions that the different models give us. Figure @ref(fig:indiv-prediction) shows what a mixed-effects model would predict should be the effect of frequency on RT for different children in the CP study. The predictions vary depending on the nature of the random effects we specify in the model.

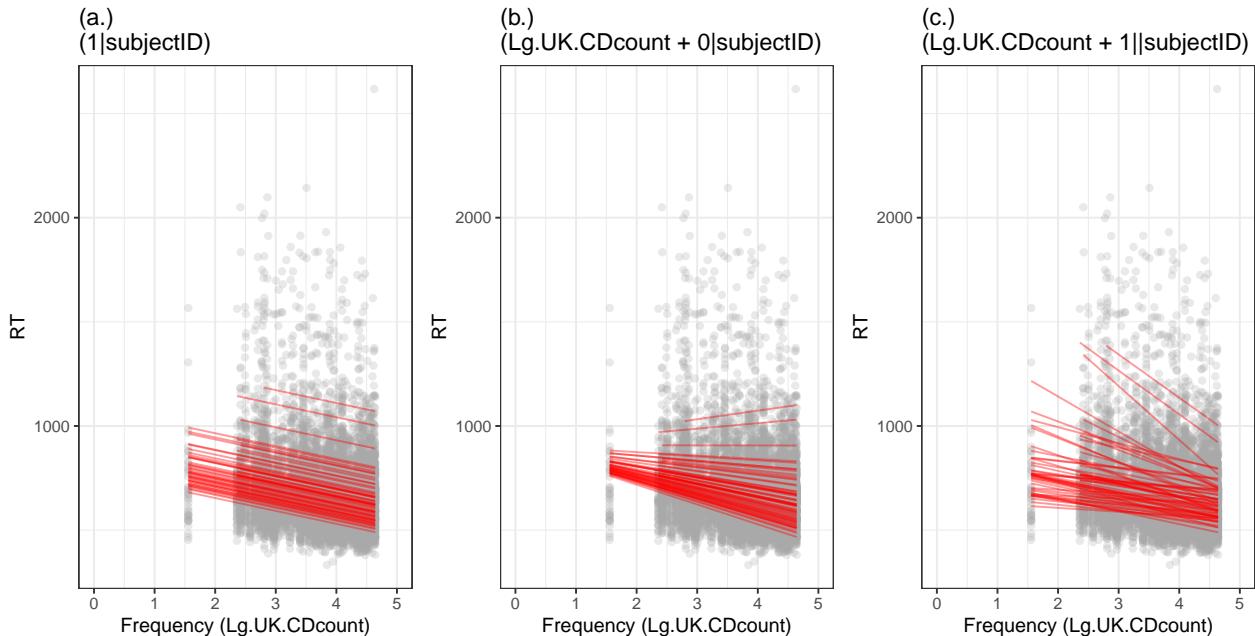


Figure 4.7: Plot showing model predictions of the effect, for each individual, of word frequency on reading reaction time – predictions vary between models incorporating (a.) random effect of participants on intercepts only; (b.) random effect of participants on slopes only and (c.) random effect of participants on intercepts and on slopes

We can see that:

1. If the model includes the random effect of *participants on intercepts only* then all the slopes are the same (the lines in the figure are parallel) because this model assumes that the only differences between participants are differences in the intercepts.

2. If the model includes the random effect of *participants on slopes only* then the slopes vary but they all have the same intercept. The plot does not show this but you can see how all the slopes are converging on one point somewhere on the left. This happens because this model assumes that the only differences between participants are differences in the slopes.
3. If the model includes the random effect of *participants on intercepts and on slopes* then we can see how the intercepts and the slopes vary. Given what we saw when we looked at the relation between frequency and RT for each participant considered separately we might argue that this model is much more realistic about the data.

4.8.6.2 Code tip

It is very important to learn to make effective use of the warnings and error messages R can produce.

Note: you do not have to just believe me when I say that `||` is in the model code to stop a problem appearing. Experiment – and see what happens when you change the code. Try this.

```
lmer.all.1 <- lmer(RT ~ Lg.UK.CDcount + (Lg.UK.CDcount + 1|subjectID),  
                    data = long.all.noNAs)  
summary(lmer.all.1)
```

Do you get an error message?

A very useful trick is to learn to copy the error message you get into a search engine on your web browser. Do this and you will find useful help, as here.

https://rstudio-pubs-static.s3.amazonaws.com/33653_57fc7b8e5d484c909b615d8633c01d51.html

4.9 Variation between stimuli: the “language as fixed-effect fallacy”

Experimental psychologists will often collect data in studies where they present some stimuli to a sample of participants (as CP did in her study). Clark (1973) showed that the appropriate analysis of experimental effects for such data requires the researcher to take into account the error variance due to unexplained or random differences between sampled participants *and* to random differences between sampled stimuli. This is true in the context of psycholinguistics but it is also true in the context of work in any field where the presented stimuli can be understood

to constitute a sample from a wider population of potential stimuli (Judd, Westfall, & Kenny, 2012).

If we were to estimate the average latency of the responses made by different children to each word, we would see that there is considerable variation between words (Figure @ref(fig:pitemsints)). Some words elicit slower and some elicit faster responses on average. We can also see that there is, again, variation in the uncertainty of estimates, as reflected in differences in the lengths of the error bars corresponding to the standard errors of the estimates.

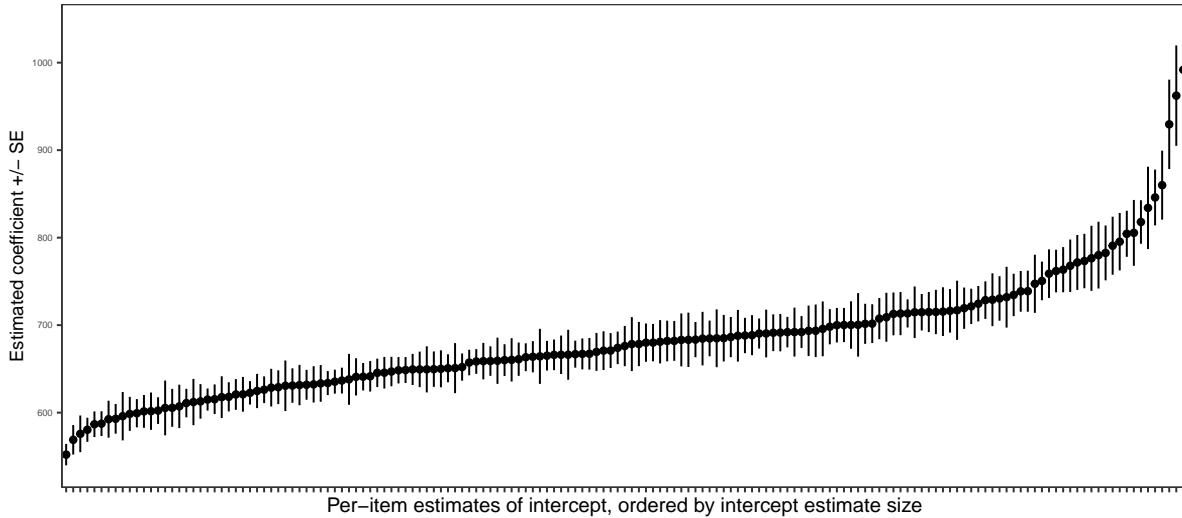


Figure 4.8: Estimated intercepts (with SEs) calculated for each stimulus word, with coefficients ordered by average latency for each word

In general, psychologists have been aware since Clark (1973, if not earlier) that responses to experimental stimuli can vary because of random or unexplained differences between the stimuli: whether the stimuli are words, pictures or stories, etc. And researchers have been aware that if we did not take such variation into account, we might mistakenly detect an experimental effect, for example, as a significant difference between mean response in different conditions, simply because different stimuli presented in different conditions varied in some unknown way, randomly, in relative difficulty.

For many years, psychologists tried to take random differences between stimuli into account, alongside random differences between participants, using a variety of strategies with important limitations (see Baayen et al., 2008, for discussion). Clark (1973) suggested that researchers could calculate $\min F'$ (not F) when doing Analyses of Variance of experimental data

This involves a series of steps.

1. You start by *aggregating* your data

- By-subjects data – for each subject, take the average of their responses to all the items
 - By-items data – for each item, take the average of all subjects' responses to that item
2. You do separate ANOVAs, one for by-subjects (F1) data and one for by-items (F2) data
 3. You put F1 and F2 together, calculating $\min F'$

Averaging data by-subjects or by-items is relatively simple. And you will often see, in the literature, psychological reports in which F1 and F2 analysis results are presented.

Calculating $\min F'$ is also relatively simple:

$$\min F' = \frac{MS_{effect}}{MS_{random-subject-effects} + MS_{random-word-differences}} = \frac{F_1 F_2}{F_1 + F_2} \quad (4.6)$$

However, after a while, psychologists stopped doing the extra step of the $\min F'$ calculation (Raaijmakers et al., 1999). They carried on calculating and reporting F1 and F2 ANOVA results but, as Baayen et al. (2008) discuss, this approach risks a high potential false positive error rate.

Psychologists also found that while the $\min F'$ approach allowed them to take into account between-participant and between-stimulus differences it could not be applied where ANOVA could not be used. This stopped researchers from taking a comprehensive approach to error variance where they wanted to conduct multiple regression analyses. You will often see multiple regression analyses of by-items data, where a sample of participants has been asked to respond to a sample of stimuli, and the analysis is of the effects of stimulus properties on outcomes averaged (over participants' responses) to the mean outcome by item. But analyzing data only by-items ensures that we lose track of participant differences. Lorch and Myers (1990) warn that analyzing only by-items mean RTs just assumes wrongly that *subjects are a fixed effect*. This approach, again, risks a higher rate of false positive errors.

4.9.1 Include the random effect of stimulus

We now no longer need to tolerate these problems.

In the context of our working example, with our analysis of the CP study data, we can build up our mixed-effects model by adding a random effect to capture the impact of unexplained differences between stimuli. We model the random effect of items on intercepts by modeling the intercept as two terms:

$$\beta_{0j} = \gamma_0 + W_{0j} \quad (4.7)$$

- where γ_0 is the average intercept and W_{0j} represents the deviation, for each word, between the average intercept and the per-word intercept.

Our model can now incorporate the additional random effect of items on intercepts:

$$Y_{ij} = \gamma_0 + \gamma_1 X_j + U_{0i} + U_{1i} X_j + W_{0j} + e_{ij} \quad (4.8)$$

In this model, the outcome Y_{ij} is related to the average intercept γ_0 and the word frequency effect $\gamma_1 X_j$ plus random effects due to unexplained differences between participants in intercepts U_{0i} and the slope of the frequency effect $U_{1i} X_j$ as well as random differences between items in intercepts W_{0j} , in addition to the residual term e_{ij} .

4.9.1.1 Fitting a mixed-effect model – now with random effects of subjects and items

We can fit a mixed-effects model of the $RT \sim frequency$ relationship, taking into account the random differences between participants *and now also* the random differences between stimulus words.

```

lmer.all.2 <- lmer(RT ~ Lg.UK.CDcount +
                     (Lg.UK.CDcount + 1 || subjectID) +
                     (1 | item_name),

data = long.all.noNAs)

summary(lmer.all.2)

Linear mixed model fit by REML ['lmerMod']
Formula: RT ~ Lg.UK.CDcount + ((1 | subjectID) + (0 + Lg.UK.CDcount |
  subjectID)) + (1 | item_name)
Data: long.all.noNAs

REML criterion at convergence: 116976.7

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-4.1795 -0.5474 -0.1646  0.3058 12.9485 

Random effects:
Groups      Name           Variance Std.Dev.
item_name   (Intercept)    3397     58.29
subjectID   Lg.UK.CDcount  3624     60.20

```

```

subjectID.1 (Intercept) 112314 335.13
Residual                 20704 143.89
Number of obs: 9085, groups: item_name, 159; subjectID, 61

```

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	971.07	51.87	18.723
Lg.UK.CDcount	-72.33	10.79	-6.703

Correlation of Fixed Effects:

	(Intr)
Lg.UK.CDcnt	-0.388

This is the same mixed-effects model as the one we discussed in sections @ref(lmer-first) and @ref(lmer-results) but with one important addition.

- We add `(1|item_name)` to take into account random differences between words in intercepts

4.9.1.1.1 Reading the results

Take a look at the model results. You should notice three changes.

1. You can see that the estimate for the effect of word frequency on reading reaction time has changed again, it is now $\beta = -72.33$
2. `item_name (Intercept) 3397` there is now an additional term in the list of random effects, giving the model estimate for variance associated with random differences between words in intercepts
3. And you can see that the residual variance has changed. In the first model `lmer.all.1` it was 23734, now it is 20704

The reduction in residual variance is one way in which we can judge how good a job the model is doing in accounting for the variance in the outcome, observed response reaction time. We can see that by adding a term to account for differences between items we can reduce the amount by which the model estimates deviate from observed outcomes. This difference in error variance is, essentially, one basis for estimating how well the model fits the data, and a basis for estimating the *variance explained* by a model in terms of the R^2 statistic you have seen before. We will come back to this.

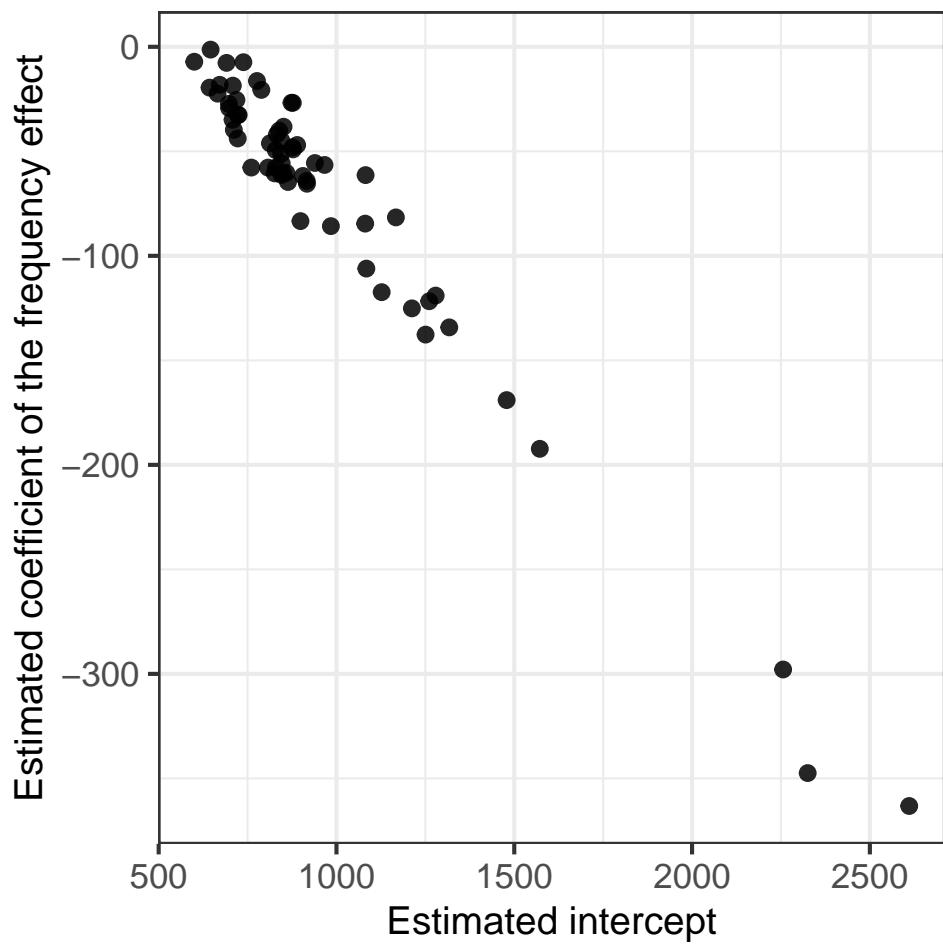


Figure 4.9: Scatterplot showing the relationship between estimated coefficients for the intercept and for the frequency effect, for each child analysed separately

4.10 Variances and covariances of random effects

As I have said, we usually do not aim to examine the specific deviation from the average intercept or the average fixed effect slope for a participant or stimulus. We estimate just the spread of deviations by-participants or by-items. A mixed-effects model like our final model includes fixed effects corresponding to the intercept and the slope of the word frequency effect plus the variances:

- $\text{var}(U_{0i})$ variance of deviations by-participants from the average intercept;
- $\text{var}(U_{1i}X_j)$ variance of deviations by-participants from the average slope of the frequency effect;
- $\text{var}(W_{0j})$ variance of deviations by-items from the average intercept;
- $\text{var}(e_{ij})$ residuals, at the response level, after taking into account all other terms.

We may expect the random effects of participants or items to covary, e.g., participants who are slow to respond may also be more susceptible to the frequency effect, as can be seen in Figure @ref(fig:covarp). Thus, we could specify the random effects of the model can incorporate terms corresponding to the covariance of random effects:

- $\text{covar}(U_{0i}, U_{1i}X_j)$

4.10.1 But remember we excluded random effects covariance

In Section @ref(double-bar), I noted how we used the `||` notation to stop the model estimating the covariance between differences between participants in intercepts and in slopes. The reason I did this is that if I had requested that the model estimate the covariance the model would have failed to converge. What this means depends on understanding how mixed-effects models are estimated. We shall have to return to a development of that understanding later. For now, it is enough to note that mixed-effects models fitted with `lmer()` often have more difficulty with random effects covariance estimates.

4.11 Reporting the results of a mixed-effects model

There is no official convention on what or how to report the results of a mixed-effects model. Lotte Meteyard and I suggest what psychologists should report in an article (Meteyard & Davies, 2020) that has been downloaded a few thousand times so, maybe, our advice will help to influence practice.

We would argue that researchers should explain what analysis they have done and, where space allows, should report both the estimates of the **fixed effects** and the estimates of the **random effects**. We think you can report the model code (maybe in an appendix, maybe in a note under a tabled summary of results).

Coefficients	Estimate	SE	t
(Intercept)	971.1	51.9	18.7
Frequency effect	-72.3	10.8	-6.7

Groups	Name	Variance	SD
item	(Intercept)	3397	58.3
participant	(Intercept)	112314	335.1
participant	Frequency	3624	60.2
residual		20704	143.9

Note: `lmer(RT ~ Lg.UK.CDcount + (Lg.UK.CDcount + 1||subjectID) + (1|item_name)`

Researchers should report their modelling in sufficient detail that their results can be reproduced by others. Barr et al. (2013) argued that choices about random effects structure affect the generalizability of the estimates of fixed effects. In particular, it seems sensible to examine the possibility that the slope of the effect of an explanatory variable may vary at random between participants or between stimuli. Correspondingly, researchers should report and explain their decisions about the inclusion of random effects.

It is normal practice in psychology to report the p-values associated with null hypothesis significance tests of effects when reporting analysis. Performing hypothesis tests using t- or F-distributions depends on the calculation of degrees of freedom yet it is uncertain how degrees of freedom should be counted when analysing multilevel data (Baayen et al., 2008). In most software applications, however, p-values associated with fixed effects may be calculated using an approximation for denominator degrees of freedom.

We will come back to how we should report the results of mixed-effects models because, here, too, we can benefit by developing our approach, in depth, step by step.

4.12 Conclusions

A large proportion of psychological studies involves scenarios in which the researcher samples both participants and some kind of stimuli. Often, the researcher will present the stimuli to the participants for response in some version of a range of possible designs: all participants see and respond to all stimuli; participants respond to different sub-sets of stimuli in different conditions (or in different groups) but they see and respond to all stimuli in a sub-set; participants are allocated to respond to stimulus sub-sets according to a counter balancing scheme (e.g., through the use of Latin squares). Whatever version of this scenario, *if* participants are responding to multiple stimuli and *if* multiple participants respond to each stimulus, then the data will have a multilevel structure such that each observation can be grouped both by

participant and by stimulus. We are interested in taking into account the random effects associated with unexplained or random differences between participants or between stimuli. We often discuss the accounting of these effects in terms of the estimation of error variances associated with the random differences, calling the effects of the differences *random effects*. Where we have to deal with both samples of participants and samples of stimuli, we can talk about *crossed random effects*.

The terms are not that important. The insight is: in general, in experimental psychological science, when we do data analysis, if we want to estimate effects of experimental variables more accurately then our models need to incorporate terms to capture the impact on observed outcomes of sampled participants and sampled stimuli. Historically, we have, as a field, learned to take into account these sampling effects. Now, and most likely, more and more commonly in the future, we are learning to use multilevel or mixed-effects models to do this.

4.12.1 Summary

We discussed the way that data are structured when they come from studies with repeated measures designs. Critically, we examined data from a common study design where a sample of stimulus items are presented for response to members of a participant sample. This means that each observation can be grouped by participant and, also, by stimulus. The possibility that observations can be grouped means that the data have a multilevel structure. The multilevel structure requires the use of linear mixed-effects models when we seek to estimate the effects of experimental variables. The fact that data can be grouped both by participant and by stimulus means that the model can incorporate random effects to capture random between-participant differences as well as between-stimulus differences. The use of mixed-effects models has meant that psychologists no longer need to adopt compromise solutions which have important limitations, like by-items and by-subjects analyses.

We reviewed the ways that experimental data can be untidy. And we outlined the steps that may be required to process untidy data into a tidy format suitable for analysis. As is typical for the data analysis we need to do for experimental psychological science, getting data ready for analysis requires a series of steps including: access; import; restructure; select variables; and filter observations.

We then developed a mixed-effects model to answer the research question:

RQ.1. What word properties influence responses to words in a test of reading aloud?

Our analysis focused on the relationship between reading response reaction time (RT, in ms) and the predictor word frequency. We examined how the effect of word frequency was estimated in a linear model ignoring the multilevel structure and then in mixed-effects models which incorporated terms to capture variance associated with random differences between participants in intercepts or in the slope of the frequency effect, and between items in intercepts.

We saw that estimates of the frequency effect differed between different models.
We considered the possibility of within-items effects.

4.12.2 Useful functions

We used a number of functions to tidy, visualize and analyze the CP study data.

- `read_csv()` and `read_csv()` to load source data files into the R workspace
- `pivot_longer()` to restructure data from wide to long
- `full_join()` to put together data from separate datasets; in our example, from datasets holding information about participant attributes, stimulus word properties, and participant behaviours
- `select()` to select the variables we need
- `filter()` to filter observations based on conditions
- `na.omit()` to remove missing values
- For visualisation, we used `facet_wrap()` to show plots of the relationship between outcome and predictor variables separately for different groups (by participant, or by item)
- We used `lmer()` to fit a multilevel model

We used the `summary()` function to get model results for both linear models and for the multilevel or liner mixed-effects model.

4.13 R code and data file access for the class

Activities in the class that goes with this chapter are associated with the following data file and .R code file:

- 402-02-mixed-effects-workbook.R
- CP study word naming rt 180211.dat
- CP study word naming acc 180211.dat
- words.items.5 120714 150916.csv
- all.subjects 110614-050316-290518.csv

A pre-tidied version of the CP study data is available as:

- long.all.noNAs.csv

You can get these materials by going to the 402 Moodle folder for week 18, and downloading the .zip (compressed) folder labeled **PSYC402-01-multilevel-resources**

Or, you can download the same folder by clicking on the link:

<https://modules.lancaster.ac.uk/mod/resource/view.php?id=1795341>

Run the code in the .R file to reproduce the results presented in this chapter and in the slides.

4.14 References

4.14.1 Recommended reading

Snijders and Bosker (2012) present a helpful overview of multilevel modelling. Baayen et al. (2008; see, also, Barr et al., 2013; Judd et al., 2012) discuss mixed-effects models with crossed random effects. Readers familiar with the book will see that I rely on it to construct the formal presentation of the models.

I wrote a tutorial article on mixed-effects models with Lotte Meteyard. We discuss how important the approach now is for psychological science, what researchers worry about when they use it, and what they should do and report when they use the method/

Meteyard, L., & Davies, R.A.I. (2020). Best practice guidance for linear mixed-effects models in psychological science, *Journal of Memory and Language*, 112, 104092, <https://doi.org/10.1016/j.jml.2020.104092>

4.14.2 References list

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, 390-412.

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

Davies, R. A., Arnell, R., Birchenough, J. M., Grimmond, D., & Houlson, S. (2017). Reading through the life span: Individual differences in psycholinguistic effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43, 1298.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103, 54.

- Lorch, R. F., Jr., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 149–157. <http://dx.doi.org/10.1037/0278-7393.16.1.149>
- 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, 416-426.
- Snijders, T.A., & Bosker, R.J. (2012). *Multilevel analysis (2nd Edition)*. London, UK: Sage.
- Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11, 702-712.

5 Developing linear mixed-effects models

5.1 Motivations: to grow in sophistication

Linear mixed-effects models are important, interesting, and sometimes challenging. We have worked through two chapters in which we have looked at three things: why multilevel or mixed-effects models are needed; what they can do; and what they involve.

We have learnt:

To recognize – The situations in research where we shall see multilevel structured data and therefore where we will need to apply multilevel or mixed-effects models.

To understand – The nature and the advantages of these models: what they are, and why they work better than other kinds of models.

To practise – How we code for mixed-effects models, and how we read or write about the results.

We now need to develop our understanding and skills further.

We have specified, run and looked at the results of mixed-effects models. We now need to examine some of the complexities that we may need to face when we are involved in mixed-effects modeling.

Our approach will continue to depend on verbal explanation, visualization and a practical code-based approach to the modeling.

5.2 The key idea to get us started

Shrinkage or regularization means that models of data should be excited by the data but not *too* excited.

This means our models work better *if* they are informed by all the data, and take into account random differences but also *if* they are not too strongly influenced by individual (participant or item) data.

5.3 Targets

We are probably now at a stage, in the development of our skills and understanding, where we can be more specific about our targets for learning: what capacities or abilities we want to have by the time we complete the course. I have held back specifying the targets in this way because, first, we had to learn the basic vocabulary. Now that we have done that, we can lay out the targets against which we can assess the progression of our learning.

We have three components of the capacity we seek to develop. These components include the capacity to understand mixed-effects models, the capacity to work with them practically in R, and the capacity to present the results.

The truth is that development of skills and understanding in relation to each component will travel at different speeds, for different people, and within any person for different components. For example, I learnt to code and report mixed-effects models as soon as I learnt to recognize the situations where they were required. But it took me longer to understand what the models are, and what they involve. Other people will follow different developmental trajectories.

I think it is also true that our internal evaluation of our understanding will not exactly match the evaluation that comes from external assessment. In other words, we might not be satisfied with our understanding but, still, our understanding might be satisfactory. It might be that we can learn to say in words what mixed-effects models are or involve, or what their results mean, *very effectively* even if we remain unsure about our understanding. For these reasons, I specify what we are aiming to develop in terms of what we can *do*.

We want to develop the capacity to **understand** mixed-effects models, the capacity to:

1. recognize where data have a multilevel structure;
2. recognize where multilevel or mixed-effects models are required;
3. distinguish the elements of a mixed-effects model, including fixed effects and random effects;
4. be able to explain how random effects can be understood in terms of random differences (or deviations), between groups or classes or individuals, in intercepts or slopes;
5. be able to explain how random effects can be understood in terms of variances, as a means to account for random differences between groups or classes or individuals in intercepts or slopes;
6. be able to explain how mixed-effects models work better than linear models, for multi-level structured data, because they take into account variances associated with random differences;
7. and be able to explain how mixed-effects models work better because they allow partial-pooling of estimates, using both information from the whole data set and information from group or class or individual specific data.

We want to develop the capacity to work practically in R with mixed-effects models, the capacity to:

1. be able to specify a mixed-effects model in `lmer()` code;
2. be able to identify how the mixed-effects model code varies, depending on the kinds of random effects that are assumed;
3. be able to identify the elements of the output or results that come from an `lmer()` mixed-effects analysis;
4. be able to interpret the fixed-effects estimates;
5. and be able to interpret the random effects estimates, both variance and covariance estimates.

We want to develop the capacity to talk about and present the results of mixed-effects models, the capacity to:

1. be able to describe in words and summary tables the results of a mixed-effects model;
2. be able to visualize the effects estimates from a mixed-effects model.

5.4 Study guide

1. Read in the example ML study data.
2. Edit example code to create alternate visualizations of variable distributions and of the relationships between critical variables.
3. Experiment with the .R code used to work with the example data.
4. Run linear mixed-effects models of demonstration data.
5. Run linear mixed-effects models of alternate data sets.

You will see that in the references list at the end, I have recommended some papers that I think provide particularly useful or readable introductions to Linear Mixed-effects Models.

5.5 The data we will work with: ML word recognition study

This week, we will be working with the **ML word recognition study** dataset. The focus of our interest is on the ways in which participant attributes (like age) or word properties (like frequency) influence the speed of response in a task measuring the ability to recognize visually presented English words.

ML examined visual word recognition in younger and older adults using the lexical decision task. Lexical decision is a very popular technique for examining word recognition, especially in adults. While not every MSc Psychology student will be interested in word recognition, or reading, everyone should understand that tasks like lexical decision are similar to a range of other tasks used in experimental psychological science. The critical features of the study are that we have an outcome – a decision response – observed multiple times (for each stimulus) for each participant. We shall be analyzing the speed of response, reaction time (RT), measured in milliseconds (ms).

Notice that where we analyze the effects of participant attributes on recognition response RTs, those attributes are recorded using a mix of survey questions (about age, etc.) and standardized ability tests. Taking individuals' scores on standardized ability tests in order to analyze the performance of the same individuals in some experimental task is a *very* important feature of psychological research in many fields.

In the lexical decision task, participants completed a series of 320 trials. In each trial, they were presented with a stimulus, a string of letters, that was either a real word (e.g., 'car') or a made-up or non-word (e.g., 'cas'). There were 160 word and 160 non-word stimuli. Each stimulus was presented one at a time on a computer screen. Participants were required to respond to the stimulus by pressing a button to indicate either that they thought the stimulus was a word (they knew) or that they thought it was a non-word. Each sequence of events, in which a stimulus was presented and a response was recorded, is known as a *trial*. The critical outcome measure was the reaction time of each response: the interval of time from the moment the stimulus was first presented (the stimulus onset) to the moment the response was made (the response onset).

The total number of participants for this study was 39, including a group of younger adults and a group of older adults. Information was collected about the participants' age, education and gender. In addition, participants were asked to complete ability measures (TOWRE sight word and phonemic tests, Torgesen et al., 1999) and a measure of reading experience (Author Recognition Test, ART, Masterson & Hayes, 2007).

In summary, ML collected data on: lexical decision task response reaction times (RTs) and accuracy; information on lexical decision stimulus items, including variables like the length or frequency of words (values taken from the English Lexicon Project, Balota et al., 2007); and information on participants, including age, reading ability and reading experience.

The ML study data includes the following variables that we will work with (as well as some you can ignore):

Identifying variables

- subjectID – identifying code for participants
- item_name – words presented as stimuli
- item_number – identifying code for words presented

Response variables

- RT – response reaction time (ms), for responses to words

Subject variables

- Age – in years
- Gender – coded M (male), F (female)
- TOWRE_wordacc – word reading skill, words read correctly (out of 104)
- TOWRE_nonwordacc – nonword reading skill, nonwords (made up words) read correctly (out of 63)
- ART_HRminusFR – reading experience score

Item variables

- Length – word length, in letters
- Ortho_N – orthographic neighbourhood size, how many other words in English a stimulus word looks like
- OLD – orthographic Levenshtein distance, how many letter edits (addition, deletion or substitution) it would take to make a stimulus word look like another English word (a measure of orthographic neighbourhood) (Yarkoni et al., 2008)
- BG_Sum, BG_Mean, BG_Freq_By_Pos – measures of how common are pairs of letters that compose stimulus words
- SUBTLWF, LgSUBTLWF, SUBLCD, LgSUBLCD – measures of how common stimulus words are, taken from the SUBTLEX corpus analysis of word frequency (Brysbaert and New, 2009)

5.5.1 Research hypotheses

Instead of posing a simple and general research question, we shall orient our work around a set of quite specific predictions. ML hypothesized:

Effects of stimulus attributes words that are shorter, that look like more other words, and that appear frequently in the language will be easier to recognize;

Effects of participant attributes older readers would be faster and more accurate than younger readers in word recognition;

Interactions between the effects of word attributes and person attributes better (older) readers will show smaller effects of word attributes.

5.5.2 Locate and download the data file

The data file can be downloaded as part of the .zip folder labelled **PSYC402-03-mixed-resources**.

You can download the folder from the Moodle section corresponding to this chapter:

<https://modules.lancaster.ac.uk/course/view.php?id=34085#section-13>

Or you can download the folder directly from:

<https://modules.lancaster.ac.uk/mod/resource/view.php?id=1801384>

The data are held in one file:

- `subjects.behaviour.words-310114.csv` information about (word) stimuli, participants, and responses in the ML study

The `.csv` file is a *comma separated values* file and can be opened in Excel.

5.6 Tidy the data

In previous classes, we have needed to combine information about responses with information about participant attributes or stimulus properties, and we have needed to restructure the data so that they are in a tidy format. For this class, many steps in the process of data tidying were completed previously. Thus, we only need to perform steps 1, 3 and 4 of the usual *tidy data* process:

1. Import the data or read the data into R, see Section @ref(import)
2. Restructure the data
3. Select or transform variables, see Section @ref(transform)

4. Filter observations, see Section @ref(filter)

We are going to first filter the observations, then transform the outcome variable. We will explain why we have to do this as we proceed.

We will use `tidyverse` library functions to do this work, as usual.

```
library(tidyverse)
```

5.6.1 Read-in the data file using `read_csv`

I am going to assume you have downloaded the data file, and that you know where it is. We use `read_csv` to read one file into R.

```
ML.all <- read_csv("subjects.beaviour.words-310114.csv", na = "-999")
```

The data file `subjects.beaviour.words-310114.csv` holds all the data about everything (behaviour, participants, stimuli) we need for our analysis exercises in one big dataset.

It is always a good idea to first **inspect what you have got** when you read a data file in to R, before you do anything more demanding.

You can inspect the first few rows of the dataset using `head()`.

```
head(ML.all)
```

```
# A tibble: 6 x 25
  item_number subjectID Test    Age Years~1 Gender TOWRE~2 TOWRE~3 ART_H~4     RT
            <dbl> <chr>   <chr> <dbl> <dbl> <chr>   <dbl> <dbl> <dbl> <dbl>
1           1 GB9     ALT     21    11 F      78     41     18    369.
2           1 NH1     TAL     52    18 M      78     56     33    725.
3           1 A15     LTA     21    16 F      95     57     9     484.
4           1 B18     TLA     69    11 M      85     54    10    518.
5           1 TC14    LAT     21    16 M      97     56     7    621.
6           1 B15     LTA     47    18 M     104     63     38    487.
# ... with 15 more variables: COT <dbl>, Subject <chr>, Trial.order <dbl>,
#   item_name <chr>, Length <dbl>, Ortho_N <dbl>, BG_Sum <dbl>, BG_Mean <dbl>,
#   BG_Freq_By_Pos <dbl>, item_type <chr>, SUBTLWF <dbl>, LgSUBTLWF <dbl>,
#   SUBTLCD <dbl>, LgSUBTLCD <dbl>, OLD <dbl>, and abbreviated variable names
#   1: Years_in_education, 2: TOWRE_wordacc, 3: TOWRE_nonwordacc,
#   4: ART_HRminusFR
```

You can examine all the variables using `summary()`.

```
summary(ML.all)
```

item_number	subjectID	Test	Age
Min. : 1.00	Length:5440	Length:5440	Min. :16.00
1st Qu.: 40.75	Class :character	Class :character	1st Qu.:21.00
Median : 80.50	Mode :character	Mode :character	Median :21.00
Mean : 80.50			Mean :36.94
3rd Qu.:120.25			3rd Qu.:53.00
Max. :160.00			Max. :73.00
Years_in_education	Gender	TOWRE_wordacc	TOWRE_nonwordacc
Min. :11.00	Length:5440	Min. : 68.00	Min. :16.00
1st Qu.:13.00	Class :character	1st Qu.: 84.00	1st Qu.:50.00
Median :16.00	Mode :character	Median : 93.00	Median :55.50
Mean :14.94		Mean : 91.24	Mean :52.41
3rd Qu.:16.00		3rd Qu.: 98.00	3rd Qu.:57.00
Max. :19.00		Max. :104.00	Max. :63.00
ART_HRminusFR	RT	COT	Subject
Min. : 1.00	Min. :-2000.0	Min. : 50094	Length:5440
1st Qu.: 7.00	1st Qu.: 498.1	1st Qu.: 297205	Class :character
Median :11.00	Median : 577.6	Median : 552854	Mode :character
Mean :15.15	Mean : 565.3	Mean : 575780	
3rd Qu.:21.00	3rd Qu.: 677.4	3rd Qu.: 810108	
Max. :43.00	Max. : 1978.4	Max. :1583651	
Trial.order	item_name	Length	Ortho_N
Min. : 21.0	Length:5440	Min. : 3.0	Min. : 0.000
1st Qu.:100.8	Class :character	1st Qu.:4.0	1st Qu.: 3.000
Median :180.5	Mode :character	Median :4.0	Median : 6.000
Mean :180.5		Mean :4.3	Mean : 7.069
3rd Qu.:260.2		3rd Qu.:5.0	3rd Qu.:11.000
Max. :340.0		Max. :6.0	Max. :24.000
BG_Sum	BG_Mean	BG_Freq_By_Pos	item_type
Min. : 3.00	Min. : 1.00	Min. : 1.0	Length:5440
1st Qu.: 81.75	1st Qu.: 67.75	1st Qu.: 74.5	Class :character
Median :151.50	Median :153.50	Median :158.0	Mode :character
Mean :155.89	Mean :153.82	Mean :149.6	
3rd Qu.:234.75	3rd Qu.:239.25	3rd Qu.:227.0	
Max. :314.00	Max. :316.00	Max. :295.0	
SUBTLWF	LgSUBTLWF	SUBLCD	LgSUBLCD
Min. : 0.57	Min. :1.477	Min. : 0.32	Min. :1.447
1st Qu.: 17.36	1st Qu.:2.947	1st Qu.: 6.67	1st Qu.:2.748
Median : 69.30	Median :3.549	Median :23.64	Median :3.298
Mean : 442.01	Mean :3.521	Mean :36.52	Mean :3.137
3rd Qu.: 290.70	3rd Qu.:4.171	3rd Qu.:65.24	3rd Qu.:3.739
Max. :6161.41	Max. :5.497	Max. :99.70	Max. :3.922

OLD

```
Min.    :1.000
1st Qu.:1.288
Median  :1.550
Mean    :1.512
3rd Qu.:1.750
Max.    :2.050
```

The summary shows some features of the dataset, or of how R interprets the dataset, that are of immediate interest to us, though we do not necessarily have to do anything about them.

1. We can see statistical summaries – showing the mean, median, minimum and maximum, etc. – of numeric variables like the outcome variable `RT`, or candidate predictor variables like `LgSUBLCD`, a measure of the frequency of occurrence of words.
2. We can see statistical summaries, also, of variables that comprise number values but which we do not want to be treated as numbers, e.g., the word stimulus coding variable `item_number`.
3. We can see that some variables are simply listed as `Class: character`. That tells us that one or more values in the columns in the datasheet that correspond to these variables are words or strings of letters or alphanumeric characters.
4. There is no sign of the presence of missing values in this dataset, no counts of `NAs`.

We do not really want R to treat a coding variable like `item_number` as numeric: it functions as a categorical or nominal variable, a factor. And we want R to treat coding variables like `subjectID` as factors. In previous chapters, we have ensured that R handles variables exactly as we require using either coercion i.e. using something like an `as.factor()` function call or, at the read-in stage, using `col_types()` specification. We are going to do neither here because we don't have to do this work; not doing it will have no impact on our analyses at this point.

What we **do need to do** is deal with a problem that is already apparent in the summary statistics: we can see that `RT` includes values as low as -2000. That can't be right.

5.6.2 Examine the distribution of raw RT data using density plots

We should examine the distribution of the outcome variable, lexical decision response reaction time (RT in ms). Observations about variable value distributions are a part of *Exploratory Data Analysis* and serve to catch errors in the dataset (e.g. incorrectly recorded scores) but also to inform the researcher's understanding of their own data.

We shall examine the distribution of the outcome variable, lexical decision response reaction time (RT in ms), using density plots. An alternative method would be to use histograms. I choose to use density plots because they allow the easy comparison of the distributions of values of a continuous numeric variable like reaction time. A density plot shows a curve. You can say that the density corresponds to the height of the curve for a given value of the variable

being depicted, and that it is related to the probability of observing values of the variable within some range of values (Howell, 2014).

Getting a density plot of RTs of responses is easy in `ggplot()`.

```
ML.all %>%
  ggplot(aes(x = RT)) +
  geom_density(size=1.5) +
  geom_rug(alpha = .2) +
  ggtitle("Raw RT") +
  theme_bw()
```

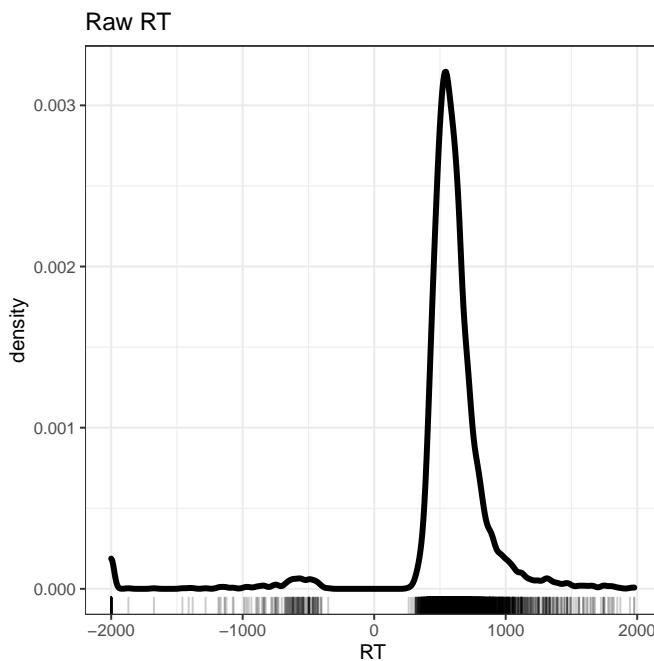


Figure 5.1: Density plot showing word recognition reaction time, correct and incorrect responses

The code delivers a plot (Figure @ref(fig:rt-all-density)) showing three peaks in the distribution of RT values. You can see that there is a peak of RT observations around 500-1000ms, another smaller peak around -500ms, and a third smaller peak around -2000ms.

The density plot shows the reaction times recorded for participants' button press 'yes' responses to word stimuli in the lexical decision task. The peaks of negative RTs represent observations that are impossible. Remember that reaction time, in a task like lexical decision, represents the interval in time between the onset of a task stimulus (in lexical decision, a word or a nonword) and the onset of the response (the button press to indicate the lexical decision).

We cannot have negative time intervals. The explanation is that ML collected her data using DMDX (Forster & Forster, 2003). DMDX records the reaction times for incorrect responses as *negative RTs*.

5.6.2.1 Code tip

The code to produce Figure @ref(fig:rt-all-density) works in a series of steps.

1. `ML.all %>%` takes the dataset, from the ML study, that we have read in to the R workspace and pipes it to the visualization code, next.
2. `ggplot(aes(x = RT)) +` creates a plot object in which the x-axis variable is specified as RT. The values of this variable will be mapped to geometric objects, i.e. plot features, that you can see, next.
3. `geom_density(size=1.5) +` first displays the distribution of values in the variable RT as a density curve. The argument `size=1.5` tells R to make the line $1.5 \times$ the thickness of the line used by default to show variation in density. Some further information is added to the plot, next.
4. `geom_rug(alpha = .2) +` with a command that tells R to add a rug plot below the density curve.
5. `ggttitle("Raw RT")` makes a plot title.

Notice that beneath the curve of the density plot, you can see a series of vertical lines. Each line represents the x-axis location of an RT observation in the ML study data set. This *rug plot* represents the distribution of RT observations in one dimension.

- `geom_rug()` draws a vertical line at each location on the x-axis that we observe a value of the variable, RT, named in `aes(x = RT)`.
- `geom_rug(alpha = .2)` reduces the opacity of each line to ensure the reader can see how the RT observations are denser in some places than others.

You can see that we have many more observations of RTs from around 250ms to 1250ms, where the rug of lines is thickest, under the peak of the density plot. This indicates what the two kinds of plots are doing.

5.6.2.2 Exercise

You should try out alternative visualisation methods to reveal the patterns in the distribution of variables in the ML dataset (or in your own data).

Take a look at the `geoms` documented in:

<https://ggplot2.tidyverse.org/reference/#section-layer-geoms>

Would a histogram or a frequency polygon provide a more informative view?

https://ggplot2.tidyverse.org/reference/geom_histogram.html

What about a dotplot?

https://ggplot2.tidyverse.org/reference/geom_dotplot.html

5.6.3 Filter observations

The density plot shows us that the raw ML lexical decision RT variable includes negative RT values corresponding to incorrect response. These have to be removed. We can do this quite efficiently by creating a subset of the original “raw” data, defined according to the RT variable using the `tidyverse` library `filter()` function.

```
ML.all.correct <- filter(ML.all, RT >= 200)
```

After we have removed negative (error) RTs, we check that the sizes of the datasets – their length, in other words, the number of rows – matches our expectations. We do this to make sure that we did the filter operation correctly.

```
length(ML.all$RT)
```

```
[1] 5440
```

```
length(ML.all.correct$RT)
```

```
[1] 5257
```

If you run the `length()` function calls then you should see that the *length* or number of observations or rows in the `ML.all.correct` dataset should be smaller than the number of observations in the `ML.all` dataset.

Having obtained a new data frame with data on just those trials where responses were correct, we can plot the distribution of RTs for just the correct responses (Figure @ref(fig:rt-correct-density)).

```
ML.all.correct %>%
  ggplot(aes(x = RT)) +
  geom_density(size=1.5) +
  geom_rug(alpha = .2) +
  ggtitle("Correct RTs") +
  theme_bw()
```

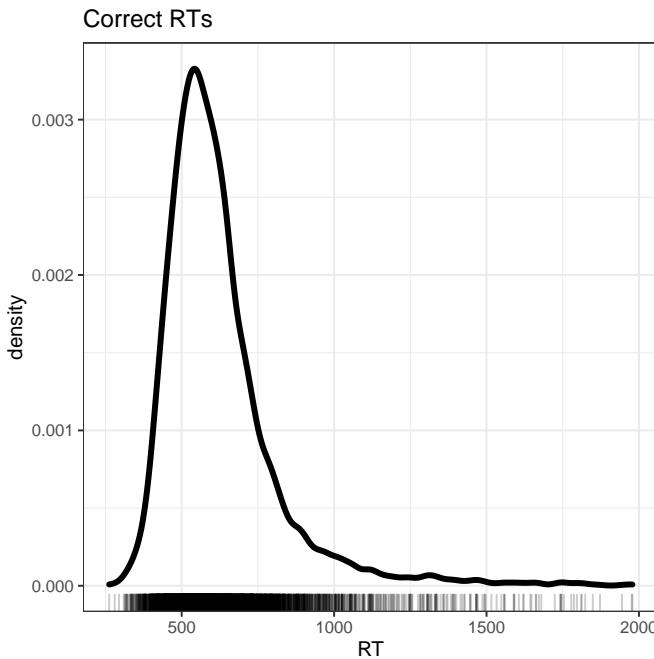


Figure 5.2: Density plot showing word recognition reaction time, correct responses only

5.6.3.1 Code tip

The filter code is written to subset the data by rows using a condition on the values of the RT variable.

`ML.all.correct <- filter(ML.all, RT >= 200)` works as follows.

1. `ML.all.correct <- filter(ML.all ...)` creates a new dataset with a new name `ML.all.correct` from the old dataset `ML.all` using the `filter()` function.
2. `filter(... RT >= 200)` specifies an argument for the `filter()` function. In effect, we are asking R to look at every value in the RT column. R will do a check through the `ML.all` dataset, row by row. *If* a row includes an RT that is greater than or equal to 200 *then* that row will be included in the new dataset `ML.all.correct`. *But if* a row includes an RT that is less than 200, then that row will not be included. We express this condition as `RT >= 200`.

The `length()` function will count the elements in whatever object is specified as an argument in the function call. This means that if you put a variable name into the function as in `length(dataset$variable)` it will count how long that variable is – how many rows there are in the column. If that variable happens to be, as here, part of a dataset, the same calculation will tell you how many rows there are in the dataset as a whole.

If you just enter `length(dataset)`, naming some dataset, then the function will return a count of the number of columns in the dataset.

5.6.3.2 Exercise

Vary the filter conditions in different ways

1. Change the threshold for including RTs from $RT \geq 200$ to something else: you can change the number, or you can change the operators from \geq to a different comparison.
2. Can you assess what impact the change has? Note that you can count the number of observations (rows) in a dataset using e.g. `length()`.

5.6.3.3 Filtering observations as a decision in the psychological research workflow

I choose to filter out or *exclude* not only error responses (where $RT < 0ms$) but also short reaction times (where $RT < 200ms$). I think that any response in the lexical decision task that is recorded as less than 200ms cannot possibly represent a real word recognition response. Participants who complete experimental psychological tasks can and do press the button before they have time to engage the psychological processes (like word recognition) that the tasks we administer are designed to probe (like lexical decision). There is some relevant literature that concerns the speed at which neural word recognition processes operate. However, I think you should note that the threshold I am setting for exclusion, here, is essentially *arbitrary*. If you think about it, I could have set the threshold at any number from 100 – 300ms or some other range. What is guiding me is experience but other researchers will have different experiences and set different thresholds. *This* is why using exclusion criteria to remove data is problematic.

Filtering or re-coding observations is an important element of the research workflow in psychological science. How we do or do not remove observations from original data may have an impact on our results (as explored by Steegen et al., 2016). It is important, therefore, that we learn how to do this reproducibly using, for example, R scripts that we can share with our research reports. I would argue that, at minimum, a researcher should report their research including:

- What exclusion criteria they use to remove data, explaining why.
- Report analyses with and without exclusions, to indicate if their results are sensitive to their decisions.

You can read further information about the practicalities of using R to do filtering here:

<https://r4ds.had.co.nz/transform.html?q=filter#filter-rows-with-filter>

I very much recommend reading the discussion by Steegen et al. (2016) of the impacts of researcher choices in dataset construction analysis results. Don

5.6.4 Select or transform the variables: the log10 transformation of RT

Figure @ref(fig:rt-correct-density) shows that we have successfully removed all errors (negative RTs) but now we see how skewed the RT distribution is. Note the *long tail* of longer RTs.

Most researchers assume that participants – healthy young adults – take about 500-1000ms to perform the task and that values outside that range correspond to either fast guesses (RTs that are too short) or to distracted or tired or bored responses (RTs that are too long). In theory, the lexical decision task should be probing automatic cognitive processes, measuring the steps from perception to visual word recognition in the time interval between the moment the stimulus is first shown and the moment the button is pressed by the participant to indicate a response. Thus, it might seem natural to exclude extreme RT values which might correspond not to automatic cognitive processes but to unknowable distraction events or boredom and inattention. However, we shall complete no further data exclusions.

For now, we can look at a commonly used method to deal with the skew that we typically see when we examine reaction time distributions. RT distributions are usually skewed with a long tail of longer RTs. You can always take longer to press the button but there is a limit to how much faster you can make your response.

Generally, we assume that departures from a model's predictions about our observations (the linear model residuals) are normally distributed, and we often assume that the relationship between outcome and predictor variables is linear (Cohen, Cohen, Aiken, & West, 2003). We can ensure that our data are compliant with both assumptions by transforming the RT distribution.

It is not *cheating* to transform variables. Transformations of data variables can be helpful for a variety of reasons in the analysis of psychological data (Cohen et al., 2003; Gelman & Hill, 2006). I do recommend, however, that you are careful to report what transformations you use, and why you do them.

Psychology researchers often take the log (often log base 10) of RT values before performing an analysis. Transforming RTs to the log base 10 of RT values has the effect of correcting the skew – bringing the larger RTs ‘closer’ (e.g., 1000 = 3 in log10) to those near the middle which do not change as much (e.g. 500 = 2.7 in log10).

```
ML.all.correct$logrt <- log10(ML.all.correct$RT)
```

We can see the effect of the transformation if we plot the log10 transformed RTs (see Figure @ref(fig:rt-correct-log-density)). We arrive at a distribution that more closely approximates the normal distribution.

```
ML.all.correct %>%
  ggplot(aes(x = logrt)) +
```

```

geom_density(size = 1.5) +
geom_rug(alpha = .2) +
ggtitle("Correct log10 RTs") +
theme_bw()

```

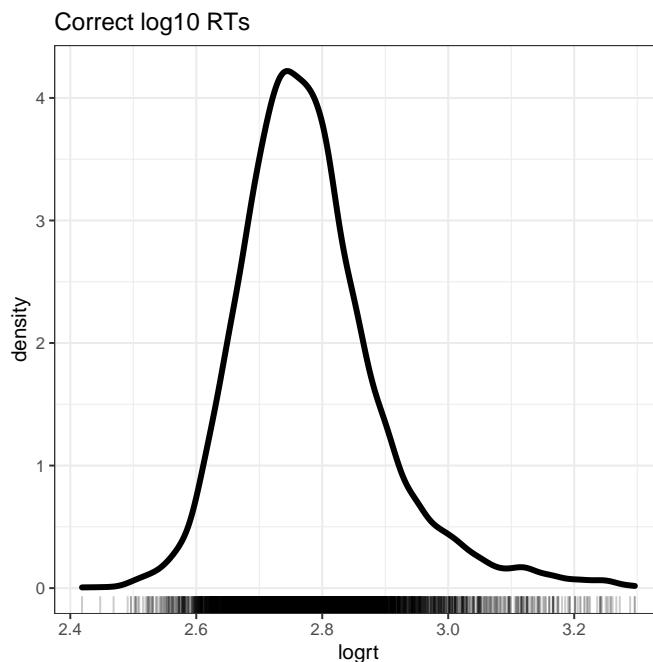


Figure 5.3: Density plot showing log10 transformed reaction time, correct responses only

5.6.4.0.1 Code tip

The `log10()` function works as follows:-

1. `ML.all.correct$logrt <- log10(...)` creates a new variable `logrt`, adding it to the `ML.all.correct` dataset. The variable is created using the transformation function `log10()`.
2. `log10(ML.all.correct$RT)` creates a new variable by transforming (to `log10`) the values of the old variable, `RT`.

There are other log transformation functions and we often see researchers using the natural log instead of the log base 10.

<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/log>

5.6.5 Data tidying – conclusions

Even when data have been structured appropriately, we will still, often, need to do some tidying before we can do an analysis. Most research work involving quantitative evidence requires a *big* chunk of data tidying or other processing before you get to the statistics.

Our data are now ready for analysis.

5.7 Repeated measures designs and crossed random effects

As we saw in the **Introduction to mixed-effects models**, many Psychologists conduct studies where it is not sensible to think of observations as being nested (Baayen et al., 2008). In this chapter, we turn to the **ML word recognition study** dataset, which has a structure similar to the CP study data we worked with previously. Again, the core concern is that the data come from a study with a **repeated-measures design** where the experimenter presented multiple stimuli for response to each participant, for several participants, so that we have multiple observations for each participant and multiple observations for each stimulus. Getting practice with this kind of data will help you to easily recognize what you have got when you see it in your own work.

ML asked all participants in a sample of people to read a selection of words (a sample of words from the language). For each participant, we will have multiple observations and these observations will not be independent. One participant will tend to be slower or less accurate compared to another. Her responses may be more or less susceptible to the effects of the experimental variables. The lowest trial-level observations can be grouped with respect to participants. However, the data can also be grouped by stimuli. For each stimulus word, there are multiple observations and these observations will not be independent. One stimulus may prove to be more challenging to all participants compared to another, eliciting slower or less accurate responses on average. In addition, if there are within-items effects, we may ask if the impact of those within-items effects is more prominent, stronger, among responses to some items compared to others. Given this common *repeated-measures* design, we can analyse the outcome variable in relation to:

fixed effects the impact of independent variables like participant reading skill or word frequency

random effects the impact of random or unexplained differences between participants and also between stimuli

5.8 Working with mixed-effects models

We are going to respond to the multilevel (or crossed random effects) structure in the data by using linear mixed-effects models to analyze the data. This week, we are going to look at what mixed-effects models do from a **new perspective**.

Our concern will be with different ways of thinking about why mixed-effects models are superior to linear models where data have a multilevel structure. Mixed-effects models tend to be more accurate in this (very common) situation because of what is called *partial pooling* and *shrinkage* or *regularization*. We use our practical example to explore these ideas.

5.8.1 Use facetting in ggplot to examine data by person

To get started, we can examine – for each individual separately – the distribution of log RT observations, in Figure @ref(fig:rt-correct-log-den-by-subj).

```
ML.all.correct %>%
  group_by(subjectID) %>%
  mutate(mean_logrt = mean(logrt, na.rm = TRUE)) %>%
  ungroup() %>%
  mutate(subjectID = fct_reorder(subjectID, mean_logrt)) %>%
  ggplot(aes(x = logrt)) +
  geom_density(size = 1.25) +
  facet_wrap(~ subjectID) +
  geom_vline(xintercept = 2.778807, colour = "red", linetype = 2) +
  scale_x_continuous(breaks = c(2.5,3)) +
  ggtitle("Plot showing distribution of logRT for each participant; red line shows mean logRT for each person")
  theme_bw()
```

Plot showing distribution of logRT for each participant; red line shows mean log10 RT

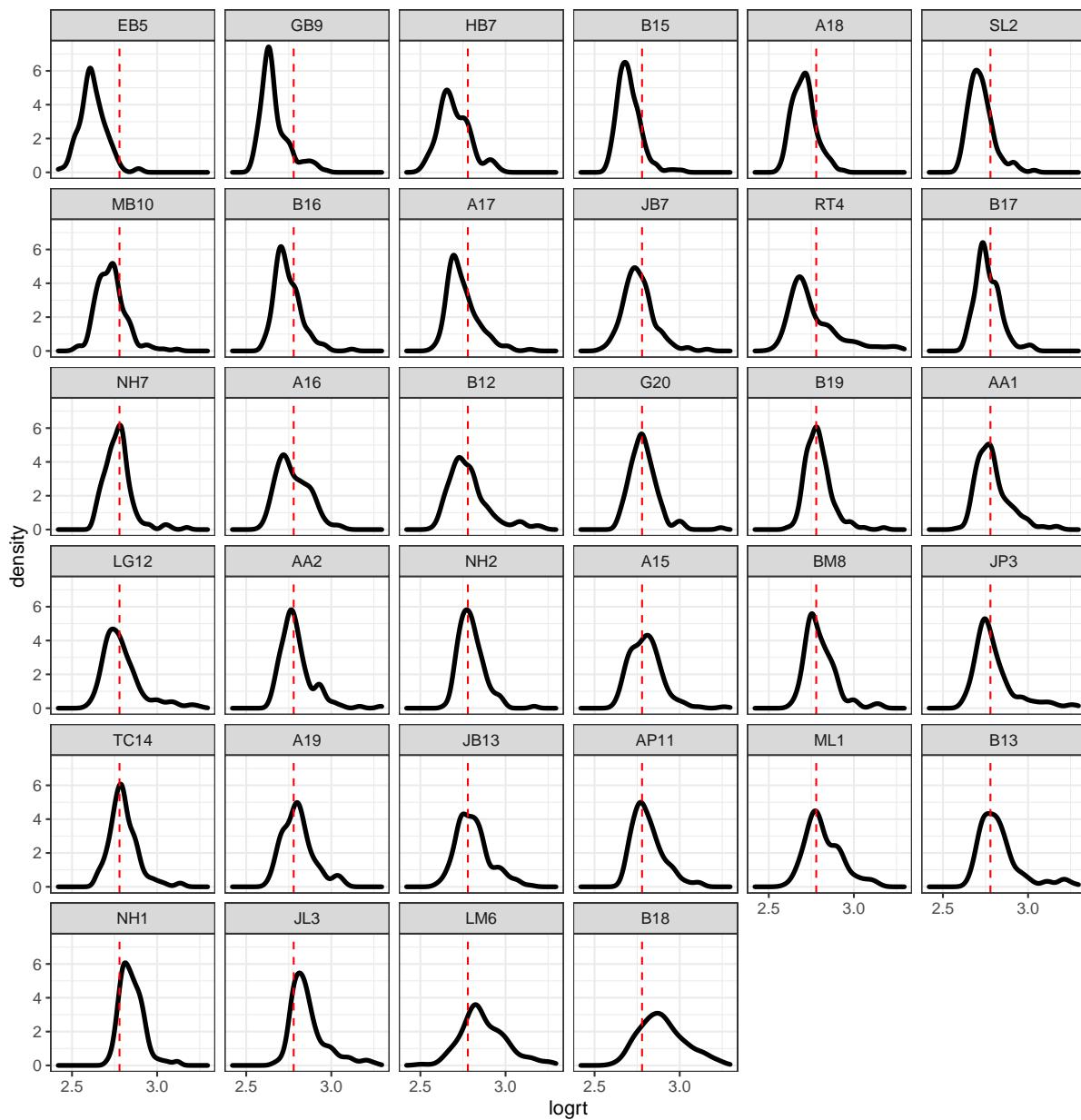


Figure 5.4: Density plot showing log10 transformed reaction time, correct responses, separately for each participant

Figure @ref(fig:rt-correct-log-den-by-subj) shows that RT distributions vary considerably between people. The plot imposes a dashed red line to indicate where the mean log10 RT is, calculated over all observations in the dataset. The plot shows the distribution of log RT for each participant, as a density drawn separately for each person. The individual plots are

ordered by the mean log RT calculated per person, so plots appear in order from the fastest to the slowest.

The grid of plots illustrates some interesting features about the data in the ML study sample. You can see how the distribution of log RT varies between individuals: some people show widely spread reaction times; some people show quite tight or narrow distributions. You can see how the shapes of the distributions varies: some people show skew; others do not. I do not see that the variation in the shapes of the distributions is related to the average speed of the person's responses.

I think the key message of the plot is that some distributions are wider (RTs are more spread out) than others. We might be concerned that people who present more variable reaction times (wider distributions) may be associated with less reliable estimates of their average response speed, or of the impact of word attributes (like word frequency) on their response speed.

5.8.1.1 Code tip

The plotting code progresses through a series of steps. This example demonstrates how you can combine data tidying and plotting steps in a single sequence, using `tidyverse` functions and the `%>%` pipe, so I will take the time to explain what is going on.

My aim is to create a grid of individual plots, showing the distribution of log RTs for each participant, so that the plots are presented in order, from the fastest participant to the slowest. Take a look at the plotting code. We can explain how it works, step by step.

```
ML.all.correct %>%
  group_by(subjectID) %>%
  mutate(mean_logrt = mean(logrt, na.rm = TRUE)) %>%
  ungroup() %>%
  mutate(subjectID = fct_reorder(subjectID, mean_logrt)) %>%
  ggplot(aes(x = logrt)) +
  geom_density(size = 1.25) +
  facet_wrap(~ subjectID) +
  geom_vline(xintercept = 2.778807, colour = "red", linetype = 2) +
  scale_x_continuous(breaks = c(2.5,3)) +
  ggtitle("Plot showing distribution of logRT for each participant; red line shows mean lo
theme_bw()
```

You will see that we present the distribution of RTs using `geom_density()` and that we present a separate plot for each person's data using `facet_wrap()`. To these elements, we add some pre-processing steps to calculate the average response speed of each individual, and to reorder the dataset by those averages.

It will make it easier to understand what is going on if we consider the code in chunks.

First, we pre-process the data before we feed it into the plotting code.

```
ML.all.correct %>%
  group_by(subjectID) %>%
  mutate(mean_logrt = mean(logrt, na.rm = TRUE)) %>%
  ungroup() %>%
  mutate(subjectID = fct_reorder(subjectID, mean_logrt)) %>%
  ...
```

1. `ML.all.correct %>%` takes the selected filtered dataset `ML.all.correct` and pipes it `%>%` to the next step.
2. `group_by(subjectID) %>%` tells R to group the data by `subject ID`. We have a set of multiple log RT observations for each `subjectID` because each participant was asked to respond to multiple word stimuli.
3. `mutate(mean_logrt = mean(logrt, na.rm = TRUE))` next calculates and stores the mean log RT for each person. We create a new variable `mean_logrt`. We calculate the average of the set of log RTs recorded for each `subjectID` and construct the new variable `mean_logrt` from these averages. We do not need to treat the data in groups so we remove the grouping, next.
4. `ungroup() %>%` having grouped the data to calculate the mean log RTs, we `ungroup` the dataset so that R can look at all observations in the next step.
5. `mutate(subjectID = fct_reorder(subjectID, mean_logrt)) %>%` asks R to look at all log RT observations in the dataset, and change the top-to-bottom order of the rows. We ask R to order observations by `subjectID` so that each person's data are listed by their average speed, from the fastest to the slowest. We then pipe these ordered data to the plotting code, next.

If you delete or comment out these first lines, you will see that R uses just a default ordering, drawing the plot for each person in the alphabetical order of their `subjectID` codes. Try it, though don't forget to start with `ML.all.correct %>%`.

Second, we draw the plots, using the data we have pre-processed.

```
ML.all.correct %>%
  ...
  ggplot(aes(x = logrt)) +
  geom_density(size = 1.25) +
  facet_wrap(~ subjectID) +
  ...
```

The key functions that create a grid of density plots are the following.

1. `ggplot(aes(x = logrt))` tells R to work with `logrt` as the x-axis variable. We shall be plotting the distribution of `logrt`.

2. `geom_density(...)` draws a density plot to show the distribution of log RT, using a thicker line `size = 1.25`
 3. `facet_wrap(~ subjectID)` creates a different plot for each level of the `subjectID` factor: we want to see a separate plot for each participant.
- `facet_wrap(~ subjectID)` works to split the dataset up by participant, with observations corresponding to each participant identified by their `subjectID`, and to then split the plotting to show the distribution of log RT separately for each participant.

I wanted to present the plots in order of the average speed of response of participants. If you look at Figure @ref(fig:rt-correct-log-den-by-subj) you can see that the position of the peak of the log RT distribution for each participant moves, from the fastest plots where the peak is around $\log RT = 2.5$ (shown from the top left of the grid), to the slowest plots where the peak is around $\log RT = 2.75$ (shown towards the bottom right of the grid)

We can then use further `ggplot` functions to edit the appearance of the plot, to make it more useful.

```
...
  geom_vline(xintercept = 2.778807, colour = "red", linetype = 2) +
  scale_x_continuous(breaks = c(2.5,3)) +
  ggtitle("Plot showing distribution of logRT for each participant; red line shows mean logRT")
  theme_bw()
```

1. `geom_vline(xintercept = 2.778807, colour = "red", linetype = 2)` draws a vertical red dashed line at the location of the mean log RT, the average of all log RTs over all participants in the dataset.
 2. `scale_x_continuous(breaks = c(2.5,3))` adjusts the x-axis labeling. The `ggplot` default might draw too many x-axis labels i.e. showing possible log RT values as tick marks on the bottom line of the plot. I want to avoid this as sometimes all the labels can be crowded together, making them harder to read.
- Drawing a vertical line at the mean calculated overall is designed to help the reader (you) calibrate their comparison of the data from different people.

5.8.1.2 Exercises

It is worthwhile experimenting with example code to figure out how it works. One way you can do this is by commenting out one line of code, at a time by putting the `#` at the start of the line. If you do this, you can see what the line of code does by, effectively, asking R to ignore it.

Another way you can experiment with code is by seeing what you can change and what effect the changes have.

- Can you work out how to adapt the plotting code to show a grid of histograms instead of density plots?
- Can you work out how to adapt the code to show a grid of plots indicating the distribution of log RT by different words instead of participants?

5.8.2 Approximations to Linear Mixed-effects models: complete pooling

As we have discussed in previous chapters, a good way to approach a mixed-effects analysis is by first estimating the effects of the experimental variables (here, frequency) using linear models, ignoring the hierarchical structure in the data. A linear model of multilevel structured data can be regarded as an **approximation** to the better analysis. We model the effects of interest, using all the data (hence, *complete pooling*) but ignoring the differences between participants. This means we can see something of the ‘true’ picture of our data through the linear model results but the linear model misses important information, which the mixed-effects model will include, that would improve its performance.

As we saw in the last chapter, we can estimate the relationship between lexical decision RTs and word frequency using a linear model:

$$Y_{ij} = \beta_0 + \beta_1 X_j + e_{ij} \quad (5.1)$$

- where Y_{ij} is the value of the observed outcome variable, the log RT of the response made by the i participant to the j item;
- $\beta_1 X_j$ refers to the fixed effect of the explanatory variable (here, word frequency), where the frequency value X_j is different for different words j , and β_1 is the estimated coefficient of the effect due to the relationship between response speed and word frequency;
- e_{ij} is the residual error term, representing the differences between observed Y_{ij} and predicted values (given the model) for each response made by the i participant to the j item.

The linear model is fit in R using the `lm()` function.

```
ML.all.correct.lm <- lm(logrt ~
                           LgSUBLCD,
                           data = ML.all.correct)

summary(ML.all.correct.lm)
```

```

Call:
lm(formula = logrt ~ LgSUBLCD, data = ML.all.correct)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.41677 -0.07083 -0.01163  0.05489  0.53411 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.885383   0.007117 405.41   <2e-16 ***
LgSUBLCD   -0.033850   0.002209 -15.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1095 on 5255 degrees of freedom
Multiple R-squared:  0.04277, Adjusted R-squared:  0.04259 
F-statistic: 234.8 on 1 and 5255 DF,  p-value: < 2.2e-16

```

We can see that, in this first analysis, the estimated effect of word frequency is $\beta = -0.033850$. I know this looks like a very small number but you should realize that the estimates for the coefficients of fixed effects like the frequency effect are scaled according to the outcome. Here, the outcome is log10 RT, where a log10 RT of 3 equals 1000ms, and, as we can calculate in R

`log10(0.925)`

[1] -0.03385827

Also, remember that frequency is scaled in logs too, so the estimate of the coefficient tells us how log10 RT changes for unit change in log frequency. The coefficient represents the estimated change in log10 RT for unit change in log frequency LgSUBLCD. So, as log frequency increases, logRT *decreases* by -0.033850 .

In this model, all the information from all participants is analyzed. In discussions of mixed-effects analyses, we say that this is a *complete pooling* model. This is because *all the data have been pooled together*, that is, we use all observations in the sample to estimate the effect of frequency.

In this model, the observations are assumed to be independent. However, we suppose that the assumption of independence is questionable given the expectation that participants will differ in their overall speed, and in the extent to which their response speed is affected by factors like word frequency.

5.8.2.1 Exercises

Vary the linear model using different outcomes or predictors

The ML study data, like the CP study data, are rich with possibility. It would be useful to experiment with it.

1. Change the predictor from frequency to something else: what do you see when you visualize the relationship between outcome and predictor variables using scatterplots?
2. Specify linear models with different predictors: do the relationships you see in plots match the coefficients you see in the model estimates?

I would recommend that you both estimate the effects of variables *and* visualize the relationships between variables using scatterplots. If you combine reflection on the model estimates with evaluation of what the plots show you then you will be able to see how reading model results and reading plots can reveal the correspondences between the two ways of looking at your data.

5.8.3 Approximations to Linear Mixed-effects models: no pooling

We can examine variation between participants by analyzing the data for each participant's responses separately, fitting a *different* linear model of the effect of word frequency on lexical decision RTs for each participant *separately*. Figure @ref(fig:no-vs-complete) presents a grid or trellis of plots, one plot per person. In each plot, you can see points corresponding to the log RT of each response made by a participant to a stimulus word. In all plots, the pink or red line represents the *complete pooling* model estimate of the effect of frequency on response RTs. The line is the same for each participant because there is only one estimated effect, based on all data for all participants. In addition, in each plot, you can see a green line. You can see that the line varies between participants. This represents the effect of frequency estimated using just the data for each participant, analyzed separately. These are the *no pooling* estimates. We call them the no pooling estimates because each is based just on the data from one participant.

Note: I was able to produce the plot in Figure @ref(fig:no-vs-complete) thanks to this helpful blog post by TJ Mahr:

<https://www.tjmahr.com/plotting-partial-pooling-in-mixed-effects-models/>

Figure @ref(fig:no-vs-complete) reveals substantial differences between participants in both average response speed and the frequency effect, alongside variation in standard errors. We can predict variation in standard errors between participants given, also, the differences between participants in the spread of log RT, illustrated by Figure @ref(fig:rt-correct-log-den-by-subj). Basically, where the distribution of log RT is more widely spread out, for any one participant, there it will be harder for us to estimate with certainty the mean or the sources of variance for the participant's response speed.

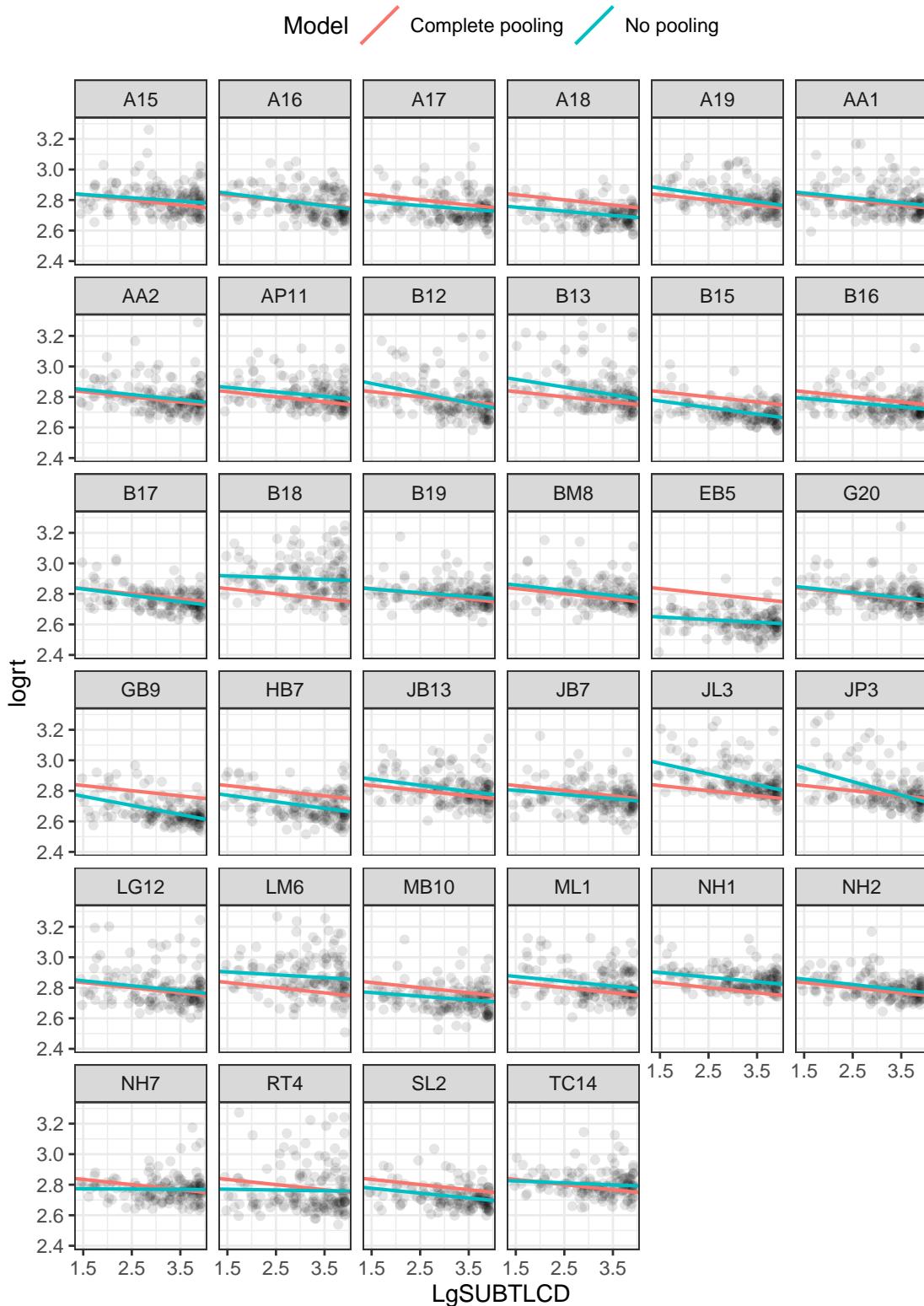


Figure 5.5: Plot showing the relationship between logRT and log frequency (LgSUBLCD) separately for each participant; red-pink line shows the complete pooling estimate, blue-green line shows the no-pooling estimate

You will notice that the *no pooling* and *complete pooling* estimates tend to be quite similar. But for some participants more than others there is variation between the estimates.

You can reflect that the *complete pooling* is unsatisfactory because it ignores the variation between the participants: some people *are* slower than others; some people *do* show a larger frequency effect than others. You can also reflect that the *no pooling* is unsatisfactory because it ignores the similarities between the participants. While there *is* variation between participants there is also similarity across the group so that the effect of frequency is similar between participants. What we need is an analytic method that is capable of *both* estimating the overall average population-level effect (here, of word frequency) and taking into account the differences between sampling units (here, participants). **That method is linear mixed-effects modeling.**

5.9 The linear mixed-effects model

5.9.1 Fixed and random effects

As you have seen before, we can account for the variation – the differences between participants in intercepts and slopes. First, we model the intercept as two terms:

$$\beta_{0i} = \gamma_0 + U_{0i} \quad (5.2)$$

- where γ_0 is the average intercept and U_{0i} is the difference for each participant between their intercept and the average intercept.

We can model the frequency effect as two terms:

$$\beta_{1i} = \gamma_1 + U_{1i} \quad (5.3)$$

- where γ_1 is the average slope and U_{1i} represents the difference for each participant between the slope of their frequency effect and the average slope.

We can then incorporate in a single model the **fixed effects** due to the average intercept and the average frequency effect, as well as the **random effects**, error variance due to unexplained differences between participants in intercepts and in frequency effects:

$$Y_{ij} = \gamma_0 + \gamma_1 X_j + U_{0i} + U_{1i} X_j + e_{ij} \quad (5.4)$$

- where the outcome Y_{ij} is related to ...
- the average intercept γ_0 and differences between i participants in the intercept U_{0i} ;
- the average effect of the explanatory variable frequency $\gamma_1 X_j$ and differences between i participants in the slope $U_{1i} X_j$;

- in addition to residual error variance e_{ij} .

5.9.2 Variance and covariance

As we first saw in the last chapter, in conducting mixed-effects analyses, we do not aim to examine the specific deviation (here, for each participant) from the average intercept or the average effect or slope. We estimate just the spread of deviations by-participants. A mixed-effects model like our final model actually includes fixed effects corresponding to the intercept and the slope of the word frequency effect plus the variances:

- $\text{var}(U_{0i})$ variance of deviations by-participants from the average intercept;
- $\text{var}(U_{1i}X_j)$ variance of deviations by-participants from the average slope of the frequency effect;
- $\text{var}(e_{ij})$ residuals, at the response level, after taking into account all other terms.

We may expect the random effects of participants or items to covary, e.g., participants who are slow to respond may also be more susceptible to the frequency effect. Thus our specification of the random effects of the model can incorporate terms corresponding to the covariance of random effects:

- $\text{covar}(U_{0i}, U_{1i}X_j)$

5.9.3 Random effects of differences between stimuli

As we know, some words elicit slower and some elicit faster responses on average. As we discussed in the last chapter, if we did not take such variation into account, we might spuriously identify an experimental effect actually due just to unexplained between-items differences in intercepts (Clark, 1973; Raaijmakers et al., 1999), committing an error: *The language as fixed effect fallacy*.

We can model the random effect of items on intercepts by modeling the intercept as two terms:

$$\beta_{0j} = \gamma_0 + W_{0j}$$

- where γ_0 is the average intercept and W_{0j} represents the deviation, for each word, between the average intercept and the per-word intercept.

Note that I ignore the possibility, for now, of differences between items in the slopes of fixed effects but I *do* come back to this.

Remember that the *The language as fixed effect fallacy* implies that thinking about the random effects of stimulus differences applies only when we are looking at experiments about

language. But you should remember that we need to think about the impact of random differences between stimuli whenever we present samples of stimuli to participants, and we collect observations about multiple responses for each stimulus. This is true whatever the nature of the stimuli.

5.9.4 A model including random effects of differences between stimuli as well as participants

Our model can *now* incorporate the random effects of participants as well as items:

$$Y_{ij} = \gamma_0 + \gamma_1 X_j + U_{0i} + U_{1i} X_j + W_{0j} + e_{ij}$$

In this model, the outcome Y_{ij} is related to the average intercept γ_0 and the word frequency effect $\gamma_1 X_j$ plus random effects due to unexplained differences between participants in intercepts U_{0i} and the slope of the frequency effect $U_{1i} X_j$ as well as random differences between items in intercepts W_{0j} , in addition to the residual term e_{ij} .

5.9.5 Fitting a mixed-effect model using lmer()

We fit a mixed-effects model of the *logrt ~ frequency* relationship using the `lmer()` function, taking into account:

- the fact that the study data have a hierarchical structure – with observations sensibly grouped by participant;
- the fact that both the frequency effect, and average speed, may vary between participants;
- and the fact that the average speed of response can vary between responses to different stimuli.

The model syntax corresponds to the statistical formula and the code is written as:

```
ML.all.correct.lmer <- lmer(logrt ~
  LgSUBLCD +
  (LgSUBLCD + 1 | subjectID) +
  (1 | item_name),
  data = ML.all.correct)

summary(ML.all.correct.lmer)
```

As will now be getting familiar, the code works as follows:

1. `ML.all.correct.lmer <- lmer(...)` creates a *linear mixed-effects model* object using the `lmer()` function.
2. `logrt ~ LgSUBLCD` the fixed effect in the model is expressed as a formula in which the outcome or dependent variable `logrt` is predicted `~` by the independent or predictor variable `LgSUBLCD` word frequency.
3. If there were more terms in the model, the terms would be added in series separated as `...by + ...`
4. `(...|subjectID)` adds random effects corresponding to random differences between sample groups (subjects) coded by the `subjectID` variable,
5. `(...1 |subjectID)` including random differences between sample groups (`subjectID`) in intercepts coded 1,
6. `(LgSUBLCD... |subjectID)` and random differences between sample groups (`subjectID`) in slopes of the frequency effect coded by using the `LgSUBLCD` variable name.
7. `(1|item_name)` adds a random effect to account for random differences between sample groups (`item_name`) in intercepts coded 1.
8. `...(..., data = ML.all.correct)` specifies the dataset in which you can find the variables named in the model fitting code.
9. `summary(ML.all.correct.lmer)` gets a summary of the fitted model object.

If you run the model code as written then you would see the following results.

```
Linear mixed model fit by REML ['lmerMod']
Formula: logrt ~ LgSUBLCD + (LgSUBLCD + 1 | subjectID) + (1 | item_name)
Data: ML.all.correct

REML criterion at convergence: -9868.1

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-3.6307 -0.6324 -0.1483  0.4340  5.6132 

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 item_name (Intercept) 0.0003268 0.01808
 subjectID (Intercept) 0.0054212 0.07363
           LgSUBLCD    0.0002005 0.01416  -0.63
 Residual            0.0084333 0.09183
```

```
Number of obs: 5257, groups: item_name, 160; subjectID, 34
```

```
Fixed effects:
```

	Estimate	Std. Error	t value
(Intercept)	2.887997	0.015479	186.577
LgSUBLCD	-0.034471	0.003693	-9.333

```
Correlation of Fixed Effects:
```

	(Intr)
LgSUBLCD	-0.764

In these results, we see:

1. First, information about the function used to fit the model, and the model object created by the `lmer()` function call.
2. Then, we see the model formula `logrt ~ LgSUBLCD + (LgSUBLCD + 1 | subjectID) + (1 | item_name)`.
3. Then, we see `REML criterion at convergence` about the model fitting process, which we can usually ignore.
4. Then, we see information about the distribution of the model residuals.
5. Then the **Random Effects**. Notice that the statistics are `Variance Std.Dev. Corr.`, that is, the variance, the corresponding standard deviation, and the correlation estimates associated with the random effects.
 - We see `Residual` error variance, just like in a linear model, corresponding to a distribution or spread of deviations between the model prediction and the observed RT for each response made by a participant to a stimulus.
 - In addition, for this model, we see `Variance` terms corresponding to what can be understood as group-level residuals. Here, the variance due to random differences between the average intercept (over all data) and the intercept for each participant, and the variance due to random differences between the average slope of the frequency effect and the slope for each participant.
 - We also see the variance due to random differences between the average intercept (over all data) and the intercept for responses to each word stimulus.
 - And we see the `corr` estimate, telling us about the *covariance* between random deviations (between participants) in the intercepts and in the slopes of the frequency effect
6. Last, just as for linear models, we see estimates of the coefficients (of the slopes) of the fixed effects, the intercept and the slope of the `logrts ~ LgSUBLCD` relationship.
7. Note that we see coefficient estimates like in a linear model summary *but no p-values*

5.9.5.1 Why aren't there p-values?

We will come back to this, see Section @ref(p-values). However, note that if $t \geq 2$ we can suppose that (for a large dataset) an effect is significant at the .05 significance level.

5.10 Mixed-effects models, partial pooling, and shrinkage or regularisation of estimates

What is the impact of the incorporation of random effects – the variance and covariance terms – in mixed-effects models? Mixed-effects models can be understood, in general, as a method to compromise between ignoring the differences between groups (here, participants constitute groups of data) as in *complete pooling* or focusing entirely on each group (participant) as in *no pooling* (Gelman & Hill, 2006). In this discussion, I am going to refer to the differences between participants but you can assume that the lesson applies generally to any situation in which you have different units in a multilevel structured dataset in which the units correspond to groups or clusters of data, as in the multiple observations recorded for each participant in the ML dataset.

5.10.1 Overfitting

The problem with ignoring the differences between groups (participants), as in the *complete pooling* model (here, the linear model), has been obvious when we examined the differences between participants (or between classes) in slopes and intercepts in previous weeks. The problem with focusing entirely on each participant, as in the *no pooling* model, has not been made apparent in our discussion yet.

If we analyze each participant separately then we will get, for each participant, for our model of the frequency effect, the per-participant estimate of the intercept and the per-participant estimate of the slope of the frequency effect. These no-pooling estimates will tend to exaggerate or overstate the differences between participants (Gelman & Hill, 2006). By basing the estimates on just the data for a person, in each per-participant analysis, the no-pooling approach *overfits* the data. You could say that the no-pooling approach gives us estimates that depend too much on the sample of data we have got, and are unlikely to be similar to the estimates we would see in other samples in future studies. The no-pooling estimates are *too strongly* influenced by the data we are currently analyzing.

5.10.2 Partial pooling: shrinkage or borrowing strength

If we look closely at Figure @ref(fig:no-vs-complete), we can see that there are similarities as well as differences between participants. Our analysis must take both into account.

What happens in mixed-effects models is that we pool information, calculating the estimates for each participant, in part based on the information we have for the whole sample (all participants, in complete pooling), in part based on the information we have about the specific participant (one participant, in no pooling). Thus, for example, the estimated intercept for a participant in a mixed-effects model is given by the weighted average of:

- the intercept estimate given by an analysis of just that participant's data (no pooling estimate);
- and the intercept estimate given by analysis of all participants' data (complete pooling estimate).

The weighted average will reflect our relative level of information about the participant's responses compared to how much information we have about all participants' responses.

- For some participants, we will have less information – maybe they made many errors, so we have fewer correct responses for an analysis.
 - For these people, because we have less information, the intercept estimate will get pulled (shrunk) towards the overall (complete pooling, all data) estimate.
- For other participants, we have more information – maybe they made all correct responses
 - For these people, because we have more information, the intercept estimate will be based more on the data for each participant.

To make sense of what this means, think about the differences between participants in how much reliable information we can have, given our sample, about their average level of response speed or about how they are affected by experimental variables. Think back to my comments about Figure @ref(fig:rt-correct-log-den-by-subj), about the differences between participants in how spread out the distributions of their log RT values are. Recall that I said that where participants' responses are more spread out – just as where we have less observations for some participants than for others – we shall inevitably have less certainty about our estimates for the effects that influence their performance if we base our account on just their data. Mixed-effects models perform better – as prediction models – than *no pooling* approaches because they are not relying, for any participant, on just their sometimes unreliable data.

We can look again at a plot showing the data for each participant. Figure @ref(fig:no-vs-complete-vs-partial) presents a grid or trellis of plots, one plot per person. In each plot, you can see points corresponding to the RT of each response made by a participant to a stimulus word. In all plots, the pink line represents the *complete pooling* data model estimate of the effect of frequency on response RTs. In each plot, the green line represents the effect of frequency estimated using just the data for each participant, the *no pooling* estimates. Now, we also see blue lines that represent the mixed-effects model *partial pooling* estimates.

It is quite difficult to identify, in this sample, where the partial pooling and no pooling estimates differ. We can focus on a few clear examples. Figure @ref(fig:no-vs-complete-vs-partial-zoom)

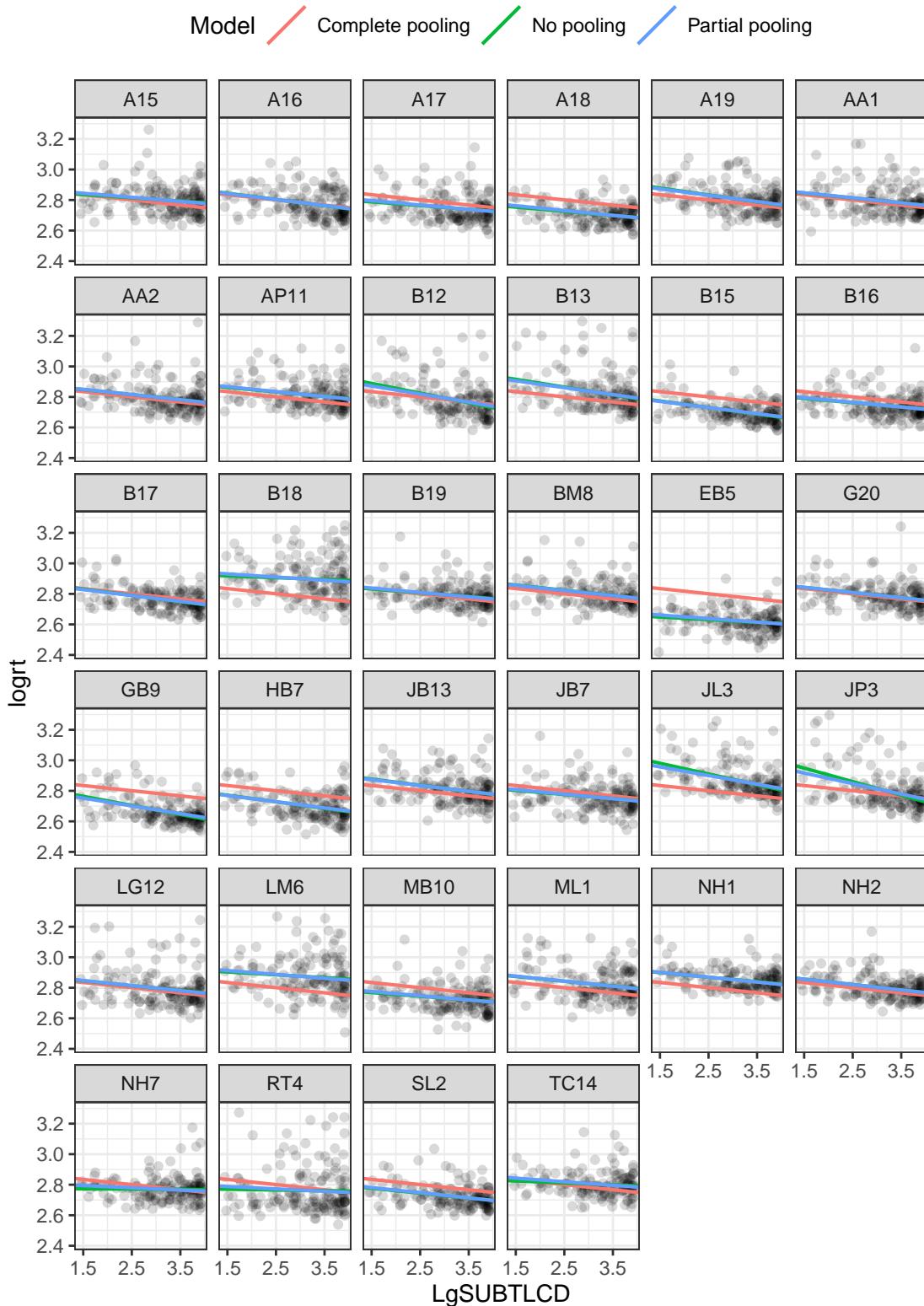


Figure 5.6: Plot showing the relationship between logRT and log frequency (LgSUBLCD) separately for each participant; pink line shows the complete pooling estimate; green line shows the no-pooling estimate; and blue line shows the linear mixed-effects model partial pooling estimate

presents a grid of plots for just four participants. I have picked some extreme examples but the plot illustrates how: (1.) for some participants e.g. AA1 all estimates are practically identical; (2.) for some participants EB5 JL3 JP3 the no-pooling and complete-pooling estimates are really quite different and (3.) for some participants JL3 JP3 the no-pooling and partial-pooling estimates are quite different.

In general, partial pooling will apply both to estimates of intercepts and to estimates of the slopes of fixed effects like the influence of word frequency in reaction time. Likewise, if we consider this idea in general, we can see how it should work whether we are talking about groups or clusters of data grouped by participant or by stimulus or by school or class or clinic
...

Formally, whether an estimate for a participant (in our example) is pulled more or less towards the overall estimate will depend not just on the number of data-points we have for that person. The optimal combined estimate for a participant is termed the *Empirical Bayes* ‘estimate’ and the weighting – the extent to which the per-participant ‘estimate’ depends on the participant’s data or the overall data – depends on the reliability of the estimate (of the intercept or the frequency effect) given by analyzing that participant’s data (Snijders & Bosker, 2012). If you think about it, smaller samples – e.g. where a participant completed less correct responses – will give you less reliable estimates (and so will samples that show more variation).

What we are looking at, here, is a form of *regularization* in which we use all the sources of information we can to ensure we take into account the variability in the data while not getting over-excited by extreme differences (McElreath, 2015). We want to see estimates pulled towards an overall average where we have little data or unreliable estimates. We can see how strongly estimates can be shrunk in a plot like Figure @ref(fig:shrinkage).

Figure @ref(fig:shrinkage) illustrates the shrinkage effect. I plotted a scatterplot of intercept and slope parameters from each model (models with different kinds of pooling), and connect estimates for the same participant. The plot uses arrows to connect the different estimates for each participant, different estimates from no-pooling (per-participant) compared to partial-pooling (mixed-effects) models. The plot shows how more extreme estimates are shrunk towards the global average estimate.

We can see how estimates are pulled towards the average intercept and frequency effect estimates. The shrinkage effect is stronger for more extreme estimates like JL3 JP3. It is weaker for estimates more (realistically) like the overall group estimates like AA1.

5.11 Estimation methods – An intuitive account of estimation in mixed-effects models

Before we move on, we can think briefly about how the mixed-effects models are estimated (Snijders & Bosker, 2012). Where do the numbers come from? I am happy to stick to a

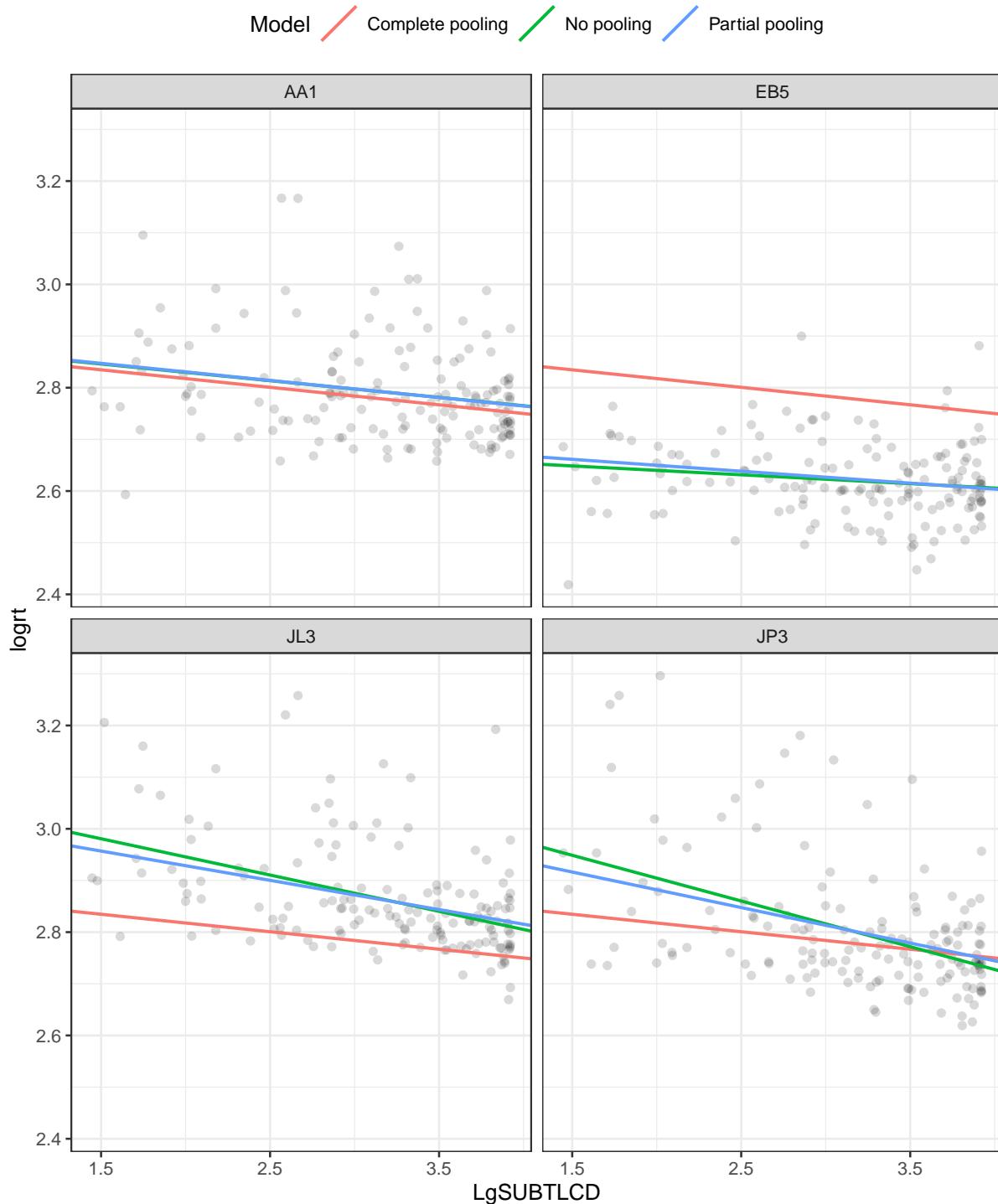


Figure 5.7: Plot showing the relationship between logRT and log frequency (LgSUBLCD) separately for each participant – for participants AA1, EB5, JL3 and JP3; pink line shows the complete pooling estimate green line shows the no-pooling estimate; and blue line shows the linear mixed-effects model partial pooling estimate

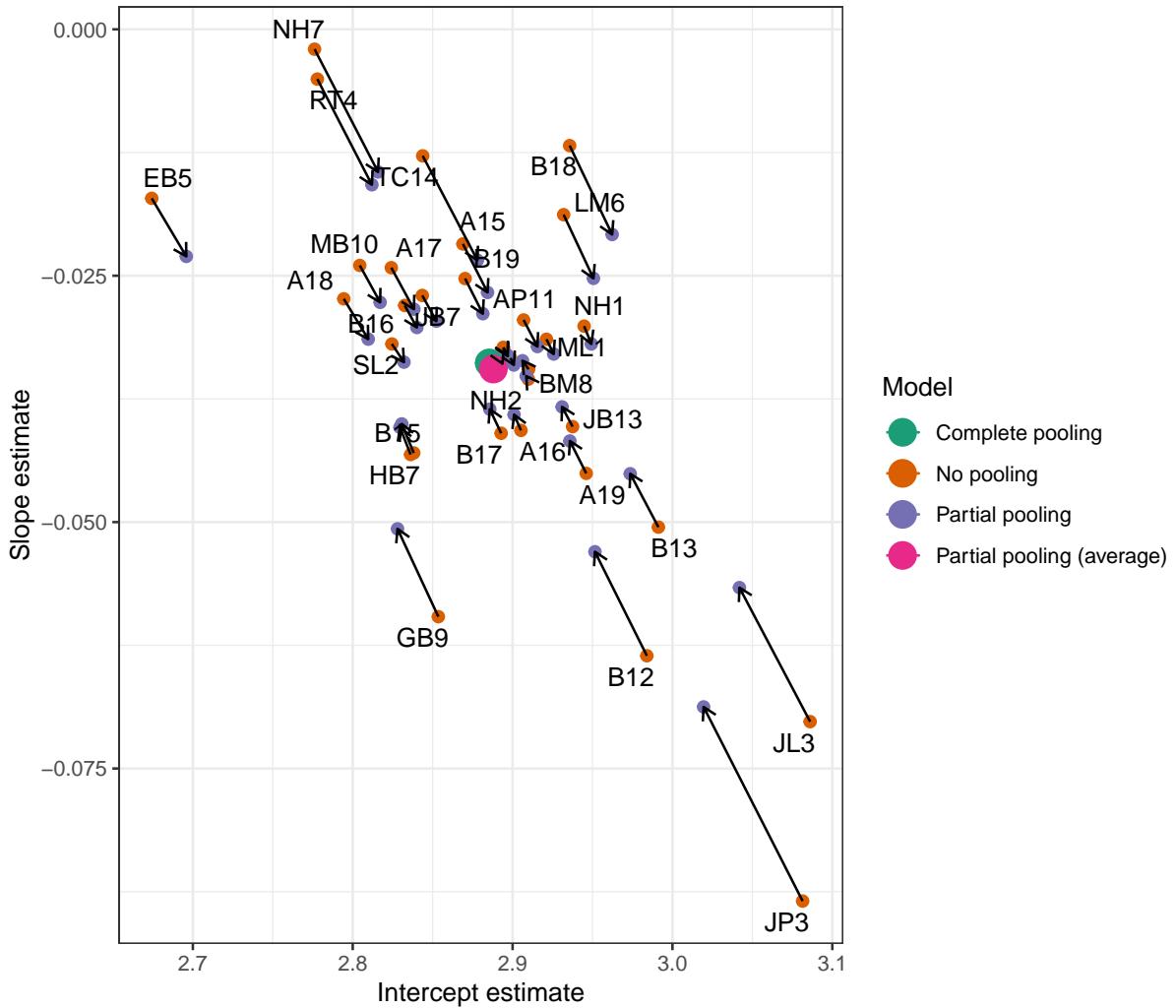


Figure 5.8: Plot illustrating shrinkage: big green and pink points show the complete pooling and partial pooling (average) estimates for the slope and intercept; orange and purple points show the no pooling (orange) and partial pooling (purple) estimates for each person; estimates for a person are connected by arrows to show the direction towards which no pooling estimates are pulled or shrunk

fairly non-technical intuitive explanation of the computation of LMEs but others, wishing to understand things more deeply, can find computational details in Pinheiro & Bates (2000), among other places. Mixed-effects models are estimated *iteratively* ...

- If we knew the random effects, we could find the fixed effects estimates by minimizing differences – like linear modeling
- If we knew the fixed effects – the regression coefficients – we could work out the residuals and the random effects
- At the start, we know neither, but we can move between partial estimation of fixed and random effect in an *iterative approach*
 - Using provisional values for the fixed effects to estimate the random effects
 - Using provisional values for the random effects to estimate the fixed effects again
 - To *converge* on the maximum likelihood estimates of effects – when the estimates stop changing

In mixed-effects models, the things that are estimated are the fixed effects (the intercept, the slope of the frequency effect, in our example), along with the variance and correlation terms associated with the random effects. Previously, I referred to the partial-pooling mixed-effects ‘estimates’ of the intercept or the frequency effect for each person, using the quotation marks because, strictly, these estimates are actually predictions, *Best Unbiased Linear Predictions (BLUPs)*, based on the estimates of the fixed and random effects.

5.11.1 Convergence problems

Mostly, our main concern, in working with mixed-effects models, is over what effects we should include, what model we should specify. But we should prepare for the fact sometimes happens that models *fail to converge*, which is to say, the model fitting algorithm fails to settle on some set of parameter estimates but has reached the limit in the number of iterations over which it has attempted to find a satisfactory set of estimates. In my experience, convergence problems do arise, typically, if one is analyzing categorical outcome data (e.g accuracy) where there may be not enough observations to distinguish satisfactory estimates given a quite complex hypothesized model. In other words, you might run into convergence problems but it will not happen often and only where you are already dealing with quite a complex situation. We take a look at this concern in more depth, in the next chapter (but see Jaeger, 2008, for a discussion of **Generalized Linear Mixed-effects models, GLMMs**; see Baayen, 2008, for a how-to guide). We need to start preparing our understanding now.

5.12 Fitting and evaluating Linear Mixed-effects models

Up to this point, we have discussed the empirical or conceptual reasons we should expect to take into account, in our model, the effects on the outcome due to systematic differences in the experimental variables, e.g., in stimulus word frequency frequency, or to random differences between participants or between stimuli. We can now think about how we should statistically evaluate the relative usefulness of these different fixed effects or random effects, where usefulness is judged in relation to our capacity to explain outcome variance, or to improve model fit to sample data. We shall take an approach that follows the approach set out by Baayen, Bates and others (Baayen et al., 2008; Bates et al., 2015; Matuschek et al., 2017).

In this approach, we shall look at the choices that psychology researchers have to make. Researchers using statistical models are always faced with choices. As we have seen, these choices begin even before we start to do analyses, as when we make decisions about dataset construction (Steegen et al., 2016). The need to make choices is always present for all the kinds of models we work with. This may not always be obvious because, for example, in using some data analysis software, researchers may rely on defaults with limited indication that that is what they are doing. Just because we are making choices does not mean we are operating subjectively in a non-scientific fashion, rather, provided we work in an appropriate mode of transparency or reflectiveness, it means we are working with an awareness of our options and the context for the data analysis (see discussion in Gelman & Hennig, 2017).

5.12.1 Model comparison approach

It is very common to see researchers using a process of model comparison to try to identify an account for their data in terms of estimates of fixed and random effects. A few key concepts are relevant to taking this approach effectively.

We will focus on building a series of models up to the most complex model supported by the data.

What does *model complexity* mean here? I am talking about something like the difference between a model including just main effects (simpler) and a model including both main effects and the interaction between the effects (more complex) or a model included just fixed effects (simpler) and a model including fixed effects as well as random effects (more complex).

Researchers may engage in comparing models to examine if one or more random effects should be included in their linear mixed-effects model. They may not be sure if they should include all random effects, that is, all random effects that could be included, given a range of grouping variables, like participant, class or stimulus, and given a range of possible effects, such as whether slopes or intercepts might vary.

Researchers may do model comparison to check if adding the effect of an experimental variable is justified. Maybe they are conducting an exploratory study in which they want to investigate if using some measurement variable helps to explain variation in the outcome. Perhaps they are

conducting an experimental study in which they want to test if the experimental manipulation, or the difference between conditions, has an impact on the outcome.

Across these scenarios, we can test if an effect *should be included* or if its inclusion in a model *is justified* by comparing models with versus without the term that corresponds to the effect. In some studies, researchers conduct model comparisons like this in order to obtain null hypothesis significance tests for the effects of the experimental variables. Typically, the model comparisons are focused on whether some measurement of model fit is or is not different when we do versus do not include the effect in question in the model.

5.12.1.1 Exercise – Model comparison questions

As our discussion progresses, I think it would be helpful to reflect on some of the questions that you may be asking yourself.

What about multiple comparisons?

You might well ask yourself: if we engage in a bunch of comparisons to check if we should or should not include a variable, isn't this just exploiting *researcher degrees of freedom*? Or, you might ask: if we are conducting multiple tests on the same data, aren't we running the risk of raising the Type I error (false positive) rate because we are doing *multiple comparisons*? I think these are good questions but, here, my task is to explain what people do, why they do it, and how it helps in your data analysis.

Is any model the best?

So you are looking at models with varying fixed effects (fitted using ML) or models with varying random effects (fitted using REML). How do you decide which model is better? Some researchers argue that trying to decide which model is better or best is inappropriate (see e.g. Gelman et al., 2014) – all models are wrong (that is, all are approximations to the data, given your assumptions), but some are more useful than others (explain or predict outcomes better, depending on your criteria, and the cost-benefit analysis). In this class, I will explain the model comparison process while acknowledging this point, because researchers are increasingly using model comparison techniques to evaluate the relative usefulness of different alternate models.

5.12.2 Model comparison using information criteria, AIC and BIC

You will often encounter, in the psychological research literature, Information Criteria statistics like BIC: they are understood within an approach: *Information-theoretic methods*. They are grounded in the insight that you have reality and then you have approximating models. The distance between a model and reality corresponds to the information lost when we use a model

to approximate reality. Information criteria – AIC or BIC – are estimates of *information loss*. The process of model selection aims to minimize information loss.

I will not discuss information criteria methods of model evaluation in detail here, because psychologists frequently use the Likelihood Ratio Test method (Meteyard & Davies, 2020), see following, but take a look at, e.g., Burnham and Anderson (2004) for a readable discussion, if you are interested. However, you should have some idea of what information criteria statistics (like AIC and BIC mean) because you will see these statistics in the outputs from model comparisons using the `anova()` function, which we shall review a bit later (Section @ref(LRT)).

In summary, Akaike showed you could estimate information loss in terms of the likelihood of the model given the data – Akaike Information Criteria, *AIC*:

$$AIC = -2\ln(l) + 2k \quad (5.5)$$

- where $-2\ln(l)$ is -2 times the log of the likelihood of the model given the data
- (l) the likelihood
 - Is proportional to the probability of observed data conditional on some hypothesis being true

You want a more likely model – less information loss, closer to reality – you want more negative or lower AIC. You can identify models that are more likely – closer to reality – with models with less wide errors, i.e. smaller residuals.

You could better approximate reality by including lots of predictors, specifying a more complex model. Models with more parameters may fit the data better but some of those effects may be spurious. Adding $+2k$ penalizes complexity, speaking crudely, and so helps us to focus on the more parsimonious less complex model that best fits the data.

Schwartz proposed an alternative estimate – Bayesian Information Criteria: *BIC*:

$$BIC = -2\ln(l) + k\ln(N) \quad (5.6)$$

- $-2\ln(l)$ -2 times the log of the likelihood of the model given the data
- $+k\ln(N)$ is the number of parameters in the model times the log of the sample size
- Thus the penalty for greater complexity is heavier in BIC

So AIC and BIC differ in the second term. A deeper difference is that AIC estimates information loss when the true model may not be among the models being considered while BIC assumes that the true model is within the set of models being considered.

At this point we just need to think about *Model selection and judgment* using AIC and BIC.

- Compare a simpler model: model 1, just main effects; model 2, main effects plus interactions
- If the more complex model better approximates reality then it will be more likely given the data
 - BIC or AIC will be closer to negative infinity: $-2\ln(l)$ will be larger
 - e.g. 10 is better than 1000, -1000 better than -10

AIC and BIC should move in the same direction. They usually will. AIC will tend to allow more complex models and that may be necessary when the researcher is engaged in a more exploratory study or wants more accurate predictions (that would be better supported by maximising the information going into the model). Using the BIC will tend to favour simpler models and that may be necessary when the researcher seeks models that replicate over the long run. Maybe a simpler model will less likely include predictors estimated because they are needed to fit noise or random outcome variation.

5.12.3 Model comparison using the Likelihood Ratio Test

Pinheiro & Bates (2000; see also Barr et al., 2013; Matuschek et al., 2017) recommend that models of varying predictor sets can be compared using Likelihood Ratio Test comparison (LRTs) where the simple model is *nested* inside the more complex model. This means, the predictors in the simpler model are a subset of the predictors in the more complex model. For example, you might have just main effects in the simpler model but both main and interaction effects in the more complex model. Or, in another example, you might have just random effects of subjects or items on intercepts in the simpler model but both random effects on intercepts and random effects on slopes of fixed effects in the more complex model.

When you compare models using the *Likelihood ratio test*, *LRT*, you are comparing alternate models of the *same data*.

Barr et al. (2013) note that we can compare models varying in the fixed effects (but constant in the random effects) or models varying in the random effects (but constant in the fixed effects) using LRTs. I have frequently reported model comparisons using the *Likelihood ratio test*, *LRT*. In part, this is for analytic reasons: I can compare simple and complex models getting multiple information criteria statistics for the models being compared in one function call, `anova([model1], [model2])`. In part, it is for social pragmatic reasons: the LRT comparison yields a significance p-value so that I can say, using the comparison, something like “The more complex model provided a significantly better fit to observation (LRT comparison, ...p = ...”.

In a *Likelihood ratio test*, *LRT*, the test statistic is the comparison of the likelihood of the simpler model with the more complex model. Fortunately for us, we can R to calculate the model likelihood and do the model comparison (Section @ref(anova)).

The comparison of models works by division: we divide the likelihood of the more complex model by the likelihood of the simpler model, calculating a *likelihood ratio*.

$$\chi^2 = 2\log \frac{\text{likelihood} - \text{complex}}{\text{likelihood} - \text{simple}} \quad (5.7)$$

The likelihood ratio is compared to the χ^2 distribution for a significance test. In this significance test, we assume the null hypothesis that the simpler model is adequate as an account of the outcome variance. We calculate the p-value for the significance test using a number for the degrees of freedom equal to the difference in the number of parameters of the models being compared.

5.13 Modeling steps recommendations

How should you proceed when you decide to use mixed-effects models? I think the answer to that question depends on whether you are doing a study that is **confirmatory** or **exploratory**.

In short, *if* you have **pre-registered** the design of your study and, as part of that registration, you recorded the hypotheses you plan to test, and the analysis method you plan to use to test your hypotheses, then the answer is simple: **fit the model you said you were going to use.**

But *if* you are doing an **exploratory** study, then you will need to make some choices, in part, depending on the nature of the sample you are working with, and other aspects of the research context, but it will help to **keep things simple**. These days, if you have not pre-registered your analysis plans, you are practically-speaking engaged in exploratory work.

In an exploratory study, I would keep things simple by comparing a series of models, fitted with different sets of predictor variables (fixed effects). (Note: if you are running mixed-effects models in R you cannot run `lmer()` models with just fixed effects.) What I do is this: for a dataset like the ML study data, where the data were collected using a repeated-measures design, so that all participants saw all stimuli, and both participants and stimuli were sampled (from the wider populations of readers or words), I would run a series of models so that the different models have varying sets of fixed effects but all models in the series have the same random effects: the random effects of subjects and items on intercepts.

In my experience, the estimates (and associated significance levels) associated with fixed effects can vary quite a bit depending on what other variables are included in the model. This has led me to take an approach where I am *not* varying too much how predictors are included in the model. As noted, this won't really apply if you are doing an *confirmatory* study in which you are obliged to include the manipulated variables. However, if you are doing something a

bit more *exploratory* than you might have to think about the kinds of predictors you include in your model, and how or when you include them.

In what order should you examine the usefulness of different sets of fixed effects? This is a difficult question to answer and the difficulty is one reason why I think we need to be cautious when we engage in model comparison to try to get to a model of our data. My advice would be to plan out in advance a sequence of model comparisons. You should begin with simpler models with fewer effects. You should begin with those effects whose impacts are well established and well understood by you. If there is a whole set of well established effects typically included in an analysis in the field in which you are working, it might be sensible to include all the effects in a single step. Then, I would use subsequent incremental steps to increase model complexity by adding effects that are theoretically justified, i.e., hypothesized, but which may be new, or may depend on the experimental manipulation you are testing out.

Having established a model with some set of sensible fixed effects (guided by information criteria or LRT statistics), I would then turn my attention to the random effects component of the model. As noted, we may expect to see random differences between subjects (and possibly between items) in both the level of average performance – random effects of subjects or items on intercepts – and in the slopes of fixed effects – random effects of subjects or items on slopes. What I do is this: for a dataset like ML’s, I examine firstly if both random effects of subjects and items on intercepts are required. I then check if random effects of subjects or items on slopes are *additionally* required in the model.

The distinction between *exploratory* and *confirmatory* studies breaks down, in my experience, when we start thinking about what random effects should be included in a model (see Barr et al., 2013; Matuschek et al., 2017, for an interesting discussion, and contrasting approaches).

5.13.1 Maximum Likelihood and Restricted Maximum Likelihood

Before we go any further, we need to briefly discuss one key choice that we face in working with mixed-effects models. This concerns the difference between Restricted Maximum Likelihood (REML) and Maximum Likelihood (ML) estimation methods. Both methods are iterative. The `lmer()` function has defaults, like any analysis function, so we often do not need to make the choice explicit. We do when we compare models that vary in fixed effects, or in random effects.

Restricted maximum likelihood in R: `REML=TRUE` is stated in the `lmer()` function call

- REML estimates the variance components while taking into account the loss of degrees of freedom resulting from the estimation of the fixed effects: *REML estimates vary if the fixed effects vary.*
- Therefore it is not recommended to compare the likelihood of models *varying in fixed effects* and *fitted using REML* (Pinheiro & Bates, 2000)

- The REML method is recommended for comparing the likelihood of models with *the same fixed effects* but *different random effects*
- REML is more accurate for random effects estimation.

Maximum likelihood in R: REML=FALSE is stated in the `lmer()` function call

- ML estimation methods can be used to fit models with varying fixed effects but the same random effects.
- ML estimation: a good place to start when building-up model complexity – adding parameters to an empty model.
- Pinheiro & Bates (2000) advise that the approach is anti-conservative (will sometimes indicate effects where there are none there) but Barr et al. (2013) argue that their analyses suggest that that is not so.

5.13.2 Comparing models of varying random effects but constant fixed effects

As noted, it is recommended (Pinheiro & Bates, 2000) that we compare models of varying random effects using Restricted Maximum Likelihood (REML) fitting. We might be comparing different models with different sets of random effects if we are in the process of working out whether our model should include random intercepts and random slopes, that is, model parameters accounting for random differences between subjects or between items in the average level of performance or in the slope of the fixed effects. I think it is sensible to build up model complexity in the random component so that we are working through a series of model comparisons, comparing more simple with more complex models where the more complex model includes the same terms as the simpler model but adds some more.

In analyzing the effect of frequency on log RT for the ML study data, we can examine whether the random effects of subjects or of items on intercepts are necessary. Then we can examine if we should take into account random effects of subjects on the slope of the fixed effect of frequency, in addition to the random effects on intercepts.

To begin with, we can look at a simpler model. We can run model with just the fixed effects of intercept and frequency, and the random effects of participants or items on intercepts only. We exclude the `(LgSUBLCD + ... | subjectID)` specification for the random effect of participants on the slope of the frequency `LgSUBLCD` effect. We use REML fitting, as follows:

```
ML.all.correct.lmer.REML.si <- lmer(logrt ~ LgSUBLCD +
                                         (1|subjectID) + (1|item_name),
                                         data = ML.all.correct, REML = TRUE)
```

```

summary(ML.all.correct.lmer.REML.si)

Linear mixed model fit by REML ['lmerMod']
Formula: logrt ~ LgSUBLCD + (1 | subjectID) + (1 | item_name)
Data: ML.all.correct

REML criterion at convergence: -9845.1

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-3.5339 -0.6375 -0.1567  0.4364  5.5851 

Random effects:
Groups   Name        Variance Std.Dev.
item_name (Intercept) 0.0003204 0.01790
subjectID (Intercept) 0.0032650 0.05714
Residual            0.0085285 0.09235
Number of obs: 5257, groups: item_name, 160; subjectID, 34

Fixed effects:
            Estimate Std. Error t value
(Intercept)  2.887697  0.013253 217.9
LgSUBLCD    -0.034390  0.002774 -12.4

Correlation of Fixed Effects:
  (Intr) 
LgSUBLCD -0.658

```

Notice:

- REML = TRUE the only change to the code, requiring the change in model fitting method
- Notice that I changed the model name to be able to distinguish the maximum likelihood from the restricted maximum likelihood model.

Following Baayen (2008) we can then run a series of models with just one random effect. Firstly, just the random effect of items on intercepts:

```

ML.all.correct.lmer.REML.i <- lmer(logrt ~
  LgSUBLCD + (1|item_name),

```

```

  data = ML.all.correct, REML = TRUE)

summary(ML.all.correct.lmer.REML.i)

Linear mixed model fit by REML ['lmerMod']
Formula: logrt ~ LgSUBLCD + (1 | item_name)
Data: ML.all.correct

REML criterion at convergence: -8337

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-3.7324 -0.6455 -0.1053  0.4944  4.8970 

Random effects:
 Groups   Name        Variance Std.Dev. 
item_name (Intercept) 0.0002364 0.01537 
Residual            0.0117640 0.10846 
Number of obs: 5257, groups: item_name, 160

Fixed effects:
            Estimate Std. Error t value
(Intercept) 2.886765  0.009047 319.07 
LgSUBLCD   -0.034206  0.002811 -12.17 

Correlation of Fixed Effects:
  (Intr) 
LgSUBLCD -0.977

```

Secondly, just the random effect of subjects on intercepts:

```

ML.all.correct.lmer.REML.s <- lmer(logrt ~
  LgSUBLCD + (1|subjectID),
  data = ML.all.correct, REML = TRUE)

summary(ML.all.correct.lmer.REML.s)

```

Linear mixed model fit by REML ['lmerMod']

```
Formula: logrt ~ LgSUBLCD + (1 | subjectID)
Data: ML.all.correct
```

```
REML criterion at convergence: -9786.3
```

```
Scaled residuals:
```

Min	1Q	Median	3Q	Max
-3.5843	-0.6443	-0.1589	0.4434	5.5266

```
Random effects:
```

Groups	Name	Variance	Std.Dev.
subjectID	(Intercept)	0.003275	0.05723
Residual		0.008837	0.09401

Number of obs: 5257, groups: subjectID, 34

```
Fixed effects:
```

	Estimate	Std. Error	t value
(Intercept)	2.885751	0.011561	249.60
LgSUBLCD	-0.033888	0.001897	-17.87

```
Correlation of Fixed Effects:
```

(Intr)
LgSUBLCD -0.517

If we now run Likelihood Ratio Test comparisons of these models, we are effectively examining if one of the random effects can be dispensed with: if its inclusion makes no difference to the likelihood of the model then it is not needed. Is the random effect of subjects on intercepts justified? Compare models, first, with `ML.all.correct.lmer.REML.si` versus without `ML.all.correct.lmer.REML.i` the random effect of subjects on intercepts. Then compare models with `ML.all.correct.lmer.REML.si` versus without `ML.all.correct.lmer.REML.s` the random effect of items on intercepts.

```
anova(ML.all.correct.lmer.REML.si, ML.all.correct.lmer.REML.i, refit = FALSE)
anova(ML.all.correct.lmer.REML.si, ML.all.correct.lmer.REML.s, refit = FALSE)
```

5.13.2.1 Code tip

We compare models using the `anova()` function. We can list as many models as we like for comparison.

Notice:

- `anova()` does the model comparison, for the models named in the list
- We specify `refit = FALSE` because otherwise R will compare ML fitted models – in practice, results are the same if compare REML vs. ML models – the refitting occurs by default to stop users from trying to compare REML models varying in fixed effects

Pinheiro & Bates (2000) advised that if one is fitting models with random effects the estimates are more accurate if the models are fitted using restricted maximum likelihood, that is achieved in the `lmer()` function call by adding the argument `REML=TRUE`. Pinheiro & Bates (2000; see e.g. pp.82-) recommended that if you compare models with the same fixed effects but varying random effects the models should be fitted using restricted maximum likelihood. We can do this for the foregoing series of models but what you will notice is that when we run the `anova()` function call, without the `refit = FALSE` argument, we get the warning `refitting model(s) with ML (instead of REML)`. Why? it's basically about stopping users from comparing REML-fitted models with varying fixed effects see e.g.

<https://stat.ethz.ch/pipermail/r-sig-mixed-models/2013q4/021040.html>

As Ben Bolker points out, in ...

<http://stackoverflow.com/questions/22892063/do-i-need-to-set-refit-false-when-testing-for-random-effects-in-lmer-models-wi>

... analyses of simulated data analyses suggest that it does not make much difference whether we use REML or ML when we are comparing models with the same fixed effects but varying random effects. But it does matter *very much* that we fit models using ML when we are comparing models with the same random effects but differing fixed effects.

When we run the `anova()` function call, it can be seen that the random effects of subjects on intercepts is required.

```
anova(ML.all.correct.lmer.REML.si, ML.all.correct.lmer.REML.i, refit = FALSE)

Data: ML.all.correct
Models:
ML.all.correct.lmer.REML.i: logrt ~ LgSUBLCD + (1 | item_name)
ML.all.correct.lmer.REML.si: logrt ~ LgSUBLCD + (1 | subjectID) + (1 | item_name)
      npar      AIC      BIC logLik deviance   Chisq Df
ML.all.correct.lmer.REML.i     4 -8329.0 -8302.7 4168.5   -8337.0
ML.all.correct.lmer.REML.si     5 -9835.1 -9802.3 4922.6   -9845.1 1508.1   1
                           Pr(>Chisq)
ML.all.correct.lmer.REML.i
ML.all.correct.lmer.REML.si  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If you look at the results of the model comparison then you should notice:

- The `ML.all.correct.lmer.REML.si` model is more complex than the `ML.all.correct.lmer.REML.i` model
 - `ML.all.correct.lmer.REML.si` includes `LgSUBLCD + (1 | subjectID) + (1 | item_name)`
 - `ML.all.correct.lmer.REML.i` includes `LgSUBLCD + (1 | item_name)`
- The more complex model `ML.all.correct.lmer.REML.si` has AIC (-9835.1) and BIC (-9802.3) numbers that are larger or more negative, and has a likelihood (4922.6) that is larger
- Than the simpler model `ML.all.correct.lmer.REML.i` which has AIC (-8329.0) and BIC (-8302.7) and likelihood (4168.5).
- The $\chi^2 = 1508.1$ statistic, on 1 Df has a p-value of `Pr(>Chisq) <2.2e-16`

You can say that the comparison of a model `ML.all.correct.lmer.REML.si` with versus a model `ML.all.correct.lmer.REML.i` without the random effect of participants on intercepts shows that the inclusion of the random effect of participants on intercepts is *warranted by a significant difference in model fit*. Notice I highlight the language you can use in your reporting.

The second model comparison shows that the random effects of items on intercepts is also justified.

```
anova(ML.all.correct.lmer.REML.si, ML.all.correct.lmer.REML.s, refit = FALSE)
```

```
Data: ML.all.correct
Models:
ML.all.correct.lmer.REML.s: logrt ~ LgSUBLCD + (1 | subjectID)
ML.all.correct.lmer.REML.si: logrt ~ LgSUBLCD + (1 | subjectID) + (1 | item_name)
      npar      AIC      BIC logLik deviance Chisq Df
ML.all.correct.lmer.REML.s     4 -9778.3 -9752.0 4893.2 -9786.3
ML.all.correct.lmer.REML.si     5 -9835.1 -9802.3 4922.6 -9845.1 58.825  1
                                 Pr(>Chisq)
ML.all.correct.lmer.REML.s
ML.all.correct.lmer.REML.si  1.723e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If you look at the results of the model comparison then you should notice:

- The `ML.all.correct.lmer.REML.si` model is more complex than the `ML.all.correct.lmer.REML.s` model
 - `ML.all.correct.lmer.REML.si` includes `LgSUBLCD + (1 | subjectID) + (1 | item_name)`
 - `ML.all.correct.lmer.REML.i` includes `LgSUBLCD + (1 | subjectID)`

- The more complex model `ML.all.correct.lmer.REML.si` has AIC (-9835.1) and BIC (-9802.3) numbers that are larger or more negative, and has a likelihood (4922.6) that is larger
- Than the simpler model `ML.all.correct.lmer.REML.s` which has AIC (-9778.3) and BIC (-9752.0) and likelihood (4893.2).
- The $\chi^2 = 58.825$ statistic, on 1 Df has a p-value of $\text{Pr}(>\text{Chisq}) 1.723e-14$

You can say that the comparison of a model `ML.all.correct.lmer.REML.si` with versus a model `ML.all.correct.lmer.REML.s` without the random effect of items on intercepts shows that the inclusion of the random effect of items on intercepts is warranted by a significant difference in model fit.

I would conclude that both random effects of subjects and items on intercepts are required. We can draw this conclusion because the difference between the model including just the random effect of items on intercepts `anova-ML-all-correct-lmer-REML-i`, or the model including just the random effect of subjects on intercepts `anova-ML-all-correct-lmer-REML-s`, compared to the model including both the random effect of items on intercepts and of subjects on intercepts `anova-ML-all-correct-lmer-REML` is significant. This tells us that the absence of the term accounting for the random effect of subjects on intercepts is associated with a significant decrease in model fit to data, in model likelihood.

5.13.3 Evaluating random effects of subjects or items on slopes

We should next consider whether it is justified or warranted to include in our model a term capturing the random effect of participants in the slope of the frequency effect. Of course, we have seen how in theory it seems to make sense to include this effect. Some researchers might ask: does the inclusion of the random effect seem justified by improved model fit to data?

I should acknowledge, here, that there is an on-going discussion over what random effects should be included in mixed-effects models (see Meteyard & Davies, 2020, for a discussion). The discussion can be seen from a number of different perspectives. Key articles include those published by Bates et al. (2015), Barr et al. (2013), and Matuschek et al. (2017). You could be advised that a mixed-effects model should include all random effects that make sense a priori, so, the random effects of participants on intercepts and on the slopes of all fixed effects that are in your model (variances and covariances) as well as all the random effects of items on intercepts and on slopes. This is characterized as the *keep it maximal* approach, associated with Barr et al. (2013) though the discussion in that article is more nuanced than this sounds. Or, you could be advised that a mixed-effects model should only include those random effects that appear to be justified or warranted by their usefulness in accounting for the data. In practice, this may mean, you should include only those random effects that appear justified by improved model fit to data, as indicated by a model comparison (see e.g. Bates et al., 2015; Matuschek et al., 2017).

I think, in practice, that maximal models can run into convergence problems. This means that many researchers adopt an approach which you could call:

Maximum justifiable Fit a model including all the random effects that make sense, that are justified by improved model fit to data (given a significance test, model comparison) for a model that actually converges.

I think it makes sense to account for all the potential random variation that you can predict should have an impact. This is not really a matter of whether a more complex model fits the data better or (as Barr et al., 2013, discuss) more complex models with more comprehensive random effects appear to control the Type I (false positive) error rate better (but, as Matuschek et al., 2017) argue, at the cost of increasing the Type II (false negative) error rate. I would argue that it is more because you should try to account for what you observe with the information available to you.

In practice you may have insufficient data or inadequate measures to enable you to fit a model that converges with all the random effects, or to enable you to fit a model that converges that can estimate what may, in fact, be very small random effects variances or covariances. But this is why some researchers are moving to adopt *Bayesian* mixed-effects modeling methods, as discussed by the developmental Psychologist, Michael Frank, for example, here:

<https://babieslearninglanguage.blogspot.com/2018/02/mixed-effects-models-is-it-time-to-go.html>

And as exemplified by my work here:

<https://peerj.com/articles/9511/>

But this discussion raises a question.

Random slopes of what? In general, and simplifying things a bit, if an effect is manipulated within-units then both random effects of those units on intercepts and slopes should be examined but if it is manipulated between units then only random effects of those units on intercepts may be required.

Examine the utility of random effects by comparing models with the same fixed effects but varying random effects. You can add a fixed effect term inside the specification of random effects to examine random slopes, as we saw earlier in this chapter.

```
ML.all.correct.lmer.REML.slopes <- lmer(logrt ~ LgSUBLCD +  
                                         (LgSUBLCD + 1 | subjectID) + (1 | item_name),  
                                         data = ML.all.correct, REML = TRUE)
```

- $(\text{LgSUBLCD} + 1 | \text{subjectID})$ we specify a random effect of subjects on intercepts and on the slope of the frequency effects
- We do not specify – it happens by default – the estimation of the covariance of random differences among subjects in intercepts and random differences among subjects in the slope of the frequency effect

And as before, we can use `anova()` to check whether the increase in model complexity associated with the addition of random slopes terms is justified by an increase in model fit to data.

```
anova(ML.all.correct.lmer.REML.si, ML.all.correct.lmer.REML.slopes, refit = FALSE)
```

```
Data: ML.all.correct
Models:
ML.all.correct.lmer.REML.si: logrt ~ LgSUBLCD + (1 | subjectID) + (1 | item_name)
ML.all.correct.lmer.REML.slopes: logrt ~ LgSUBLCD + (LgSUBLCD + 1 | subjectID) + (1 | item_name)
                                 npar      AIC      BIC logLik deviance Chisq Df
ML.all.correct.lmer.REML.si      5 -9835.1 -9802.3 4922.6   -9845.1
ML.all.correct.lmer.REML.slopes    7 -9854.1 -9808.1 4934.0   -9868.1 22.934  2
Pr(>Chisq)

ML.all.correct.lmer.REML.si
ML.all.correct.lmer.REML.slopes 1.047e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Inspection of the results shows us that, in fact, adjusting the model to include random effects of subjects in the slopes of the fixed effects due to differences in word frequency *does* improve model fit to data. We can say that the inclusion of the random effect is *warranted by improved model fit to data* ($\chi^2(1df) = 22.9, p < .001$) The model with the random slopes is a better model of the data ML observed.

5.13.4 Effects estimates and *significance* or p-values

If you look at the fixed effects summary, you can see that we do not get p-values by default. Bates and colleagues (see online lmer discussions, which you will find if you search for “Bates, Bolker, lmer, p-values”) have basically decided that because it is not really sensible to estimate the residual degrees of freedom for a model in terms of the number of observations, given that the number of observations concerns one level of a multilevel dataset that might also include some number of subjects, some number of items, then one cannot (without such degrees of freedom) accurately calculate p-values to go with the t-tests on the coefficients estimates, therefore they don’t.

While this makes sense to me (see comments earlier on Bayesian methods), Psychologists will need p-values. This is now relatively easy. We can run mixed-effects models with p-values from significance tests on the estimates of the fixed effects coefficients using the library(lmerTest).

```
library(lmerTest)

ML.all.correct.lmer.REML.slopes <- lmer(logrt ~ LgSUBLCD +
                                         (LgSUBLCD + 1|subjectID) + (1|item_name),
                                         data = ML.all.correct, REML = TRUE)

summary(ML.all.correct.lmer.REML.slopes)

Linear mixed model fit by REML. t-tests use Satterthwaite's method [lmerModLmerTest]
Formula: logrt ~ LgSUBLCD + (LgSUBLCD + 1 | subjectID) + (1 | item_name)
Data: ML.all.correct

REML criterion at convergence: -9868.1

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-3.6307 -0.6324 -0.1483  0.4340  5.6132 

Random effects:
Groups      Name        Variance Std.Dev. Corr
item_name   (Intercept) 0.0003268 0.01808
subjectID   (Intercept) 0.0054212 0.07363
           LgSUBLCD   0.0002005 0.01416 -0.63
Residual            0.0084333 0.09183
Number of obs: 5257, groups: item_name, 160; subjectID, 34

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)    
(Intercept) 2.887997  0.015479 47.782839 186.577 < 2e-16 ***
LgSUBLCD   -0.034471  0.003693 60.338787 -9.333 2.59e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
```

(Intr)
LgSUBLCD -0.764

Basically, the call to access the `lmerTest` library ensures that when we run the `lmer()` function we get a calculation of an approximation to the denominator degrees of freedom that enables the calculation of the p-value for the t-test for the fixed effects coefficient. An alternative, as I have noted (Section @ref(LRT)) is to compare models with versus without the effect of interest.

5.13.4.1 Exercises

It will be useful for you to examine model comparisons with a different set of models for the same data.

You could try to run a series of models in which the fixed effects variable is something different, for example, the effect of word `Length` or the effect of orthographic neighbourhood size `Ortho_N`.

I would consider the model comparisons in the sequence shown in the foregoing, one pair of models at a time, to keep it simple. When you look at the model comparison, ask: is the difference between the models a piece of complexity (an effect) whose inclusion in the more complex model is justified or warranted by improved model fit to data?

5.14 Reporting results

5.14.1 Reporting comparisons of ML and REML models

If you look at an example mixed-effects analysis report, like Davies et al. (2013), you can see a few features of the reporting:

1. Because it was an exploratory study, I started by reporting the comparison of models varying in fixed effects.
 - I explain what predictors are included in each model.
 - I explain how I make decisions about which model to select.
2. I then go on to discuss the comparison of models varying in random effects.

We stepped through a series of models. Firstly, assuming the same random effects of subjects and items on intercepts, we compared models differing in fixed effects: a model (model 1) with just initialstress factors; a model (model 2) with initialstress factors plus linear effects due to the orthographic.form, frequency, semantic, and bigram.frequency factors; and lastly a model (model 3) with the same factors as model 2 but adding restricted cubic splines for the frequency and orthographic.form factors to examine the evidence for the presence of curvilinear effects of frequency and length (the orthographic.form factor loads heavily on length).

Notice also that I try to standardize the language and structure of the paragraphs – that kind of repetition or rhythm helps the reader, I think, by making what is not repeated – the model specifications – more apparent. Your style may differ, however, and that's alright.

We evaluated whether the inclusion of random effects was necessary in the final model (model 3) using LRT comparisons between models with the same fixed effects structure but differing random effects. Here, following Pinheiro & Bates (2000; see, also, Baayen, 2008), models were fitted using the REML=TRUE setting in lmer. We compared models that included: (i.) both random effects of subjects and items, as specified for model 3; (ii.) just the random effect of subjects; (iii.) just the random effect of items.

I want you to notice something more, concerning the predictors included in each different model:

1. I do not include predictors one at a time, I include predictors in sets
 - for the Davies et al. (2013) data set, I include first phonetic coding variables then psycholinguistic variables
 - I include linear effects then additional terms allowing the effects to be curvilinear

Finally, you can see that I report model comparisons in terms of Likelihood ratio test comparisons, firstly, considering the basis for selecting one model out of the models varying in fixed effects:

Comparing models 1 and 2, models with initialstress factors but differing in whether they did or did not include key psycholinguistic factors like orthographic.form, the LRT statistic was significant ($\chi^2 = 1,007, 4df, p = 2 * 10^{-16}$). Comparing models 2 and 3, i.e. models with initialstress and key psycholinguistic components but differing in whether they did or did not use restricted cubic splines to fit the orthographic.form and frequency effects, the LRT statistic was significant ($\chi^2 = 23, 2df, p = 1 * 10^{-5}$).

then I report the selection of models varying in random effects:

We compared models that included: (i.) both random effects of subjects and items, as specified for model 3; (ii.) just the random effect of subjects; (iii.) just the random effect of items. The difference between models (i.) and (ii.) was significant ($\chi^2 = 185, 1df, p = 2 * 10^{-16}$) indicating inclusion of an item effect was justified. The difference between models (i.) and (iii.) was significant ($\chi^2 = 17, 388, 1df, p = 2 * 10^{-16}$) indicating inclusion of a subject effect was justified.

5.14.2 Reporting the model: summary

You should include in your report:

- A summary of fixed effects – just like in linear models, with coefficient estimates, standard errors, t and p (if you use it)
- Report random effects variance and covariance (if applicable)
- In text, report likelihood comparisons

I like to present the final model summary in a table that is structured like a multiple regression model summary table showing the random and the fixed effects.

5.15 Summary

We examined another example of data from a repeated measures design study, this time, from a study involving adults responding to the lexical decision task, the ML study dataset.

We explored in more depth why linear mixed-effects models are more effective than other kinds of models when we are analyzing data with multilevel or crossed random effects structure. We discussed the critical ideas: pooling, and shrinkage. And we looked at how mixed-effects models employ partial-pooling so as to be more effective than alternative approaches dependent on complete pooling or no pooling estimates. Mixed-effects models work better because they use both information from the whole dataset and information about each group (item or participant). This ensures that model estimates take into account random differences but are regularized so that they are not dominated by less reliable group-level information.

We considered, briefly, how mixed-effects models are estimated. Then we examined, in depth, how mixed-effects models are fitted, compared and evaluated. The model comparison approach was set out, and we looked at both practical steps and at some of the tricky questions that, in practice, psychologists are learning to deal with.

We discussed how to compare models with varying random or fixed effects. We focused, especially, on the comparison of models with varying random effects. Methods for model comparison, including the use of information criteria and the Likelihood Ratio Test, were considered.

We discussed p-values, questions about calculating them, and a simple method for getting them when we need to report significance tests. The final discussion concerned how mixed-effects models should be reported.

5.15.1 Useful functions

We used two functions to fit and evaluate mixed-effects models.

- We used `lmer()` to fit a mixed-effects model
- We used `anova()` to compare two or more models using AIC, BIC and the Likelihood Ratio Test

We used the `lmerTest` library to get significance tests for coefficient estimates of fixed effects.

5.16 R code and data file access for the class

Activities in the class that goes with this chapter are associated with the following data file and .R code file:

- `402-03-mixed-workbook.R`
- `subjects.behaviour.words-310114.csv`

The data and .R workbook files can be downloaded as part of the .zip folder labelled **PSYC402-03-mixed-resources**.

You can download the folder from the Moodle section corresponding to this chapter:

<https://modules.lancaster.ac.uk/course/view.php?id=34085#section-13>

Or you can download the folder directly from:

<https://modules.lancaster.ac.uk/mod/resource/view.php?id=1801384>

Run the code in the .R file to reproduce the results presented in this chapter and in the slides.

5.17 References

5.17.1 Recommended reading

Snijders and Bosker (2012) present a helpful overview of multilevel modelling. Baayen et al. (2008; see, also, Barr et al., 2013; Judd et al., 2012) discuss mixed-effects models with crossed random effects.

I wrote a tutorial article on mixed-effects models with Lotte Meteyard. We discuss how important the approach now is for psychological science, what researchers worry about when they use it, and what they should do and report when they use the method/

Meteyard, L., & Davies, R.A.I. (2020). Best practice guidance for linear mixed-effects models in psychological science, *Journal of Memory and Language*, 112, 104092, <https://doi.org/10.1016/j.jml.2020.104092>

5.17.2 A very useful FAQ

Can be found here:

<https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html>

5.17.3 References list

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R.* CUP.

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, 390-412.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39, 445– 459. <http://dx.doi.org/10.3758/BF03193014>

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

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33, 261–304.

- Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments & Computers*, 35, 116–124. <http://dx.doi.org/10.3758/BF03195503>
- Gelman, A. (2014). The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective. *Journal of Management*, DOI: 0149206314525208.
- Gelman, A., & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society*, 180, 967-1033.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511790942>
- Howell, D. C. (2004). *Fundamental statistics for the behavioural sciences 5th ed*. Belmont: Thomson Brookes/Cole.
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103, 54.
- Masterson, J., & Hayes, M. (2007). Development and data for UK versions of an author and title recognition test for adults. *Journal of Research in Reading*, 30, 212–219. <http://dx.doi.org/10.1111/j.1467-9817.2006 .00320.x>
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305-315.
- Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer-Verlag.
- 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, 416-426.
- Snijders, T.A., & Bosker, R.J. (2012). *Multilevel analysis (2nd Edition)*. London, UK: Sage.
- Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11, 702-712.
- Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *TOWRE Test of word reading efficiency*. Austin, TX: Pro-ed.

6 Introduction to Generalized Linear Mixed-effects Models

We have been discussing how we can use Linear Mixed-effects models to analyze multilevel structured data, the kind of data that we commonly acquire in experimental psychological studies, for example, when our studies have repeated measures designs. The use of Linear Mixed-effects models is appropriate where the outcome variable is a continuous numeric variable like reaction time. In this chapter, we extend our understanding and skills by moving to examine data where the outcome variable is categorical: this is a context that requires the use of **Generalized Linear Mixed-effects Models (GLMMs)**.

We will begin by looking at the motivations for using GLMMs. We will then look at a practical example of a GLMM analysis, in an exploration in which we shall reveal some of the challenges that can arise in such work. The R code to do the modeling is very similar to the code we have used before. The way we can understand the models is also similar but *with one critical difference*. We start to understand that difference here.

6.1 The key idea to get us started

Categorical outcomes cannot be analyzed using linear models (ANOVA or t-test or linear models or linear mixed-effects models) without having to make some important compromises.

You need to do something about the categorical nature of the outcome.

6.2 Targets

In this chapter, we look at Generalized Linear Mixed-effects Models (GLMMs): we can use these models to analyze outcome variables of different kinds, including outcome variables like response accuracy that are coded using discrete categories (e.g. correct vs. incorrect). Our aims are to:

1. Understand the reasons for using GLMMs when we analyze discrete outcome variables
2. Recognize the limitations of alternative methods for analyzing such outcomes

3. Practice running GLMMs with varying random effects structures
4. Practice reporting the results of GLMMs, including through the use of model plots

6.3 Study guide

1. Play with the .R file used to create examples for the lecture.
2. Edit example code to create alternate visualizations of variable distributions and of the relationships between critical variables.
3. Run generalized linear mixed-effects models of demonstration data.
4. Run generalized linear mixed-effects models of alternate data sets.

You will see that in the references list at the end, I have recommended some papers that I think provide particularly useful or readable introductions to GLMMs.

6.4 Motivations

6.4.1 Our focus is on the analysis of categorical outcome variables

We often need to analyze outcome or dependent variables which comprise observations of responses that are discrete or categorical. We need to learn to recognize research contexts that require GLMMs. Categorical outcome variables can include any of the following.

- The accuracy of responses: is a response correct or incorrect?
- The membership of one group out of two possible groups: e.g., is a participant impaired or unimpaired; e.g., was a recorded eye movement, a fixation, to the left or to the right visual field?
- The membership of one group out of multiple possible groups: e.g., is a participant a member of one out of some number of groups, say, a member of a religious or ethnic group; e.g., is an incorrect response one out of some number of possible error types?
- Responses that can be coded in terms of ordered categories: e.g., a response on a (Likert) ratings scale.
- Outcomes like the frequency of occurrence of an event, e.g., how many arrests are made at a particular city location?

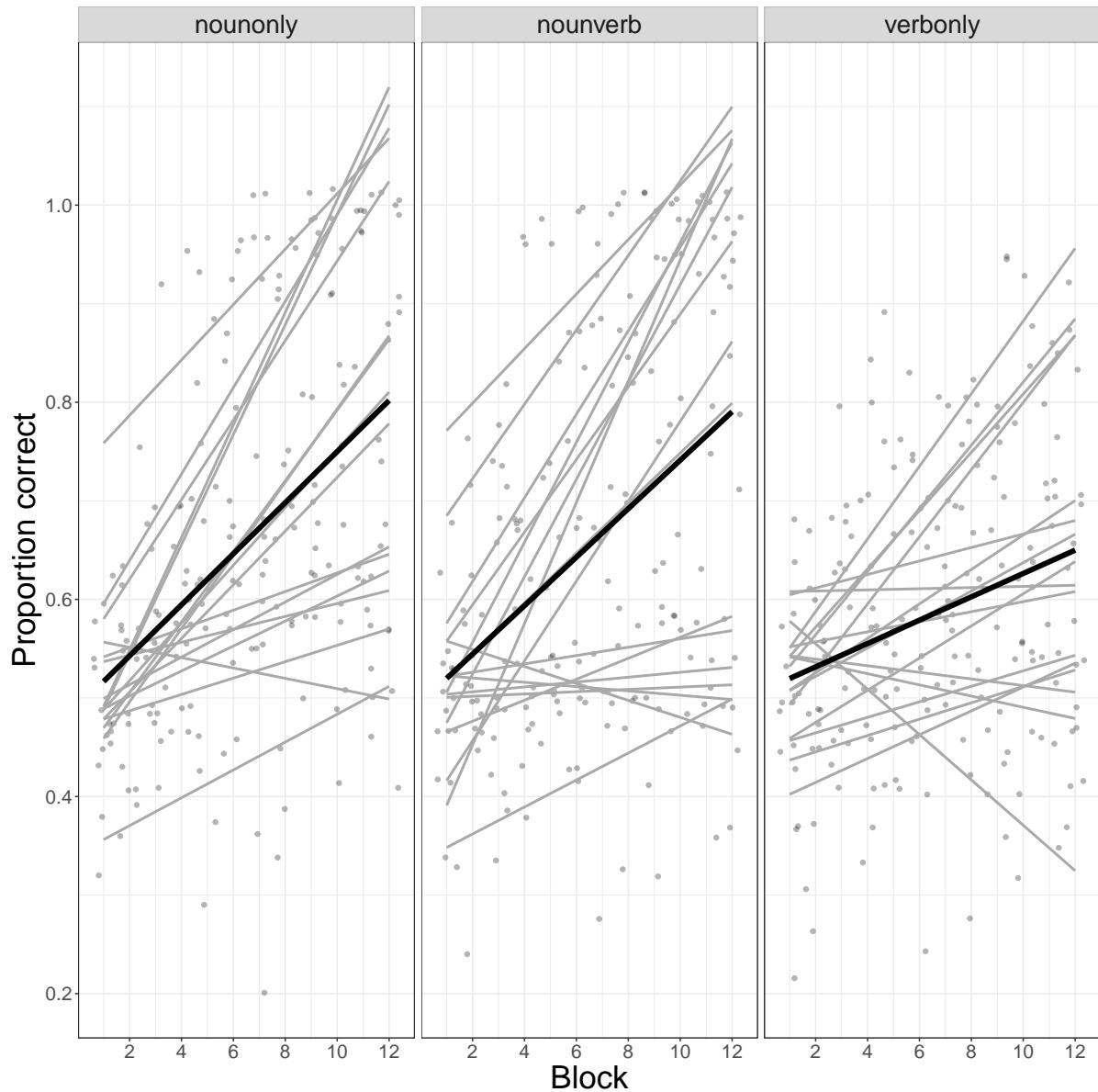


Figure 6.1: Monaghan et al. (2015) artificial word learning study: plot showing the proportion of responses correct for each participant, in each of 12 blocks of 24 learning trials, in each learning condition; each grey line shows the linear model prediction of the proportion correct, for each person, by learning block, in each condition; black lines show the average prediction of the proportion correct, by learning block, in each condition. The position of points has been jittered.

In this chapter, we will focus on **accuracy data**: where the outcome variable consists of responses observed in a behavioural task, the accuracy of responses was recorded, and responses could either be correct or incorrect. The accuracy of a response is, here, coded under a **binary** or dichotomous classification though we can imagine situations when a response is coded in multiple different ways. Other binary outcome variables may include, for example, eye fixations when a fixation could either be to a region of interest or outside the region of interest.

Those interested in analyzing outcome data from ratings scales, that is, ordered categorical outcome variables, often called ordinal data, may wish to read about ordinal regression analyses, which you can do in R using functions from the *ordinal* library.

https://cran.r-project.org/web/packages/ordinal/vignettes/clmm2_tutorial.pdf

Those interested in analyzing outcome data composed of counts may wish to read about poisson regression analyses in Gelman and Hill (2007).

It will be apparent in our discussion that researchers have used, and will continue to use, a number of traditional methods to analyze categorical outcome variables when really they should be using GLMMs. We will talk about these alternatives first, so that you recognize what is being done when you read research articles. Critically, we will discuss the limitations of such methods because these limitations explain why we bother to learn about GLMMs.

6.4.2 Recognize the limitations of alternative methods for analyzing response accuracy

If you want to analyze data from a study where responses can be either correct or incorrect but not both (and not anything else), then your outcome variable is categorical, and your analysis approach ought to respect that. However, if you read enough psychological research articles then you will see many reports of data analyses in which the researchers collected data on the accuracy of responses but then present the results of analyses that *ignored* the binary or dichotomous nature of accuracy. We often see response accuracy analyzed using an approach that looks something like the following:

- The accuracy of responses (correct vs. incorrect) is counted, e.g., as the number of correct responses or the number of errors;
- The percentage, or the proportion, of responses that are correct or incorrect is calculated, for each participant, for each level of each experimental condition or factor;
- The percentage or proportion values are then entered as the outcome or dependent variable in ANOVA or t-test or linear model (multiple regression) analyses of response accuracy.

You will see many reports of ANOVA or t-test or linear model analyses of accuracy. Why can't we follow these examples, and save ourselves the effort of learning how to use GLMMs? The reason is that these analyses are, at best, approximations to more appropriate methods. Their results can be expected to be questionable, or misleading, for reasons that we discuss next.

6.4.2.1 Accuracy is bounded between 1 and 0, linear model predictions or confidence intervals are not

To illustrate the problems associated with using traditional analysis methods (like ANOVA or multiple regression), when working with accuracy as an outcome, we start by looking at data from an artificial vocabulary learning study (reported by Monaghan, Mattock, Davies, & Smith, 2015). Monaghan et al. (2015) recorded responses made by participants to stimuli in a test where the response was correct (coded 1) or incorrect (coded 0). In our study, we directly compared learning of noun-object pairings, verb-motion pairings, and learning of both noun and verb pairings simultaneously, using a cross-situational learning task. (Those interested in this dataset can read more about it at the online repository associated with this chapter.) The data will have a multilevel structure because you will have multiple responses recorded for each person, and for each stimulus. But what concerns us is that if you attempt to use a linear model to analyze the effects of the experimental variables then you will see some paradoxical results that are easily demonstrated.

Let's imagine that we wish to estimate the effects of experimental variables like learning condition: learning trial block (1-12); or vocabulary condition (noun-only, noun-verb, verb-only). We can calculate the proportion of responses correct made by each person for each condition and learning trial block. We can then plot the regression best fit lines indicating how proportion of responses correct varies by person and condition. Figure @ref(fig:word-learning-lm-per-subject-for-report) shows the results. **Look at where the best fit lines go.**

Figure @ref(fig:word-learning-lm-per-subject-for-report) shows how variation in the outcome, here, the proportion of responses that are correct, is bounded between the y-axis limits of 0 and 1 while the best fit lines exceed those limits. Clearly, if you consider the accuracy of a person's responses in any set of trials, for any condition in an experiment, the proportion of responses that can be correct can vary only between 0 (no responses are correct) and 1 (all responses are correct). There is no inbuilt or intrinsic limits to the proportion of responses that a *linear model* can predict would be correct. According to linear model predictions, if you follow the best fit lines in Figure @ref(fig:word-learning-lm-per-subject-for-report), there are conditions, or there are participants, in which the proportion of a person's responses that could be correct will be *greater than 1*. That is impossible.

6.4.2.2 ANOVA or regression require the assumption of homogeneity of variance but for binary outcomes like accuracy the variance is proportional to the mean

The other fundamental problem with using analysis approaches like ANOVA or regression to analyze categorical outcomes like accuracy is that we cannot assume that the variance in accuracy of responses will be homogenous across different experimental conditions.

The logic of the problem can be set out as follows ...

- Given a binary outcome, e.g., where the response is correct or incorrect,
- For every trial, there is a probability p that the response is correct.
- The variance of the proportion of trials (per condition) with correct responses is dependent on p and greater when $p \sim .5$, the probability that a response will be correct.

Jaeger (2008; p. 3) then explains the problem like this. If the probability of a binomially distributed outcome like response accuracy differs between two conditions (call them conditions 1 and 2), the variances will only be identical if p_1 (the proportion of correct responses in condition 1) and p_2 (the proportion of correct responses in condition 1) are equally distant from 0.5 (e.g. $p_1 = .4$ and $p_2 = .6$). The bigger the difference in distance from 0.5, comparing the conditions, the less similar the variances will be.

Differences close to 0.5 will matter less than differences closer to 0 or 1. Even if p_1 and p_2 are unequally distant from 0.5, as long as they are close to 0.5, the variances of the sample proportions will be similar. Sample proportions between 0.3 and 0.7 are considered close enough to 0.5 to assume homogeneous variances (Agresti, 2002).

Unfortunately, we usually cannot determine a priori the range of sample proportions in our experiment. In general, variances in two binomially distributed conditions will not be homogeneous but, as you will recall, in both ANOVA and regression analysis, we assume homogeneity of variance in the outcome variable when we compare the effect of differences (in the mean outcome) between the different levels of a factor. This means that if we design a study in which the outcome variable is the accuracy of responses in different experimental conditions and we plan to use ANOVA or regression to estimate the effect of variation in experimental conditions on response accuracy then *unless we get lucky* our estimation of the experimental effect will take place under circumstances in which the application of the analysis method (ANOVA or regression) and thus the analysis results will be invalid.

6.4.2.3 Summary: Recognize the limitations of traditional methods for analyzing response accuracy

The application of traditional (parametric) analysis methods like ANOVA or regression to categorical outcome variables like accuracy is very common in the psychological literature. The problem is that these approaches can give us misleading results.

Traditionally, researchers have recognized the limitations attached to using methods like ANOVA or regression to analyze categorical outcomes like accuracy and have applied remedies, transforming the outcome variables, e.g. the arcsine root transformation, to render them ‘more normal’. However, as Jaeger (2008) demonstrates, the remedies like the arcsine transformation that have traditionally been applied are often not likely to succeed.

6.5 Understanding the Generalized part of the Generalized Linear Mixed-effects Models in practical terms

What we need, then, is a method that allows us to analyze categorical outcomes. We find the appropriate method in Generalized Linear Models, and in Generalized Linear Mixed-effects Models for repeated measures or multilevel structured data. We can understand these methods, as their name suggests, as *generalizations* of linear models or linear mixed-effects models, generalizations that allow for the categorical nature of some outcome data. You can understand how Generalized Linear Mixed-effects Models work by seeing them as analyses of categorical outcome data like accuracy where the outcome variable is transformed, as I explain next (see Baguley, 2012, for a nice clear explanation).

Our problems begin with the need to estimate effects on a bounded outcome like accuracy with a linear model which, as we have seen, will yield unbounded predictions.

The logistic transformation takes p the probability of an event with two possible outcomes, and turns it into a **logit**: the natural logarithm of the odds of the event. The effect of this transformation is to turn a discrete binary bounded outcome into a continuous unbounded outcome.

- Transforming a probability to odds $o = \frac{p}{1-p}$ is a partial solution.
- Odds are, for example, the ratio of the probability of the occurrence of an event compared to the probability of the non-occurrence of an event, or, in terms of a response accuracy variable, the ratio of the probability of the response being correct compared to the probability of the response being incorrect.
- And odds are continuous numeric quantities that are scaled from zero to infinity.
 - You can see how this works if you run the calculations using the equation $o = \frac{p}{1-p}$ in R as `odds <- p/(1-p)`, replacing p with various numbers (e.g. $p = 0.1, 0.01, 0.001$).
- We can then use the (natural) logarithm of the odds $\text{logit} = \ln \frac{p}{1-p}$ because using the logarithm removes the boundary at zero because log odds ranges from negative to positive infinity.
 - You can see how this works if you run the calculations using the equation $\text{logit} = \ln \frac{p}{1-p}$ in R as `logit <- log(p/(1-p))`, replacing p with smaller and smaller numbers (e.g. $p = 0.1, 0.01, 0.001$) gets you increasing negative log odds.

When we model the log odds (logit) that a response will be correct, the model is called a *logistic regression* or logistic model. We can think of logistic models as working like linear models with log-odds outcomes.

$$\ln \frac{p}{1-p} = \text{logit} p = \beta_0 + \beta_1 X_1 \dots \quad (6.1)$$

- We can describe the predicted log odds of a response of one type as the sum of the effects
- log odds range from negative to positive infinity (logit of 0 corresponds to proportion of .5)

Baguley (2012) notes that it is advantageous that odds and probabilities are both directly interpretable. We are used to seeing and thinking in everyday life about the chances that some event will occur.

6.6 The data we will work with: the Ricketts word learning study

We will be working with data collected for a study investigating word learning in children, reported by Ricketts, Dawson, and Davies (2021). You will see that the study design has both a repeated measures aspect because each child is asked to respond to multiple stimuli, and a longitudinal aspect because responses are recorded at two time points. Because responses were observed to multiple stimuli for each child, and because responses were recorded at multiple time points, the data have a multilevel structure. These features require the use of mixed-effects models for analysis.

We will see, also, that the study involves the factorial manipulation of learning conditions. This means that, when you see the description of the study design, you will see embedded in it the 2×2 factorial design beloved of psychologists. You will be able to generalize from our work this week to many other research contexts where psychologists conduct experiments in which conditions are manipulated according to a factorial design.

However, our focus here is on the fact that the outcome for analysis is the accuracy of the responses made by children to word targets in a spelling task. The categorical nature of accuracy as an outcome is the reason why we now turn to use Generalized Linear Mixed-effects Models.

6.6.1 Study information

I am going to present the study information in some detail, in part, to enable you to make sense of the analysis aims and results and, in part, so that we can simulate results reporting in a meaningful context.

6.6.1.1 Introduction: the background for the study

Vocabulary knowledge is essential for processing language in everyday life and it is vital that we know how to optimize vocabulary teaching. One strategy with growing empirical support is orthographic facilitation: children and adults are more likely to learn new spoken words that are taught with their orthography (visual word forms; for a systematic review, see Colenbrander, Miles & Ricketts, 2019). Why might orthographic facilitation occur? Compared to spoken inputs, written inputs are less transient across time and less variable across contexts. In addition, orthography is more clearly marked (e.g., the ends of letters and words) than the continuous speech stream. Therefore, orthographic forms may be more readily learned than phonological forms, providing a more effective ‘anchoring device’ (Ehri, 2014; Krepel, de Bree, & de Jong, 2020), or hook, on which to hang semantic information.

Ricketts et al. (2021) investigated how school-aged children learn words. We conducted two studies in which children learned phonological forms and meanings of 16 polysyllabic words in the same experimental paradigm. To test whether orthographic facilitation would occur, half of the words were taught with access to the orthographic form (orthography present condition) and the other half were taught without orthographic forms (orthography absent condition). In addition, we manipulated the instructions that children received: approximately half of the children were told that some words would appear with their written form (explicit group); the remaining children did not receive these instructions (incidental group). Finally, we investigated the impact of spelling-sound consistency of word targets for learning, by including words that varied continuously on a measure of pronunciation consistency (after Mousikou, Sadat, Lucas, & Rastle, 2017).

The quality of lexical representations was measured in two ways. A cuing hierarchical response task (definition, cued definition, recognition) was used to elicit semantic knowledge from the phonological forms, providing a fine-grained measure of semantic learning. A spelling task indexed the extent of orthographic learning for each word. We focus on the analysis of the spelling task responses.

Ricketts et al. (2021) reported two studies. We focus on Study 1, in which Ricketts et al. measured knowledge of newly learned words at two intervals: first one week and then, again, eight months after training. Longitudinal studies of word learning are rare and this is the first longitudinal investigation of orthographic facilitation.

We addressed three research questions.

6.6.1.2 Research questions

1. Does the presence of orthography promote greater word learning? We predicted that children would demonstrate greater orthographic learning for words that they had seen (orthography present condition) versus not seen (orthography absent condition).

2. Will orthographic facilitation be greater when the presence of orthography is emphasized explicitly during teaching? We expected to observe an interaction between instructions and orthography, with the highest levels of learning when the orthography present condition was combined with explicit instructions.
3. Does word consistency moderate the orthographic facilitation effect? For orthographic learning, we expected that the presence of orthography might be particularly beneficial for words with higher spelling-sound consistency, with learning highest when children saw and heard the word, and these codes provided overlapping information.

6.6.1.3 Design

Children were taught 16 novel words in a 2×2 factorial design. The presence of orthography (orthography absent vs. orthography present) was manipulated within participants: for all children, eight of the words were taught with orthography present and eight with orthography absent. Instructions (incidental vs. explicit) were manipulated between participants such that children in the explicit condition were alerted to the presence of orthography whereas children in the incidental condition were not.

6.6.1.4 Participants

In Study 1, 41 children aged 9-10 years completed the word learning task and completed semantic and orthographic assessments one week after learning (Time 1), and eight months later (Time 2). We tested children from one socially mixed school in the South-East of England ($M_{age} = 9.95$, $SD = .53$).

6.6.1.5 Stimulus materials

Stimuli comprised 16 polysyllabic words, all of which were nouns. We indexed consistency at the whole word level using the H uncertainty statistic (after Mousikou et al., 2017; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995). An H value of 0 would indicate a consistent item (all participants producing the same pronunciation), with values > 0 indicating greater inconsistency (pronunciation variability) with increasing magnitude.

6.6.1.6 Procedure

A pre-test was conducted to establish participants' knowledge of the stimulus words before i.e. *pre-* training was administered. Then, each child was seen for three 45-minute sessions to complete training (Sessions 1 and 2) and post-tests (Session 3). In Study 1, longitudinal post-test data were collected because children were post-tested at two time points. (Here, we

refer to “post-tests” as the tests done to test learning, after i.e. *post* training.) Children were given post-tests in Session 3, as noted: this was Time 1. They were then given post-tests again, about eight months later at Time 2.

6.6.1.7 Outcome (dependent) variables – Orthographic post-test

This post-test was used to examine orthographic knowledge after training. Children were asked to spell each word to dictation and spelling productions were transcribed for scoring. For the purposes of our learning in Week 20, we focus on the accuracy of responses. Each response made by a child to a target word was coded as correct or incorrect.

Note that a more sensitive outcome measure of orthographic knowledge was also taken. Responses were also scored using a Levenshtein distance measure, using the `stringdist` library (van der Loo, 2019). This score indexes the number of letter deletions, insertions and substitutions that distinguish between the target and child’s response. In the published report (Ricketts et al., 2021) we focus our analysis of the orthographic outcome on the Levenshtein distance measure of response spelling accuracy, and further details on the analysis approach (Poisson rather than Binomial Generalized Linear Mixed-effects Models) can be found in the paper.

6.6.2 Locate and download the data file

Go to the 402 Moodle folder for Week 20, find and download the data file we need from the Week 20 resources folder:

<https://modules.lancaster.ac.uk/course/view.php?id=34085#section-14>

Or download the resources folder, including the data file and associated scripts by clicking on the link:

<https://modules.lancaster.ac.uk/mod/resource/view.php?id=1809329>

We will be working with the data about the orthographic post-test outcome for the longitudinal study:

- `long.orth_2020-08-11.csv`

Where **long** indicates the longitudinal nature of the data-set. The `.csv` file is a *comma separated values* file and can be opened in Excel.

6.7 Tidy the data

The data are already *tidy*: each column in `long.orth_2020-08-11.csv` corresponds to a variable and each row corresponds to an observation.

We will nevertheless be using `tidyverse` functions to prepare the data for analysis. We load the `tidyverse` and other helpful libraries here.

```
library(broom)
library(effects)
library(gridExtra)
library(here)
library(lattice)
library(knitr)
library(lme4)
library(MuMIn)
library(sjPlot)
library(tidyverse)
```

6.7.1 Read-in the data file using `read_csv`

I am going to assume you have downloaded the data file, and that you know where it is. We use `read_csv` to read the data file into R.

```
long.orth <- read_csv("long.orth_2020-08-11.csv",
                      col_types = cols(
                        Participant = col_factor(),
                        Time = col_factor(),
                        Study = col_factor(),
                        Instructions = col_factor(),
                        Version = col_factor(),
                        Word = col_factor(),
                        Orthography = col_factor(),
                        Measure = col_factor(),
                        Spelling.transcription = col_factor()
                      )
                    )
```

You can see, here, that within the `read_csv()` function call, I specify `col_types`, instructing R how to treat a number of different variables. You can read more about this here:

<https://readr.tidyverse.org/articles/readr.html>

It is always a good to inspect what you have got when you read a data file in to R. (I leave it as an exercise for you to do.)

```
summary(long.orth)
```

Some of the variables included in the .csv file are listed, following, with information about value coding or calculation.

Participant Participant identity codes were used to anonymize participation.

Time Test time was coded 1 (time 1) or 2 (time 2). For the Study 1 longitudinal data, it can be seen that each participant identity code is associated with observations taken at test times 1 and 2.

Instructions Variable coding for whether participants undertook training in the *explicit* or *incidental* conditions.

Word Letter string values showing the words presented as stimuli to the children.

Orthography Variable coding for whether participants had seen a word in training in the orthography *absent* or *present* conditions.

Consistency-H Calculated orthography-to-phonology consistency value for each word.

zConsistency-H Standardized Consistency H scores

Score Outcome variable – for the orthographic post-test, responses were scored as:

- 1 – correct, if the target spelling was produced in full
- 0 – incorrect, if the target spelling was not produced

The summary will show you that we have a number of other variables available, including measures of individual differences in reading or reading-related abilities or knowledge, but we do not need to pay attention to them, for our exercises. If you are interested in the dataset, you can find more information about the variables in the Appendix for this chapter and, of course, in Ricketts et al. (2021).

6.7.2 Code categorical factors

The data are tidy but we need to do a bit of work, before we can run any analyses, to fix the coding of the categorical predictor (or independent) variables, the factors Orthography, Instructions, and Time. By default, R will *dummy code* observations at different levels of a factor. So, for a factor or a categorical variable like **Orthography** (present, absent), R will code one level name e.g. **absent** as 0 and the other e.g. **present** as 1. The 0-coded level is termed the *reference level*, which you could call the baseline level, and by default R will code the level with the name appearing earlier in the alphabet as the reference level.

All this is usually not important. When you specify a model in R where you are asking to estimate the effect of a categorical variable like `Orthography` (present, absent) then, by default, what you will get is an estimate of the average difference in outcome, when all other factors are set to zero, estimated as the difference in outcomes comparing the reference level and the other level or levels of the factor. This will be presented, for example, like the output shown following, for a Generalized Linear Model (i.e., a logistic regression) analysis of the effect of Orthography condition, ignoring the random effects:

```
summary(glm(Score ~ Orthography, family = "binomial", data = long.orth))
```

```
Call:
glm(formula = Score ~ Orthography, family = "binomial", data = long.orth)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-0.8510 -0.8510 -0.6763  1.5436  1.7818 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.35879   0.09871 -13.765 < 2e-16 ***
Orthographypresent 0.52951   0.13124   4.035 5.47e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1431.9 on 1262 degrees of freedom
Residual deviance: 1415.4 on 1261 degrees of freedom
AIC: 1419.4

Number of Fisher Scoring iterations: 4
```

You can see that you have an estimate, in the summary, of the effect of orthographic condition shown as:

`Orthographypresent` 0.52951.

This model (and default coding) gives us an estimate of how the log odds of a child getting a response correct changes if we compare the responses in the `absent` condition (here, treated as the baseline or reference level) with responses in the `present` condition. (Notice that R tells us about the estimate by adding the name of the factor level that is *not* the reference level, here, `present` to the name of the variable `Orthography` whose effect is being estimated.)

We can see that the log odds of a correct response increase by 0.52951 when the orthography (visual word form or spelling) of a word is present during learning trials.

However, as Dale Barr explains here:

<http://talklab.psy.gla.ac.uk/tvw/catpred/>

It is better not to use R's default dummy coding scheme if we are analyzing data where the data come from a study involving two or more factors, and we want to estimate not just the main effects of the factors but also the effect of the interaction between the factors.

In our analyses, we want the coding that allows us to get estimates of the main effects of factors, and of the interaction effects, somewhat like what we would get from an ANOVA. This requires us to use effect coding.

We can code whether a response was recorded in the absent or present condition using numbers. In dummy coding, for any observation, we would use a column of zeroes or ones to code condition: i.e., absent (0) or present (1). In effect coding, for any observation, we would use a column of ones or minus ones to code condition: i.e., absent (-1) or present (1). (With a factor with more than two levels, we would use more than one column to do the coding: the number of columns we would use would equal the number of factor condition levels minus one.) In effect coding, observations coded -1 are in the reference level.

With effect coding, the constant (i.e., the intercept for our model) is equal to the grand mean of all the observed responses. And the coefficient of each of the effect variables is equal to the difference between the mean of the group coded 1 and the grand mean.

You can read more about effect coding here:

<https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-effect-coding/>

6.7.2.1 Category coding practicalities

We follow recommendations to use sum contrast coding for the experimental factors. Further, to make interpretation easier, we want the coding to work so that for both orthography and presentation conditions, doing something is the “high” level in the factor – hence:

- Orthography, absent (-1) vs. present (+1)
- Instructions, incidental (-1) vs. explicit (+1)
- Time, test time 1 (-1) vs. time 2 (+1)

We use a modified version of the `contr.sum()` function (provided in the `memisc` library) that allows us to define the base or reference level for the factor manually:

<https://www.rdocumentation.org/packages/memisc/versions/0.99.17.2/topics/contr>

```
library(memisc)
```

Note: we have seen in some classes that we cannot appear to load `library(memisc)` and `library(tidyverse)` at the same time without getting weird warnings. So I would load `library(memisc)` after I have loaded `library(tidyverse)` and maybe unload it afterwards: just click on the button next to the package or library name in R-Studio to detach the library (i.e., stop it from being available in the R session).

In the following sequence, I first check how R codes the levels of each factor by default, I then change the coding, and check that the change gets me what I want.

We want effects coding for the orthography condition factor, with orthography condition coded as -1, +1. Check the coding.

```
contrasts(long.orth$Orthography)
```

```
present  
absent      0  
present     1
```

You can see that Orthography condition is initially coded, by default, using dummy coding: absent (0); present (1). We want to change the coding, then check that we have got what we want.

```
contrasts(long.orth$Orthography) <- contr.sum(2, base = 1)  
contrasts(long.orth$Orthography)
```

```
2  
absent  -1  
present   1
```

We want effects coding for the presentation condition factor, with presentation condition coded as -1, +1. Check the coding.

```
contrasts(long.orth$Instructions)
```

```
incidental  
explicit      0  
incidental    1
```

Change it.

```
contrasts(long.orth$Instructions) <- contr.sum(2, base = 2)
contrasts(long.orth$Instructions)
```

```
1
explicit 1
incidental -1
```

We want effects coding for the Time factor, with Time coded as -1, +1 Check the coding.

```
contrasts(long.orth$Time)
```

```
2
1 0
2 1
```

Change it.

```
contrasts(long.orth$Time) <- contr.sum(2, base = 1)
contrasts(long.orth$Time)
```

```
2
1 -1
2 1
```

6.7.2.2 Code tip

I use `contr.sum(a, base = b)` to do the coding, where a is the number of levels in a factor, and b tells R which level to use as the baseline or reference level. I usually need to check the coding before and after I specify it.

6.8 Working with GLMMs in R

A small change in R `lmer` code allows us to extend what we know about linear mixed-effects models to conduct *Generalized Linear Mixed-effects Models*. We change the function call from `lmer()` to `glmer()`. However, we have to make some other changes, as we detail in the following sections.

We will be examining the impact of the experimental effects, that is, the *fixed effects* associated with the impacts on the outcome **Score** (accuracy of response in the word spelling test) associated with the following comparisons:

- Time: time 1 versus time 2
- Orthography: present versus absent conditions
- Instructions: explicit versus incidental conditions
- Standardized spelling-sound consistency
- Interaction between the effects of Orthography and Instructions
- Interaction between the effects of Orthography and consistency

We will begin by keeping the random effects structure simple.

6.8.1 Specify a random intercepts model

In our first model, we will specify just random effects of participants and items on intercepts.

```
long.orth.min.glmer <- glmer(Score ~  
    Time + Orthography + Instructions + zConsistency_H +  
    Orthography:Instructions +  
    Orthography:zConsistency_H +  
    (1 | Participant) +  
    (1 | Word),  
  
    family = "binomial",  
    glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=2e5)),  
  
    data = long.orth)  
  
summary(long.orth.min.glmer)
```

The code works as follows.

- `glmer()` the function name changes because now we want a *generalized* linear mixed-effects model of accuracy
- With `(1 | Participant)` we include random effects of participants on on intercepts

- With `(1 | Word)` we include random effects of stimulus on on intercepts
- `family = binomial` accuracy is a binary outcome variable (correct, incorrect) so we assume a binomial probability distribution
- `glmerControl(optimizer="bobyqa", ...)` we change the underlying mathematical engine (the optimizer) to cope with greater model complexity, and we allow the model fitting functions to take longer to find estimates with `optCtrl=list(maxfun=2e5)`

Notice how we specify the fixed effects. We want `glmer()` to estimate “main effects and interactions” that we hypothesized.

We specify the *main* effects with:

```
Time + Orthography + Instructions + zConsistency_H +
```

We specify the *interaction* effects with:

```
Orthography:Instructions +
```

```
Orthography:zConsistency_H +
```

Where we ask for estimates of the fixed effects associated with:

- `OrthographyInstructions`: the interaction between the effects of Orthography and Instructions
- `OrthographyzConsistency-H`: the interaction between the effects of Orthography and consistency

6.8.1.1 Code tip

There are two forms of notation we can use to specify interactions in R. The simplest form is to use something like this:

```
Orthography*Instructions
```

This will get you estimates of:

- `Orthography` present versus absent conditions
- `Instructions` explicit versus incidental conditions
- `Orthography x Instructions` the interaction between the effects of Orthography and Instructions

So, in general, if you want estimates of the effects of variables A, B and the interaction A x B, then you write A*B.

We can also use the colon symbol to specify only the interaction, i.e., ignoring main effects, so if you specify A:B then you will get an estimate of the interaction A x B but not the effects A, B. With the coding

```
Score ~ Orthography + Instructions + Orthography:Instructions
```

I would be making explicit that I want estimates for the effects of Orthography, Instruction and the interaction between the effects of Orthography and Instructions.

6.8.2 Read the results

If you run the model code, you will get the results shown in the output.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
Family: binomial ( logit )
Formula: Score ~ Time + Orthography + Instructions + zConsistency_H +
Orthography:Instructions + Orthography:zConsistency_H + (1 |
Participant) + (1 | Word)
Data: long.orth
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))

      AIC      BIC    logLik deviance df.resid
1040.4   1086.7   -511.2    1022.4     1254

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-5.0994 -0.4083 -0.2018  0.2019  7.4940 

Random effects:
Groups      Name        Variance Std.Dev.
Participant (Intercept) 1.840     1.357
Word        (Intercept) 2.224     1.491
Number of obs: 1263, groups: Participant, 41; Word, 16

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.878462  0.443957 -4.231 2.32e-05 ***
Time2        0.050136  0.083325  0.602    0.547  

```

```

Orthography2          0.455009   0.086813   5.241 1.59e-07 ***
Instructions1        0.042289   0.230336   0.184    0.854
zConsistency_H       -0.618093   0.384005  -1.610    0.107
Orthography2:Instructions1  0.005786   0.083187   0.070    0.945
Orthography2:zConsistency_H  0.014611   0.083105   0.176    0.860
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Correlation of Fixed Effects:

	(Intr)	Time2	Orthg2	Instr1	zCns_H	Or2:I1
Time2	0.008					
Orthogrpby2	-0.059	0.008				
Instructns1	0.014	0.025	0.001			
zCnsstnc_H	0.016	-0.002	-0.029	-0.001		
Orthgrp2:I1	-0.002	-0.001	0.049	-0.045	0.000	
Orthgr2:C_H	-0.027	0.001	0.179	0.000	-0.035	-0.007

In these results, we see:

1. First, information about the function used to fit the model, and the model object created by the `glmer()` function call
2. Then the model formula including main effects `Score ~ Time + Orthography + Instructions + zConsistencyH`
3. As well as interactions `Orthography:Instructions + Orthography:zConsistencyH`
4. And the random effects `(1 | Participant) + (1 |Word)`
5. Then we see `REML criterion at convergence` about the model fitting process, which we can usually ignore
6. Then we see information about the model algorithm
7. Then we see model fit statistics, including `AIC BIC logLik`
8. Then we see information about the distribution of residuals
9. Then the **Random Effects**. Notice that the statistics are `Variance Std.Dev.`, that is, the variance and the corresponding standard deviation associated with the random effects
 - The variance due to random differences between the average intercept (over all data) and the intercept for each participant
 - And the variance due to random differences between the average intercept (over all data) and the intercept for responses to each word stimulus

10. Last, just as for linear models, we see estimates of the coefficients of the fixed effects, the intercept and the slopes of the experimental variables.
11. Note that we get p-values for what are called Wald null hypothesis significance tests on the coefficients.

We can see that one effect is significant. The estimate for the effect of the presence compared to the absence of Orthography is 0.455009. The positive coefficient tells us that the log odds that a response will be correct is higher when Orthography is present compared to when it is absent.

We can also see an effect of consistency that can be considered to be marginal or near-significant by some standards. The estimate for the effect of `zConsistency_H` is -0.618093 indicating that the log odds of a response being correct decrease for unit increase in the standardized H consistency measure.

6.8.3 GLMMs and hypothesis tests

As we discussed in the last chapter, we can conduct null hypothesis significance tests by comparing models that differ in the presence or absence of a fixed effect or a random effect, using the Likelihood Ratio Test. In the results output for a GLMM by the `glmer()` function, you can see that alongside the estimates of the coefficients (and standard error) for the fixed effects we also have z and p-values. Wald z tests for GLMMs test the null hypothesis of no effect by comparing the effect estimate with their standard error, and comparing the resulting test statistic to zero (Bolker et al., 2009).

6.8.4 Presenting and visualizing the effects

We usually want to do more than just report whether experimental effects are or are not significant. It helps us to present and interpret the estimates from a model if we can visualize the model prediction. There are a variety of tools that help us to do this.

6.8.4.1 sjPlot library

We can use the `plot_model` function from the `sjPlot` library. The following sequence of code takes information from the model we have just run, then generates model predictions, of change in the probability of a correct response (Score) for different levels of the Orthography factor and the consistency variable. I chose these variables because they are the significant or near-significant effects.

```

porth <- plot_model(long.orth.min.glmer,
                     type="pred",
                     terms = "Orthography") +
  theme_bw() +
  ggtitle("Predicted probability") +
  ylim(0,1)

pzconsH <- plot_model(long.orth.min.glmer,
                      type="pred",
                      terms = "zConsistency_H") +
  theme_bw() +
  ggtitle("Predicted probability") +
  ylim(0,1)

grid.arrange(porth, pzconsH,
             ncol=2)

```

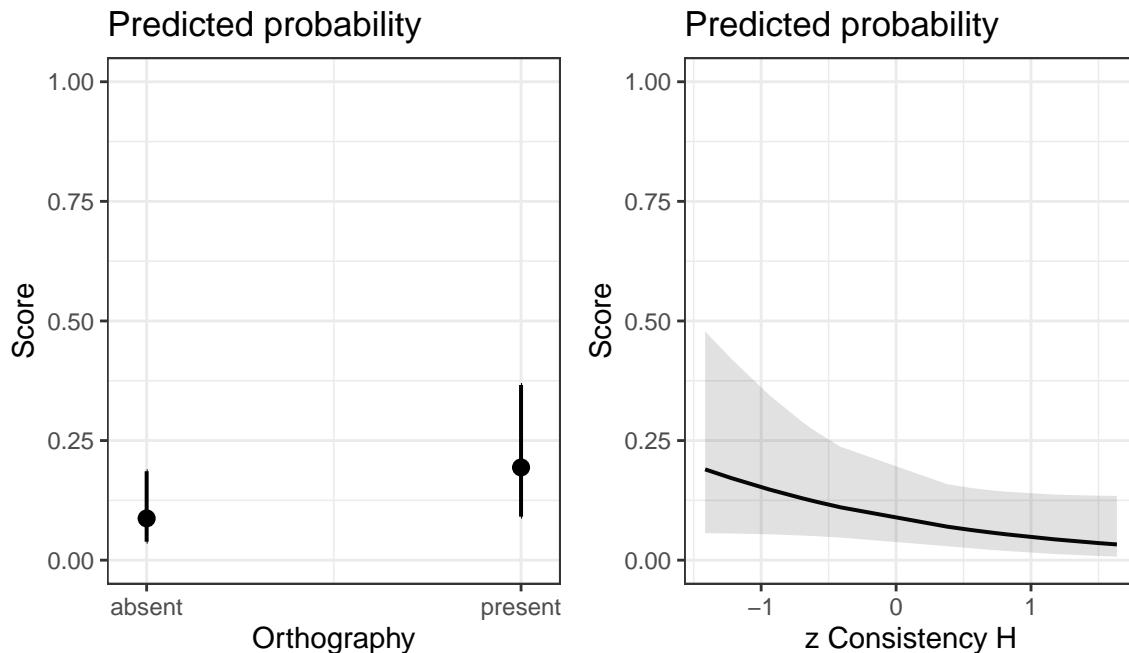


Figure 6.2: Effect of orthography condition (present versus absent) on probability of a response being correct

The plots in Figure @ref(fig:sjplot-orthography) show clearly how the probability of a correct response is greater for the conditions where Orthography had been present (versus absent) during the word learning phase of the study. We can also see a trend such that the probability of

a response being correct decreases as the (in-)consistency of a target word tends to increase.

6.8.4.1.1 Code tip

You will need to install the `sjPlot` library first, and then run the command `library(sjPlot)` before creating your plot.

Notice:

- `plot_model()` produces the plot
- `plot_modellong.orth.min.glmer`) specifies that the plot should be produced given information about the previously fitted model `long.orth.min.glmer`
- `type = "pred"` tells R that you want a plot showing the model predictions, of the effect of, e.g., Orthography condition

The function outputs an object whose appearance can be edited as a `ggplot` object.

See the following set of blog posts for detailed advice and explanation:

https://cran.r-project.org/web/packages/sjPlot/vignettes/plot_marginal_effects.html

https://cran.r-project.org/web/packages/sjPlot/vignettes/plot_interactions.html

https://cran.r-project.org/web/packages/sjPlot/vignettes/plot_model_estimates.html

And see the library manual for detailed technical information:

<https://cran.r-project.org/web/packages/sjPlot/sjPlot.pdf>

6.9 Examining if we should include random effects

So far, we have been considering the results of a *random intercepts* model in which we take into account the random effects of participants and stimulus word differences on intercepts. We have ignored the possibility that the slopes of the experimental variables might vary between participants or between words. We now need to examine the question: What random effects should we specify?

I must warn you that the question and the answer are complex but the coding is quite simple, and the approaches you can take to address the question are, now, quite well recognized by the psychological community. In other words, the community recognizes the nature of the problem, and recognizes the methods you can potentially follow to solve the problem.

The complexity, for us, lies not in the mathematics: the coding is simple and the `glmer()` function does the work. I think the complexity lies, firstly, in how we have to think about the study design, what gets manipulated or allowed to vary. I find it very helpful to sketch out,

by hand, what the study design means in relation to who does what in an experiment. And when I work with collaborators, I typically ask them to explain the design verbally while I try to write in pencil what the dataset looks like (for a small number of participants or stimuli). The complexity lies, secondly, and in how we have to translate our understanding of the study design to a specification of random effects. We can master *that* aspect of the challenge through practice.

6.9.0.1 Getting started

So, what random effects should we include? If you go back to the description of the study design, then you will be able to see that a number of possibilities follow, theoretically, from the design.

A currently influential set of recommendations (Barr et al., 2013; but see discussions in Bates et al., 2015; Matuschek et al., 2017; Meteyard & Davies, 2020) has been labeled *Keep it maximal*:

- If you are testing effects manipulated according to a pre-specified design then you should:
 - Test random intercepts – due to random differences between subjects or between items (or other sample grouping variables)
 - Test random slopes for all within-subjects or within-items (or other) fixed effects

This means that specification of the random effects structure requires two sets of information:

1. What are the fixed effects?
2. What are the grouping variables: did you test multiple participants using multiple stimuli (e.g., words ...) or did you test participants under multiple different conditions (e.g., levels of experimental condition factors)?

Our answers to these questions then dictate potentially how we specify the random effects structure for our model.

1. We can consider that we should certainly include a random effect of each grouping variable (e.g., participants, stimulus word) on intercepts.
2. We can reflect that we should also include a random effect of a grouping variable e.g. (`\participant`) on the slope of each variable that was manipulated *within* the units of that variable.

- When we specify models, we should remember that by default if you specify `(1 + something | participant)` then you are specifying that you want to take into account variance due to the random differences between participants in intercepts (`1 ... | participant`), plus variance due to the random differences between participants in slopes (`... something | participant`), plus the covariance (or correlation) between the random intercepts and the random slopes.

You will have become familiar with the practice of referring to effects as *within-subjects* or *between-subjects* previously, in working with ANOVA. Here, whether an effect is *within-subjects* or *within-items* or not has relevance to whether we can or should specify a random effect of subjects or of items on the slope of a fixed effect.

In deciding what random effects we should specify, we need to think about what response data we have recorded, for each of the experimental variables, given the study design. This is because if we want to specify a random effect of participants (or stimulus words) on the slope of an experimental condition then we need to have data, for each person, on their responses under all levels of the condition.¹ If we want to estimate the effect of the experimental manipulation of learning condition, for example, the impact of the presence of orthography, for a person, we need to have data for both levels of the condition (orthography absent and orthography present) for that person. If you think about it, we cannot estimate the slope of the effect of the presence of orthography without response data recorded under both the orthography absent condition and the orthography present condition. If we can estimate the slope of the effect *then* we can estimate how the slope of the effect deviates between participants.

We can spell out how the experimental conditions were manipulated for the example study, as follows. (Writing out this kind of account, for yourself, will be helpful perhaps when you are planning an analysis and have to work out what the random effects could be.)

- The effect of Orthography was manipulated *within* participants and *within* stimulus words. This is because the presence of orthography (orthography absent versus orthography present) was manipulated so that we have data about test responses to each word under both Orthography conditions, and data about responses from each child under both conditions.
- The effect of Instructions was manipulated *between* participants. This is because Instructions (incidental vs. explicit) were manipulated between participants such that children in the explicit condition were alerted to the presence of orthography whereas children in the incidental condition were not.

¹Note that, the strengths of mixed-effects models mean we do not need complete data for responses to all stimuli, under all conditions, for each participant; the method can tolerate imbalances in the data. Likewise, if we are interested in, say, the slope of a numeric variable like, as in the example data, word spelling-sound consistency, we do not need to worry about having complete data, for each person, for responses at each level of the consistency variable. We just need enough data, observed responses at different levels of the variable, to be able to estimate the slope of the variable.

3. We can say that the effects of Orthography and of Instructions are both manipulated *within* words. Items were counterbalanced across instruction and orthography conditions, with all words appearing in both orthography conditions for approximately the same number of children within the explicit and incidental groups.
4. The effect of spelling-sound consistency varies *between* words because different words have different consistency values but the effect of consistency varies *within* participants because we have response data for each participant for their responses to words of different levels of consistency.
5. We recorded responses for all participants and all words so we can say that the effect of Time (test time 1 versus time 2) can also be understood to vary *within* both participants and words. This means that, for each person's response to each word on which they are tested, we have response data recorded at both test times.

These considerations suggests that we should specify a model with the random effects:

1. The random effects of participants on intercepts, and on the slopes of the effects of Time, Orthography and spelling-sound consistency, as well as all corresponding covariances.
2. The random effects of stimulus words on intercepts, and on the slopes of the effects of Time, Orthography and Instructions, as well as all corresponding covariances.

This is simple to do using the code shown following.

```
long.orth.max.glmer <- glmer(Score ~
  Time + Orthography + Instructions + zConsistency_H +
  Orthography:Instructions +
  Orthography:zConsistency_H +
  (Time + Orthography + zConsistency_H + 1 | Participant) +
  (Time + Orthography + Instructions + 1 |Word),
  family = "binomial",
  glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=2e5)),
  data = long.orth)

summary(long.orth.max.glmer)
```

Where we have the critical terms:

1. (Time + Orthography + zConsistency_H + 1 | Participant) to account for the random effects of participants on intercepts, and on the slopes of the effects of Time, Orthography and spelling-sound consistency, as well as all corresponding covariances.
2. (Time + Orthography + Instructions + 1 | Word) the random effects of stimulus words on intercepts, and on the slopes of the effects of Time, Orthography and Instructions, as well as all corresponding covariances.

If you run this code, however, you will see that you get warnings along with your estimates.

```
boundary (singular) fit: see help('isSingular')

Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
Family: binomial ( logit )
Formula: Score ~ Time + Orthography + Instructions + zConsistency_H +
Orthography:Instructions + Orthography:zConsistency_H + (Time +
Orthography + zConsistency_H + 1 | Participant) + (Time +
Orthography + Instructions + 1 | Word)
Data: long.orth
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))

      AIC      BIC      logLik deviance df.resid
1053.6   1192.4    -499.8     999.6     1236

Scaled residuals:
    Min      1Q  Median      3Q      Max
-4.1014 -0.4027 -0.1723  0.2037  7.0331

Random effects:
Groups      Name        Variance Std.Dev. Corr
Participant (Intercept) 2.043126 1.42938
              Time2       0.005675 0.07533   0.62
              Orthography2 0.079980 0.28281   0.78 -0.01
              zConsistency_H 0.065576 0.25608   0.49  0.99 -0.16
Word        (Intercept) 2.793447 1.67136
              Time2       0.046736 0.21618   0.14
              Orthography2 0.093740 0.30617  -0.68 -0.81
              Instructions1 0.212706 0.46120  -0.74 -0.05  0.38
Number of obs: 1263, groups: Participant, 41; Word, 16

Fixed effects:
Estimate Std. Error z value Pr(>|z|)
```

```

(Intercept) -2.10099 0.49588 -4.237 2.27e-05 ***
Time2 0.02077 0.12285 0.169 0.865741
Orthography2 0.52480 0.15496 3.387 0.000708 ***
Instructions1 0.24467 0.27281 0.897 0.369801
zConsistency_H -0.67818 0.36310 -1.868 0.061802 .
Orthography2:Instructions1 -0.05133 0.10004 -0.513 0.607908
Orthography2:zConsistency_H 0.05850 0.11634 0.503 0.615064
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Correlation of Fixed Effects:

	(Intr)	Time2	Orthg2	Instr1	zCns_H	Or2:I1
Time2	0.097					
Orthogrpphy2	-0.280	-0.187				
Instructns1	-0.278	0.023	0.109			
zCnsstncy_H	0.062	0.005	-0.064	0.120		
Orthgrp2:I1	0.024	0.005	-0.071	0.212	-0.065	
Orthgr2:C_H	-0.062	0.049	0.246	-0.031	-0.440	0.001
optimizer (bobyqa)						
convergence code: 0 (OK)						
boundary (singular) fit: see help('isSingular')						

Notice, especially, the warning:

```
boundary (singular) fit: see ?isSingular
```

If you put the warning message text into a search engine, then you will get directed to a variety of discussions about what they mean and what you should do about them.

A highly instructive blog post by (I think) Ben Bolker provides some very useful advice:

https://rstudio-pubs-static.s3.amazonaws.com/33653_57fc7b8e5d484c909b615d8633c01d51.html

Where we see this advice:

Check singularity

If the fit is singular or near-singular, there might be a higher chance of a false positive (we're not necessarily screening out gradient and Hessian checking on singular directions properly); a higher chance that the model has actually misconverged (because the optimization problem is difficult on the boundary); and a *reasonable argument that the random effects model should be simplified*.

The definition of singularity is that some of the constrained parameters of the random effects theta parameters are on the boundary (equal to zero, or very very close to zero ...)

(Emphases added.)

I am going to take his advice and simplify the random effects part of the model. We know that the random intercepts model converges fine and now we know that the maximal model does not. Thus, our task is now to identify a model that includes random effects of participants or items on slopes and still converges without warnings.

6.9.1 Examine the utility of random effects by comparing models with the same fixed effects but varying random effects

I am just going to assume we need both random effects of subjects and of items on intercepts so I focus on *random slopes* here. (This assumption may not always be true but is often useful.)

We can fit a series of models as follows. Note that I will not show the results for every model, to save space, but you should run the code to see what happens. Look out for convergence or singularity warnings, where they appear.

6.9.1.1 Random effects of subjects and stimulus items on intercepts

In the first model, we have just random effects of participants or items on intercepts. This is where we started.

```
long.orth.min.glmer <- glmer(Score ~
                                Time + Orthography + Instructions + zConsistency_H +
                                Orthography:Instructions +
                                Orthography:zConsistency_H +
                                (1 | Participant) +
                                (1 | Word),
                                family = "binomial",
                                glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=2e5)),
                                data = long.orth)

summary(long.orth.min.glmer)
```

We saw that the model converged, and we looked at the results previously.

What next? A simple approach we can take is to see if we can add each fixed effect to be included in our random effects terms, one effect at a time.

6.9.1.2 Random effects of subjects and stimulus words on the slope of the Orthography effect

In our second model, we change the random effects terms so that we can account for the random effects of participants and of items on intercepts as well as on the slopes of the Orthography effect. (The Orthography effect is both within-subjects and within-items.)

```
long.orth.2.glmer <- glmer(Score ~  
    Time + Orthography + Instructions + zConsistency_H +  
    Orthography:Instructions +  
    Orthography:zConsistency_H +  
    (dummy(Orthography) + 1 || Participant) +  
    (dummy(Orthography) + 1 || Word),  
    family = "binomial",  
    glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=2e5)),  
    data = long.orth)  
  
summary(long.orth.2.glmer)
```

This model converges without warnings.

6.9.1.2.1 Code tip

Notice that we specify that we `||` do not want random covariances; we are keeping things simple in each step.

Note the use of `dummy()` inside the random effects terms. The ‘dummy’ is a mis-leading name; we are not talking about dummy coding (as above). Here, the `dummy()` stops R from mis-interpreting the requirement to estimate the effect of the differences between category levels, within random effects.

The reason is explained here:

<https://rdrr.io/cran/lme4/man/dummy.html>

You can see what impact it has by specifying, instead, the naked random effect:

e.g. (Orthography + 1 || Participant).

6.9.1.3 Random effects of subjects and stimulus words on the slope of the Instructions effect

Next we can add **Instructions** to take into account random differences between words in the slope of this effect. We show the results for this model as they are instructive.

```
long.orth.3.glmer <- glmer(Score ~
                           Time + Orthography + Instructions + zConsistency_H +
                           Orthography:Instructions +
                           Orthography:zConsistency_H +
                           (dummy(Orthography) + 1 || Participant) +
                           (dummy(Orthography) + dummy(Instructions) + 1 || Word),
                           family = "binomial",
                           glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=2e5)),
                           data = long.orth)

summary(long.orth.3.glmer)
```

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [glmerMod]
Family: binomial (logit)
Formula: Score ~ Time + Orthography + Instructions + zConsistency_H +
 Orthography:Instructions + Orthography:zConsistency_H + (dummy(Orthography) +
 1 || Participant) + (dummy(Orthography) + dummy(Instructions) +
 1 || Word)
Data: long.orth
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))

AIC	BIC	logLik	deviance	df.resid
1036.5	1098.2	-506.2	1012.5	1251

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.9951	-0.4068	-0.1920	0.1838	5.9308

Random effects:

Groups	Name	Variance	Std.Dev.
Participant	(Intercept)	1.64393	1.2822
Participant.1	dummy(Orthography)	0.55604	0.7457
Word	(Intercept)	1.94313	1.3940
Word.1	dummy(Orthography)	0.01608	0.1268
Word.2	dummy(Instructions)	0.86694	0.9311

Number of obs: 1263, groups: Participant, 41; Word, 16

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.9621258	0.4400570	-4.459	8.24e-06	***
Time2	0.0507347	0.0843236	0.602	0.54740	
Orthography2	0.4263703	0.1114243	3.827	0.00013	***
Instructions1	0.1907424	0.2632579	0.725	0.46873	
zConsistency_H	-0.6270892	0.3669873	-1.709	0.08750	.
Orthography2:Instructions1	-0.0265727	0.1048392	-0.253	0.79991	
Orthography2:zConsistency_H	-0.0006303	0.0878824	-0.007	0.99428	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:

	(Intr)	Time2	Orthg2	Instr1	zCns_H	Or2:I1
Time2	0.007					
Orthogrphy2	0.028	0.008				
Instructns1	-0.127	0.022	0.019			
zCnsstncy_H	0.017	-0.002	-0.026	0.011		
Orthgrp2:I1	0.013	0.001	0.003	0.068	-0.009	
Orthgr2:C_H	-0.030	0.002	0.182	0.002	-0.026	-0.018

This model also converges without warnings.

Take a look at the random effects summary. You can see:

- Participant (Intercept) 1.64393 estimated variance due to random effect of participants on intercepts
- Participant.1 dummy(Orthography) 0.55604 variance due to random effect of participants on the slope of the Orthography effect
- Word (Intercept) 1.94314 variance due to random effect of words on intercepts

- Word.1 dummy(Orthography) 0.01607 variance due to random effect of words on the slope of the Orthography effect
- Word.2 dummy(Instructions) 0.86694 variance due to random effect of words on the slope of the Instructions effect

We do *not* see correlations (random effects covariances) because we use the `||` notation to stop them being estimated. We want to stop them being estimated because we want to see what we gain from adding just the requirement, first, to estimate the variance associated with random effects of participants or words on the slopes of the experimental variables.

Also, we can suspect that adding the requirement to estimate covariances will blow the model up for two reasons. The maximal model, including random slopes variances *and* covariances clearly did not converge. Secondly, at least two of the correlations listed in the random effects, for the maximal model, were pretty extreme with `Corr = -0.01` and `= 0.99`; such extreme values ($r \sim \pm 1$) are bad signs; see the discussion in @ref(bad-signs).

6.9.1.4 Adding effects a bit at a time

In the following models, because we can see that we can get a model to converge with random effects of participants or items on Orthography, and random effect of participants on Instructions, I am going to keep these random effects in the model. I will check if adding further effects is OK too, in terms of successful convergence. I am going to treat all the following models as variations on a theme, the theme being: can we add anything else to:

```
(dummy(Orthography) + 1 || Participant) +
(dummy(Orthography) + dummy(Instructions) + 1 || Word),
```

6.9.1.5 Random effects of subjects on the slope of the consistency effect

Next we see if we can add `zConsistency_H`.

```
long.orth.4.a.glmer <- glmer(Score ~
  Time + Orthography + Instructions + zConsistency_H +
  Orthography:Instructions +
  Orthography:zConsistency_H +
  (dummy(Orthography) + zConsistency_H + 1 || Participant) +
```

```

  (dummy(Orthography) + dummy(Instructions) + 1 || Word),

  family = "binomial",
  glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=2e5)),

  data = long.orth)

boundary (singular) fit: see help('isSingular')

summary(long.orth.4.a.glmer)

Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
Family: binomial ( logit )
Formula: Score ~ Time + Orthography + Instructions + zConsistency_H +
   Orthography:Instructions + Orthography:zConsistency_H + (dummy(Orthography) +
   zConsistency_H + 1 || Participant) + (dummy(Orthography) +
   dummy(Instructions) + 1 || Word)
Data: long.orth
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))

      AIC      BIC    logLik deviance df.resid
1038.5   1105.3   -506.2    1012.5     1250

Scaled residuals:
    Min      1Q  Median      3Q      Max
-3.9952 -0.4068 -0.1920  0.1838  5.9309

Random effects:
Groups      Name        Variance Std.Dev.
Participant (Intercept) 1.644e+00 1.282e+00
Participant.1 dummy(Orthography) 5.560e-01 7.456e-01
Participant.2 zConsistency_H 1.259e-10 1.122e-05
Word        (Intercept) 1.943e+00 1.394e+00
Word.1      dummy(Orthography) 1.604e-02 1.266e-01
Word.2      dummy(Instructions) 8.669e-01 9.311e-01
Number of obs: 1263, groups: Participant, 41; Word, 16

Fixed effects:
Estimate Std. Error z value Pr(>|z|)
```

```

(Intercept) -1.9621298 0.4400610 -4.459 8.24e-06 ***
Time2 0.0507346 0.0843237 0.602 0.54740
Orthography2 0.4263731 0.1114190 3.827 0.00013 ***
Instructions1 0.1907320 0.2632610 0.724 0.46876
zConsistency_H -0.6270931 0.3669852 -1.709 0.08749 .
Orthography2:Instructions1 -0.0265706 0.1048369 -0.253 0.79992
Orthography2:zConsistency_H -0.0006353 0.0878784 -0.007 0.99423
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Correlation of Fixed Effects:

```

(Intr) Time2 Orthg2 Instr1 zCns_H Or2:I1
Time2 0.007
OrthogrpHy2 0.028 0.008
Instructns1 -0.127 0.022 0.019
zCnsstncy_H 0.017 -0.002 -0.026 0.011
Orthgrp2:I1 0.013 0.001 0.003 0.068 -0.009
Orthgr2:C_H -0.030 0.002 0.182 0.002 -0.026 -0.018
optimizer (bobyqa) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')

```

We see two useful pieces of information when we run this model:

- We get the warning **boundary (singular) fit** see `?isSingular`: that tells us the model algorithm could not converge on effects estimates, given the model we specify, given the data
- Note, also, we see the random effects variance estimate `Participant.2 zConsistency_H 1.259e-10`

These two things are possibly connected: the singularity warning; and the estimated variance of $1.259e-10$ (i.e. a very very small number) associated with the random effect of participants on the slope of the `zConsistency_H`. We can expect that the model fitting algorithm is going to have difficulty estimating nothing, or something close to nothing: here, the very very small variance associated with the between-participant differences in the slope of the non-significant effect of word spelling-sound consistency on response accuracy.

6.9.1.6 Random effects of subjects and stimulus words on the slope of the time effect

What about the random effects of participants or of words on the slope of the effect of Time?

```

long.orth.4.b.glmer <- glmer(Score ~
  Time + Orthography + Instructions + zConsistency_H +

```

```

Orthography:Instructions +
Orthography:zConsistency_H +
(dummy(Orthography) + dummy(Time) + 1 || Participant) +
(dummy(Orthography) + dummy(Instructions) + dummy(Time) + 1 || Word)

family = "binomial",
glmerControl(optimizer="bobyqa", optCtrl=list(maxfun=2e5)),

data = long.orth)

boundary (singular) fit: see help('isSingular')

summary(long.orth.4.b.glmer)

Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
Family: binomial ( logit )
Formula: Score ~ Time + Orthography + Instructions + zConsistency_H +
Orthography:Instructions + Orthography:zConsistency_H + (dummy(Orthography) +
dummy(Time) + 1 || Participant) + (dummy(Orthography) + dummy(Instructions) +
dummy(Time) + 1 || Word)
Data: long.orth
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e+05))

      AIC      BIC    logLik deviance df.resid
1040.5   1112.5   -506.2    1012.5     1249

Scaled residuals:
    Min      1Q  Median      3Q      Max
-3.9952 -0.4068 -0.1920  0.1838  5.9309

Random effects:
Groups           Name        Variance Std.Dev.
Participant     (Intercept) 1.644e+00 1.2821915
Participant.1   dummy(Orthography) 5.560e-01 0.7456312
Participant.2   dummy(Time)    0.000e+00 0.0000000
Word            (Intercept) 1.943e+00 1.3939636

```

```

Word.1      dummy(Orthography) 1.604e-02 0.1266478
Word.2      dummy(Instructions) 8.669e-01 0.9310620
Word.3      dummy(Time)        2.161e-07 0.0004648
Number of obs: 1263, groups: Participant, 41; Word, 16

```

Fixed effects:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.9621302	0.4400578	-4.459	8.24e-06 ***
Time2	0.0507347	0.0843253	0.602	0.54740
Orthography2	0.4263731	0.1114192	3.827	0.00013 ***
Instructions1	0.1907329	0.2632605	0.725	0.46876
zConsistency_H	-0.6270910	0.3669905	-1.709	0.08750 .
Orthography2:Instructions1	-0.0265706	0.1048370	-0.253	0.79992
Orthography2:zConsistency_H	-0.0006351	0.0878784	-0.007	0.99423

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1	' '	1	

Correlation of Fixed Effects:

```

(Intr) Time2 Orthg2 Instr1 zCns_H Or2:I1
Time2    0.007
Orthogrphy2  0.028  0.008
Instructns1 -0.127  0.022  0.019
zCnsstncy_H  0.017 -0.002 -0.026  0.011
Orthgrp2:I1  0.013  0.001  0.003  0.068 -0.009
Orthgr2:C_H -0.030  0.002  0.182  0.002 -0.026 -0.018
optimizer (bobyqa) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')

```

Nope.

We again see two useful pieces of information when we run this model:

- We get the warning **boundary (singular) fit** see `?isSingular`: that tells us the model algorithm could not converge on effects estimates, given the model we specify, given the data
- Note, also, we see the random effects variance estimate `Participant.2 dummy(Time) 0.000e+00`
- And we see the variance estimate `Word.3 dummy(Time) 2.161e-07`

Again, we can surmise that a mixed-effects model will get into trouble – and we will see convergence warnings – where we have a fixed effect with little impact (like, here, Time, or, before consistency) and we are asking for the estimation of variance associated with random differences in slopes where there may be, in fact, little random variation. Possibly, these two

things are connected too: we are perhaps unlikely to see random differences in the slope of the effect of an experimental variable if the effect is at or near zero. Possibly, we may see the effect of an experimental variable which is very very consistent. Think back to Week 16 and the effect associated with the relation between maths and physics scores where there seemed to be little variation between classes in the slope representing the relation.

6.9.2 Bad signs

We can see that a model has difficulty if we see things like:

1. Convergence warnings, obviously
2. Very very small random effects variances
3. Extreme random effects correlations of ± 1.00

If we see a warning that the model fitting algorithm nearly failed to converge: `boundary (singular) fit: see ?glimmerControl` or failed to converge then this tells us that, given the data, the mathematical engine (optimizer) underlying the `lmer()` function model fitting got into trouble because, in short, it was trying to find estimates for effects that were close to not being there at all.

If the variances for the random effects of participants or stimulus items on the slopes of an experimental variable are very small this suggests that the level of complexity in the model cannot really be justified or that the model will have difficulty estimating it. Extreme correlations (near 0 or 1) between random effects on intercepts and on slopes of fixed effects suggest the level of complexity in the model cannot really be justified (see also the discussion in Bates et al., 2015; Matuschek et al., 2017).

6.9.3 Comparison of models varying in random effects

There is no point comparing the models that do not converge, so we focus on those that do.

Does the addition of random slopes improve model fit? We can compare the model in pairs, as follows, to test whether each addition in model complexity improves model fit. We run the code for the model comparisons as follows.

First, we compare the model `long.orth.min.glmer` (just random intercepts) with `long.orth.2.glmer` to check if increasing model complexity, by accounting for random differences between participants or words in the slope of the Orthography effect improves model fit to data.

```
anova(long.orth.min.glmer, long.orth.2.glmer)
```

```

Data: long.orth
Models:
long.orth.min.glmer: Score ~ Time + Orthography + Instructions + zConsistency_H + Orthography:
long.orth.2.glmer: Score ~ Time + Orthography + Instructions + zConsistency_H + Orthography:
npar      AIC      BIC  logLik deviance Chisq Df Pr(>Chisq)
long.orth.min.glmer    9 1040.4 1086.7 -511.20   1022.4
long.orth.2.glmer     11 1041.0 1097.6 -509.51   1019.0 3.3909  2      0.1835

```

Second, we compare the model `long.orth.min.glmer` (just random intercepts) with `long.orth.3.glmer` to check if increasing model complexity, by accounting for random differences between participants or words in the slope of the Instructions effect improves model fit to data.

```
anova(long.orth.min.glmer, long.orth.3.glmer)
```

```

Data: long.orth
Models:
long.orth.min.glmer: Score ~ Time + Orthography + Instructions + zConsistency_H + Orthography:
long.orth.3.glmer: Score ~ Time + Orthography + Instructions + zConsistency_H + Orthography:
npar      AIC      BIC  logLik deviance Chisq Df Pr(>Chisq)
long.orth.min.glmer    9 1040.4 1086.7 -511.20   1022.4
long.orth.3.glmer     12 1036.5 1098.2 -506.24   1012.5 9.9115  3      0.01933 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

I think this justifies reporting the model `long.orth.3.glmer`

If you look at the outputs, the addition of the random effects of participants and words on the slope of the effect of Orthography cannot be justified ($\chi^2 = 3.3909, 2df, p = .1835$) by improved model fit, in comparison to a model with the random effects of participants and words on intercepts.

The model comparison summary indicates that the addition of a random effect of words on the slope of the Instructions effect is justified by significantly improved model fit to data ($\chi^2 = 9.9115, 3df, p = 0.01933$).

Notice I take a bit of a short-cut here, by adding both random effects for Orthography and Instructions. Logically, if adding the random effect for Orthography does not improve model fit but adding the random effects for both random effects for Orthography and Instructions does then it is the addition of Instructions that is doing the work. However, I should prefer to see as complex a random effects structure as possible, provided a model converges.

While adding the random effect of Orthography does not improve model fit significantly, you see researchers allowing a generous p-value threshold for inclusion of terms (i.e. it is ok to add

variables up to where $p < .2$). Matuschek et al. (2017) argue that when we are engaged in model selection – here, this is what we are doing because we are trying to figure out what model (with what random effects) we should use – then we should resist the reflex to choose the model that seems justified because the LRT $\chi^2, p < .05$. The χ^2 alpha level cannot be interpreted as the expected model selection Type I error rate (in Null Hypothesis Significance Test terms) but rather as the relative weight of model complexity and goodness-of-fit (look back at the discussion of model comparison in the previous chapter). In this sense, setting a threshold such that we include an effect only if $\chi^2, p < .05$ will always tend to penalize model complexity, and tend therefore to lead us to choose simpler (perhaps too simple) models.

6.9.4 Addressing convergence problems

Sometimes (g)lmer() has difficulty finding estimates for effects in a model given a dataset. If it encounters problems, the problems are expressed as warnings about convergence failures. Convergence failures typically arise when the model is too complicated for the data (see the discussion in Bates et al., 2015; Eager & Roy, 2017; Matuschek et al., 2017; Meteyard & Davies, 2020). As we have seen, problems can occur if you are trying to estimate (or predict) random effects terms that are very small – that do not really explain much variance in performance. Problems can also occur if variables have very large or very small ranges.

We can detect and address these problems in a number of ways:

- Sometimes convergence problems can be fixed by switching the optimizer used to fit the model – we can do this by adding the argument: `glmerControl(optimizer = "bobyqa")` to the `glmer()` function call, as I did for the class example. Switching optimizers is a quick solution to a common problem: models can fail to converge for a number of different reasons. In short, there may not be enough data for the model fitting process to settle on appropriate estimates for fixed or random effects.
- Sometimes, a warning message advises us to consider rescaling the continuous numeric predictor variable. For this reason, and others, I usually standardize numeric predictor variables, as a default, before the analysis.
- Sometimes, the warnings tell us that we need to simplify the random effects part of the model. We can simplify the random effects structure of a mixed-effects model in a number of ways: As we examine the estimates resulting from a model fit we can consider whether the variance and covariance terms are small or large. Ultimately, I decided that the effects of subjects and items on intercepts was important, as was, to some extent, the effect of words on the slope of the effect of Instructions.

The approach we have progressed through is widely used (see discussions in Baayen et al., 2008; Barr, et al., 2013; Luke, 2017; Matuschek et al., 2017; Meteyard & Davies, 2019).

See, especially, the troubleshooting guide by Ben Bolker:

https://rstudio-pubs-static.s3.amazonaws.com/33653_57fc7b8e5d484c909b615d8633c01d51.html

6.9.4.1 Exercises

To demonstrate the impact of these adjustments, you could refit the models discussed in the foregoing:

1. Without using the standardized consistency variable as a predictor
2. Without using the modification to the model fitting code, i.e. deleting the line that includes `glmerControl(optimizer = "bobyqa")`
3. Making the last most complex model more complex by adding further random effects of subjects or items on the slopes of both main effects and the interaction

6.9.4.2 Summary advice

My advice, then, is to consider whether random effects should be included in a model based on

1. Theoretical reasons, in terms of what your understanding of a study design allows and requires, with respect to random differences between groups (classes, participants, stimuli etc.) or stimuli;
2. Model convergence, as when models do or do not converge;
3. Over a series of model comparisons, an evaluation of whether model fit is improved by the inclusion of the random effect.

This sounds like it involves work, judgment, and a process. It also sounds like people may disagree on the judgment or the process so that you shall have to share data and code, to enable others to check if the results vary depending on different decisions or different approaches. And it sounds like you will need to not only figure out what to do but also justify the approach you take when you report the results. I think all these things are true.

This is why Lotte Meteyard and I advise that researchers need to explain their approach, and share their data and code, when they report their analyses. Our best practice guidance for reporting mixed-effects models includes, among other things, the advice that in reports ...

1. Random effects are explicitly specified according to sampling units (e.g., participants, items), the data structure (e.g., repeated measures) and anticipated interactions between fixed effects and sampling units (e.g., intercepts only or intercepts and slopes). Fixed effects and covariates are specified from explicitly stated research questions and/or hypotheses.

2. Report the size of the sample analysed in terms of total number of data points and of sampling units (e.g., number of participants, number of items, number of other groups specified as random effects, such as classes of children).
3. A clear statement of the methods by which models are compared/selected; e.g., simple to complex, covariates first, random effects first, fixed effects first etc.
4. Report comparison method (LRT, AIC, BIC) and justify the choice.
5. A complete report of all models compared (e.g., in appendices/supplementary data/analysis scripts) with model equations and the result of comparisons.
6. If models fail to converge, the approach taken to manage this should be comprehensively reported. This should include the formula for each model that did or did not converge and a rationale for a) the simplification method used and b) the final model reported. This may be most easily presented in an analysis script.

This looks like a lot of work.

Why bother?

I think it is always worth asking this question.

The first answer is that it is all relative. In my own experience, a lot of the effort spent in the research workflow used to be occupied by face-to-face data collection: weeks or months of testing; now all data get collected online, and it gets finished overnight. Considerable time and effort (as Hadley Wickham's joke runs, 80% of analysis effort) was spent on tidying the data before analysis; I still do this work but it is now much faster and less effortful, thanks to `tidyverse`. A lot of effort used to be spent by me or colleagues on the literature review, the power analysis, or the stimulus preparation: that still happens. And a lot of effort used to be spent on doing the analysis and figuring out what the results mean: that, too, still happens. It is up to you if you want to spend ten months on data collection and five minutes on data analysis (as another joke has it, a million bucks on the data and a nickel on the statistics).

I think we *do* need to work at understanding the most appropriate analysis for our data, based on both our theoretical expectations and a data-driven evaluation. No-one is going to help us unsee multilevel structure in the data, or save us from the obligation to take into account random effects. Matuschek et al. (2017; p.312) argue that “The goal of model selection is not to obtain a significant p-value; the goal is to identify the most parsimonious model that can be assumed to have generated the data.” This makes sense to me. And their analyses show that determining a parsimonious model with a standard model selection criterion is a defensible choice, a way to take into account random effects, while controlling for both the risk of false positives, and the risk of false negatives.

6.10 Reporting model results

We have discussed how to report the results of mixed-effects models previously. The same conceptual structure, and similar language, can be used to report the results of Generalized Linear Mixed-effects Models (GLMMs).

I think you need to think about reporting the analysis as a task in which you first prepare the reader, explaining the motivation for using GLMMs, then present the analysis you did (the process, in outline), then present the results you shall later discuss.

- Start by explaining the study design – outline the fixed effects that have to be estimated
- Explain how random effects structure was selected – be prepared to present a short version of the story in the main part of the report – sharing your code, in an appendix, to illustrate the steps in full

You can see that I followed this progression of steps in the report I wrote for the published version of the analysis discussed, as an example, for this class (Monaghan et al., 2015; see following).

Lotte Meteyard and I recommend that results reporting should:

- Provide equation(s) that transparently define the reported model(s). An elegant way to do this is providing the model equation with the table that reports the model output.
- And that final model(s) should be reported in a table that includes all parameter estimates for fixed effects (coefficients, standard errors and/or confidence intervals, associated test statistics and p-values if used), random effects (standard deviation and/or variance for each random effect, correlations/covariances if modelled) and some measure of model fit (e.g. R-squared, correlation between fitted values and data).
- While researchers should be able to share the coding script used to complete the analysis and, wherever possible, share data that generated the reported results.

For the word learning study we have been working through, the Results section for the report would include the following elements:

Explain approach We used mixed-effects models to analyse data because this approach permits modelling of both participant- and item-level variability simultaneously, unlike more traditional approaches such as ANOVA. In this study, multiple participants responded to multiple items, meaning that both participants and items were sources of nonindependence in our data (i.e. responses from the same participant are likely to be correlated, as are responses to the same item). Compared to ANOVA, mixed-effects models offer a more flexible approach, and are better able to handle missing data without significant loss of statistical power (Baayen, Davidson, & Bates, 2008).

Explain how you get from the study design to the model you use to test or estimate key effects

We took a hypothesis driven approach, estimating the fixed effects of time (Time 1 versus Time 2), Orthography (absent versus present), Instructions (incidental versus explicit) and consistency (standardized H), as well as the interaction between orthography and instructions and the interaction between orthography and consistency. Different levels of the three binary fixed effects were sum coded. Consistency H, as a numeric predictor variable, was standardized to z scores before entry to models as a predictor.

Outline the model comparison or model selection work The models were initially fitted specifying just random effects to account for variation by participants and stimuli in accuracy (random intercepts) plus terms to estimate the fixed effects of the experimental conditions ([name them]), and the interactions [name them]. Following the recommendations of Barr, Levy, Scheepers, and Tily (2013; see also Baayen, 2008; Matuschek et al., 2017), we fitted further models adding both random intercepts and random slopes for the random effects. Likelihood ratio test comparison of models showed that a model with both random intercepts and slopes ... fit the data better than a model with just random intercepts ($\chi^2(df) = ..., p =)$.

Use appendices or supplementary materials To give the reader full information on models fit, model comparisons

Help the reader with a concise summary of estimates As I have advised for reporting linear models, I included a tabled summary of coefficient estimates, presenting fixed and random effects (see e.g. Davies et al., 2013; Monaghan et al., 2015)

Show and tell Use figures – model prediction plots, as seen – to help the reader to see what the fixed effects estimates imply.

Which model do we report?

Note that given the model comparison results we have seen, I would probably report the estimates from `long.orth.3.glmer`. The model appears to include the most comprehensive account of random effects while still being capable of converging.

6.11 Summary

We focused on the need to use Generalized Linear Mixed-effects Models (GLMMs). We identified the kind of outcome data (like response accuracy) that requires analysis using GLMMs. Alternative methods, and their limitations, were discussed.

We examined a study that incorporates repeated measures (participants respond to multiple stimuli), a 2 x 2 factorial design, and a longitudinal aspect (participants tested at two time points), the word learning study (Ricketts et al., in press).

We discussed the need to use effect coding for categorical predictor variables (factors). We work through example code to set factor level coding as required.

We worked through a random intercepts GLMM, and identified the critical elements of the model code, and of the results summary, including hypothesis test p-values. We examined how to present visualizations of fixed effects estimates (model predictions) using different libraries.

We then moved on to considering the question of what random effects we should include in the model. We considered the study design in some depth, and explored what random effects we could, in theory, expect to require. We then worked through a model comparison approach. We looked at some warning signs, what they indicate, and how to deal with them.

We considered how to report the model selection (or comparison, or building) process, and how to report the model for presentation of results.

6.11.1 Useful functions

We used two functions to fit and evaluate mixed-effects models.

- We used `glmer()` to fit a mixed-effects model
- We used `anova()` to compare two or more models using AIC, BIC and the Likelihood Ratio Test

6.12 R code and data file access for the class

Activities in the class that goes with this chapter are associated with the following data file and .R code file:

- `04-glmm-workbook.R`
- `long.orth_2020-08-11.csv`

You can download the data and the .R code files from here:

<https://modules.lancaster.ac.uk/mod/resource/view.php?id=1809329>

Run the code in the .R file to reproduce the results presented in this chapter and in the slides.

You can also see resources that you can use, optionally, to extend your practice and deepen your understanding, by exploring the analysis of the noun-verb learning “Gavagai” study data (Monaghan et al., 2015). These resources can be downloaded as part of the same folder and comprise a dataset with an extensively commented .R analysis script.

- `402-04-GLMM-exercise-gavagai-data-analysis-notes.R`

- noun-verb-learning-study.csv

6.13 References

6.13.1 Recommended reading

The example studies referred to in this chapter are published in (Monaghan et al., 2015; Ricketts et al., in press).

Ben Bolker provides a very readable introduction to Generalized Linear Mixed-effects Models (Bolker et al., 2009; see also Jaeger, 2008).

Baayen et al. (2008; see, also, Barr et al., 2013) discuss mixed-effects models with crossed random effects.

The issue of model comparison or model selection, and the appropriate choice of random effects structure is discussed by Baayen (Baayen et al., 2008; Barr et al., 2013; Bates et al., 2015; Eager & Roy, 2017; Matuschek et al., 2017).

I wrote a tutorial article on mixed-effects models with Lotte Meteyard (Meteyard & Davies, 2020). We discuss how important the approach now is for psychological science, what researchers worry about when they use it, and what they should do and report when they use the method.

Book length introductions are provided by Snijders and Bosker (2012; good introduction in simple language), Gelman and Hill (2006; very influential), and Pinheiro and Bates (2000; critical for lme4).

6.13.2 A very useful FAQ

Can be found here:

<https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html>

6.13.3 References list

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R.* CUP.

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, 390-412.

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences.* Macmillan International Higher Education.

- Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39, 445– 459. <http://dx.doi.org/10.3758/BF03193014>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255-278.
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.
- Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24, 127–135.
- Chambré, S. J., Ehri, L. C., & Ness, M. (2017). Orthographic facilitation of first graders' vocabulary learning: does directing attention to print enhance the effect? *Reading and Writing*, 1-20. doi:10.1007/s11145-016-9715-z
- Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Colenbrander, D., Miles, K. P., & Ricketts, J. (2019). To See or Not to See: How Does Seeing Spellings Support Vocabulary Learning? *Language, Speech, and Hearing Services in Schools*, 50(4), 609-628. doi:10.1044/2019_LSHSS-VOIA-18-0135
- Eager, C., & Roy, J. (2017). Mixed effects models are sometimes terrible. *arXiv preprint*
- Ehri, L. C. (2014). Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning. *Scientific Studies of Reading*, 18(1), 5-21. doi:10.1080/10888438.2013.819356
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511790942>
- 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.
- Krepel, A., de Bree, E. H., & de Jong, P. F. (2020). Does the availability of orthography support L2 word learning? *Reading and Writing*. doi:10.1007/s11145-020-10078-6
- Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior research method, 49, 1494–1502.
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305-315.
- Mengoni, S. E., Nash, H., & Hulme, C. (2013). The benefit of orthographic support for oral vocabulary learning in children with Down syndrome. *Journal of Child Language*, 40 (Special Issue 01), 221-243. doi:10.1017/S0305000912000396

Monaghan, P., Mattock, K., Davies, R. A., & Smith, A. C. (2015). Gavagai is as gavagai does: Learning nouns and verbs from Cross-Situational statistics. *Cognitive Science*, 39, 1099-1112.

Mousikou, P., Sadat, J., Lucas, R., & Rastle, K. (2017). Moving beyond the monosyllable in models of skilled reading: Mega-study of disyllabic nonword reading. *Journal of Memory and Language*, 93, 169-192. doi:<http://dx.doi.org/10.1016/j.jml.2016.09.003>

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer-Verlag.

Raijmakers, 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, 416-426.

Ricketts, J., Dawson, N., & Davies, R. (2021). The hidden depths of new word knowledge: Using graded measures of orthographic and semantic learning to measure vocabulary acquisition. *Learning and Instruction*, 74, 101468.

Snijders, T.A., & Bosker, R.J. (2012). *Multilevel analysis (2nd Edition)*. London, UK: Sage.

6.14 Appendix: Example dataset variable information

Further information about the variables in the `long.orth_2020-08-11.csv` dataset.

Participant Cell values comprise character strings coding for participant. Participant identity codes were used to anonymise participation. Children included in studies 1 and 2 – participants in the longitudinal data collection – were coded "EOF[number]". Children included in Study 2 only (i.e., the older, additional, sample) were coded "ND[number]".

Time Test time was coded 1 (time 1) or 2 (time 2). For the Study 1 longitudinal data, it can be seen that each participant identity code is associated with observations taken at test times 1 and 2.

Study Observations taken for children included in studies 1 and 2 – participants in the longitudinal data collection – were coded "Study1&2". Children included in Study 2 only (i.e., the older, additional, sample) were coded "Study2".

Instructions Variable coding for whether participants undertook training in the *explicit* or *incidental* conditions.

Version Experiment administration coding

Word Letter string values show the words presented as stimuli to children.

Consistency-H Calculated orthography-to-phonology value for each word.

Orthography Variable coding for whether participants had seen a word in training in the orthography *absent* or *present* conditions.

Measure Variable coding for the post-test measure: `Sem_all` if the semantic post-test; `Orth_sp` if the orthographic post-test.

Score Variable coding for response category. For the semantic (sequential or dynamic) post-test, responses were scored as corresponding to:

- 3 – correct response in the definition task
- 2 – correct response in the cued definition task
- 1 – correct response in the recognition task
- 0 – if the item wasn't correctly defined or recognised

For the orthographic post-test, responses were scored as:

- 1 – correct, if the target spelling was produced in full
- 0 – incorrect

WASI_mRS Raw score – Matrix Reasoning subtest of the Wechsler Abbreviated Scale of Intelligence

TOWREsweRS Raw score – Sight Word Efficiency (SWE) subtest of the Test of Word Reading Efficiency; number of words read correctly in 45 seconds.

TOWREpdeRS Raw score – Phonemic Decoding Efficiency (PDE) subtest of the Test of Word Reading Efficiency; number of nonwords read correctly in 45 seconds.

CC2regRS Raw score – Castles and Coltheart Test 2; number of regular words read correctly

CC2irregRS Raw score – Castles and Coltheart Test 2; number of irregular words read correctly

CC2nwRS Raw score – Castles and Coltheart Test 2; number of nonwords read correctly

WASI_vRS Raw score – vocabulary knowledge indexed by the Vocabulary subtest of the WASI-II

BPVSRS Raw score – vocabulary knowledge indexed by the British Picture Vocabulary Scale – Third Edition

Spelling.transcription Transcription of the spelling response produced by children in the orthographic post-test

Levenshtein.Score Children were asked to spell each word to dictation and spelling productions were transcribed for scoring. Responses were scored using a Levenshtein distance measure, using the `stringdist` library (van der Loo, 2019). This score indexes the number of letter deletions, insertions and substitutions that distinguish between the target and child's response. For example, the response 'epegram' for target 'epigram' attracts a Levenshtein score of 1 (one substitution). Thus, this score gives credit for partially correct responses, as well as entirely correct responses. The maximum score is 0, with higher scores indicating less accurate responses.

zTOWREsweRS We standardized TOWREsweRS values, calculating the z score as $z = \frac{x - \bar{x}}{sd_x}$, over all observations in the *longitudinal* (Study 1) or *concurrent* (Study 2) data-set, using the “`scale()`“ function in R.

zTOWREpdeRS Standardized TOWREpdeRS scores

zCC2regRS Standardized CC2regRS scores

zCC2irregRS Standardized CC2irregRS scores

zCC2nwRS Standardized CC2nwRS scores

zWASIvRS Standardized WASIvRS scores

zBPVSRs Standardized BPVSRs scores

Part III

WRITING ABOUT RESEARCH

7 Introduction: the why

The research report assignment requires students to locate, access, analyse and report previously collected data. This introduction is intended to answer the first question anybody might ask.

- Why: what is the motivation for the assignment?

In following materials, I will answer the questions.

- How can the assignment be done?
- What do we expect students to do?

It is going to appear, at first, that I am going a *long* way away from telling you what you need to do for the assignment. I hope you will agree that the discussion that follows is worth your time in reading it. It will help you to understand *why* we are asking you to do the assignment, and *why* we are looking for what we are looking for. It will help you to understand *how* this work will aid your development. And it will help to show *how* doing the assignment furnishes the opportunity for research experience that will help you later in your working life.

For those who are more eager to start the work, here are the links to the what information in Chapter 8 and to the how information in Chapter 9.

7.1 The key ideas

There are two ideas motivating our approach. It will be helpful to you if I sketch them out early, here. We can demonstrate the usefulness of these ideas as we progress through our work.

The first key idea is expressed clearly in sociological discussions of science. This is that there is a difference between science “...being done, science in the making, and science already done, a finished product ...” [Bourdieu (2004); p.2]. The awareness we want to develop is that there are two things: there is the story that may be presented in a textbook or in a lecture about scientific work or scientific claims; and there is the work we do in practice, as we develop graduate skills, and as we exercise those skills professionally in the workplace.

The second key idea connects to the first. This idea is that reported analyses are not *necessary* or *sufficient* to the data or the question. What does this mean? It means that the same data

can reasonably be analysed in different ways. There is no *necessary* way to analyse some data though there may be conventions or normal practices (Kuhn, 1970). It means that it is unlikely that any one analysis will do all the work that could be done (a sufficiency) to get you from your data to useful or reasonable answers to your questions.

These ideas may be unsettling but they are realistic. Stating them will better prepare you for professional work. In the workplace, the accuracy of these ideas will emerge when you see how a team in any sector (health, marketing ...) gets from its data to its product. If we talk about the ideas now, we can get you ready for dealing with the practical and the ethical concerns you will confront when that happens.

We will begin by discussing psychological research, and research *about* psychological research, to answer the question: **Why: what is the motivation for the assignment?** We will then move to answering the **what** and the **how** questions.

7.2 Why: what is the motivation for the assignment?

7.2.1 The wider context: crisis and revolution

We are here because we are interested in humans and human behaviour, and because we are interested in scientific methods of making sense of these things. Some of us are aware that science (including psychological science) has undergone a rolling series of crises: the replicability or replication crisis (Pashler & Harris, 2012; Pashler & Wagenmakers, 2012); the statistical crisis (A. Gelman & Loken, 2014b); and the generalizability crisis (Yarkoni, 2022). And that science is undergoing a response to these crises, evidenced in the advocacy of pre-registration (Nosek et al., 2018, 2019), and of registered reports (Nosek & Lakens, 2014), the use of open science badges (e.g., [for the journal Psychological Science](#)), the completion of large-scale replication studies (Aarts et al., 2015), and the identification of open science principles (Munafò et al., 2017). We may usefully refer, collectively, to the crises and the responses, as the *credibility revolution* (Vazire, 2018)

We could teach a course on this (in Lancaster, we do) but I must be brief, here, and invite you to follow the references, if you are interested. Before going on, I want to call your attention to the fact that important elements of the hard work in trying to make science work better has been led by PhD students and by junior researchers (e.g., Herndon et al., 2014). Graduate students may, at first, assume that the fact that a research article has been published in a journal means the findings that are reported must be *true*. Most of the time, some educated skepticism is more appropriate. An important driver of the realization that there are problems evident in the literature, and that there are changes we can make to improve practice, comes from independent **post-publication review work** exposing the problems in published work (see, e.g., [this account by Andrew Gelman](#))

Tip

- Allow yourself to feel skeptical about the reports you read *then* work with the motivation this feeling provides.

In brief, then, most practicing scientists now understand *or should* understand that many of the claims we encounter in the published scientific literature are unlikely to be supported by the evidence (Ioannidis, 2005), whether we are looking at the evidence of the results in the reports themselves, or evidence in later attempts to find the same results (e.g., Aarts et al., 2015). We suspect that this may result from a number of causes. We understand that researchers may engage in questionable research practices (John et al., 2012). We understand that researchers may exploit the potential for flexibility in doing and reporting analyses (Simmons et al., 2011a). We understand that there are problems in how psychologists use or talk about the measurement of psychological constructs (Flake & Fried, 2020). We understand that there are problems in how psychologists sample people for their studies, both in where we recruit (Bornstein et al., 2013; Henrich et al., 2010; Wild et al., 2022), and in how many we recruit (Button et al., 2013; Cohen, 1962; Sedlmeier & Gigerenzer, 1989; Vankov et al., 2014). We understand that there are problems in how psychologists specify or think about their hypotheses or predictions (Meehl, 1967; Scheel, 2022). And we understand that there are problems in how scientists do, or rather do not, comply with good practice recommendations designed to fix these problems (discussed further in the following).

This discussion could (again) be unsettling. This list of problems could make you angry or sad. I, like others, think it is exciting. It is exciting because these problems have probably existed for a long time (e.g., Cohen, 1962; Meehl, 1967) and now, having identified the problems, we can hope to do something about it. It is exciting because if you care about people, the study of people, or the applications in clinical, education and other domains of the results of the study of people, then you might hope to see better, more useful, science in the future (Vazire, 2018).

As someone who teaches graduate and undergraduate students, I want to help you to *be the change you want to see in the world*¹. We cannot solve every problem but we can try to do better those things that are within our reach. I am going to end this introduction with a brief discussion of some ideas we can use to guide our better practices.

7.2.2 The specific context: what we need to look at, conceptually and practically

In this course, for this assignment, we are going to focus on:

1. multiverse analyses
2. kinds of reproducibility

¹This encouragement is often attributed to Gandhi but is attributed ([\(here\)](#)) to a Brooklyn school teacher, Ms Arleen Lorrance, who led a transformative school project in the 1970s.

3. the current state of the match between open science ideas and practices

In the classes on the linear model, we will discuss:

4. the links between theory, prediction and analysis
5. psychological measurement
6. samples
7. variation in results

7.2.3 Multiverse analyses: multi- what?

7.2.3.1 A first useful metaphor: the pipeline

I am going to link this discussion to a metaphor (see Figure Figure 7.1) or a description you will find useful: **the data analysis pipeline** or **workflow**.

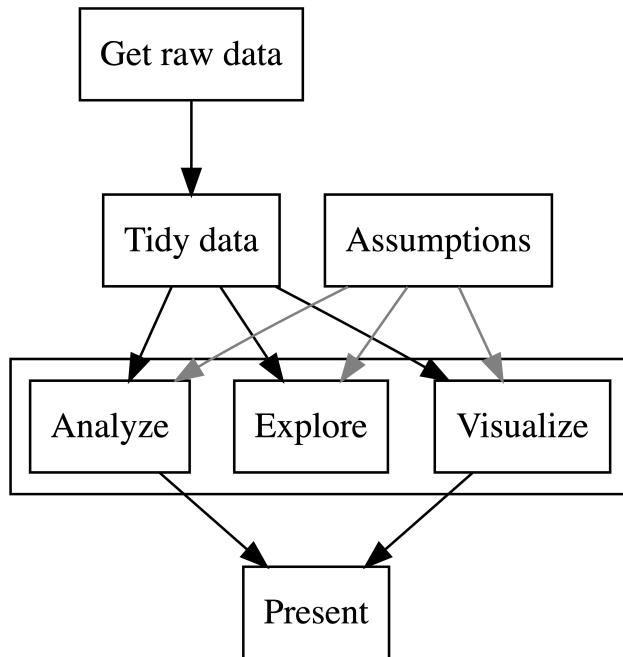


Figure 7.1: The data analysis pipeline or workflow

This metaphor or way of thinking is very common ([take a look at the diagram in Wickham and Grolemund's 2017 book "R for Data Science"](#)) and you may see the words "data pipeline" used in job descriptions, or you may benefit from saying, in a job application, something like: *I am skilled in designing and implementing each stage of the quantitative data analysis pipeline,*

from data tidying to results presentation. I say this because scientists I have mentored got their jobs because they can do these things – and successfully explained that they can do these things – in sectors like educational testing, behavioural analysis, or public policy research.

The reason this metaphor is useful is that it helps us to organize our thinking, and to manage what we do when we do data analysis, we:

- get some data;
- process or tidy the data;
- explore, visualize, and analyze the data;
- present or report our findings.

We introduce the idea that your analysis work will flow through the stages of a *pipeline* from getting the data to presenting your findings because, next, we will examine how pipelines can *multiply*.

💡 Tip

- As you practice your data analysis work, try to identify the elements and the order of your work, as the parts of a *workflow*.

7.2.3.2 A second useful metaphor: the garden of forking paths

What researchers have come to realize: because we started looking ... The open secret that has been well kept (Bourdieu, 2004): because everybody who does science knows about it, yet we may not teach it; and because we do not write textbooks revealing it ... Is that at each stage in the analysis workflow, we can and do make choices where multiple alternative choices are possible. A. Gelman & Loken (2014a) capture this insight as the “garden of forking paths”² (see Figure 7.2).

The general idea is that it is possible to have **multiple potential different paths** from the data to the results. The results will vary, depending on the path we take. In an analysis, we could take multiple different paths simply because at point A we decide to do B1, B2 or B3, maybe we choose B1, and then at point B1, we may decide to do C1, C2 or C3. Here, maybe we have our raw data at point A. Maybe we could do one of two different things when we tidy the data: action B1 or B2. Then, when we have our tidy data, maybe we can choose to do our analysis in one of six ways. Where we are at each step *depends* on the choices we made at the previous steps.

In the end, it may appear to us that we took one path or that **only one path was possible**. When we report our analysis, in a dissertation or in a published journal article, we may report the analysis **as if only one analysis path had been considered**. But, critically, our

²The term is taken from the name of a short story by Jorge Luis Borges, “El jardín de senderos que se bifurcan”.

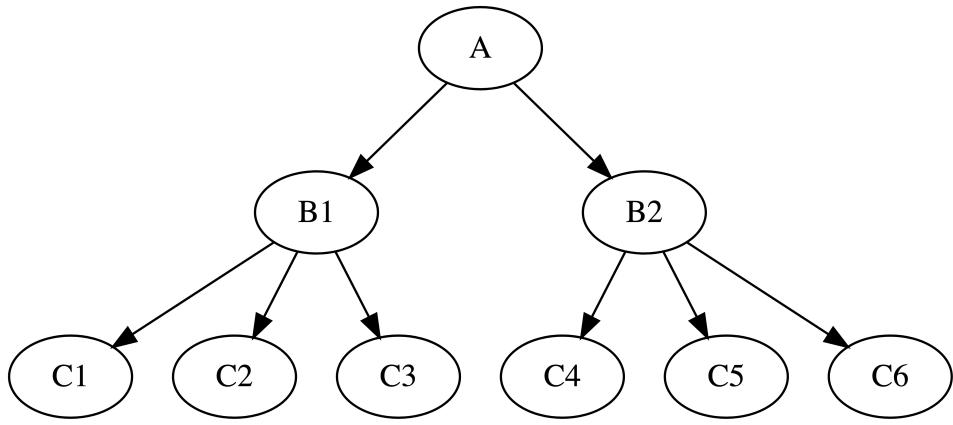


Figure 7.2: Forking paths in data analysis

findings may depend on the choices we made and this variation in results may be hidden from view.

I am talking about forking paths because the *multiplicity* of paths has consequences, and we discuss these next.

 Tip

- It is about here, I hope, that you can start to see why it would make sense to access data from a published study and to examine if you can get the same results as the study authors.

7.2.4 Multiverse analyses

I am going to discuss, now, what are commonly called *multiverse analyses*. Psychologists use this term, having been introduced to it in an influential paper by Steegen et al. (2016a), but it comes from theoretical physics ([take a look at wikipedia](#)).

I explain this because I do not want you to worry. The ideas themselves are within your grasp whatever your background in psychology or elsewhere. It is the implications for our data analysis practices that are *challenging*. They are challenging because what we discuss should increase your skepticism about the results you encounter in published papers. And they are challenging because they *reveal your freedom* to question whether published authors could have done their analysis in a different way.

We are going to look at:

1. dataset construction
2. analysis choices

7.2.4.1 The link between the credibility revolution and the multiverse

In first discussing the wider context (of crisis and revolution), then discussing the specific context (of multiverses and, in the following, of reproducibility), I should be clear about **the link between the two things**. The finding that some results may not be supported by the evidence is probably due to a mix of causes. But one of those causes will be the combination of uncertainty over data processing or the uncertainty over analysis methods revealed in multiverse analyses, as we see next, combined with the limitations of data and code sharing, and the incompleteness of results reporting (as we see later).

7.2.4.2 The data multiverse

When you collect or access data for a research study, the complete raw dataset you receive is almost never the complete dataset you analyze or whose analysis you report. This is not a story about deliberately cheating. It is a story about the normal practice of science (Kuhn, 1970).

Picture some common scenarios. You did a survey, you got responses from a 100 participants on 10 questions, and you asked people to report their education, ethnicity and gender. You did an experimental study, you tested two groups of 50 people each in 100 trials (imagine a common task like the Stroop test), and you observed the accuracy and the timing of their responses. You tested 100 children, 20 children in each of five different schools, on a range of educational ability measures.

In these scenarios, the psychologist or the analyst of behavioural data *must* process their data. In doing so, you will ask yourself a series of questions like:

- how do we code for **gender, ethnicity, education**?
- what do we do about reaction times that are very short, e.g., $RT < 200ms$ or very long, e.g., $RT > 1500ms$?
- if we present multiple questions measuring broadly the same thing (e.g. how confident are you that you understand what you have read? how easy did you find what you read?) how do we summarize the scores on those questions? do we combine scores?
- what do we do about people who may not appear to have understood the task instructions?

Typically, the answers to these questions will be given to you by your supervisor, a colleague or a textbook example. For example, we might say:

- “We excluded all reaction times greater than 1500ms before analysis.”

Typically, the explanation for these answers are rarely explained. We might say:

- “*Consistent with common practice in this field*, we excluded all reaction times greater than 1500ms before analysis.”

But the reader of a journal article typically **will not see** an explanation for why, as in the example, we exclude reaction times greater than 1500ms and not 2000ms or 3000ms, etc. We typically do not see an explanation for why *we* exclude all reaction times greater than 1500ms but *other* researchers exclude all reaction times greater than 2000ms. (I do not pick this example at random: there are serious concerns about the impact on analyses of exclusions like this (Ulrich & Miller, 1994).)

What Steegen et al. (2016a) showed is that a dataset can be processed for analysis in multiple different ways, with a number of reasonable alternate choices that can be applied, for each choice point: construction choices about classifying people or about excluding participants given their responses. If a different dataset is constructed for each combination of alternatives then many different datasets can be produced, all starting from the same raw data. (For their example study, Steegen et al. (2016a) found they could construct 120 or 210 different datasets, based on the choice combinations.) Critically, for us, Steegen et al. (2016a) showed that if we apply the same analysis method to the different datasets then our results will vary.

Let me spell this out, bit by bit:

- we approach our study with the same research question, and the same verbal prediction;
- we begin with the exact same data;
- we then construct different datasets depending on different *but equally reasonable* processing choices;
- we then apply the same analysis analysis, to test the same prediction, using each different dataset;
- we will see different results for the analyses of the different datasets.

Alternate constructions of the same data may cause variation in the results of statistical tests. Some kinds of data processing choices may be more influential on results than others. It seems unlikely that we can identify, in advance, which choices matter more.

Steegen et al. (2016a) suggest that we can *deflate* (shrink) the multiverse in different ways. I want to state their suggestions, here, because we will come back to these ideas in the classes on the linear model.

1. Develop better theories and improved measurement of the constructs of interest.
2. Develop more complete and precise theory for why some processing options are better than others.

But you will be asking yourself: **what do I need to think about, for the assignment?**

💡 Tip

- When you read a psychological research report, identify where the researchers talk about how they process their data: classification, coding, exclusion, transformation, etc.
- If you can access the raw data, ask yourself: could different choices change the results of the same analysis?

7.2.4.3 Analysis multiverses

Even if we begin with the same research question and, critically, the *same dataset*, the results of a series of studies show that different researchers will often (reasonably) make *different choices about the analysis* they do to answer the research question. We often call these studies (analysis or model) **multiverse** studies. In these studies, we see variation in analysis and this variation is also associated with variation in results.

An influential example, in psychology, is reported by Silberzahn and colleagues (Silberzahn et al., 2017; Silberzahn & Uhlmann, 2015) who asked 29 teams of researchers to answer the same question (“Are (soccer) referees more likely to give red cards to players with dark skin than to players with light skin?”) with the same dataset (data about referee decisions in football league games). The teams made their own decisions about how to answer the question in doing the analysis. The teams shared their plans, and commented on each others’ ideas. The discussion did not lead to a consensus about what analysis approach is best. In the end, the different teams did different analyses and, critically, the **different analyses had different results**. The results varied in whether the test of the effect of players skin colour (on whether red cards were given) was significant or not, and on the strength of the estimated association between the darkness of skin colour (lighter to darker) and the chances (low to high) of getting a red card.

There have now been a series of multiverse or multi-analyst studies which demonstrate that, under certain conditions, different researchers may adopt different analysis approaches – which will have *different results* – in answering the same research question with the same data. This demonstration has been repeated in studies in health, medicine, psychology, neuroscience, and sociology, among other research fields (e.g., Parsons (n.d.); Breznau et al. (2022); Klau et al. (n.d.); Klau et al. (2021); Wessel et al. (2020); Poline et al. (2006); Maier-Hein et al. (2017); Starns et al. (2019); Fillard et al. (2011); Dutilh et al. (2019); Salganik et al. (2020); Bastiaansen et al. (2020); Botvinik-Nezer et al. (2020); Schweinsberg et al. (2021); Patel et al. (2015); see, for reviews, and some helpful guidance, Aczel et al. (2021); Del Giudice & Gangestad (2021); Hoffmann et al. (n.d.); Wagenmakers et al. (2022)).

In these studies, we typically see variation in how psychological constructs are operationalized (e.g., how do we measure or code for social status?), how data are processed or datasets constructed (as in Steegen et al. (2016b)), plus variation in *what* statistical techniques are

used, and in *how* those techniques are used. This variation can be understood to reflect kinds of **uncertainty** (Klau et al., n.d.; Klau et al., 2021): uncertainty about how to process data, and uncertainty about the model or methods we should use to test or estimate effects. Further research makes it clear that we should be aware, if we are not already, of the variation in results that can be expected because different researchers may choose to design studies, and construct stimulus materials, in different ways given the same research hypothesis information (Landy et al., 2020).

But you will be asking yourself: **what do I need to think about, for the assignment?**

 Tip

- When you read a psychological research report, identify where the researchers talk about how they analyse their data: the hypothesis or prediction they test; the method; their assumptions; the variables they include; the checks or the alternate analyses they did or did not do.
- If you can access the data and analysis code, ask yourself: could different methods change the results of the same analysis?

7.2.4.4 What can we conclude – the story so far?

This is a good place to look at what we have discussed, and present an evaluation of the story so far.

This is not a story where everybody or nobody is right or where everything or nothing is true³. Instead, we can be guided by the advice (Meehl, 1967; Scheel, 2022; Steegen et al., 2016a) that we should (1.) seek better and more complete theorizing about the constructs of interest and how we measure them, and (2.) seek more complete and more precise theory so that some options are theoretically superior than others, and should be preferred, when constructing datasets or specifying analysis methods.

Not all research questions and not all hypothesis information will allow an equally wide variety of potential reasonable approaches to the analysis. As Paul Meehl argued a long time ago (Meehl, 1967, 1978), and researchers like Anne Scheel (Scheel et al., 2021; Scheel, 2022) argued more recently, the complexity of the thing we study – people, and what they do – and the still early development of our understanding of this thing, mean that what we *want* but what we do not see, in psychology, are scientifically productive tests of *falsifiable theories*. (See, consistent with this perspective, discussions by Auspurg & Brüderl (2021) and by Del Giudice & Gangestad (2021) about the range of analysis possibilities that may or may not be allowed, in multiverse analyses, by more or less clear research questions or well-developed causal theories.)

³There could be a story where the hero (us) ultimately learns to reject binary (present, absent; significant, non-significant) choices, and embrace variation, or embrace uncertainty (a. Gelman, 2015; Vasishth & Gelman, 2021).

Our concern should not so much be with being able to do statistical analysis, or with finding significant or not significant results. It would be more useful to do analyses to test concrete, inflexible, precise predictions that *can be wrong*.

Nor is this a story, I think, about the potential for *cheating*. While we may refer to subjective choices or to researcher flexibility, the differences that we see do not resemble the *researcher degrees of freedom* (Simmons et al., 2011b) some may exploit, consciously or unconsciously, to change results to suit their aims. Instead, the multiverse results show us the impact of the reasonable differences in approach that different researchers may sensibly choose to take when they try to answer a research question with data.

Not all alternates, at a given point of choosing, in the data analysis workflow, will have equal impact. Work by Young (Young, 2018; Young & Holsteen, 2017) indicates that if we deliberately examine the impact of method or model uncertainty, over different sets of possible choices — about what variables or what observations we include in an analysis, for example — we may find that some results are robust to an array of different options, while other results are highly susceptible to different choices. This work suggests another way in which uncertainty about methods or variation in results can be turned into progress in understanding the phenomena that interest us: through systematic, informed, interrogation of the ways that results can vary.

In general, in science, the acceptance of research findings must always be negotiated (Bourdieu, 2004). Here, we see that the grounds of negotiation should often include an analysis of the impact on the value of evidence of the different analysis approaches that researchers can or do apply to the data that underly that evidence.

But you will be asking yourself: **what do I need to think about, for the assignment?**

Tip

- The results of multiverse analyses show us that if we see one analysis reported in a paper, or one workflow, that does not mean that only one analysis can reasonably be applied.
- If you read the methods or results section of a paper, you should reflect: what other analysis methods could be used here? How could variation in analysis method — in what or how you do the analysis — influence the results?

Making you aware of the potential for analysis choices is useful because developing researchers, including graduate students, are often not aware of the room for choice in the data analysis workflow. Developing researchers — you — may be instructed that “this is how we do things” or “you should follow what researchers did previously”. Following convention is not necessarily a bad thing: it is a feature of the normal practice of science (Kuhn, 1970). However, you can now see, perhaps, that there likely will be alternative ways to process or to analyse data than the approach a supervisor, lab or field normally adopts.

This understanding or awareness has three implications for practice, it means:

1. When we talk about the analysis we do, we should explain our choices.
2. We should check, or enable others to check, what impact making different choices would have on our results.
3. Most importantly: **we can allow ourselves the freedom to critically evaluate** the choices researchers make, even the choices researchers make in published articles.

7.2.5 From the multiverse to kinds of reproducibility

Multiverse analyses and post-publication analyses, in general, show that we can and should question or critically evaluate the analyses we encounter in the literature. This work can usefully detect problems in original published analyses (e.g., A. Gelman & Weakliem, 2009; Herndon et al., 2014; Wagenmakers et al., 2011). It can demonstrate where original published claims are or are not robust to variation of analysis method or approach.

Given these lessons, and the implications we have identified, we should expect or hope to see open science practices (Munafò et al., 2017; Nosek et al., 2022):

- share data and code;
- publish research reports in ways that enable others to check or query analyses.

As we discuss, following, these practices are now common but the quality of practice can sometimes be questioned. This matters for you because it makes it more challenging – in specific identifiable locations – to locate, access, analyse and report previously collected data.

The discussion of current practices identifies where or how the assignment may be more challenging, but also identifies some of the exact places where the assignment provides **a real opportunity to do original research work**.

First, I am going to introduce some ideas that will help you to think about what you are doing when you do this work. We focus on the concept of *reproducibility*.

Gilmore et al. (2017; following Goodman et al., 2016) present three kinds of reproducibility:

- methods reproducibility
- results reproducibility
- inferential reproducibility

In looking at reproducibility, here, we are considering how much, or in what ways, the results or the claims that are made in a published study can be found or *repeated* by someone else.

7.2.5.1 Methods reproducibility

As Gilmore et al. (2017) discuss, **methods reproducibility** means that another researcher should be able to get the same results if they use the same tools and analysis methods to analyse the same dataset [some researchers also refer to *analytic reproducibility* or *computational reproducibility*; see e.g. Crüwell et al. (n.d.); Hardwicke et al. (2018); Hardwicke et al. (n.d.); Laurinavichyute et al. (2022); Minocher et al. (n.d.)].

In neuroimaging, the multiplicity of possible implementations of the data analysis pipeline (Carp, 2012a), and the fact that important elements or information about the pipeline deployed by researchers may be missing from published reports (Carp, 2012b), can make it challenging to identify how results can be reproduced.

In psychological science, in evaluating reports of results from analyses of behavioural data collected through survey or experimental work, in principle, we should expect to be able to access the data collected by the study authors, follow the description of their analysis method, and reproduce the results they report.



Tip

- For an assignment in which we ask students to locate, access, analyse and report previously collected data, we are directly concerned with *methods reproducibility*.

7.2.5.2 Results reproducibility

Results reproducibility means that if another researcher completes a new study with new data they are able to get the same results as the results reported following an original study: this often referred to as *replication*. The replication studies that have been reported (e.g., Aarts et al., 2015), and continue to be reported (see, for example, the studies discussed by Nosek et al. (2022)), in the last several years, present attempts to examine the results reproducibility of published findings.

In the classes on the linear model, we will examine if similar or different results are observed in a series of studies using the same procedure and the same materials. We shall discuss, in those classes, in more depth, what results reproducibility (or study replication) can or cannot tell us about the behaviours that interest us.

7.2.5.3 Inferential reproducibility

Inferential reproducibility means that if a researcher repeats a study (aiming for results reproducibility) or re-analyzes an original dataset (aiming for methods reproducibility) then they can come to the same or similar conclusions as the authors of the report of an original study.

How is inferential reproducibility not methods or results reproducibility? Goodman et al. (2016) explain that researchers can make the same conclusions from different sets of results and *can reach different conclusions from the same set of results*.

How is it possible to reach different conclusions from the same results? We can imagine two scenarios.

First, we have to think about the wider research field, the research context, within which we consider a set of results. It may be that two different researchers will come to look at the same results with different expectations about what the results *could* tell us (in Bayesian terms, with different prior expectations). Given different expectations, it is easy to imagine different researchers looking at the same results and, for example, one researcher being more skeptical than another about what conclusion can be taken from those results. (In the class on graduate writing skills, I discuss in some depth the importance of reviewing a research literature in order to get an understanding of the assumptions, conventions or expectations that may be shared by the researchers working in the field.)

Second, imagine two different researchers looking at the same results — picture the original authors of a published study, and someone doing a post-publication re-analysis of their data — you can expect that the re-analysis or the reproducibility analysis could identify reasons to value the evidence differently, or to reach more skeptical conclusions, through critical evaluation of:

- data processing choices;
- the choice of the method used to do analysis;
- choices in how the analysis method is used.

Where that critical evaluation involves an analysis of the choices the original researchers made, perhaps involving an analysis of other choices they could have made, perhaps reflecting on how effectively the analyses address a given research question or test a given prediction.

Tip

- We can think about the work we do, when we analyse previously reported data, in terms of the need to identify the *reproducibility* of results, methods and inferences.
- In psychological science, determining that someone can get the same results, by analyzing the same data, or will reach the same conclusions from the same results, are important – potentially, original – research contributions.

7.2.6 The current state of the match between open science ideas and practices

I have said that we should expect or hope to see open science practices (Munafò et al., 2017; Nosek et al., 2022) where researchers:

- share data and code;
- publish research reports in ways that enable others to check or query analyses.

This raises an important question: **What exactly do we see, when we look at current practices?** The question is important because answering it helps to identify where the challenges are located when you complete your work to locate, access, analyse and report previously collected data.

I break the discussion of what we see into two parts. Firstly, I look at the results of audits of data and code sharing (see Section 7.2.6.2): are data shared and can we access the data? Secondly, I discuss analyses of methods reproducibility, and shared data and code usability (see Section 7.2.6.3): can others reproduce the results reported in published articles, given shared data? can others access and run shared analysis code? can others use the shared code to reproduce the reported results? Again, I need to be brief but reference sources that you can follow-up.

7.2.6.1 The link between the credibility revolution and the reproducibility of results

I should be clear, before we go on, about **the link** between the *credibility revolution* in science, and the effort to examine reproducibility of results. Many elements of the credibility revolution emerged out of the observation that it has often been difficult to repeat the results of published studies when we conduct new studies (replication studies or results reproducibility; e.g., Aarts et al. (2015)). However, it is clearly difficult to know *what* to replicate or reproduce if we cannot reproduce the results presented in a study report (methods reproducibility), given the study data (Artner et al., 2021; Laurinavichyute et al., 2022; Minocher et al., n.d.).

7.2.6.2 Data and code sharing

Research on data and code sharing practices suggest that practices have improved, from earlier low levels.

In an important early report, Wicherts et al. (2006) observed that it was very difficult to obtain data reported in psychological research articles from the authors of the articles. They asked for data from the lead authors of 141 articles published in four leading psychology journals, for about 25% of the studies. This low response rate was found despite the fact that authors in these journals must agree to the principle that data can be shared with others wishing to verify claims.

Practice *has* changed: how?

One change to practice has involved the use of **open science badges**. In journals like **Psychological Science** authors of articles may be awarded badges — Open Data, Open Materials, Preregistration badges — by the editorial team. Authors can apply for and earn the badges

by providing information about open practices, and journal articles are published with the badges displayed near the front of the articles.

In theory, initiatives like encouraging authors to earn open science badges should mean that data sharing practices improve, enabling access to data and code for those, like you, who would like to re-analyze previously published data. In theory, all you should need to do — to locate and access data — is just search articles in the journal *Psychological Science* for studies with open data badges, and follow links from the published articles to then access study data at an open repository like the *Open Science Framework (OSF)*. What do we see in practice?

Analyses reported by Kidwell et al. (2016) as well as analyses reviewed by Nosek et al. (2022) indicate that more articles have claimed to make data available in the time since badges were introduced. When they did their analysis, Kidwell et al. (2016) found that a substantial proportion, but not all, of the articles in *Psychological Science* can be found to actually provide access to shared data. However, critically, many but not all the articles with open data badges provide access to data available through an open repository, data that are correct, complete and usable (Kidwell et al., 2016). In their later report, the analyses reviewed by Nosek et al. (2022) suggest that the use of repositories like OSF for data sharing may be accelerating but that, over the last few years, the rate at which open science practices like sharing data, overall, appears to be substantial but not yet reported or observed in a majority of the work of researchers.

Many journals now require the authors of articles to include a **Data Availability Statement** to locate their data. Analyses by Federer (2022) indicate that Data Availability Statements for articles published in the open access⁴ journal PLOS ONE often, helpfully, include Digital Object Identifiers (DOIs) or Universal resource locators (URLs) enabling direct access to shared data (i.e., without having to contact authors). Of those DOIs or URLs, most appeared to be associated with resources that could successfully be retrieved. In contrast, analyses reported by Gabelica et al. (2022) that where article authors state that “data sets are available on reasonable request” (the most common availability statement), most of the time, the authors did not respond or declined to share the data (see similar findings, across fields, by Tedersoo et al., 2021). Clearly, in the analyses of open science practices we have seen so far, data sharing is more effective where sharing does not have to work through authors.

Tip

- When you are looking for a study in order to get data that you can then reanalyze, it makes sense to look, first, for studies focusing on research questions that interest you.
- When you are looking for published reports where the authors share data, look for articles with open science badges or where you can see a Data Availability Statement.

⁴Open access journals publish articles that are free to read or download.

- Choose articles where the authors provide a direct link to their data, where the data are located on an open repository like the Open Science Framework (there are other repositories).

7.2.6.3 Enabling others to check or query analyses

Research on data and code sharing practices suggest that practices have improved but that there are concerns about the quality of the sharing. Here, the critical concern relates to the word *enable* in the objective: that we should publish research reports in ways that *enable* others to check or query analyses.

John Towse and colleagues (Towse et al., 2021) examined the quality of open datasets to assess their quality in terms of their completeness and reusability (see also Roche et al., 2015).

- completeness:** are all the data and the data descriptors supporting a study's findings publicly available?
- reusability:** how readily can the data be accessed and understood by others?

For a sample of datasets, they found that about half were incomplete, and about two-thirds were shared in a way that made them difficult to use. Practices tended to be slightly better in more recent publications. (Broadly similar results are reported by (Hardwicke et al., 2018).)

Where data were found to be incomplete, this appeared to be, in part, because participants were excluded in the processing of the data for analysis but this information was not in the report, or because data were shared without a guide or “readme” file or data dictionary (or codebook) explaining the structure, coding or composition of the shared data.

Potentially important for future open science practices, (Towse et al., 2021; also Roche et al., 2015) found that sharing data as *Supplementary materials* may appear to carry risks that, in the long term, mean that data may become inaccessible.

💡 Tip

- When you locate open data you can access, look for a guide, “readme” file, codebook or data dictionary explaining the data: you need to be able to understand what the variables are, what the observations relate to (observations per person, per trial?) and how variables are coded.
- Locate and examine carefully the parts of the published report, or the data guide, where the authors explain how they processed their data.

A number of studies have been conducted to examine whether shared data and analysis code can be reused by others to reproduce the results reported in papers (e.g., Artner et al., 2021; Crüwell et al., n.d.; Hardwicke et al., n.d.; Hardwicke et al., 2018; Laurinavichyute et al., 2022;

Minocher et al., n.d.; Obels et al., 2020; see Artner et al., 2021 for a review of reproducibility studies). In critical respects, the researchers doing this work are doing work similar to the work we are helping students to do, locating, accessing, and analyzing previously collected data. In these studies, typically, the researchers progressed through a series of steps.

1. Searched the articles published in a journal (e.g., *Cognition*, the *Journal of Memory and Language*, *Psychological Science*), published in a topic area across multiple journals (e.g., social learning, psychological research), or associated with a specific practice (e.g., registered reports).
2. Selected a subset of articles where it was identified that data could be accessed.
3. Identify a target result or outcome to reproduce, for each article. In their analyses, Hardwicke and colleagues (Hardwicke et al., n.d.; Hardwicke et al., 2018) focused on attempting to reproduce primary or *straightforward and substantive* outcomes: substantive – if emphasized in the abstract, or presented in a table or figure; straightforward – if the outcome could be calculated using the kind of test one would learn in an introductory psychology course (e.g., t-test, correlation).
4. Attempted to reproduce the results reported in the article, using the description of the data analysis presented in the article, and the analysis code (if provided), in some cases asking for information from the original study authors, in other cases working independently of original authors.

What the reproducibility studies appear to show is that, for many published reports, *if* data are shared and *if* the shared data are accessible and reusable *then*, most of the time, the researchers **could reproduce** the results presented by the original study authors (Hardwicke et al., n.d.; Hardwicke et al., 2018; Laurinavichyute et al., 2022; Minocher et al., n.d.; Obels et al., 2020; but see Crüwell et al., n.d.). This is great. But what is interesting, for us, is where the reproducibility researchers encountered challenges. You may encounter the same or similar challenges.

I list some challenges that the researchers describe, following. Before you look at the list, I want to assure you: you will not find *all* these challenges present for any one article you look at. Most likely, you will find one or two challenges. Obviously, some challenges will be more difficult than others.

Tip

- When you find a study you are interested in, with open data and maybe open analysis code, your main challenge will often be to identify exactly what analysis the original study authors did to answer their research question.
- Locate and examine carefully the parts of the published report where the authors explain how they did the analysis that gave them their key result. Usually that key result should be identified in the abstract or in the conclusion.

7.2.6.3.1 Data challenges

1. Data Availability Statements or open science badges indicate data are shared but data are not directly accessible through a link to an open repository.
2. The data are shared and accessible but there is missing or incorrect information *about* the data. The documentation, codebook or data dictionary is missing or incomplete. There is unclear or missing information about the variables or the observations, or about the coding of variable values, responses.
3. Original study authors may share raw and processed data or just processed or just raw data. It may not be clear how raw data were processed to construct the data analysed for the report. It may not be clear how variables were transformed or calculated or processed.
4. There may be mismatches between the variables referred to in the report and the variables named in the data file. It may be unclear how a data file corresponds to a study described in a report, where there are multiple studies and multiple data files.

7.2.6.3.2 Analysis challenges

1. The original report includes a description of the analysis but the description of the analysis procedure is incomplete or ambiguous.
2. There may be a mismatch, in the report, between a hypothesis, and the analysis specified to test the hypothesis (maybe in the Methods section), compared to a long sequence of results reported in the Results section. This makes it difficult to identify the key analysis.
3. It is easier to reproduce results if both data and code are shared because the presentation of the analysis code usually (not always) makes clear what analysis was done to get the results presented in the report.
4. Sometimes, analysis code is shared but it is difficult to use because it requires proprietary software (e.g., SPSS) or because it requires function libraries that are no longer publicly available.
5. Sometimes, there are errors in the analysis. Sometimes, there are errors in the presentation of the results, where results have been incorrectly copied into reports from analysis outputs.

7.3 This is why

The research report assignment requires students to locate, access, analyse and report previously collected data. At the start of the introduction, I said I would explain the answer to the question:

- Why: what is the motivation for the assignment?

I summarize, following, the main points of the answer I have given. When you review these points, I want you to think about two things, returning to the ideas of Bourdieu (2004) and Kuhn (1970) I sketched at the start.

Often what we do in science is guided by convention, the assumptions and habits of *normal practice* (Kuhn, 1970). These conventions can work in our minds so that if we encounter an *anomaly* or *discrepancy* between what we expect and what we find, in our work, we may usually blame ourselves: it was something wrong that we did or failed to do. It can cause us anxiety if we do not reproduce a result we think we should be able to reproduce (Lubega et al., n.d.). But I want you to understand, from the start, that sometimes, if you think you have found an error or a problem in a published analysis or a shared dataset, **you may be right**.

If there is anything we have learned, through the findings of replication studies, multiverse analyses, and reproducibility audits it is that people make mistakes, different choices are often reasonable, and we *always* need to check the evidence.

7.3.1 Summary: this is why

1. We are in the middle of a credibility revolution. The lessons we have learned so far oblige us to think about and to teach good open science practices that safeguard the value of evidence in psychology.
2. This matters, even if we do not care about scientific methods, because if we care about the translation into policy or practice – in clinical psychology, in education, health, marketing and other fields – what we do will depend on the value of the research evidence that informs policy ideas or practice guides.
3. Focusing on data analysis, it is useful to think about the whole *data pipeline* in analysis, the workflow that takes us from data collection to raw data to data processing to analysis to the presentation of results.
4. At every stage of the data pipeline, there are choices about what to do. There are not always reasons why we make one choice instead of another. Sometimes, we are guided by convention, example or instruction.
5. The existence of choices means the path we take, when we do data analysis, can be one path among multiple different *forking paths*.
6. For some parts of the pipeline – dataset construction, data analysis choices – reasonable people might make different decisions to sensibly answer the same research question, given the same data. This variation between pathways can be more or less important in influencing the results we see.
7. If results tend to stay similar across different ways of doing analysis, we might conclude that the results are reasonably robust across contexts, choices, or other variation in methods.
8. To *enable* others to see what we did (versus what we could have done), to see how we got to our results from our data, it is important to share our data and code.

9. Everyone makes mistakes and we should make it easy for others, and ourselves, to find those mistakes by sharing our data and code in accessible, clear, usable ways.
10. We need to **teach and learn** how to share effectively the data and the code that we used to answer our research questions.

In constructing the assignment – in asking *and supporting* students to locate, access, analyse and report previously collected data – we are presenting an opportunity to really investigate and evaluate existing practices.

You may find that this work is challenging, in some of the places that reproducibility research has identified there can be challenges. Where the challenges cannot be fixed – if you have found an interesting study but the study data are inaccessible or unusable – we will advise you to move on to another study. Where the challenges can be fixed – if data require processing, or if analysis information requires clarification – we will provide you with help or enabling information so that you fix the problems yourself.

Tip

- Maybe the main lesson from this exercise is a reminder of the *Golden rule*: **treat others as you would like to be treated**.
- If it is frustrating when it is difficult to understand information about an analysis or about data, or when it is difficult to access and reuse shared data and code.
- When it is your turn, do better, reflecting on what frustrated you.

One last question: why not just do less demanding or challenging tasks? Because this is part of what makes graduate degree valuable, what will make you more skilled in the workplace. Most of the time, we work in teams, we inherit problems or data analysis tasks, or are given results with partial information. The lessons you learn here will help you to effectively navigate those situations.

8 What

8.1 PSYC401 Project – research report – what you are expected to do

We present the following guidelines to help you to complete the coursework assessment. If you have any questions, email Padraig Monaghan at: p.monaghan@lancaster.ac.uk

Note the information mirrors exactly the information provided on Moodle:

<https://modules.lancaster.ac.uk/mod/page/view.php?id=1921399>

8.1.1 What data can I analyse?

Reports will concern, usually, findings from analyses of data-sets we have provided to you. Some students may wish to analyse data collected in previous studies or data accessed from online sources: they should correspond with Padraig Monaghan or Rob Davies if they wish to do so.

The evaluation of reports will focus on clarity, read the following for discussion of what is required.

We expect students to use one of the analysis methods taught in the module. Marks will be awarded depending:

- on how appropriate the method is to the context, to the study design, to answering the research question, and to the features of the data; the appropriateness of methods to contexts will be taught in class;
- on how effectively the analysis is explained; students must explain the motivations for their decisions, explain their methods, and explain their findings effectively to gain points.

8.1.2 What structure should reports take?

1. The reports should include abstract, introduction, methods, results, discussion and references sections, like a short research article in the journal *Psychological Science*. You can view examples of articles here

<https://journals.sagepub.com/toc/PSS/current>

2. Word count limit: no more than 1500 words are allowed for all materials.
3. Unlike a published research article, for PSYC401, the Results and Discussion sections must be written in full, but the Introduction and Methods sections can be written in the form of notes.

8.1.3 What content should reports present?

8.1.3.1 Introduction and Method sections

The focus of marking will be on the quality of the Results and Discussion sections. This means you can write your notes in the Introduction and Methods sections as short answers to the following questions:-

8.1.3.1.1 Introduction

- What did the researchers do and why did the researchers do it?
- What was the question addressed in the study and why is it interesting?
- What were the hypotheses?
- What results were expected and how would they relate to the hypotheses?

How can you write this as a set of notes? We require main points of information on the hypotheses concerning expected results. We will ignore the absence of citations, or of explanations of critical previous experimental work, in the Introduction.

8.1.3.1.2 Method

Note the origin of the data at the start of the method section. As for the Introduction, your method section writing needs to furnish answers to questions like the following:-

- What was done to collect the data?
- Who were tested (Participants)?
- What materials were used in testing (Materials)?
- What was the design of the study?
- What procedure was used?

How can you write this as a set of notes? We require main points of information, especially the main features of the data analyzed – what were the variables, how many observations were recorded, what exclusions or other data treatment steps were applied?

8.1.3.2 Results and Discussion sections

The focus of marking will be on the quality of the Results and Discussion sections. This means you must write in complete sentences in full paragraphs in a style appropriate for a research article appearing in a journal like *Psychological Science*. You must not use notes for these sections. You must write text that explains to the reader the analysis you did, why you did it, the results you found, and the implications of those results. You should write the text for the sections so that the questions listed following are answered fully.

If you use a data set that is already published in a journal such as *Psychological Science*, then your presentation of the results must differ from that in the article in ways that highlight new features of the data.

8.1.3.2.1 Results

Be clear on what the outcome measure or dependent variable for analysis was, and on what factors or predictor variables were brought into the analysis of that outcome. You then need to ensure the results section answers the following questions:-

- What hypotheses were tested?
- What methods were used to test the hypotheses?
- Why are they appropriate?
- What were the results? What were the direction and relative size of effects?

Do what seems reasonable using one or more of the analysis methods practiced in class, or practiced in association with the workbooks, and explain your reasoning.

8.1.3.2.2 Discussion

What the reader must be able to do, given your report, is understand the answer to the following questions:

- What are the theoretical implications of the study findings?
- What are the practical implications?

Reports should present **enough information that the reader can understand**: the background and motivation for a study; the features of the data analyzed and the methods of data collection; the approach taken in analysis, the analysis steps, and the results; the relationship between the observed results and the expected results, and the interpretation of findings in relation to previous work. To be clear about clarity: explain, spell things out (decisions, reasoning, interpretations) as if you were explaining them to a reasonably intelligent reader, a Psychologist who is not a specialist in the area of study occupied by the study reported, i.e. me. The main point is that you should keep in mind what the reader should get out of (what benefit) reading your report.

8.1.4 What format?

8.1.4.1 Statistics, tables and figures should follow APA guidelines. See here for a free guide:

For general APA formatting of reports:

https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_style_introduction.html

And for APA formatting of statistics and numbers:

https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/apa_numbers_statistics.html

Though the APA guidelines are the authoritative guide.

8.1.4.2 Add a link to the data analysed for the report

9 How

The research report assignment requires students to locate, access, analyse and report previously collected data. Here, we answer the question:

- How can the assignment be done?

We outline the workflow you can follow, proceeding through a series of steps to complete the essential tasks. Look at this outline, make a plan, and then follow the advice, taking it **one step at a time**.

9.1 The variety of things students do

Students have taken a variety of approaches to the assignment.

- Some students choose to complete an analysis of a publicly available dataset, analyzed previously, data for which the report has been published in a journal article.
- Some students choose to complete an analysis of a publicly available dataset that has been made available (for a report published as a data journal) but has not been analysed previously.
- Some students choose to complete an analysis of one of the data-sets used for practical exercises in class: the example or demonstration data we collect together as the *curated data*.

Ask in class or on the discussion forum for advice about any one of these approaches.

Here, I offer guidance on what to do if you want to locate, access, and analyse previously collected data where those data are presented in a journal article. I consider, first, working with datasets where an analysis of the data has been presented in the article (see Section 9.2). I then look at working with datasets where the data are presented without an analysis (see Section 9.3). Our advice on working with datasets presented without an analysis will overlap in key respects with our advice on working with curated data.

9.2 Working with data associated with a published analysis

In the following, I split our guidance into two parts. I look next at the task of locating, accessing and checking the data (Section 9.2.1). Then I look at the task of figuring out what analysis you can do with the data (see Section 9.2.2). Obviously, you cannot consider an analysis if you cannot be sure that you can work with the data (Mincher et al., n.d.).

9.2.1 Locate, access and check the data

At the start of your work on the assignment, you will need to (1.) locate then (2.) access data for analysis, and then you will need to (3.) check that the data are usable. I set out advice on doing each step, following. Work through the steps: **one step at a time**.

9.2.1.1 Locate

It is usually helpful to find a dataset where the data have been collected in a study within a topic area you care about, or could be interested in. It is helpful because you will need to work with the data and it will be motivating if you are interested in what the data concern. And it is helpful because, often, you will need to do a bit of reading on related research to learn about the context for the data collection, and you will usually want to read research sources that interest you.



Tip

The task here is:

- Do a search: look for an article with usable data in a topic area that interests you.

There are at least two ways you can do this. Both should be reasonably quick methods to get to a usable dataset.

1. Do a search on [Google scholar](#)).
2. Do a search on the webpages of a journal.

Most psychological research is published in journals like *Psychological Science*. If you want, you can look at a list of psychology journals [here](#).

In a journal like *Psychological Science* you can look through lists of previously published articles (in issues, volumes, by year) on the journal webpage. [Here is the list of issues for Psychological Science..](#)

9.2.1.1.1 Key words

In both methods, you are looking for an article associated with data (and maybe analysis code) you can access and that you are sure you can use. In both methods, you need to first think about some **key words** to use in your search. Ask yourself:

- What are you interested in? What population, intervention or effect, comparison, or outcome?

Then:

- What words do people use, in articles you have seen, when they talk about this thing?

You can use these words, and maybe consider alternate terms. For example, I am interested in **reading comprehension** or **development reading comprehension** but researchers working on **reading development** might also refer to **children reading comprehension**.

You want to be as efficient as possible so combine your search for articles in an interesting topic area with your search for accessible data. We can learn from the research we discussed on data sharing practices (see Section 7.2.6.2) by looking for specific markers that data associated with an article should be accessible.

If you are doing a search (1.) on [Google scholar](#), I would use the key words related to your topic plus words like: **open data badge**; **open science badge**. So, I would do a search for the words: **reading comprehension open data badge**. I have done this: you can try it. The search results will list articles related to the topic of reading comprehension, where the authors claim to have earned the open data badge because they have made data available.

If you are doing a search (2.) in a journal list of articles, then what you are looking for are articles that interest you and which are listed with open data badges. In the listing for *Psychological Science* ([here](#)) a quick read of the journal issue articles index shows that article titles are listed together with symbols representing the open science badges that authors have claimed.

In other journals (e.g., *PLOS ONE*, *PeerJ*, *Collabra*), you may be looking for interesting articles with the words **Data Availability Statement**, **Data Accessibility Statement**, **Supplementary data** or **Supplementary materials** in the article webpage somewhere. Journals like *PeerJ* or *Collabra*, in particular, make it easy to locate data associated with published articles on their web pages.

In *Collabra*, you can find published articles through the journal webpage ([here](#)). If you click on the title of any article, and look at the article webpage, then on the left of the article text, you can see an index of article contents and that index lists the **Data Availability Statement**. Click on that and you are often taken to a link to a data repository.

9.2.1.2 Access

If you have located an interesting article with evidence (an open data badge or a data accessibility statement) that the authors have shared their data, you need to check that you can access the data. Most of the time, now, you are looking for a link you can use to go directly to the shared data. The link is often presented as a hyperlink on a webpage, associated with Digital Object Identifiers (DOIs) or Universal resource locators (URLs). Or, increasingly, you are looking for a link to a data repository on a site like the Open Science Framework (OSF).

💡 Tip

The task here is:

- Access the data associated with the article you have found.

Here are some recent examples from my work that you can check, to give you a sense of where or how to find the accessible link to the shared data.

Ricketts, J., Dawson, N., & Davies, R. (2021). The hidden depths of new word knowledge: Using graded measures of orthographic and semantic learning to measure vocabulary acquisition. *Learning and Instruction*, 74, 101468. <https://doi.org/10.1016/j.learninstruc.2021.101468>

Rodríguez-Ferreiro, J., Aguilera, M., & Davies, R. (2020). Semantic priming and schizotypal personality: Reassessing the link between thought disorder and enhanced spreading of semantic activation. *PeerJ*, 8, e9511. <https://doi.org/10.7717/peerj.9511>

These are both open access articles.

If you look at the webpage for, Rodríguez-Ferreiro et al. (2020), ([here](#)), you can do a search in the article text for the keyword OSF (on the article webpage, use keys CMD-F plus OSF). You are checking to see if you can click on the link *and* if clicking on the link takes you to a repository listing the data for the article. The Rodríguez-Ferreiro et al. (2020) article is associated with a data plus analysis code repository ([OSF](#))

Notice that on the repository webpage, you can see a description of the project plus .pdf files and a folder **Dataset and Code**. If you can click through to the folders, and download the datafiles, you have accessed the data successfully.

I have guided you, here, through to the Rodríguez-Ferreiro et al. (2020) data repository, can you find the data for the Ricketts et al. (2021) repository?

9.2.1.3 Check

If you have located an interesting article with data that you can access, and if you have read the introductory notes (see Section 7.2.6.3), then you will know that you need to make sure that you can use the data.

💡 Tip

The task here is:

- Check the data and the data documentation to make sure you can understand what you have got *and* whether you can use it.

What make data usable are:

1. Information in the article, or in the data repository documentation, on the study design and data collection methods: you need to be able to understand where the data came from, how they were collected, and why.
2. Clear data documentation: you need to find information on the variables, the observations, the scoring, the coding, and whether and how the data were processed to get them from raw data state to the data ready for analysis.

Data documentation is often presented as a note or a wiki page or a miniature paper and may be called a *codebook*, *data dictionary*, *guide to materials* or something similar. You will need to check that you can find information on (examples shown are from the Rodríguez-Ferreiro et al. (2020) OSF *guide to materials*):

- what the data files are called e.g. `PrimDir-111019.csv`;
- how the named data files correspond to the studies presented in the report;
- what the data file columns are called and what variables the column data represent e.g. `relation`, `coding for prime-target relatedness condition ...`;
- how scores or responses in columns were collected or calculated e.g. `age`, `giving the age in years ...`;
- how coding was done, if coding was used e.g. `biling`, `giving the bilingualism status`;
- whether data were processed, how missing values were coded, whether participants or observations were excluded before analysis e.g. `Missing values in the rt column ... coded as NA`

If these information are not presented, or are not clear: **walk away**.

9.2.2 Plan the analysis you want to do

After you have found an interesting article, and have confirmed that you can use the associated data, you will need to plan what analysis you want to do.

Tip

The task here is:

- Identify and understand the analysis in the article.
- Work out what analysis *you* want to do.

Students have taken a variety of approaches to the assignment.

- Some students choose to complete a reanalysis of the data, in an attempt to reproduce the results presented in the article (see Section 9.2).
- Some students choose to complete an alternate analysis of the data, varying elements of the analysis (see Section 7.2.4).

Either way, you will want to first make sure you can identify exactly *what* the authors of the original study did, *how* they did it, and *why* they did it.

You can process the key article information efficiently using the *QALMRI* method we discussed in the class on graduate writing skills (Brosowsky et al., n.d.; Kosslyn & Rosenberg, 2005). You are first aiming to **locate** information on the broad and the specific question the study addresses, the methods the study authors used to collect data, the results they report, and the conclusions they present given the results. Can you find these bits of information?

9.2.2.1 Are you interested in attempting a methods reproducibility test?

Following Hardwicke and colleagues (Hardwicke et al., n.d.; Hardwicke et al., 2018) it would be sensible to focus on identifying the primary or *substantive* result for a study in an article.

- **Substantive** if emphasized in the abstract, or presented in a table or figure.

As we discussed in the class on graduate writing skills, the article authors *should* signal what they consider to be the primary result for a study by telling you that a result is critical or key or that a result is the or an answer to their research question.

Tip

- An article may present multiple studies: focus on one.
- The results section of an article, for a study, may list multiple results: identify the primary or substantive result.

If you are, then you will want to identify a result that is both substantive and *straightforward* (Hardwicke et al., n.d.; Hardwicke et al., 2018).

- **straightforward** if the outcome could be calculated using the kind of test you have been learning about or will learn about (e.g., t-test, correlation, the linear model)

Psychological science researchers use a variety of data analysis methods and not all the analyses that you read about will be analyses done using methods that you know about. The use of the methods we teach — t-test, correlation, and the linear model — are very *very* common; that is why we teach them. But you may also see reports of analyses done using methods like ANOVA, and multilevel or (increasingly) linear mixed-effects models (Meteyard & Davies, 2020b).

In the research on the reproducibility of results in the literature (see Section 7.2.6.3), the researchers attempting to reproduce results often focused on answering the research question the original authors stated using the data the original authors shared. This does not mean that they always tried to *exactly* reproduce an analysis or an analysis result. Sometimes, that was not possible.

Sometimes, you will encounter an article and a dataset you are interested in but the analysis presented in the article looks a bit complicated, or more complex than the methods you have learned would allow you to do. In this situation, don't give up. What you can do – maybe with our advice – is identify *a part* of the primary result that you *can* try to reproduce. For example, what if the original study authors report a linear mixed-effects analysis of the effects of both prime relatedness and schizotypy score on response reaction time (Rodríguez-Ferreiro et al., 2020)? Maybe you have not learned about mixed-effects models, or you have not learned about analysing the effects of two variables but you *have* (you will) learn about analysing the effect of one variable using the linear model method: OK then, do an analysis of the shared data using the method you know.

You may be helped, here, by knowing about two good-enough (mostly true) insights from statistical analysis:

1. Many of the common analysis methods you see used in psychological science can be coded as a linear model.
2. More advanced common analysis methods — (Generalized) Linear Mixed-effects Models (GLMMs) — can be understood as more sophisticated versions of the linear model. (Conversely, the linear model can be understood as an approximation of a GLMM.)

There is a nice discussion of the idea that common statistical tests are linear models [here](#).

Tip

- Identify the analysis method used to get the result you are interested in.
- If it is complex or unfamiliar, discuss whether a simpler method can be used.

- If the result is complex, discuss whether you can attempt to reproduce a part or a simpler result.

9.2.2.2 Are you interested in attempting a different analysis?

It can be interesting and important work to complete a simpler analysis of shared data. Sometimes, we learn that a simpler analysis is as good account of the behaviour we observe as other more complex analyses. This can happen if, for example, our theory predicts that two effects should work together but an analysis shows that we can explain behaviour in an account in which the two effects are independent. For example, Ricketts et al. (2021) predicted that children should learn words more effectively if they were shown the spellings of the words *and* they were told they would be helped by seeing the spelling but, in our data, we found that just seeing the spellings was enough to explain the learning we observed.

In completing analyses that *vary* from original analyses, we are engaging in the kind of work people do when they do *multiverse analyses* or *robustness checks* (see Section 7.2.4).



Tip

In planning an alternate or multiverse analysis, do not suppose that you need to do multiple analyses: you do not.

In planning an alternate or multiverse analysis, you will want to begin by critically evaluating the analysis you see described in the published article. I talk about how to do this, next.

Before we go on, note that I previously discussed an example of how to critically evaluate the results of published research in the context of Rodríguez-Ferreiro et al. (2020). Take a look at the Introduction of that article. There, we summarised the analyses researchers did previously and used the information about the analyses to explain inconsistencies in the research literature. We found limitations in the analyses that people did that had (negative) consequences for the strength of the conclusions we can take from the data.

9.2.2.2.1 Critically evaluate the analysis description

If you revisit our discussion of multiverse analyses, you will see that we discussed two things: (1.) analyses of the impact on results of varying how you construct datasets for analysis (Section 7.2.4.2) and (2.) analyses of the impact on results of varying what analysis method you use, or how you use the method (see Section 7.2.4.3). These are both good ways to approach thinking about the description of the analysis you see in a published article.

As we noted in Section 7.2.4.2, you almost always have to process the data you collect (in an experiment or a survey) before you can analyze the data. Often, this means you need to code for responses to survey questions e.g. asking people to self-report their gender, or you need to

identify and code for people making errors when they try to do the experimental task you set them, or you need to process the data to exclude participants who took too long to do the task (if taking too long is a problem). Not all of these processing steps will have an impact on the results but some might. This is why you can sometimes do **useful** and sometimes **original** research work in reanalyzing previously published data.

You can begin your analysis planning work by first identifying exactly what data processing the original study authors did then identifying what different data processing they could have done. Remember the research we discussed in relation to reproducibility studies, you need to be prepared for the possibility that it is challenging to identify what researchers did to process their data for analysis Section 7.2.6.3.1. To identify the information you need, look for keywords like `code`, `exclude`, `process`, `tidy`, `transform` in the text of the article, or look for words like this in the documentation you find in the data repository.

When you have identified this information, you can then consider three questions:

1. What data processing steps were completed before analysis?
2. What were the reasons given explaining why these processing steps were completed?
3. What could happen to the results if different choices were made?

Working through these questions can then get you to a good plan for an analysis of the data. For example, a simple but useful analysis you can do is to check what happens to the results if you do an analysis with data from all the participants tested, if participants are excluded (for some reason) in the data processing step. Obviously, if the original study authors *only* share processed data, you cannot do this kind of work. Another simple but useful analysis you can do is to check what happens to the results if you change the coding of variables. Sometimes different coding of categorical variables (e.g., ethnicity) are reasonable. For example, you can ask: what happens if you analyze the impact of the variable given a different coding? (In case you are reading these notes and thinking about recoding a factor, there are some useful functions you can use; [read about them here](#).)

 Tip

- Do you want to check the impact of varying data processing choices: check, do you need and have access to the raw data? can you see how to recode variables?

As we noted in Section 7.2.4.3, when we consider how to answer a research question with a dataset, it is often possible to imagine multiple different analysis methods: reasonable alternatives. Most often, this is most clearly apparent when we are looking at an *observational* dataset or data collected given a *cross-sectional* study design.

In *cross-sectional* or *observational* studies, we typically are not manipulating experimental conditions, and we are often analyzed data using some kind of linear model. We often collect data or have access to data on a number of different variables relevant to our interests. For example, in studies I have done on how people read (R. Davies et al., 2013; R. A. I. Davies

et al., 2017), we wanted to know what factors would predict or influence how people do basic reading tasks like reading aloud. We collected information on many different kinds of word properties and on the attributes of the participants we tested. (Note: the papers are associated with data repositories in Supplementary Materials.) It is an **open question** which variables should be included in a prediction model of the observed outcome (reading response reaction times). Therefore, if you are interested in a study like this, and can access usable data from the study, it will often be true that you are able to sensibly motivate a different analysis of the study data using a different choice of variables.

As discussed in a number of interesting analyses, over the years (e.g., Patel et al., 2015), researchers may be interested in the specific impact of one particular predictor variable (e.g., we may be interested in whether it is easier to read words we learned early in life), but will need to include in their analysis that variable plus other variables known to affect the outcome. In that situation, the effect of the variable of interest may appear to be different depending on what other variables are also analyzed. This makes it interesting and useful to check the impact of different analysis choices.

We will look at data like these, for analyses involving the linear model, in our classes on this method.

💡 Tip

- Do you want to check the impact of different analysis choices: check, do you need and have access to a choice of variables?
- Can you think of some reasons to justify using a different choice of variables in your analysis.

9.2.3 Summary: working with data associated with a published analysis

Here's a quick summary of the advice we have discussed so far.

- At the start of your work, you will need to (1.) locate then (2.) access data for analysis, and then you will need to (3.) check that the data are usable.
- Once you have confirmed you have found interesting data you can use, you should plan your analysis.
- Students do a variety of kinds of analysis. Whatever your interest, you first will want to first make sure you can identify exactly what the authors of the original study did, how they did it, and why they did it.
- If you are interested in attempting a methods reproducibility test (can you repeat a result, given shared data?) you will perhaps benefit from focusing a result that is both substantive and straightforward.
- If you are interested in doing an alternate analysis, you can critically evaluate the data processing and the data analysis choices that the original study authors made. You

can consider whether other choices would be appropriate, and might sensibly motivate a (limited) investigation of the impact of a different analysis pipeline choice on the results.

What if you access interesting data that were shared without a previous analysis? We talk about that situation, next.

9.3 Working with data that are not associated with a published analysis

A number of datasets have been published online with information about the data but with no analysis. You can look for data that may be interest you in a number of different places, now, but I would focus on one. I talk about that next. Then I offer some guidance on how you might approach analyzing such data Section 9.3.2.

9.3.1 Looking for open data

Wicherts and colleagues set up the Journal of Open Psychology Data (JOPD) to make it easier for Psychologists to share experimental data. A link to the journal webpage is [here](#)) Usually, a data paper reports a study and provides a link to a downloadable dataset.

Some datasets that I have looked at in JOPD and other places include the following.

9.3.1.1 Wicherts intelligence and personality data

Wicherts did what he recommended and put a large dataset online [here](#)

You can analyse these data in a number of different interesting ways. You can explore relationships between gender, intelligence and personality differences.

The data file and an explanatory document are located at the end of the article. Read the article, it's worth your time. Wicherts reports:

The file includes data from our freshman-testing program called “Testweek” (Busato et al., 2000, Smits et al., 2011 and Wicherts and Vorst, 2010) in which 537 students (age: $M = 21.0$, $SD = 4.3$) took the Advanced Progressive Matrices (Raven, Court, & Raven, 1996), a test of Arithmetic, a Number Series test, a Hidden Figures Test, a test of Vocabulary, a test of Verbal Analogies, and a Logical Reasoning test (Elshout, 1976).

Also included are data from a Dutch big five personality inventory (Elshout & Akkerman, 1975), the NEO-PI-R (Hoekstra, Ormel, & Fruyt, 1996), scales of social desirability and impression management (based on work by Paulhus, 1984 and Wicherts, 2002), sex of the participants, and grade point averages of the freshmen’s first trimester that may act as outcome variable.

9.3.1.2 Smits personality data

Smits and colleagues (including Wicherts) put an even larger dataset online at the Journal of Open Psychology Data [here](#))

You will need to register to be able to download the data but the process is simple.

The Smits dataset includes **Big-5** personality scores for several thousand individuals recorded over a series of years. You can analyse these data in interesting ways including examining changes in personality scores among students over different years.

9.3.1.3 Embodied terror management

Tjew A Sin and colleagues shared a dataset at the Journal of Open Psychology Data on an interesting study they did to test the idea that interpersonal touch or simulated interpersonal touch can relieve existential concerns (fear of death) among individuals with low self-esteem. The data can be found [here](#))

The Tjew A Sin can be downloaded from a link to a repository location, given at the end of the article. You will likely need to register to download the data. Note that the spreadsheets holding the study data include 999 values to code for missing data. Note also that the data spreadsheets include (in different columns) scores per participant for various measures e.g. mortality anxiety or self-esteem. The measures are explained in the paper. To use the data, you will need to work out the simple process of how to sum the scores across items to get e.g. a measure of self-esteem for each person.

9.3.1.4 Demographic influences on disgust

Berger and Anaki shared data on the disgust sensitivity of a large sample of individuals. The data are from the administration of the Disgust Scale to a set of Hebrew speakers. They can be found [here](#))

The experimenters collected data on participants' characteristics so that analyses of the way in which sensitivity varies in relation to demographic attributes is possible. You will see that the disgust scale is explained in the paper. The different disgust scores, for each item in the disgust scale, can be found in different columns. The disgust scores, for person, are calculated overall as values: `Mean_general_ds`, `Mean_core`, `Mean_AnimalReminder`, `Mean_Contamination`

When you download the dataset, you may need to change the file name, adding a suffix: `.txt` (for the tab delimited file), to be opened in Excel, or `.sav` (for the SPSS data file), to be opened in SPSS – to the file name to allow you to open it in the appropriate application.

9.3.2 Thinking about analyses of open data

The availability of rich, curated, clearly usable datasets with many variables can make it challenging to decide what to do.

I would advise beginning with an exploratory analysis of the data you have accessed. You will want to begin by using the data visualization skills we have taught you to examine:

1. The distributions of the variables that interest you using histograms, density plots or bar charts.
2. The potential relationship between variables using scatterplots.

In such *Exploratory Data Analyses*, you are interested in what the data visualization tells you about the nature of the dataset you have accessed. The papers associated with the datasets can sometimes offer only outline information: how the data were collected, coded, and processed. You may need to satisfy yourself that there is nothing odd or surprising about the distributions of scores. This stage can help you to identify problems like survey responses with implausible scores.

The work you do in exploring, and summarizing, the data variables that interest you will often constitute a substantial element of the work you can do and present for your report. You may discuss, for advice, what parts of this work will be interesting or useful to present.

Then, our advice is simple.



Tip

- When working with open datasets, consider keeping the analysis *simple*.

Note that *simple* is relative. Do what interests you. Work with the methods you have learned or will learn (the linear model).

In practice, you will find that part of the challenge is located not in using the data or in running an analysis like a linear model, it is in (1.) justifying or motivating the analysis and (2.) explaining the implications of your findings.

Working on the thinking you must develop to motivate an analysis or to explain implications requires you to do some (limited) reading of relevant research. (Relevant sources will be cited in data papers, as part of their outline of the background for their data collection.) If you consider the advice we discussed in the graduate class on developing writing skills, you will see that there I talked about how you might extract data from a set of relevant sources (papers) to get an understanding of the questions people ask, the assumptions they make. That is the kind of process you can follow to develop your thinking around the analysis you will do. What you are looking for is information you can use so that you can say something brief about, for example, why it might be interesting to analyze, say, whether personality (measured using the Big-5) varies given differences in gender or differences between population cohorts. The

reading and the conceptual development should be fairly limited, not extensive, but should be sufficient that you can write something sensible when you introduce and then when you discuss your analysis results.

9.4 Summary: how

In this chapter, I have outlined some advice on how you might approach the task of locating, accessing, and analyzing previously collected data. The main advice is to think about your workflow in stages, then progress through the work one step at a time.

You will need to begin by assuring yourself that you can find a dataset that interests you, and that you can access and use the data. The usability of data will require clear, understandable, descriptions in the published article (if any) about the research question and hypothesis, the study design, the data collection methods, the data processing steps, and the data analysis (if any). Sometimes, useful information about data processing and data analysis can be found in detail in repository documentation (e.g., in guides to materials) but only referenced in the text of the article.

If you know you can locate, access and have checked data as usable, you will want to think about what analysis you want to do the data. The approach you take depending on what aims you would like to pursue.

If you are interested in attempting a methods reproducibility test (i.e. checking if you can repeat presented results, given shared data), then you will first need to identify a substantive and straightforward result to try to reproduce. If you identify a primary result to examine, you will want to check that you can work with the data that have been shared, and then that you can use the analysis methods you have learned to reproduce some or all of the result that interests you.

If you are interested in doing an alternate or a different analysis (from what may be presented), you may need to consider the information you can locate on data processing and on data analysis choices. Did the original study authors process the data before sharing it, how? are the raw data available? What analyses did the authors do and why? When you consider this information, you may critically evaluate the choices made. In the context of this critical evaluation, you may find good reasons to justify doing a different analysis, whether to examine the impact of making different data processing choices, or to examine the impact of using a different analysis method, or of applying the same method differently (e.g., by including different variables).

In considering an analysis of data shared without a published set of results, you may want to keep your approach simple. Focus on what analysis you can do using the methods you have learned. And think about the understanding you will need to develop, to justify the analysis you do, and to make sense, in the discussion of your report of the analysis results you will present.

It is always a good idea to explore your data using visualization techniques throughout your workflow.

 Tip

- You can always get advice, do not hesitate to ask.
- We are happy to discuss your thinking, especially in class.

Part IV

END

10 Summary

To complete when book is completed.

References

- Aarts, E., Dolan, C. V., Verhage, M., & Van der Sluis, S. (2015). Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. *BMC Neuroscience*, 16(1), 1–15. <https://doi.org/10.1186/s12868-015-0228-5>
- Aczel, B., Szaszi, B., Nilsonne, G., Akker, O. R. van den, Albers, C. J., Assen, M. A. van, Bastiaansen, J. A., Benjamin, D., Boehm, U., Botvinik-Nezer, R., Bringmann, L. F., Busch, N. A., Caruyer, E., Cataldo, A. M., Cowan, N., Delios, A., Dongen, N. N. van, Donkin, C., Doorn, J. B. van, ... Wagenmakers, E.-J. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife*, 10, e72185. <https://doi.org/10.7554/eLife.72185>
- Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, 26(5), 527–546. <https://doi.org/10.1037/met0000365>
- Auspurg, K., & Brüderl, J. (2021). Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the “Many Analysts, One Data Set” Project. *Socius*, 7, 23780231211024421. <https://doi.org/10.1177/23780231211024421>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008a). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008b). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology: General*, 133(2), 283–316. <https://doi.org/10.1037/0096-3445.133.2.283>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013a). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013b). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255–278.
- Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., Chow, S.-M., Jonge, P. de, Emerencia, A. C., Epskamp, S., Fisher, A. J., Hamaker, E. L., Kuppens, P., Lutz, W., Meyer, M. J., Moulder, R., Oravec, Z., Riese, H., Rubel, J., ... Bringmann, L. F. (2020). Time to get personal? The impact of researchers choices on

- the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*, 137, 110211. <https://doi.org/10.1016/j.jpsychores.2020.110211>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using {lme4}. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Belenky, G., Wesensten, N. J., Thorne, D. R., Thomas, M. L., Sing, H. C., Redmond, D. P., Russo, M. B., & Balkin, T. J. (2003). Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. *Journal of Sleep Research*, 12(1), 1–12. <https://doi.org/10.1046/j.1365-2869.2003.00337.x>
- Bornstein, M. H., Jager, J., & Putnick, D. L. (2013). Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Developmental Review*, 33(4), 357–370. <https://doi.org/10.1016/j.dr.2013.08.003>
- Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., ... Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810), 84–88. <https://doi.org/10.1038/s41586-020-2314-9>
- Bourdieu, P. (2004). *Science of Science and Reflexivity*. Polity.
- Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H. V., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., ... Żółtak, T. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, 119(44), e2203150119. <https://doi.org/10.1073/pnas.2203150119>
- Brosowsky, N., Parshina, O., Locicero, A., & Crump, M. (n.d.). *Teaching undergraduate students to read empirical articles: An evaluation and revision of the QALMRI method*. <https://doi.org/10.31234/osf.io/p39sc>
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376.
- Carp, J. (2012a). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, 6, 149.
- Carp, J. (2012b). The secret lives of experiments: Methods reporting in the fMRI literature. *Neuroimage*, 63(1), 289–300.
- Chang, W. (2013). *R graphics cookbook*. o'Reilly Media.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65(3), 145–153. <https://doi.org/10.1037/h0045186>
- Crüwell, S., Aphorop, D., Baker, B. J., Colling, L., Elson, M., Geiger, S. J., Lobentanzer, S., Monéger, J., Patterson, A., Schwarzkopf, D. S., Zaneva, M., & Brown, N. J. L. (n.d.). *What's in a badge? A computational reproducibility investigation of the open data badge policy in one issue of psychological science*. <https://doi.org/10.31234/osf.io/729qt>
- Davies, R. A. I., Birchenough, J. M. H., Arnell, R., Grimmond, D., & Houlston, S. (2017).

- Reading through the life span: Individual differences in psycholinguistic effects. *Journal of Experimental Psychology: Learning Memory and Cognition*, 43(8). <https://doi.org/10.1037/xlm0000366>
- Davies, R., Barbón, A., & Cuetos, F. (2013). Lexical and semantic age-of-acquisition effects on word naming in spanish. *Memory and Cognition*, 41(2), 297–311. <https://doi.org/10.3758/s13421-012-0263-8>
- Del Giudice, M., & Gangestad, S. W. (2021). A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions. *Advances in Methods and Practices in Psychological Science*, 4(1), 2515245920954925. <https://doi.org/10.1177/2515245920954925>
- Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Krypotos, A.-M., Kupitz, C. N., Leite, F. P., Lerche, V., Lin, Y.-S., Logan, G. D., Palmeri, T. J., Starns, J. J., Trueblood, J. S., Maanen, L. van, ... Donkin, C. (2019). The Quality of Response Time Data Inference: A Blinded, Collaborative Assessment of the Validity of Cognitive Models. *Psychonomic Bulletin & Review*, 26(4), 1051–1069. <https://doi.org/10.3758/s13423-017-1417-2>
- Federer, L. M. (2022). Long-term availability of data associated with articles in PLOS ONE. *PLOS ONE*, 17(8), e0272845. <https://doi.org/10.1371/journal.pone.0272845>
- Fillard, P., Descoteaux, M., Goh, A., Gouttard, S., Jeurissen, B., Malcolm, J., Ramirez-Manzanares, A., Reisert, M., Sakaie, K., Tensaouti, F., Yo, T., Mangin, J.-F., & Poupon, C. (2011). Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. *NeuroImage*, 56(1), 220–234. <https://doi.org/10.1016/j.neuroimage.2011.01.032>
- Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456–465. <https://doi.org/10.1177/2515245920952393>
- Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The Science of Visual Data Communication: What Works. *Psychological Science in the Public Interest*, 22(3), 110–161. <https://doi.org/10.1177/1529100621105195>
- Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *Journal of Clinical Epidemiology*, 150, 33–41. <https://doi.org/10.1016/j.jclinepi.2022.05.019>
- Gelman, a. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A bayesian perspective. *Journal of Management*, 41(2), 632–643. <https://doi.org/10.1177/0149206314525208>
- Gelman, A., & Loken, E. (2014a). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Psychological Bulletin*, 140(5), 1272–1280.
- Gelman, A., & Loken, E. (2014b). The statistical crisis in science. *American Scientist*, 102(6), 460–465. <https://doi.org/10.1511/2014.111.460>
- Gelman, A., & Unwin, A. (2013). Infovis and Statistical Graphics: Different Goals, Different Looks. *Journal of Computational and Graphical Statistics*, 22(1), 2–28. <https://doi.org/10.1080/10618600.2012.761137>

- Gelman, A., & Weakliem, D. (2009). Of beauty, sex and power. *American Scientist*, 97(4), 310–316. <https://doi.org/10.1511/2009.79.310>
- Gilmore, R. O., Diaz, M. T., Wyble, B. A., & Yarkoni, T. (2017). Progress toward openness, transparency, and reproducibility in cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1396, 5–18. <https://doi.org/10.1111/nyas.13325>
- Goldstein, H. (1995). *Multilevel statistical models*. Edward Arnold.
- Golino, H., & Gomes, C. (2014). Psychology data from the “BAFACALO project: The Brazilian Intelligence Battery based on two state-of-the-art models – Carroll’s Model and the CHC model”. *Journal of Open Psychology Data*, 2(1), e6. <https://doi.org/10.5334/jopd.a.f>
- Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341).
- Hardwicke, T. E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M. B., Peloquin, B. N., deMayo, B. E., Long, B., Yoon, E. J., & Frank, M. C. (n.d.). Analytic reproducibility in articles receiving open data badges at the journal psychological science: An observational study. *Royal Society Open Science*, 8(1), 201494. <https://doi.org/10.1098/rsos.201494>
- Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, 5(8), 180448. <https://doi.org/10.1098/rsos.180448>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, 33(2-3). <https://doi.org/10.1017/S0140525X0999152X>
- Herndon, T., Ash, M., & Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 38(2), 257–279. <https://doi.org/10.1093/cje/bet075>
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21(5), 1157–1164. <https://doi.org/10.3758/s13423-013-0572-3>
- Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., & Boulesteix, A.-L. (n.d.). The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science*, 8(4), 201925. <https://doi.org/10.1098/rsos.201925>
- Ioannidis, J. P. a. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), 0696–0701. <https://doi.org/10.1371/journal.pmed.0020124>
- 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. <https://doi.org/10.1177/0956797611430953>
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). *Treating stimuli as a random factor in social psychology : A new and comprehensive solution to a pervasive but largely ignored problem*. 103(1), 54–69. <https://doi.org/10.1037/a0028347>
- Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost,

- effective method for increasing transparency. *PLoS Biology*, 14(5), 1–15. <https://doi.org/10.1371/journal.pbio.1002456>
- Klau, S., Hoffmann, S., Patel, C. J., Ioannidis, J. P., & Boulesteix, A.-L. (2021). Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework. *International Journal of Epidemiology*, 50(1), 266–278. <https://doi.org/10.1093/ije/dyaal64>
- Klau, S., Schönbrodt, F., Patel, C. J., Ioannidis, J., Boulesteix, A.-L., & Hoffmann, S. (n.d.). *Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology*. <https://doi.org/10.31234/osf.io/c7v8b>
- Kosslyn, S. M., & Rosenberg, R. S. (2005). *Fundamentals of psychology: The brain, the person, the world, 2nd ed.* Pearson Education New Zealand.
- Kreft, I., & Leeuw, J. de. (1998). *Introducing multilevel modeling* (D. Wright, Ed.). Sage Publications.
- Kuhn, T. S. (1970). *The structure of scientific revolutions* ([2d ed., enl]). University of Chicago Press.
- Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., Bergh, D. van den, Marsman, M., Derkx, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., ... Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, 146(5), 451–479. <https://doi.org/10.1037/bul0000220>
- Laurinavichyute, A., Yadav, H., & Vasishth, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy. *Journal of Memory and Language*, 125, 104332. <https://doi.org/10.1016/j.jml.2022.104332>
- Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 16(1), 149–157.
- Lubega, N., Anderson, A., & Nelson, N. (n.d.). *Experience of irreproducibility as a risk factor for poor mental health in biomedical science doctoral students: A survey and interview-based study*. <https://doi.org/10.31222/osf.io/h37kw>
- Maier-Hein, K. H., Neher, P. F., Houde, J.-C., Côté, M.-A., Garyfallidis, E., Zhong, J., Chamberland, M., Yeh, F.-C., Lin, Y.-C., Ji, Q., Reddick, W. E., Glass, J. O., Chen, D. Q., Feng, Y., Gao, C., Wu, Y., Ma, J., He, R., Li, Q., ... Descoteaux, M. (2017). The challenge of mapping the human connectome based on diffusion tractography. *Nature Communications*, 8(1), 1349. <https://doi.org/10.1038/s41467-017-01285-x>
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34(2), 103–115.
- Meehl, P. E. (1978). *Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology*. 46(September 1976), 806–834.
- Meteyard, L., & Davies, R. A. I. (2020a). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, 112, 104092. <https://doi.org/10.1016/j.jml.2020.104092>

- //doi.org/10.1016/j.jml.2020.104092
- Meteyard, L., & Davies, R. A. I. (2020b). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, 112. <https://doi.org/10.1016/j.jml.2020.104092>
- Minocher, R., Atmaca, S., Bavero, C., McElreath, R., & Beheim, B. (n.d.). Estimating the reproducibility of social learning research published between 1955 and 2018. *Royal Society Open Science*, 8(9), 210450. <https://doi.org/10.1098/rsos.210450>
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie Du Sert, N., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 1–9. <https://doi.org/10.1038/s41562-016-0021>
- Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, 23(10), 815–818.
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606.
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, 73, 719–748. <https://doi.org/10.1146/annurev-psych-020821-114157>
- Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, 45(3), 137–141. <https://doi.org/10.1027/1864-9335/a000192>
- Obels, P., Lakens, D., Coles, N. A., Gottfried, J., & Green, S. A. (2020). Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science*, 3(2), 229–237. <https://doi.org/10.1177/2515245920918872>
- Parsons, S. (n.d.). *Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability*. <https://doi.org/10.31234/osf.io/y6tcz>
- Pashler, H., & Harris, C. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, 7(6), 531–536. <https://doi.org/10.1177/1745691612463401>
- Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6), 528–530. <https://doi.org/10.1177/1745691612465253>
- Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68(9), 1046–1058. <https://doi.org/10.1016/j.jclinepi.2015.05.029>
- Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in s and s-plus (statistics and computing)*. Springer.

- Poline, J.-B., Strother, S. C., Dehaene-Lambertz, G., Egan, G. F., & Lancaster, J. L. (2006). Motivation and synthesis of the FIAC experiment: Reproducibility of fMRI results across expert analyses. *Human Brain Mapping*, 27(5), 351–359. <https://doi.org/10.1002/hbm.20268>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). sage.
- Ricketts, J., Dawson, N., & Davies, R. (2021). The hidden depths of new word knowledge: Using graded measures of orthographic and semantic learning to measure vocabulary acquisition. *Learning and Instruction*, 74, 101468. <https://doi.org/10.1016/j.learninstruc.2021.101468>
- Roche, D. G., Kruuk, L. E. B., Lanfear, R., & Binning, S. A. (2015). Public data archiving in ecology and evolution: How well are we doing? *PLoS Biology*, 13(11), 1–12. <https://doi.org/10.1371/journal.pbio.1002295>
- Rodríguez-Ferreiro, J., Aguilera, M., & Davies, R. (2020). Semantic priming and schizotypal personality: reassessing the link between thought disorder and enhanced spreading of semantic activation. *PeerJ*, 8, e9511. <https://doi.org/10.7717/peerj.9511>
- Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., ... McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15), 8398–8403. <https://doi.org/10.1073/pnas.1915006117>
- Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*, 31(1), e2295. <https://doi.org/10.1002/icd.2295>
- Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science*, 16(4), 744–755. <https://doi.org/10.1177/1745691620966795>
- Schweinsberg, M., Feldman, M., Staub, N., Akker, O. R. van den, Aert, R. C. M. van, Assen, M. A. L. M. van, Liu, Y., Althoff, T., Heer, J., Kale, A., Mohamed, Z., Amireh, H., Venkatesh Prasad, V., Bernstein, A., Robinson, E., Snellman, K., Amy Sommer, S., Otner, S. M. G., Robinson, D., ... Luis Uhlmann, E. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes*, 165, 228–249. <https://doi.org/10.1016/j.obhdp.2021.02.003>
- Sedlmeier, P., & Gigerenzer, G. (1989). Statistical power studies. *Psychological Bulletin*, 105(2), 309–316.
- Silberzahn, R., & Uhlmann, E. L. (2015). Crowdsourced research: Many hands make tight work. *Nature*, 526(7572), 189–191. <https://doi.org/10.1038/526189a>
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M., Dalla Rosa, A., Dam, L., Evans, M., Flores Cervantes, I., ... Nosek, B. (2017). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science*. <https://doi.org/10.1037/ampr.2017.0000001>

<https://doi.org/10.31234/osf.io/qkwst>

- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011b). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011a). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Snijders, T. A. B., & Bosker, R. J. (2004). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publications Ltd.
- Starns, J. J., Cataldo, A. M., Rotello, C. M., Annis, J., Aschenbrenner, A., Bröder, A., Cox, G., Criss, A., Curl, R. A., Dobbins, I. G., Dunn, J., Enam, T., Evans, N. J., Farrell, S., Fraundorf, S. H., Gronlund, S. D., Heathcote, A., Heck, D. W., Hicks, J. L., ... Wilson, J. (2019). Assessing Theoretical Conclusions With Blinded Inference to Investigate a Potential Inference Crisis. *Advances in Methods and Practices in Psychological Science*, 2(4), 335–349. <https://doi.org/10.1177/2515245919869583>
- Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016a). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702–712.
- Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016b). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702–712.
- Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, 8(1), 192. <https://doi.org/10.1038/s41597-021-00981-0>
- Towse, J. N., Ellis, D. A., & Towse, A. S. (2021). Opening Pandora's Box: Peeking inside Psychology's data sharing practices, and seven recommendations for change. *Behavior Research Methods*, 53(4), 1455–1468. <https://doi.org/10.3758/s13428-020-01486-1>
- Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, 123, 34–80.
- Vankov, I., Bowers, J., & Munafò, M. R. (2014). On the persistence of low power in psychological science. *Quarterly Journal of Experimental Psychology*, 67(5), 1037–1040. <https://doi.org/10.1080/17470218.2014.885986>
- Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, 59(5), 1311–1342. <https://doi.org/10.1515/ling-2019-0051>
- Vazire, S. (2018). Implications of the Credibility Revolution for Productivity, Creativity, and Progress. *Perspectives on Psychological Science*, 13(4), 411–417. <https://doi.org/10.1177/1745691617751884>
- Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, 605(7910), 423–425. <https://doi.org/10.1038/d41586-022-01332-8>
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & Maas, H. L. J. van der. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on bem (2011). *Journal of Personality and Social Psychology*, 100(3), 426–432. <https://doi.org/10.1037/a0022790>

- Wessel, I., Albers, C., Zandstra, A. R. E., & Heininga, V. E. (2020). *A multiverse analysis of early attempts to replicate memory suppression with the think/no-think task*.
- Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, 61(7), 726–728. <https://doi.org/10.1037/0003-066X.61.7.726>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'*. <https://cran.r-project.org/package=tidyverse>
- Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data.* ” O'Reilly Media, Inc.”
- Wild, H., Kyröläinen, A.-J., & Kuperman, V. (2022). How representative are student convenience samples? A study of literacy and numeracy skills in 32 countries. *PLOS ONE*, 17(7), e0271191. <https://doi.org/10.1371/journal.pone.0271191>
- Wilke, C. O. (n.d.). *Fundamentals of data visualization*. <https://clauswilke.com/dataviz/>
- Wilkinson, L. (2013). *The Grammar of Graphics*. Springer Science & Business Media.
- Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45, e1. <https://doi.org/10.1017/S0140525X20001685>
- Young, C. (2018). Model uncertainty and the crisis in science. *Socius*, 4, 2378023117737206.
- Young, C., & Holsteen, K. (2017). Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis. *Sociological Methods & Research*, 46(1), 3–40. <https://doi.org/10.1177/0049124115610347>