The research report: why, what and how

Rob Davies

02/11/2022

Table of contents

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).

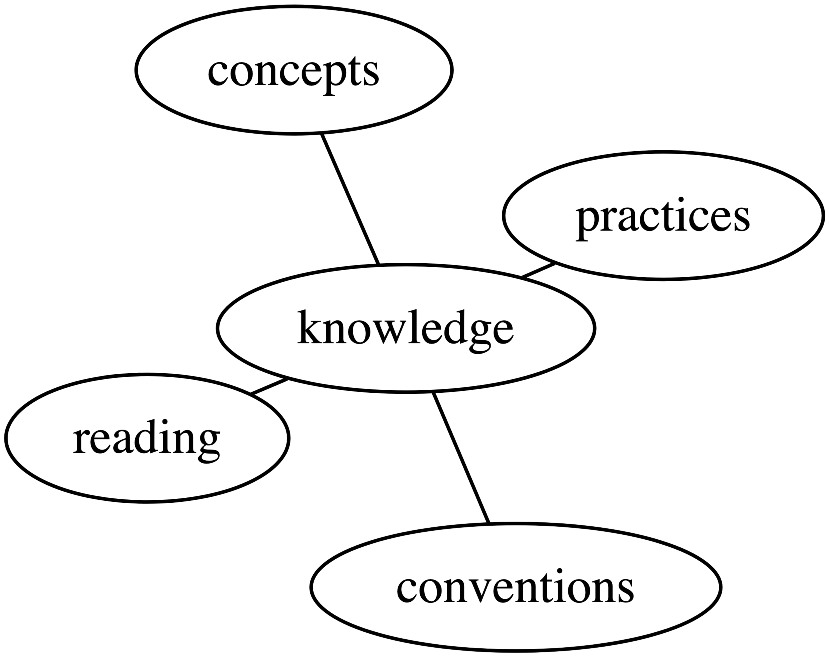


Figure 0.1: This is a simple graphviz graph.

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

The research report

1 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 in Chapter 2 and to the how information in Chapter 3.

1.1 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.

1.2 Why: what is the motivation for the assignment?

1.2.1 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.

1.2.2 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

[1] This 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

1.2.3 Multiverse analyses: multi- what?

1.2.3.1 A first useful metaphor: the pipeline

I am going to link this discussion to a metaphor (see Figure Figure 4.1) or a description you will find useful: **the data analysis pipeline** or **workflow**.

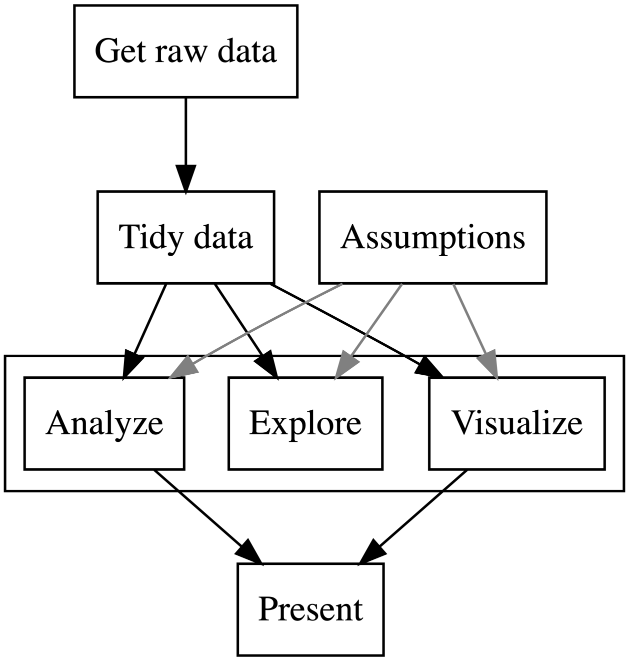


Figure 1.1: The data analysis pipeline or workflow

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*.

1.2.3.2 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 1.2).

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

[2] The term is taken from the name of a short story by Jorge Luis Borges, "El jardin de senderos que se bifurcan".

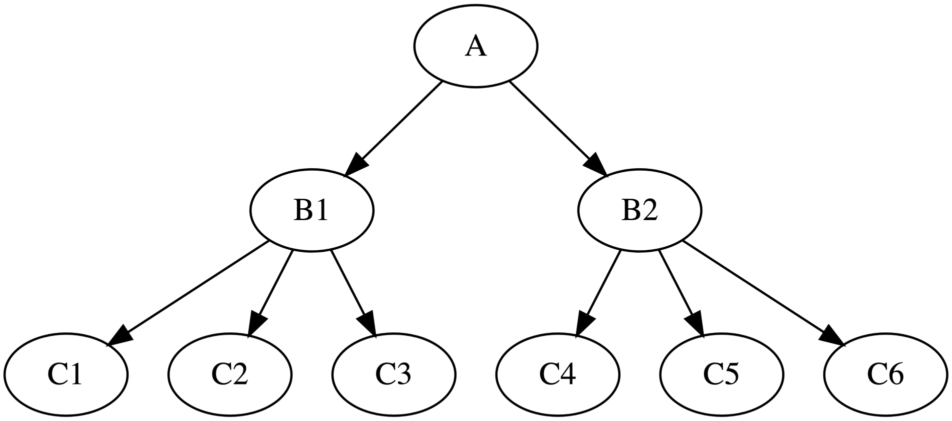


Figure 1.2: Forking paths in data analysis

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 makes sense to access data from a published study and to examine if you can get the same results as the study authors.

1.2.4 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

1.2.4.1 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).

1.2.4.2 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 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?

1.2.4.3 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, neuoscience, 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?

1.2.4.4 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 true [3]. 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.)

[3] There 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.

1.2.5 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.

1.2.5.1 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.*

1.2.5.2 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.

1.2.5.3 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.

1.2.6 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 1.2.6.2): are data shared and can we access the data? Secondly, I discuss analyses of methods reproducibility, and shared data and code usability (see Section 1.2.6.3): 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.

1.2.6.1 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.).

1.2.6.2 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 access [4] 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.

[4] Open 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).

1.2.6.3 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.

1.2.6.3.1 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.

1.2.6.3.2 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.

1.3 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.

1.3.1 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.

2 What

2.1 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 Padraic 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

2.1.1 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 Padraic 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.

2.1.2 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.

2.1.3 What content should reports present?

2.1.3.1 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:-

2.1.3.1.1 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.

2.1.3.1.2 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?

2.1.3.2 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.

2.1.3.2.1 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.

2.1.3.2.2 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.

2.1.4 What format?

2.1.4.1 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.

2.1.4.2 Add a link to the data analysed for the report

3 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**.

3.1 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 3.2). I then look at working with datasets where the data are presented without an analysis (see Section 3.3). Our advice on working with datasets presented without an analysis will overlap in key respects with our advice on working with curated data.

3.2 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 3.2.1). Then I look at the task of figuring out what analysis you can do with the data (see Section 3.2.2). Obviously, you cannot consider an analysis if you cannot be sure that you can work with the data (Minocher et al., n.d.).

3.2.1 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**.

3.2.1.1 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*..

3.2.1.1.1 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 1.2.6.2) 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.

3.2.1.2 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* 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?

3.2.1.3 Check

If you have located an interesting article with data that you can access, and if you have read the introductory notes (see Section 1.2.6.3), 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**.

3.2.2 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 3.2).
- Some students choose to complete an alternate analysis of the data, varying elements of the analysis (see Section 1.2.4).

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?

3.2.2.1 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, 2020).

In the research on the reproducibility of results in the literature (see Section 1.2.6.3), 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.

3.2.2.2 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 1.2.4).

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.

3.2.2.2.1 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 1.2.4.2) and (2.) analyses of the impact on results of varying what analysis method you use, or how you use the method (see Section 1.2.4.3). 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 1.2.4.2, 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 1.2.6.3.1. 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 1.2.4.3, 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.

3.2.3 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.

3.3 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 3.3.2.

3.3.1 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.

3.3.1.1 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.

3.3.1.2 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.

3.3.1.3 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.

3.3.1.4 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_Animal_reminder`, `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.

3.3.2 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.

3.4 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 II

Making the most of your skills

4 Data visualization

4.1 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 4.9.

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.

4.2 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 1.2.3.1, 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.

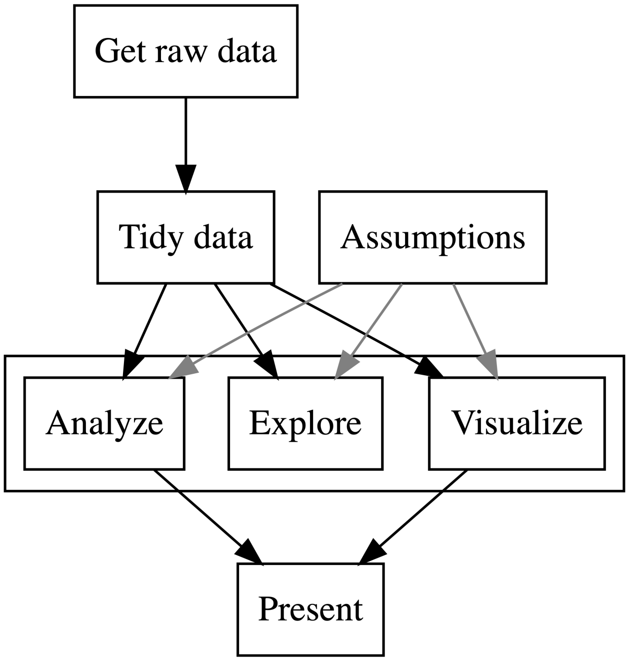


Figure 4.1: The data analysis pipeline or workflow

4.3 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 exerimental 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.

4.4 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.

4.4.1 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.

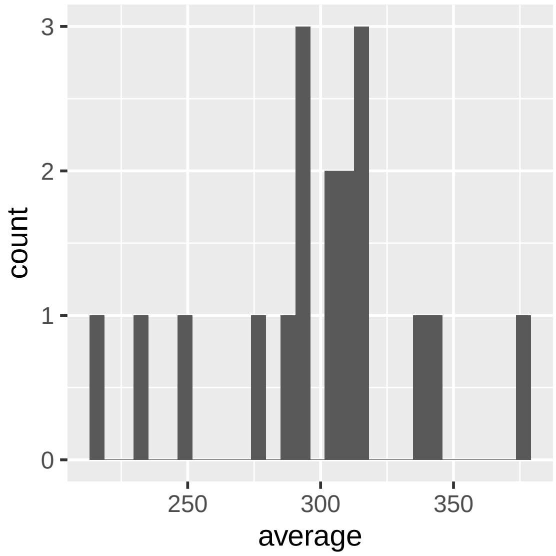
4.4.2 Pipes

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:

```r
sleepstudy %>%
  group_by(Subject) %>%
  summarise(average = mean(Reaction)) %>%
  ggplot(aes(x = average)) +
  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 commmenting 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.

4.5 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?

4.5.1 Purposes

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.

4.5.2 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.

4.6 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.

4.6.1 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`

4.6.2 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 4.2 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.

4.6.2.1 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()
```

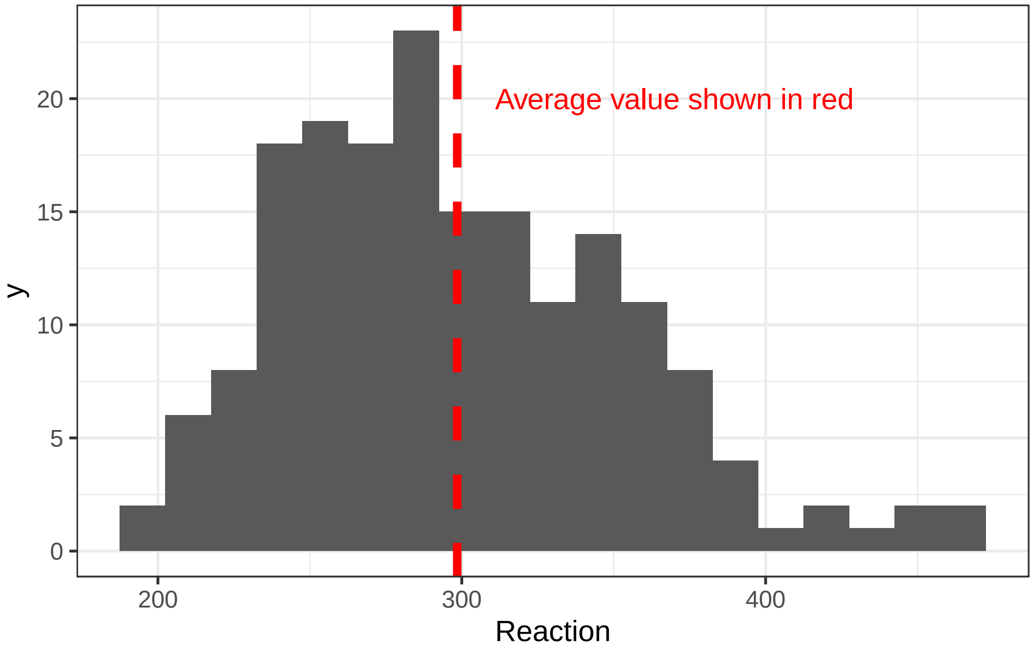


Figure 4.2: Figure showing a histogram of `sleepstudy` reaction time data

4.6.2.2 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 4.2 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.

4.6.3 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 4.3 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.

4.6.3.1 Plot

```r
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()
```

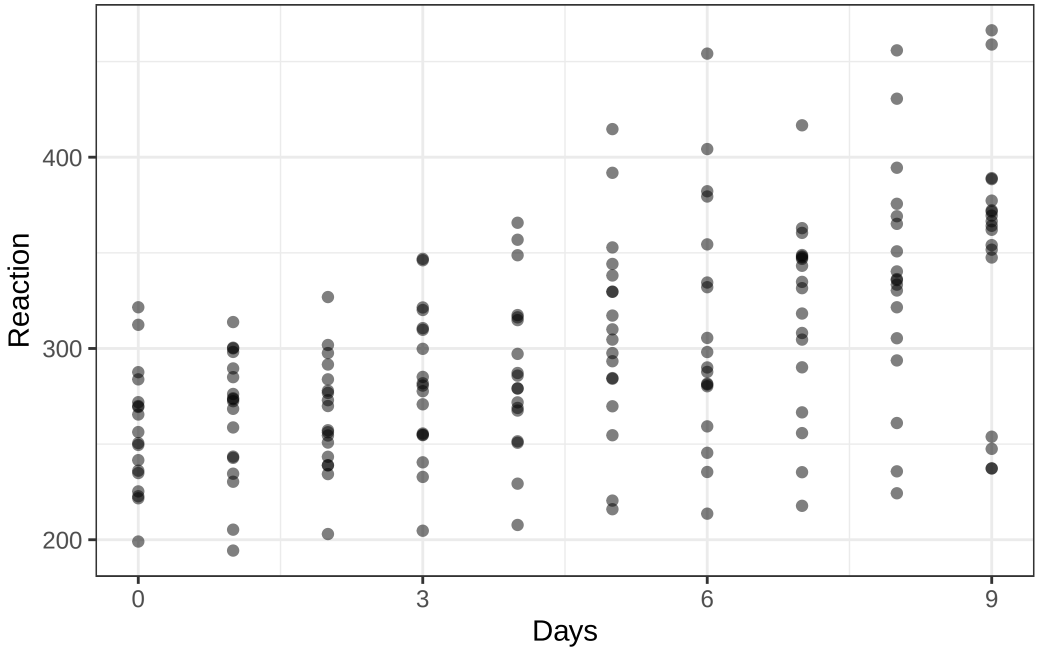


Figure 4.3: Figure showing a scatterplot of the relation between reaction time and days in the `sleepstudy` data

4.6.3.2 Notes

Notice the numbered steps in producing this plot.

```r
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-xis 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 4.4 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.

4.6.3.3 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()
```

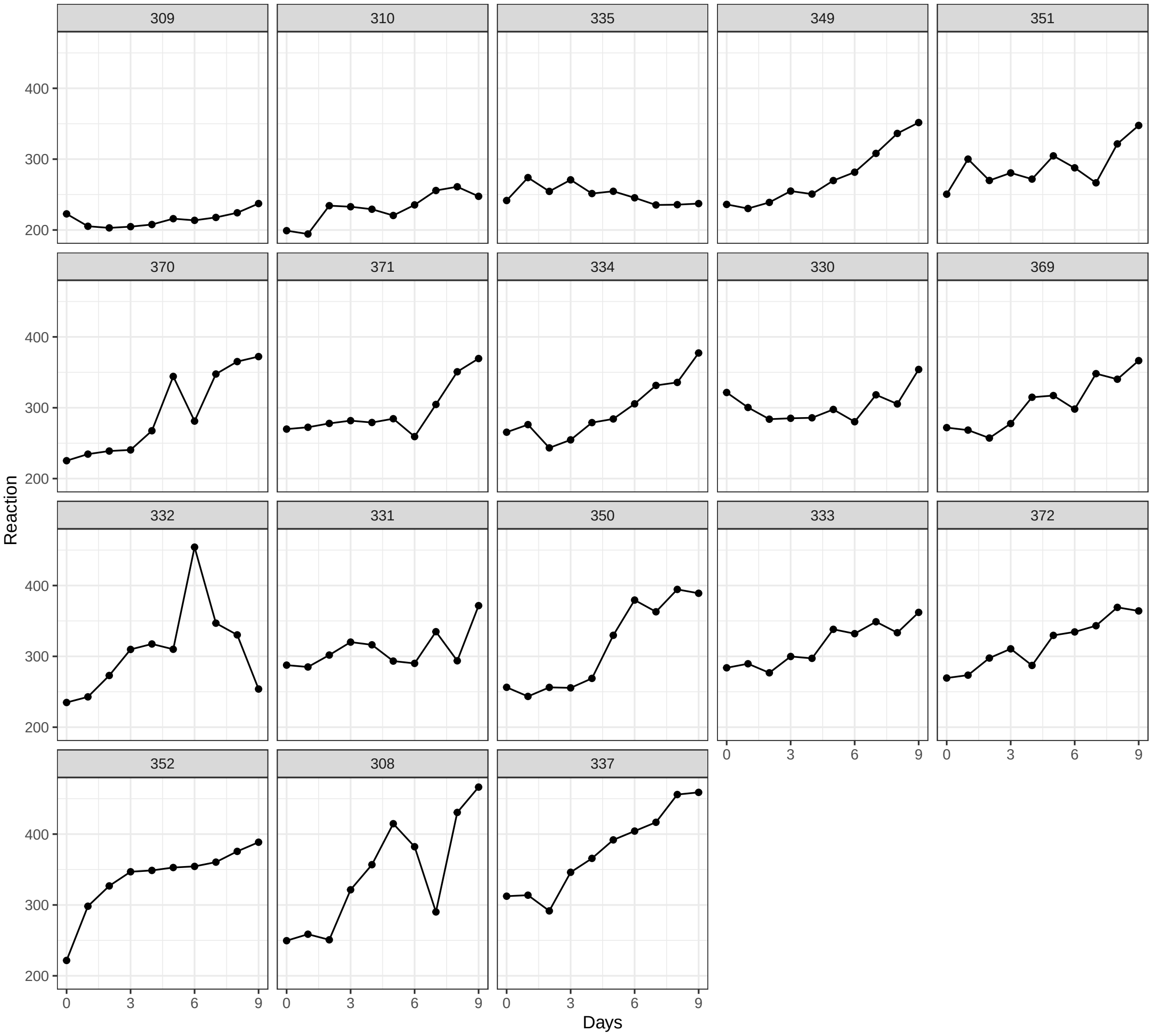


Figure 4.4: Figure showing a scatterplot of the relation between reaction time and days: here, we plot the data for each participant separately

4.6.3.4 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 4.4 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 (Vasishth & 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.

4.6.4 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.

4.7 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.

4.7.1 Set up for coding

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

4.7.1.1 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)
```

4.7.1.2 Get the data

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

4.7.1.2.1 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.

4.7.1.2.2 Get the data through a downloadable archive

Download the data.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.

4.7.2 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 x 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.

4.7.2.1 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).

- `WASImRS` — 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.

4.7.3 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 4.7.4.1).

```
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           WASImRS     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`).

4.7.4 Process the data

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

4.7.4.1 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.

```
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()

                        )
                      )
```

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)
```

4.7.4.1.1 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")
```

4.7.4.1.2 Further information

You can read more about `read_csv()` here

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

4.7.4.2 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 here [1] but it does require us to do some data processing, as I explain next.

[1] As 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:BPVSRS) %>%
  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
BPVSRS        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?

```
conc.orth.subjs <- conc.orth %>%
  group_by(Participant) %>%
  mutate(mean.score = mean(Levenshtein.Score)) %>%
  ungroup() %>%
  distinct(Participant, .keep_all = TRUE) %>%
  select(WASImRS:BPVSRS, 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:BPVSRS, mean.score, Participant)` then tells R to select just the columns with information about participant attributes. `(WASImRS:BPVSRS` tells R to select every column between `WASImRS` and `BPVSRS` 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
BPVSRS        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 tabled 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”.

4.7.4.2.1 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?

4.7.4.2.2 Further information

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

4.7.5 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.

4.7.6 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.

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

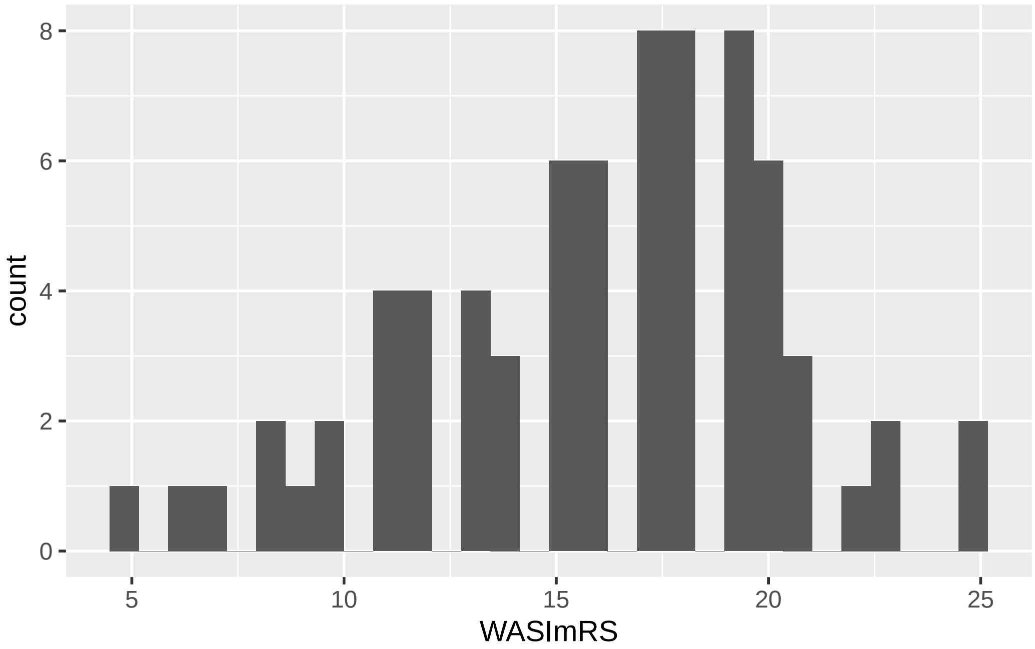


Figure 4.5: Distribution of WASImRS intelligence scores

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 4.5 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.

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

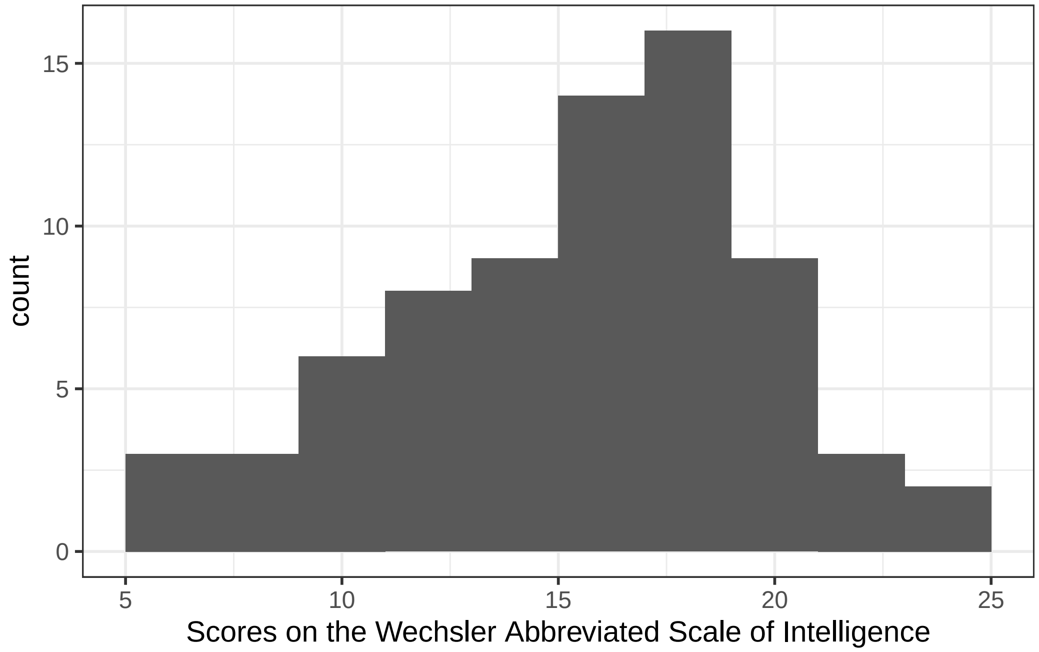


Figure 4.6: Distribution of WASImRS intelligence scores

Figure 4.6 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.

4.7.6.1 Exercise

We could, if we wanted, add a line and annotation to indicate the mean value, as you saw in Figure 4.2.

- 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?

4.7.7 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.

4.7.7.1 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 4.7 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).

```r
p.WASImRS <- ggplot(data = conc.orth.subjs, aes(x = WASImRS)) +
  geom_histogram(binwidth = 2) +
  labs(x = "WASI matrix") +
  theme_bw()

p.TOWREsweRS <- ggplot(data = conc.orth.subjs, aes(x = TOWREsweRS)) +
  geom_histogram(binwidth = 5) +
  labs(x = "TOWRE words") +
  theme_bw()

p.TOWREpdeRS <- ggplot(data = conc.orth.subjs, aes(x = TOWREpdeRS)) +
  geom_histogram(binwidth = 5) +
  labs(x = "TOWRE phonemic") +
  theme_bw()

p.CC2regRS <- ggplot(data = conc.orth.subjs, aes(x = CC2regRS)) +
  geom_histogram(binwidth = 2) +
  labs(x = "CC regular words") +
  theme_bw()

p.CC2irregRS <- ggplot(data = conc.orth.subjs, aes(x = CC2irregRS)) +
  geom_histogram(binwidth = 2) +
  labs(x = "CC irregular words") +
  theme_bw()

p.CC2nwRS <- ggplot(data = conc.orth.subjs, aes(x = CC2nwRS)) +
  geom_histogram(binwidth = 2) +
  labs(x = "CC nonwords") +
  theme_bw()

p.WASIvRS <- ggplot(data = conc.orth.subjs, aes(x = WASIvRS)) +
  geom_histogram(binwidth = 2) +
  labs(x = "WASI vocabulary") +
  theme_bw()

```

```
p.BPVSRS <- ggplot(data = conc.orth.subjs, aes(x = BPVSRS)) +
  geom_histogram(binwidth = 3) +
  labs(x = "BPVS vocabulary") +
  theme_bw()

p.mean.score <- ggplot(data = conc.orth.subjs, aes(x = mean.score)) +
  geom_histogram(binwidth = .25) +
  labs(x = "Mean orthographic test score") +
  theme_bw()

p.mean.score + p.BPVSRS + p.WASIvRS + p.WASImRS +
  p.CC2nwRS + p.CC2irregRS + p.CC2regRS +
  p.TOWREpdeRS + p.TOWREsweRS + plot_layout(ncol = 3)
```

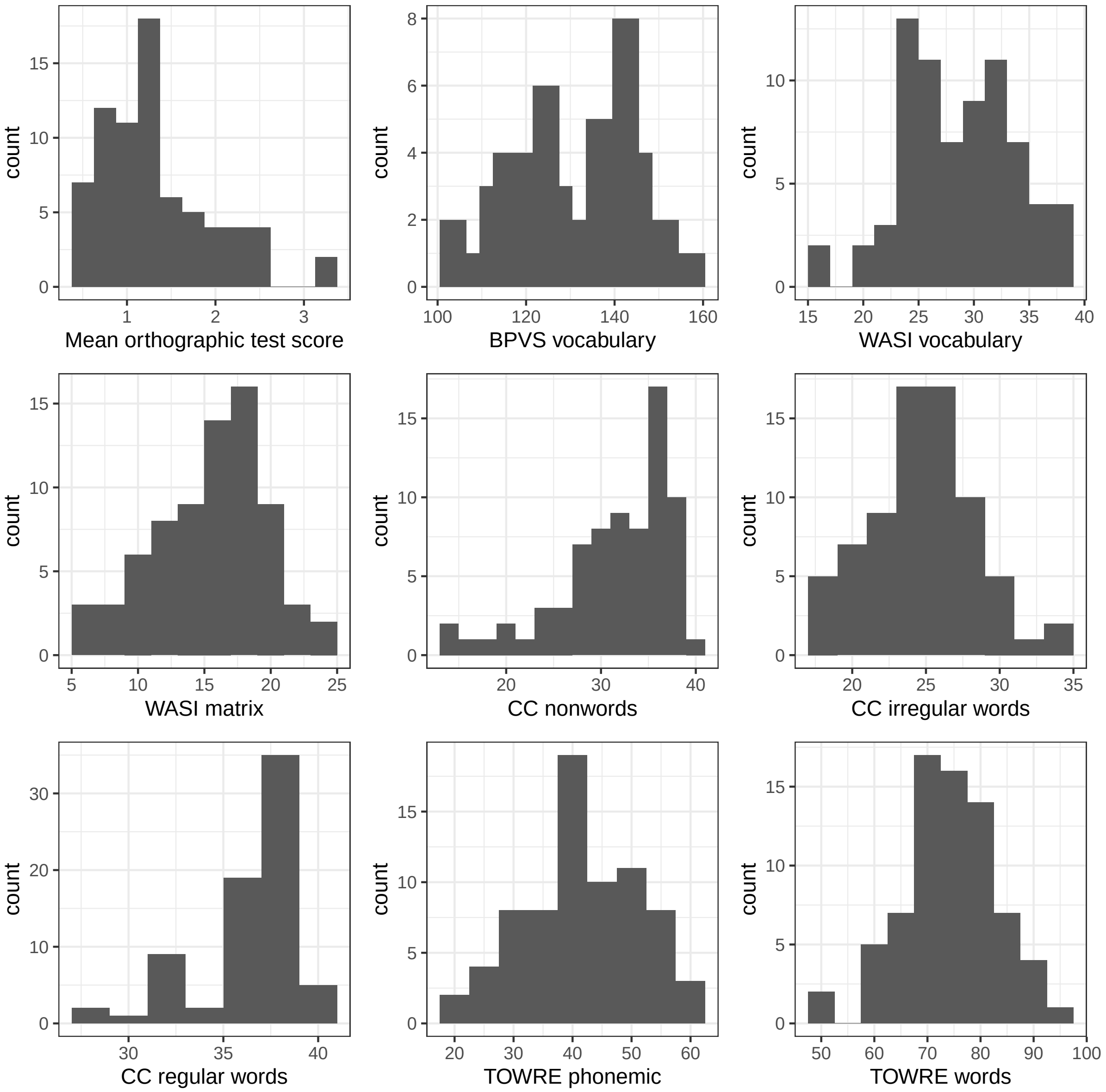


Figure 4.7: Distribution of childrens' scores on ability measures

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.BPVSRS + 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.BPVSRS + 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 4.7 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 the 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 4.7, but this style allows you to gain an impression, quickly, of how for 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.

4.7.7.2 Plot

```r
library(ggridges)
conc.orth.subjs %>%
  pivot_longer(names_to = "task", values_to = "score", cols = WASImRS:mean.score) %>%
  ggplot(aes(y = task, x = score)) +
  geom_density_ridges(stat = "binline", bins = 20, scale = 0.95, draw_baseline = FALSE) +
  theme_ridges()
```

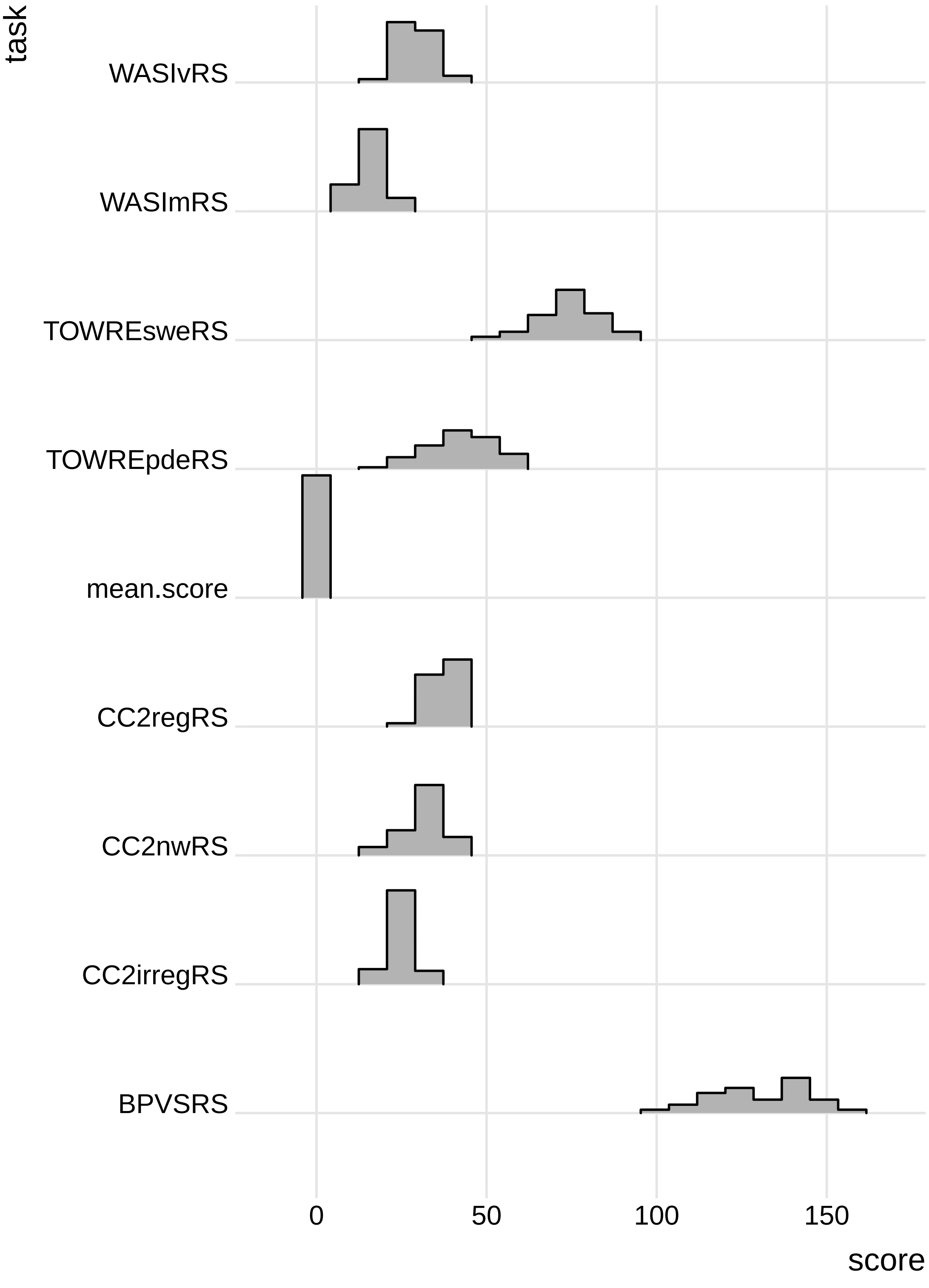


Figure 4.8: Distribution of childrens' scores on ability measures

4.7.7.3 Notes

1. `library(ggridges)` 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" wwith 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 `ggridges` here.

4.7.7.4 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 4.7.4.2) 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.

4.7.7.4.1 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 4.7.4.2 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(WASImRS: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.

4.7.7.4.2 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 4.9 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. facetted 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.

4.7.7.5 Plot

4.7.7.6 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 =
  geom_histogram(binwidth = 5) +
  labs(x = "BPVS vocabulary score", title = "c. Superimposed histograms") +
  theme_bw()
```

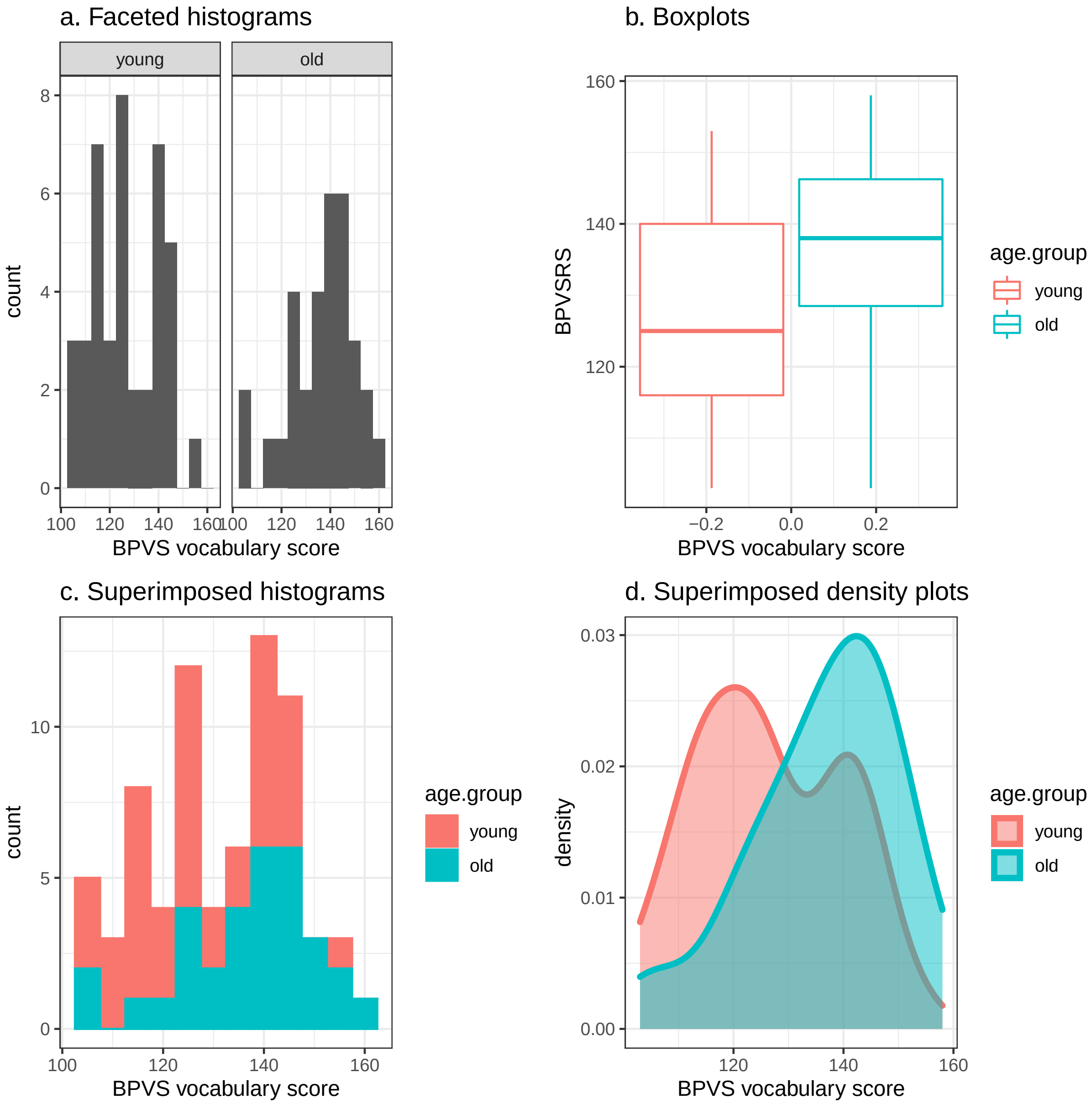


Figure 4.9: Distribution of childrens' scores on the BPVS vocabulary measure: distributions are compared between the younger and older age groups

```
p.colour.density <- ggplot(data = conc.orth.subjs, aes(x = BPVSRS, colour = age.group, fil
  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 = BPVSRS, 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 = BPVSRS, 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 = BPVSRS, 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 4.10.

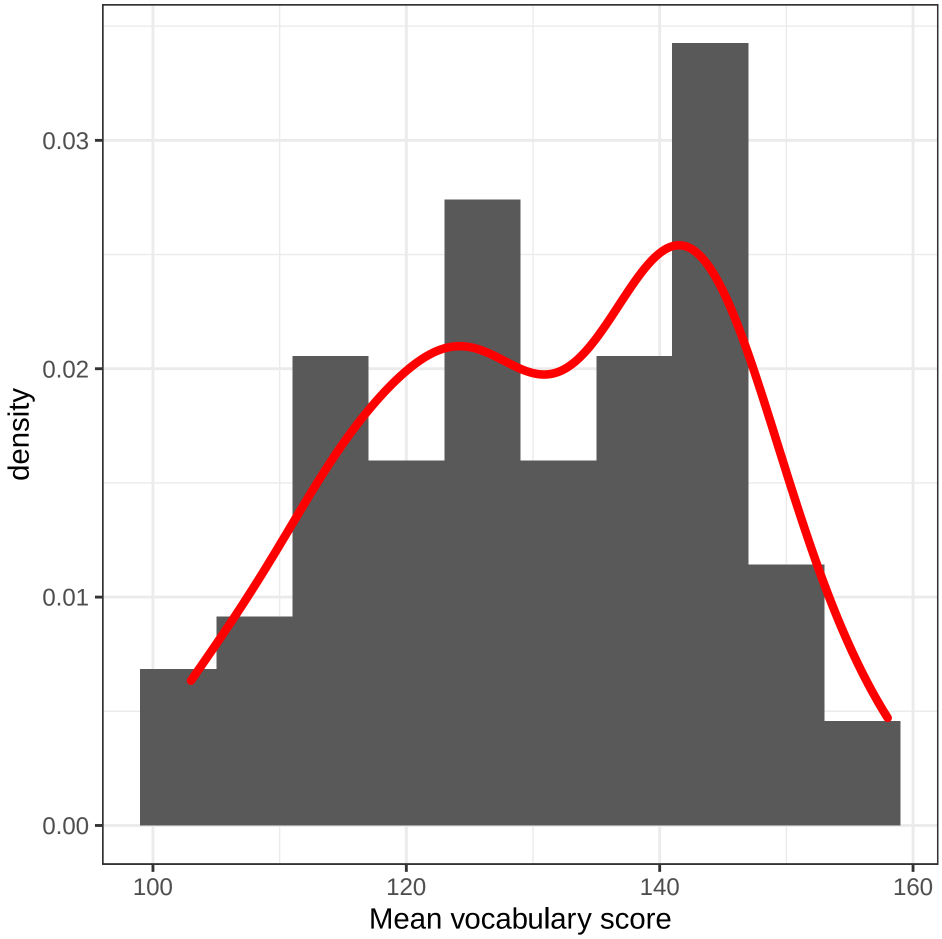


Figure 4.10: Distribution of childrens' scores on the BPVS vocabulary measure. The figure shows the histogram versus density plot representation of the same data distribution

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 4.11 shows the distribution of scores on both the BPVS and WASI vocabulary knowledge measures.

```r
p.BPVSRS.density <- ggplot(data = conc.orth.subjs, aes(x = BPVSRS, colour = age.group, fil
  geom_density(alpha = .5, size = 1.5) +
  geom_rug(alpha = .5) +
  geom_vline(xintercept = 120, linetype = "dashed") +
  geom_vline(xintercept = 140, linetype = "dotted") +
  labs(x = "BPVSRS vocabulary score") +
  theme_bw()

p.WASIvRS.density <- ggplot(data = conc.orth.subjs, aes(x = WASIvRS, colour = age.group, f
  geom_density(alpha = .5, size = 1.5) +
  geom_rug(alpha = .5) +
  labs(x = "WASI vocabulary score") +
  theme_bw()

p.BPVSRS.density + p.WASIvRS.density + plot_layout(guides = 'collect')
```

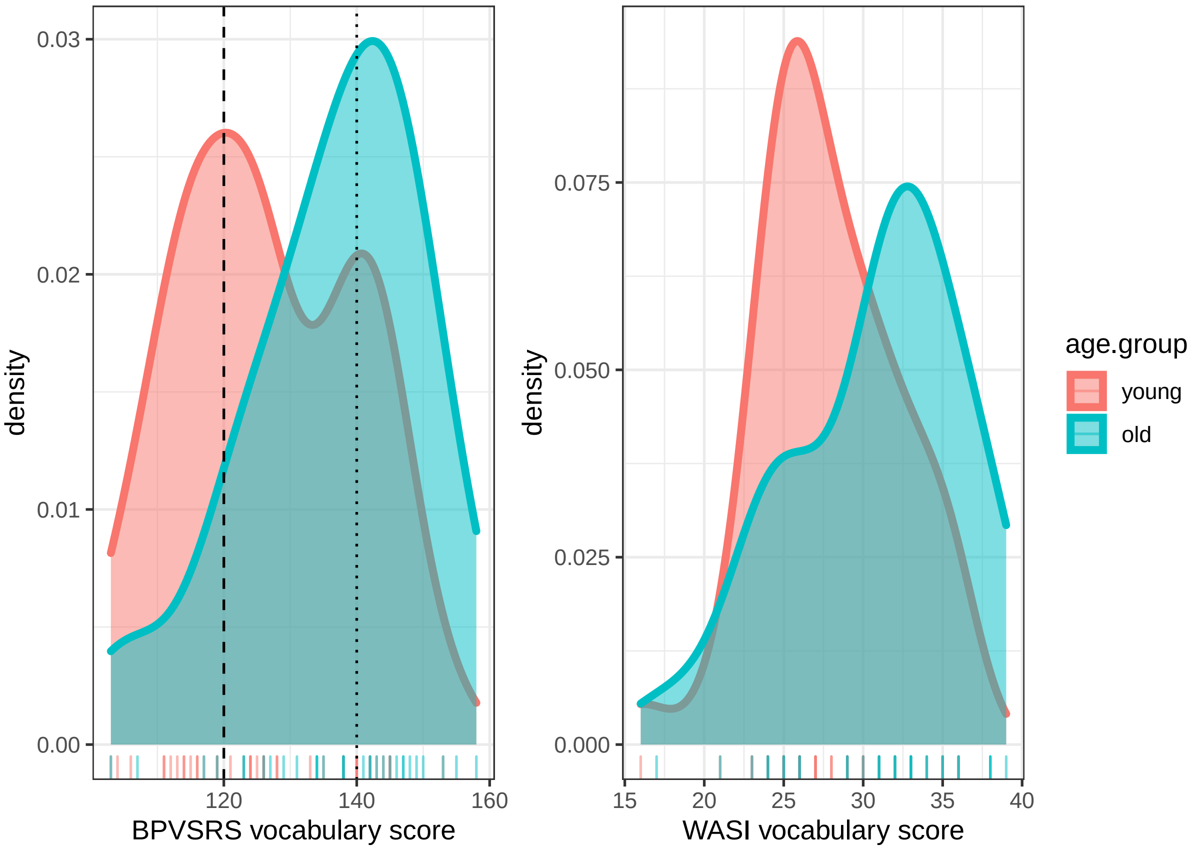


Figure 4.11: Distribution of childrens' scores on the BPVS and WASI vocabulary measures.

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.

4.7.7.6.1 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 4.11, 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.

4.7.7.6.2 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.

4.7.8 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.

4.7.9 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.

4.7.9.1 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.

```
ggplot(data = conc.orth.subjs, aes(x = WASIvRS, y = BPVSRS)) +
  geom_point() +
  labs(x = "WASI vocabulary score",
       y = "BPVSRS vocabulary score",
       title = "Are WASI and BPVS vocabulary scores associated?") +
  theme_bw()
```

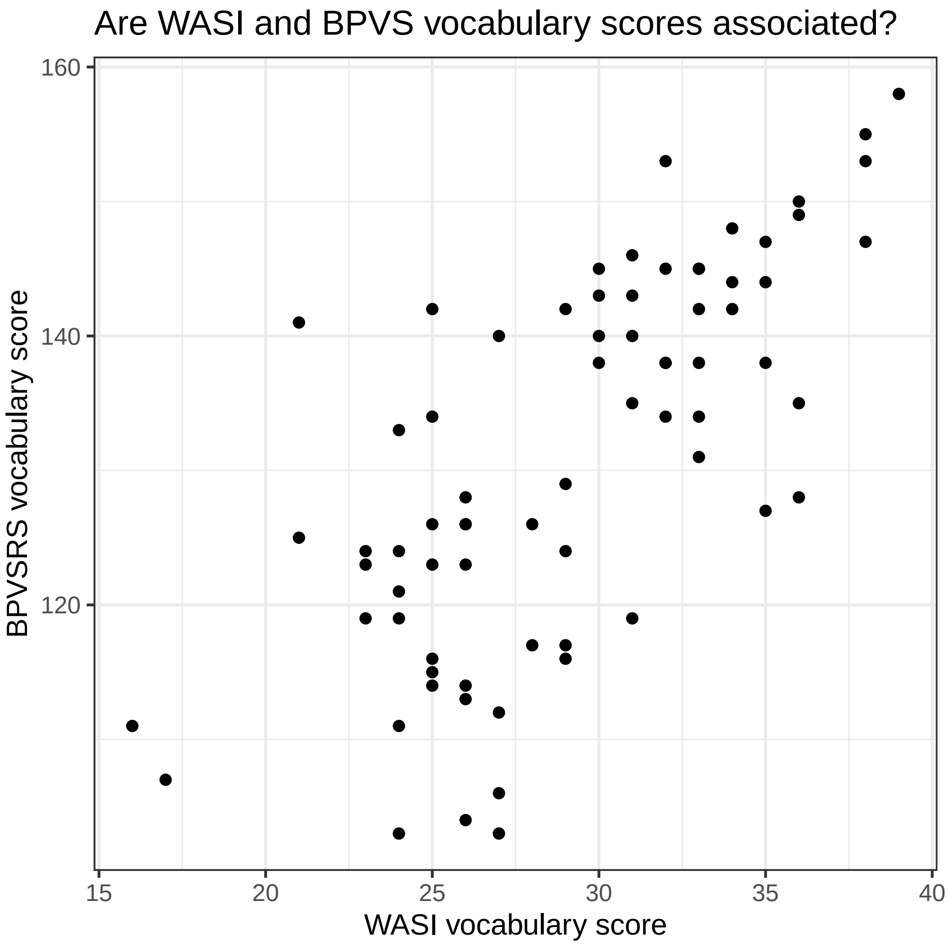


Figure 4.12: Scatterplot indicating the potential association of childrens' scores on the BPVS and WASI vocabulary measures.

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 4.12 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.

4.7.9.2 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.

```r
ggplot(data = conc.orth.subjs, aes(x = WASIvRS, y = BPVSRS)) +
  geom_point() +
  geom_smooth() +
  labs(x = "WASI vocabulary score",
       y = "BPVSRS vocabulary score",
       title = "Are WASI and BPVS vocabulary scores associated?") +
  theme_bw()
```

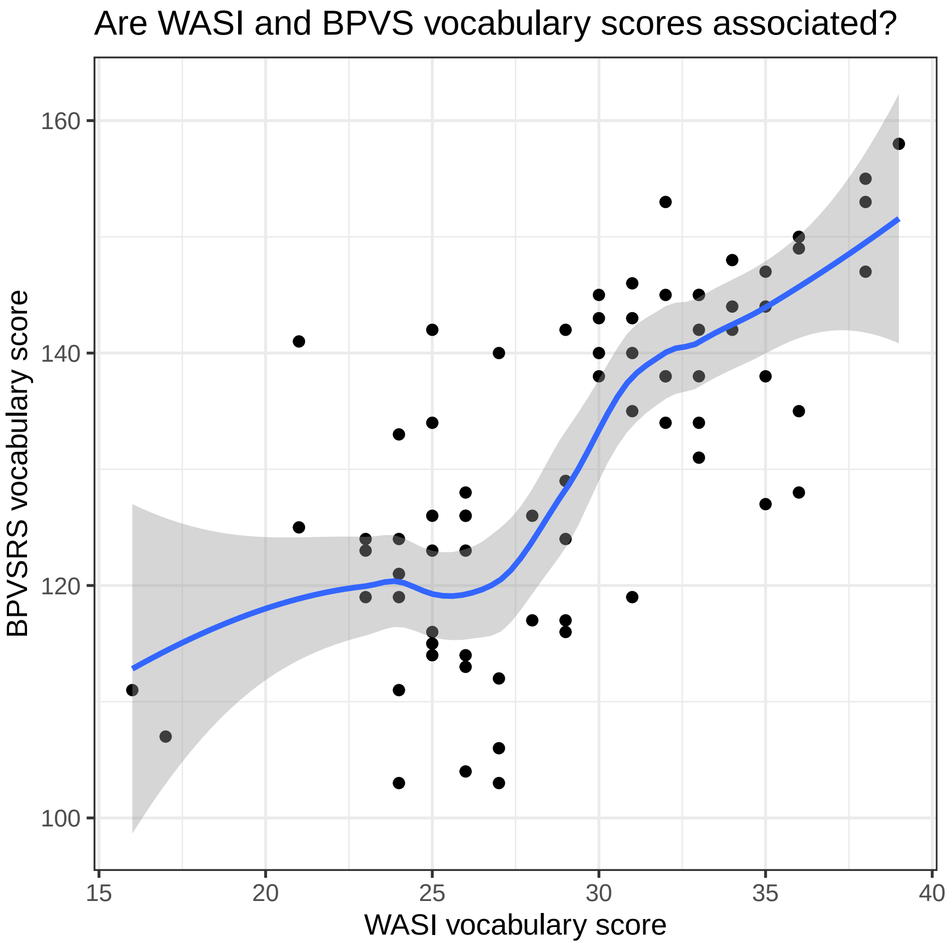


Figure 4.13: Scatterplot indicating the potential association of childrens' scores on the BPVS and WASI vocabulary measures.

The only coding difference between this plot Figure 4.13 and the previous plot Figure 4.12 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 4.13. 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 4.13 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.

```r
ggplot(data = conc.orth.subjs, aes(x = WASIvRS, y = BPVSRS)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(x = "WASI vocabulary score",
       y = "BPVSRS vocabulary score",
       title = "Are WASI and BPVS vocabulary scores associated?") +
  theme_bw()
```

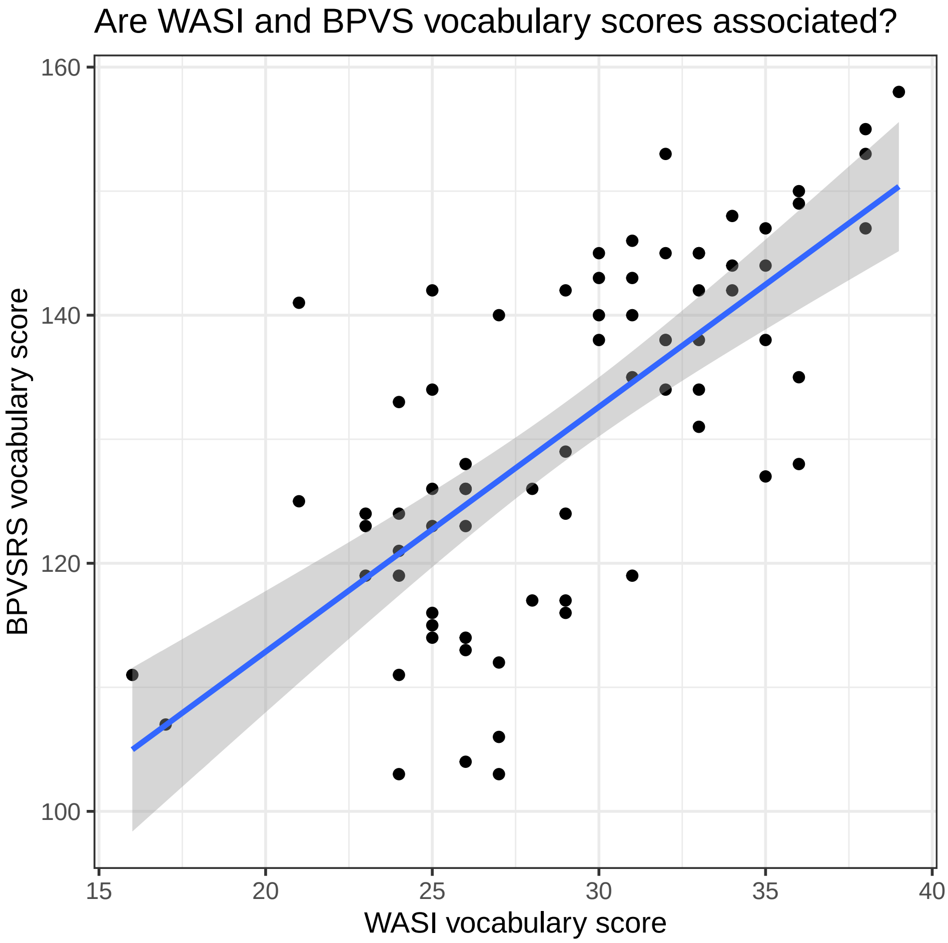


Figure 4.14: Scatterplot indicating the potential association of childrens' scores on the BPVS and WASI vocabulary measures.

Notice the difference between Figure 4.13 and Figure 4.14:

- `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.

4.7.9.3 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 4.14. 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 there are points indicating that different children have the same WASI score of 25 but BPVS scores ranging from about 115 to 140.

4.7.9.4 Polish the appearance of a plot for presentation

Figure 4.14 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.

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

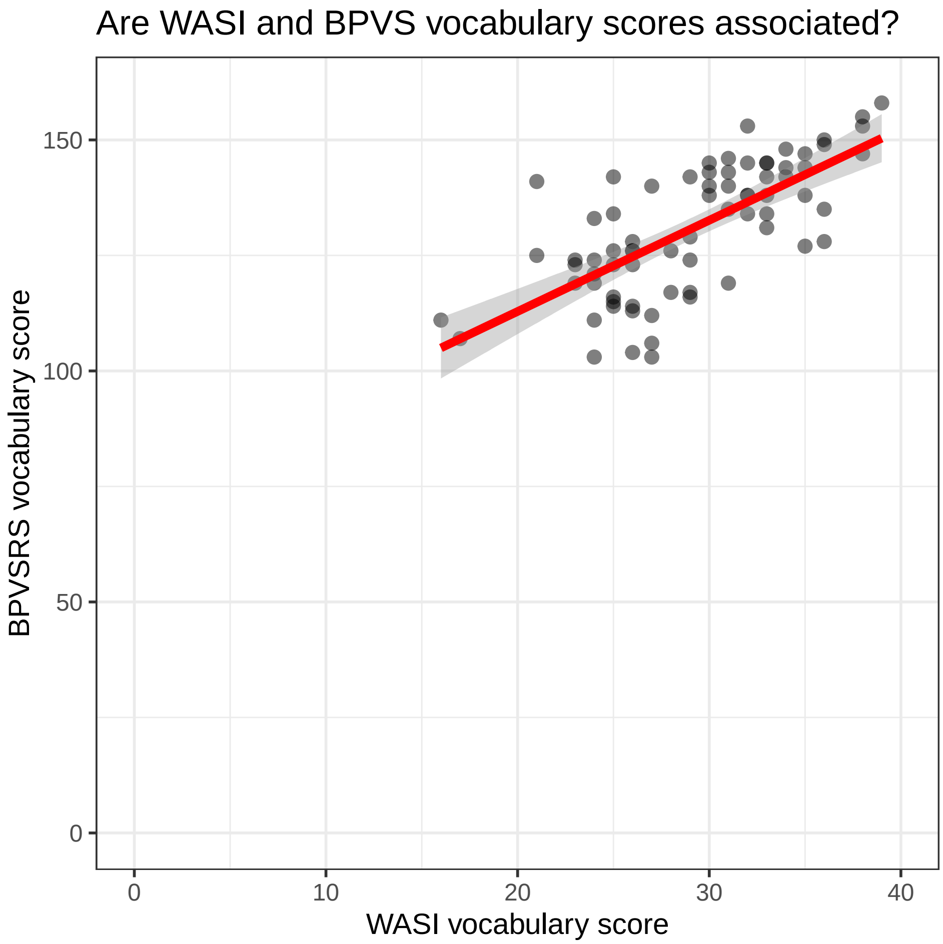


Figure 4.15: Scatterplot indicating the potential association of childrens' scores on the BPVS and WASI vocabulary measures.

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 trans-

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.

4.7.9.5 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 4.16 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 4.7.2, 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 4.16 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.

4.7.9.6 Plot

4.7.9.7 Notes

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

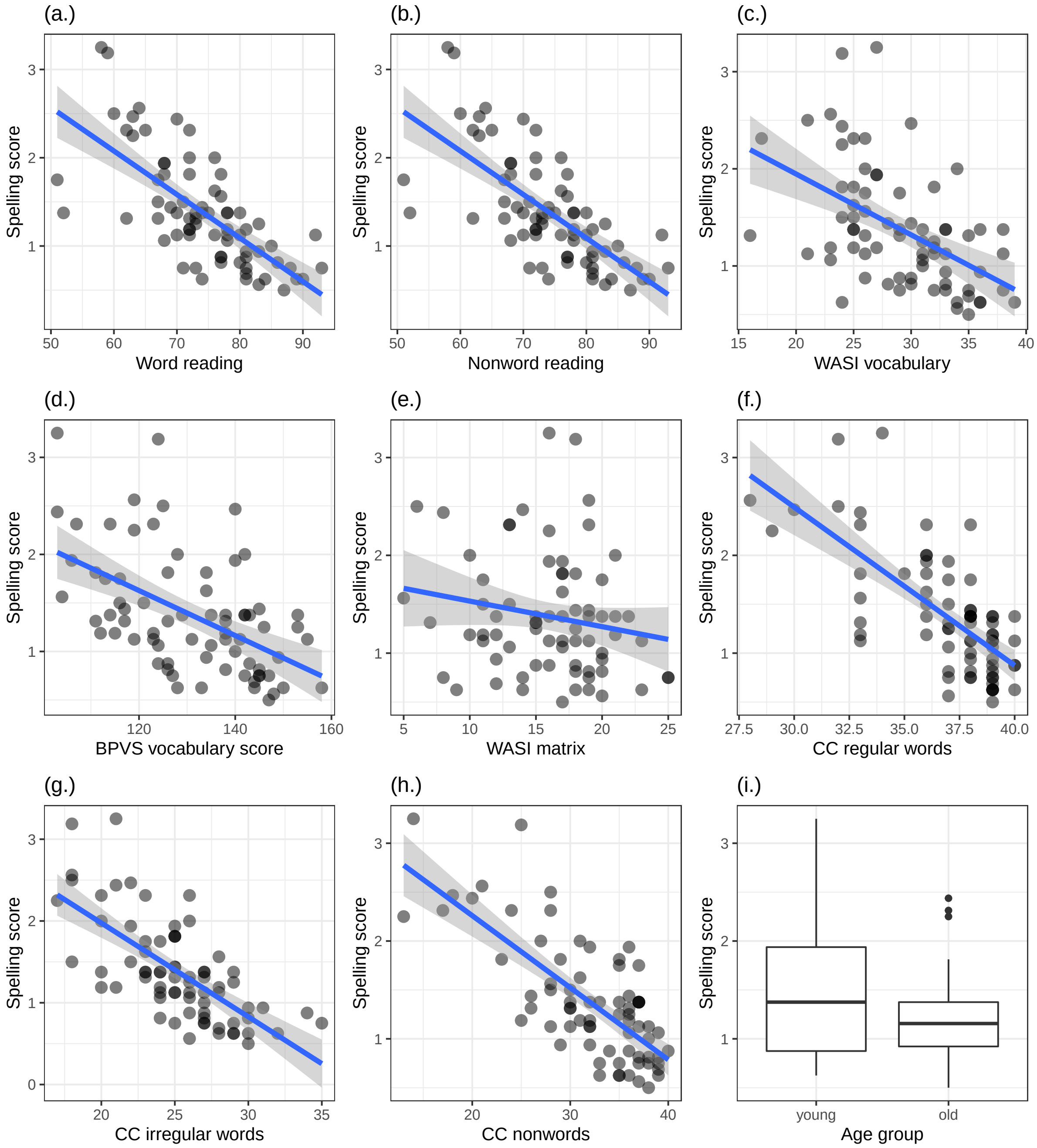


Figure 4.16: Grid of scatterplots showing the potential association between mean spelling score, for each child, and variation in the Castles and Coltheart (CC) and TOWRE measures of word or nonword reading skill, WASI and BPVS measures of vocabulary knowledge, the WASI matrix measure of intelligence, 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.BPVSRSvsmean.score <- ggplot(data = conc.orth.subjs,
                              aes(x = BPVSRS,
                                  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.BPVSRSvsmean.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.BPVSRSvsmean.score + p.WASImRSvsmean.score + p.CC2regRSvsmean.score +
  p.CC2irregRSvsmean.score + p.CC2nwRSvsmean.score + p.age.groupvsmean.score
```

Figure 4.16 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.

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

As explained in Section 4.7.2, in the Ricketts et al. (2021) study, we taught children taught 16 novel words in a study with a *2 x 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 4.1: 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:

```r
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 4.17.

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

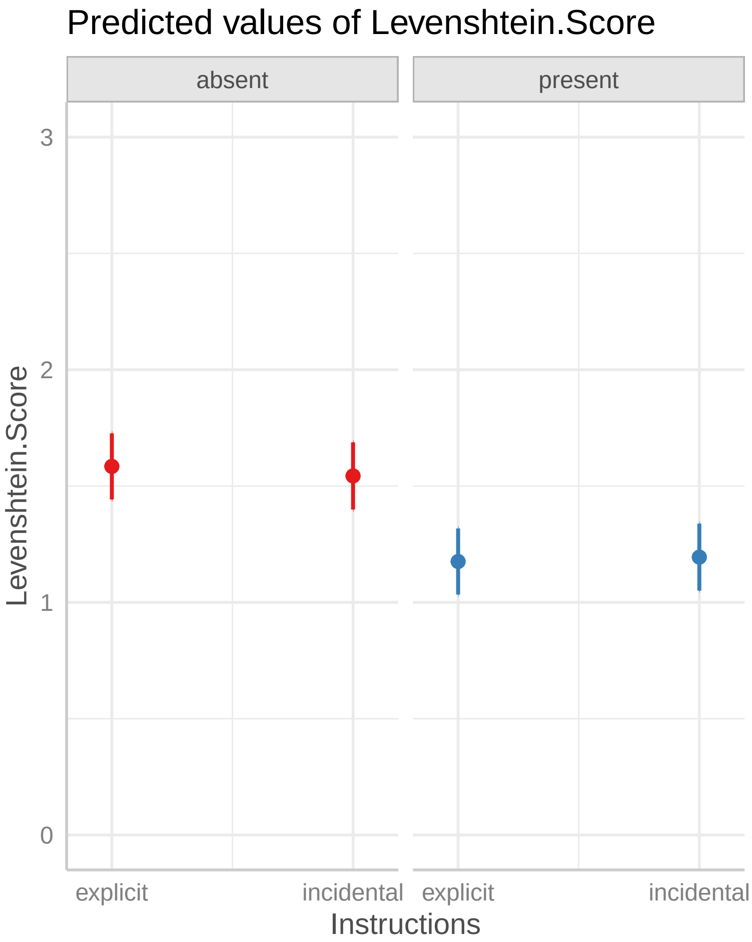


Figure 4.17: Dot and whisker plots showing the predicted effect on outcome spelling (Levenshtein) score, given different experimental conditions: Orthography (present, absent) x Instruction (explicit, incidental).

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 4.17, 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 4.17, 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 4.18.

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

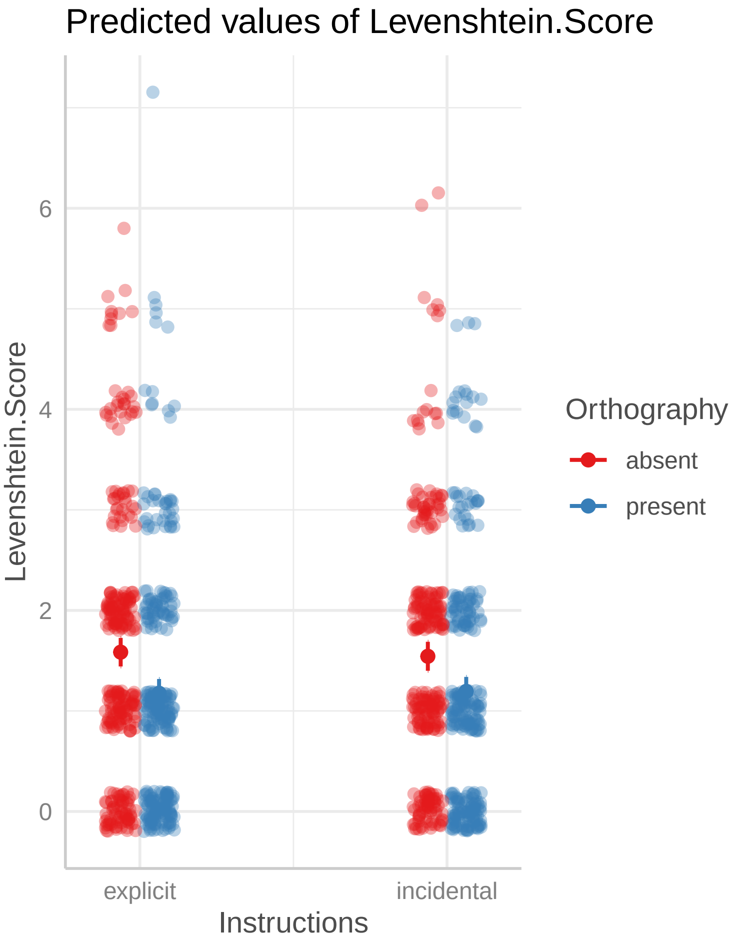


Figure 4.18: Dot and whisker plots showing the predicted effect on outcome spelling (Levenshtein) score, given different experimental conditions: Orthography (present, absent) x Instruction (explicit, incidental). The estimates are shown as dot-whisker points. In addition, the plot shows as points the spelling score observed for each child for each response recorded in the `conc.orth` dataset.

Figure 4.18 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.

```r
score.by.subj <- conc.orth %>%
  group_by(Participant, Orthography) %>%
  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.

```
score.by.subj.diff <- score.by.subj %>%
  pivot_wider(names_from = Orthography, values_from = mean.score) %>%
  mutate(difference.score = absent - present) %>%
  pivot_longer(cols = c(absent, present),
               names_to = 'Orthography',
               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 4.19 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).

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

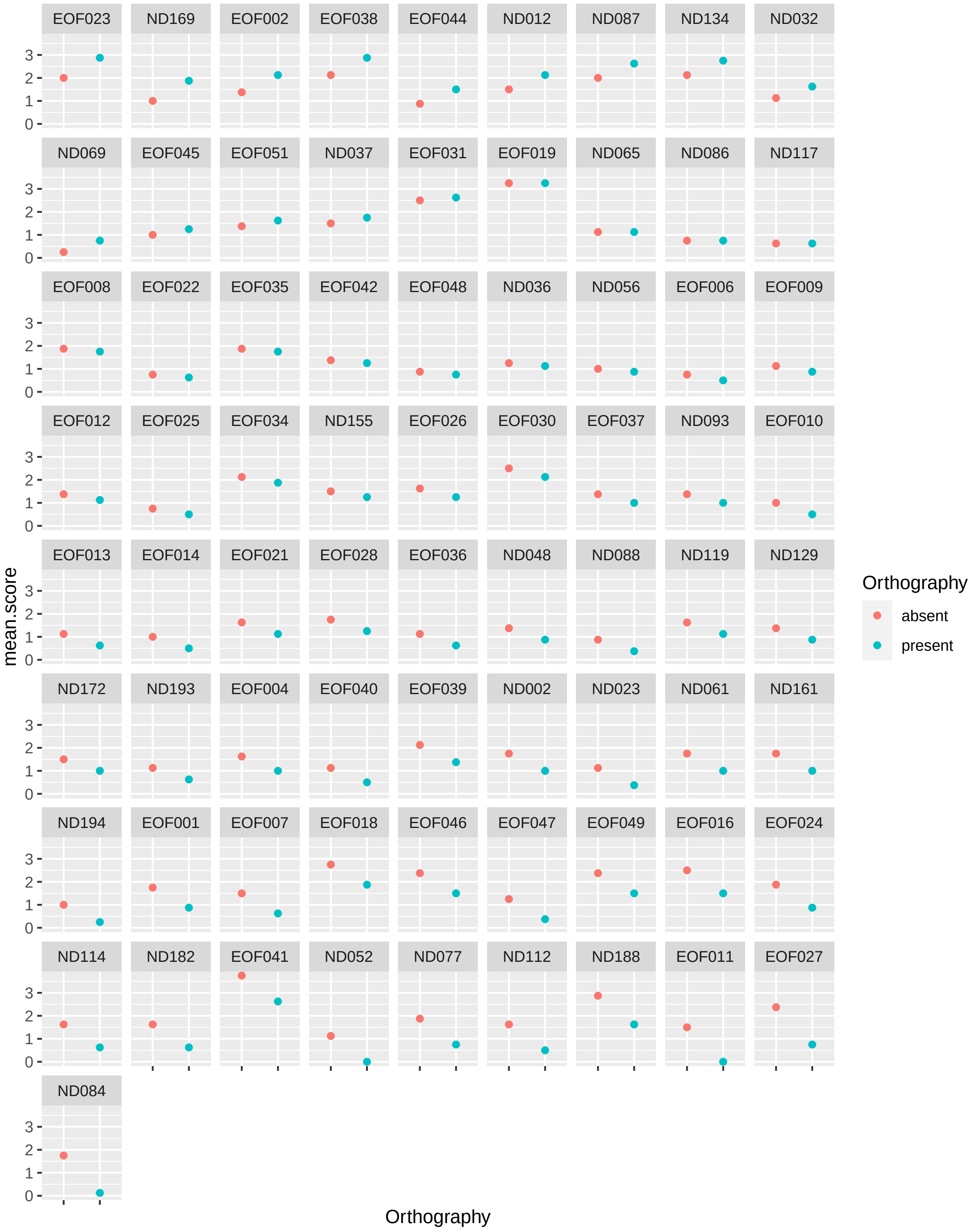


Figure 4.19: Dot plots showing the different outcome spelling (Levenshtein) scores, for each participant, in the different experimental conditions: Orthography (present, absent). Plots are ordered, from top left to bottom right, by the difference between mean spelling scores in the absent versus present 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.

4.7.11 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.

4.8 Next steps for development

To take your development further, take a look at the resources listed in Section 4.9.

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 4.7.1.2.1, 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 schizotypyal traits) and experimental manipulations (of semantic priming) so you can do similar things with those data as we have explored here.

4.9 Helpful resources

4.9.1 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)/

- 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

4.9.2 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/

5 R knowledge

Part III

End

6 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

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., Oravecz, 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., Apthorp, 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., & Houlson, 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/10.1177/15291006211051956

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

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

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/dyaa164

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.

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., Derks, 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

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. (2020). 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

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/10.31234/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

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