Operations Manual

For Reber's Research Methods 301

Table of Contents

Preface

These are chapters that include content specific to Research Methods as taught by Prof. Paul Reber at Northwestern University.

This includes:

- **Experiment 1**: a "levels of processing" design that can be completed quickly and used as examples of experimental design, analysis and reporting
- Specific scripts for running a t-test in R on these data as formatted
- Specific methods information for this experiment to guide the reporting process
- Additional background on levels of processing theory to support writing up the Experiment 1 design

- Hands-on practice scripts for running and interpreting 2x2 designs using ANOVA to statistically evaluate the outcomes

- **Experiment 2**: an extended 2x2 design on levels of processing for which the students can ask as co-experimenters
- Specific scripts for running an ANOVA in R on these data as formatted
- Specific methods information for reporting this experiment in APA format

- **Final Projects**: Guidelines for small-group class projects to be run by students as the final project for the class
- How to prepare the final project proposal
- Guidlines for presenting the final project design to the class

Note that this document is intended to accompany specific set of files made available through educational software (currently Canvas).

On Canvas will be the data files used for analysis. Data files with the ".xlsx" extension are intended to be used with Microsoft Excel. Northwestern University provides a site license for all Microsoft products to students. You will likely need to download the full Excell app as the online versions of the software do no include the full set of options for preparing data visualizations (the same tools are missing from the Google Sheets software at the current time). Files that have the ".csv" extension are of a format called comma separated values, which are suitable for input to analysis programs like R. Files that have the ".R" extension are script files to be used with the free software program R within the RStudio software package.

Additional publications will also be available on Canvas. Some of these are background for Experiments 1 and 2 to help prepare reports of our in-class experiments. Others are example research reports used to practice analysis of methodology, drawing conclusions from data and interpreting experimental hypotheses.

Experiment 1

Learning Research with a Hands-on Approach

- Participate in a short psychology experiment using the QR code or the link below. When you have finished you will get a Completion Code to enter as the answer to the first assignment for the class.
- *Note: the experiment and questions/discussion below will be covered on the first day of class. Review the Q&A below if you want a refresher for that discussion.*
- If you are trying to catch up on the content related to Experiment 1 because you did not have a chance to work through the experiment and then the analysis questions in class, use the QR code or URL to go to the reberlab.org site to participate in Experiment 1. It is not strictly necessary to contribute data to the class process of working up to the analysis and then written report of the results, but it is useful to have the perspective of the experiment participant as a starting point for explaining the methodology.

Or use the following link:
https://tinyurl.com/Reber205

What was this experiment about?

Refer to Chapter 1 in the main text for a discussion of the design and implementation of this simple experiment and introduction of a set of key terms for describing experimental design. The chapter is written using a brief piece of the experiment as a design example, but after completing the online version, you have a fuller sense of the experience of being a participant.

In particular, after participating in the experiment online, you will have experienced a common feature of psychological studies that even after completing the study, it is not entirely clear what the experimenter was trying to do. You were not told at the beginning of the study that memory would be tested at the end and you only experienced one of the two different conditions used in the study (rating liking or counting vowels). Over the next several chapters, the reasons for these aspects of experimental design will be made clear as important parts of obtaining reliable data to draw conclusions from.

Experiment 1 as an example

Not only will the design of this experiment be used as an example of how psychological studies are constructed, but we will also be examining the data collected and analyzing these data in class. The statistical approach here is based on providing recipes for organizing experiment data and carrying out analysis using the program R/RStudio. You should have some prior experience with the mathematical foundations of statistics. However, in this class we will practically focus on selecting the correct analysis based on the design, running the analysis and then correctly interpreting the analysis output and reporting the results in standard APA format.

Experiment 1 Methods

We will be preparing a written report of the results of our in-class experiment later in the class. For this, you will need some additional information about the methodology in order to report that correctly.

Materials

A set of 60 words was used for the study and test stimuli. Words were seleted to have a written frequency of 30-80 per million and to be 5-8 letters in length.

There were the two different conditions in which participants rated the words in the first part, but as much as possible, everything else about the experience of being in the study was kept constant across conditions. For both rating tasks, the words were shown on screen for exactly 4 seconds, regardless of when the response was made. The total time in the trivia section was 3 minutes for all participants. All participants completed the same recognition test with the 30 words that had been rated and the same 30 new words (the order of the words was randomized across all participants).

Both Groups See the Same 30 Words

POCKET, PAINT, PRISON, QUARTER, CITIZEN, VEHICLE, ROUGH, BRAIN, TEMPLE, PRINCE, MEDICINE, FILLING, GUARD, JOURNAL, ENGINE, PALACE, GRAVE, BRANCH, CONCRETE, DANCER, SALARY, BASEMENT, MATCH, NATIVE, STABLE, FENCE, SWIMMING, QUEEN, OCEAN, FACTORY

In Chapters 3 and 4, we will review why these elements of design are important so that the results can support valid conclusions about how the IV affected the DV. Chapter 5 will disuccss sampling methodologies and how they influence the intepretation of restuls. Chapter 6 and the Chapter here on Experiment 1 Analysis will provide concrete guidance on carrying out the statistical analysis of the results. Chapter 7 will describe how to format a written scientific report to standard, APA (American Psychological Association) style. When carrying out that project, you will need to refer back to this section for these methodological

details to report.

For the Participants subsection of the written report, you will need information about the number of participants whose data is included in the analysis. Typically, we have data from at least 2 sections of Research Methods that are combined to provide a robust sample for testing our hypothesis about memory.

For the Procedure subsection, you need to include all the key details about the implementation of the procedure. That includes the instructions provided to the participants during the rating tasks. As a participant, you only saw one set of instructions and half the participants saw the same instructions as you did. The other half saw instructions described in Chapter 1 of the main text. For both groups, the word to be responded to was on screen for 4 seconds and advanced automatically at that time whether a response was made or not (keeping the presentation time constant across words, conditions and participants). After rating all 30 of the words, there was a 3 minute delay where you were asked to do random trivia questions. That was followed by the surprise recognition test of 60 words (30 old, 30 new) which is also explained in Chapter 1.

Experiment Hypothesis

Stated in terms of the operational definitions of the independent variable and the dependent variable, the question being evaluated by this simple study is whether rating liking of words will lead to better memory than counting the vowels in a word. That description is closely tied to the exact details of the methodology and procedure used in the study. We can also say that we are testing whether deeper processing of words leads to better memory than shallow processing. That more theoretical description can be tied back to the background research that has tested that hypothesis previously in several studies.

The next chapter in the Operations Manual provides a guided review of

the background literature most relevant for preparing a written report of this study, which you will complete as the first writing assignment for the Research Methods class. The two key references are Craik & Lockhart (1972) and Craik & Tulving (1975). While the findings still stand today, these published reports are quite old and the writing style is not one that you should follow exactly in preparing your report of our results. In addition, you should review a more recent retrospective by Craik (2000) reflecting on the original theoretical ideas, how it has held up and what limitations might still be seen in this framework.

As you review that background research, you should try to identify similarities among the methodology reported there and used in our study. For example, did those studies use a similar procedure for implementing shallow encoding? How about deep encoding? An alternate hypothesis about time spent during study is considered in Craik & Tulving (1975). Does that concern apply to our study?

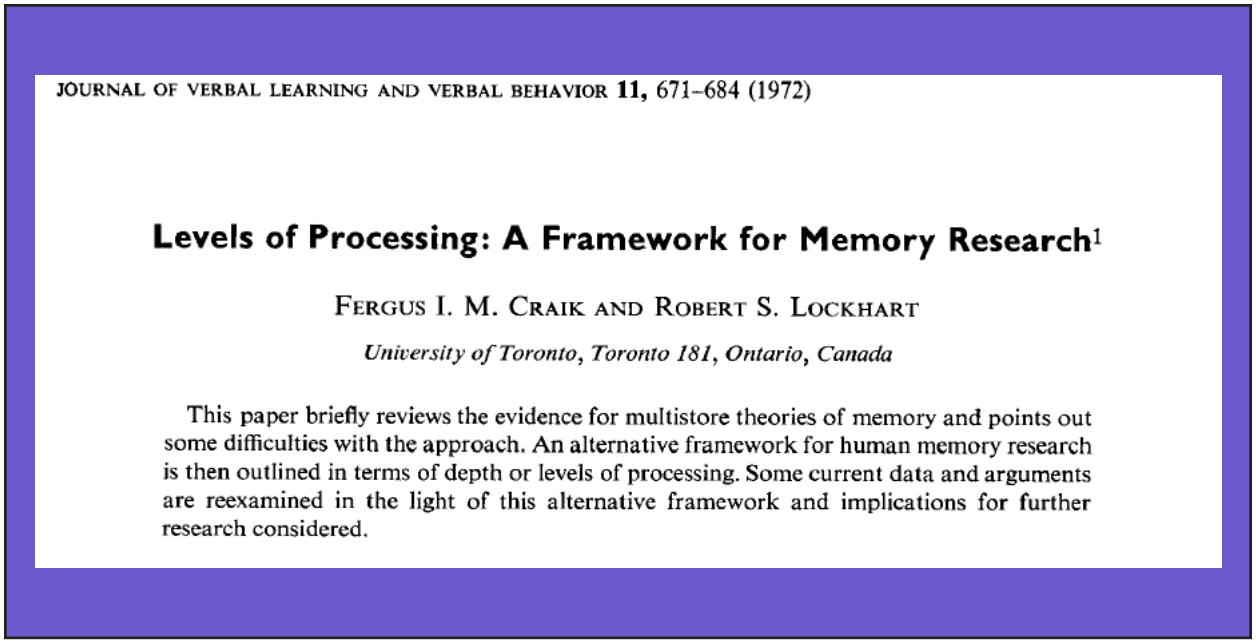
Once we have prepared a report of the findings of our in-class experiment, we will begin on a follow-up study in which the students will act as experimenters instead of participants. That study is anticipated to be another simple experiment utilizing the levels of processing framework and extending it to another novel domain. Although you will need some familiarity with this background theoretical research, remember that the goal of these projects in this class is the understanding of the methodology, design and conclusions (and not an expectation to learn a lot of theory about memory operation).

Experiment 1 Background

The first experiment is based on an early idea in the development of the modern understanding of how memory works in the human brain. Four background papers will be used to explain the underlying theoretical ideas across the first two experiments that we will carry out together in the class.

The core idea of **Levels of Processing** was first articulated in a publication by Craik & Lockhard (1972) that explains the background theory in the context of memory theories at that time. This paper also reviews a foundational idea that *memory* itself is not a simple unitary concept but can be better thought of as depending on both short-term and long-term components.

The **Short Tern Store (STS)** referred to in this paper is a concept commonly called **Working Memory** in more modern accounts of memory function and is actually conceptualized as more of a "workspace" for holding small amount of information in mind briefly. This cognitive function is very limited. It is the source of the famous 7 ± 2 capacity limit for holding things like a series of digits in mind. Modern studies have refined this idea to show that the specific number 7 is not the most accurate description but that the core idea of some small number of *chunks* of information can be held is correct (3 is a more common estimate, although a *chunk* itself can hold a lot of information in some contexts). In addition to the capacity limit, information in working



JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR **11,** 671–684 (1972)

Levels of Processing: A Framework for Memory Research[1]

FERGUS I. M. CRAIK AND ROBERT S. LOCKHART

University of Toronto, Toronto 181, Ontario, Canada

This paper briefly reviews the evidence for multistore theories of memory and points out some difficulties with the approach. An alternative framework for human memory research is then outlined in terms of depth or levels of processing. Some current data and arguments are reexamined in the light of this alternative framework and implications for further research considered.

memory fades over a matter of seconds and needs to be continually refreshed to maintain in awareness. For this reason, this cognitive function is sometimes considered more of a function of *attention* than *memory*, which is more usually expected to last more than a few seconds.

Our Experiment 1 is a study of **long-term memory (LTM)**, which is what most people expect when they think of memory. While you might think that the study of memory refers to experiences you had hours, days or years ago, it turns out that the same brain systems and cognitive functions support memory from about 30s ago through the whole rest of your prior life. As a result, the brief delay where you completed trivia questions before getting the memory test ensured that we are looking at the operation of long-term memory function in our study here (and in Experiment 2 later).

Once we have an understanding of that framework, an AI-generated summary of levels of processing theory can be seen to be an accurate and useful description.

The *Levels of Processing* theory, proposed by Craik and Lockhart in 1972, is a foundational concept in cognitive psychology that explains how depth of processing affects memory.

Core Idea

Craik and Lockhart challenged the dominant "modal model" of memory (which focused on separate short-term and long-term stores) by suggesting that memory is not determined by how long information is held, but by how deeply it is processed. They proposed that deeper, more meaningful processing leads to stronger, longer-lasting memory traces.

Types of Processing

They described a continuum of processing levels, from shallow to deep:
Shallow processing involves surface features, such as:
Structural encoding (e.g., noticing if a word is in capital letters)
Phonemic encoding (e.g., focusing on how a word sounds)

Deep processing involves *semantic analysis*:
Semantic encoding (e.g., thinking about a word's meaning, making connections to other concepts)
The deeper the level of analysis, the greater the likelihood that the information will be remembered.

Supporting Evidence

In their classic experiments, participants were asked to process words under different conditions. For example:

One group might judge whether a word was printed in uppercase (shallow).

Another might decide whether the word fit in a sentence (deep).

Participants remembered words much better in the deep processing condition, supporting the theory.

Implications

Memory is enhanced by elaborative rehearsal (thinking deeply about meaning), not just maintenance rehearsal (repeating items).

Educational strategies benefit from promoting active learning that encourages connections and meaningful engagement with material.

The core ideas at the basis of the levels of processing theory were revisited in a retrospective report, Craik (2000), which provides some more modern perspective on the idea nearly 30 years later together with some autobiographical discussion by the author. In this review, it is acknowledged that the concept of depth is one that is not very precisely defined. We can think of this as a challenge with the operational definition in this study that is meant to capture something related to how content to be remembered gets connected with existing knowledge. But a precise description of exactly what the means and how it would work in the brain is an area of active research in the cognitive neuroscience of memory. In spite of not being fully understood at the level of neural mechanism, the effect of manipulating study by asking participants to engage more or less with the meaning of the material is an extremely robust and reliable effect, making it a useful example of experimental design.

Exercises Part 1

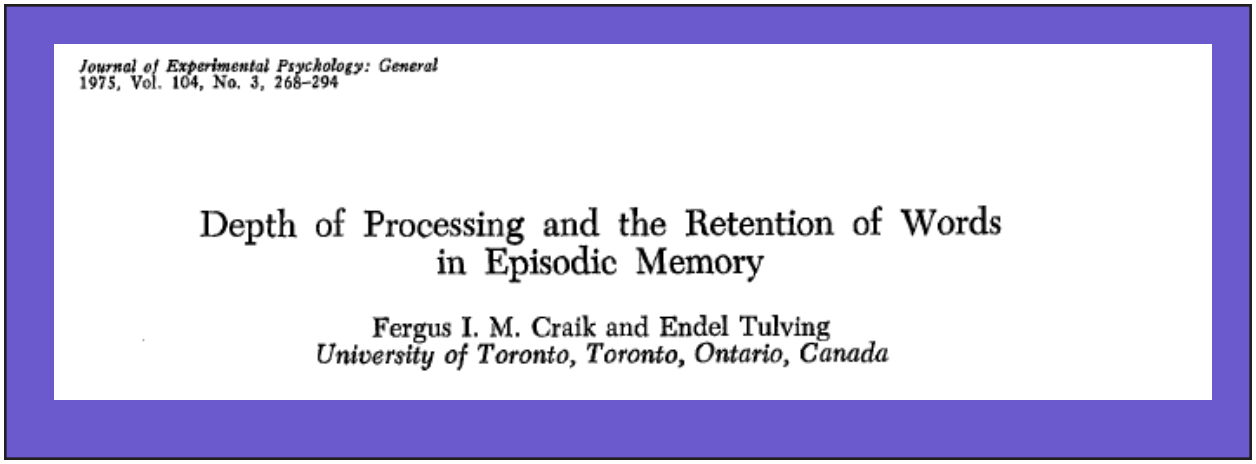
Read Craik & Lockhart (1972) to orient you to the background theory behind our hypothesis for Experiment 1.

It is worth noting that this is a fairly old paper that reflects the theoretical understanding at that time. The "levels of processing" theory is presented as an alternative to "multistore models." In modern memory research, elements of both theoretical ideas turn out to be true and the two approaches are not seen as inconsistent with each other.

The description and data of the multistore models reflects studies done prior to 1972. It is a useful overview, but if you are interested in the general topic of studies of memory, be aware that is a historical overview from a very long time ago. Characterization of the new ideas related to 'levels of processing' comes after this review in the paper.

Answer the following questions from the reading:

1. What is *depth of processing* and why might it lead to better memory?
2. In our study, how would our definition of *deep encoding* connect to this theoretical idea?
3. In our study, how does our definition of *shallow encoding* provide a control comparison?
4. From the prior work cited (e.g., p 677), give an example of how researchers have implemented a different procedure to create shallow encoding.
5. Give another example of a procedure to create deep encoding from the briefly reviewed prior work.



Journal of Experimental Psychology: General
1975, Vol. 104, No. 3, 268–294

Depth of Processing and the Retention of Words in Episodic Memory

Fergus I. M. Craik and Endel Tulving
University of Toronto, Toronto, Ontario, Canada

Experimental Findings

For further background, the publication by Craik & Tulking (1975) reports a series of studies using experimental design to test the levels of processing hypothesis. Across 10 separate studies, manipulations that increased the depth of processing of words were consistently found to increase scores on memory tests that occurred afterwards.

This report contains an excellent set of findings related to hypotheses about how to influence memory by different kinds of depth-related manipulations. However, it is a long and complex report and a full discussion of all of the reported research would go beyond the goals here of illustrating experimental methodology. In addition, the reporting style used at the time of publication (mid-1970s) is not consistent with modern scientific reporting to APA style. This paper should not be used an example of modern scientific writing for the class-based writing assignments. Better (and shorter) examples will be provided.

The annotated pdf of this report provides some simplifying guidance to the key points within this publication that we will focus on in class. The annotations are further provided here for reference.

Annotations

- This article opens with a "Summary" where we would normally find an "Abstract" in modern scientific publishing. The goals of this section are similar, but this Summary is much longer than a properly formatted Abstract section would be. Of course, the whole paper is longer than most scientific publications that are reported today. In the modern style, these studies would likely have broken up the series of experiments here across several publications to disseminate the work earlier and provide other researchers the opportunity to do similar parallel research.
- The Introduction is also slightly on the longer side, but the content overlaps a lot with the Craik & Lockhart (1972) paper you just read. Since most of these results are quite old now, we can observe the style of the new research presented as grounded in and building on prior work across years of related studies.
- The General Method section (p. 271) is one that you will typically only see in papers that are presenting a series of studies with similar methodologies. To save space and avoid redundancy, some of the common operational definitions, materials, and experimental apparatus are explained here in one place. For our purposes, this section is quite useful in clearly explaining a series of related operational definitions of various levels of deep processing and the specific procedures used in the studies to follow, especially in Table 1.
- This section also introduces some archaic apparatus for this kind of work done before the advent of computers. A "tachistoscope" is an enclosed box participants would look into through eyeholes. It was initially kept completely dark, but a light could be turned on at a precise moment that usually also started a precise timing circuit. Typically, a printed card with stimuli, like a picture or word, was in the box that could only be seen when the light activated. For studies of perceptual cognition, it provides excellent control over every aspect of the perceptual process enabling very rigorous controls.
- At this level of control, we will see times described in milliseconds, msec, which is thousandths of a second. That is, 1000 msec = 1 second. The text described presenting words for 200 msec, which is about a fifth of a second, which is about as much time as the average person would need to comfortably and confidently read

a printed word. Many cognitive psychology studies will use timing given in msec or ms, which can be unfamiliar at first. You can use rules of thumb for reading these reports such as anything less than 200 ms is moving or happening very quickly. If participants are responding to stimuli in 1500 ms, they are taking about a second and half, which does not leave a lot of time for complicated calculation. Presenting a word for 4000 ms, is 4 seconds, which is more than enough time for reading and provides some time to think about the word after.

- We also see several operational definitions of the measure of memory used for the dependent variable in these studies. In different experiments, different types of memory tests will be used, described here. Note that these are all slightly different operational definitions of the same idea of how much memory did the participants have of the word lists studied earlier.
- The section labelled Experiment 1 (p. 272) is where you will find the description of the procedure that gives how the independent variable of processing depth was implemented and how many levels were used.
- Table 2 (p. 273) is where we can see the main summary of the dependent variable values, the data from the study.
- Experiments 2-4 replicate and extend this basic result. It's ok to skim over these here and jump ahead to Experiment 5.
- The section labeled Processing Time Versus Encoding Operations explains a key extraneous variable that the authors want to control to be able to strengthen their conclusion that depth of processing improves memory. Processing time can end up accidentally confounded with depth, potentially challenging the validity of the conclusion.
- Experiment 5 provides the explanation of a method aimed to control for the problematic extraneous variable.
- Table 4 (p. 281) shows the data obtained from using this approach. You should be able to unpack from these average numbers if the new methodology was effective at controlling processing time and what happened to the memory measure.
- Experiments 6-8 further explore some other aspects of depth, but this can be skimmed to jump ahead to Experiment 9.

- For Experiment 9, the rationale provided falls under an idea here that we will refer to in class as "external validity." That is, do the results observed work outside the laboratory in less perfectly controlled conditions. The observation that the depth manipulation works in a classroom is partly why our in-class Experiment 1 is our example study.

For each of the 10 reported studies, you should be able to identify the independent variable that was manipulated by the experimenters and the dependent variable(s) measured. Most of these studies are good examples of simple experimental design where conditions are contrasted to show differences in the outcome measures. The statistical measures are largely using ANOVA as the main method, which we will discuss in class in later chapters as the main tool for complex designs. For our first class experiment, the design is a simple two-group comparison that can be evaluated using a two independent sample t-test.

The focus of this class will be mainly on design and interpretation of research methods. While statistics are an important part of our ability to draw conclusions from research, the main goal of that part of our quantitative approach is to have confidence that the differences observed were not simply due to chance. Beyond the mathematical aspect of our statistical tools, successfully replicating a result by obtaining it with another study is an excellent and intuitive way to establish the reliability of an experimental effect. By containing a large number of replications of the result of manipulating the dept of processing, we can have a great deal of confidence in this finding even beyond any details of the statistics being reported.

Exercises Part 2

6. Craik & Tulving (1975) reports a series of studies examining the effect of various approaches to deep and shallow processing on memory. Review this publication and answer the following questions about specific experiments reported there comparing the procedure to our Experiment 1.

- In their Experiment 1, how many levels of the IV were used? What was the DV measure of memory?
- Their Experiment 5 is carefully designed to address what confounding alternative hypothesis? To do so, what aspect of the IV is made as constant as possible?
- In what way was their Experiment 9 similar to our in-class experiment? Identify some methodological differences

7. Briefly answer the following questions about experimental control from in-class Experiment 1:

- Why have both groups read the same words?
- Why have 1-5 scales for responding for both conditions?
- Why require the word to be on screen for minimum 3 s?
- Does it matter if the trivia questions use words from the study list?

Experiment 1: Analysis

Data Analysis for Experiment 1

This chapter pairs with **Chapter 6** in the main text to provide a specific recipe for carrying out the t-test needed to analyze the data from Experiment 1 produced by students on the first day of class.

The first section provides a step-by-step walk through of installing and running the analysis program R and using a provided R *script* to do the t-test analysis on the data provided by the instructors.

The second section explains how to make a data visualization to present the results of the data as a bar plot. Instructions for making the graph are provided based on using Microsoft Excel. Note that one key element of the graph, the representation of the *standard error of the mean* on the average data, takes several steps and this option is not available on all software packages (e.g., Google Sheets and the online Microsoft Excel do not provide this option and cannot be used unfortunately). Chapter 6 in the text explains how to format the graph as a Figure to be included in an APA format manuscript.

Running a t-test in R

Start by installing the R program and the RStudio suite (in 2 steps)

- Use the link below to go to R download page and choose the version that is compatible with your computer's operating system: https://cran.r-project.org
- Once R has downloaded, install it on your computer.
    - It requires permissions.
    - Accept the license.
    - Install all the default components.
    - Don't customize startup options.
    - Default additional tasks are fine.
- After R has been installed use this to download the RStudio version that is compatible with your operating system:
  https://rstudio.com/products/rstudio/download/#download
    - If you are coming through the RStudio site, go to products, then RStudio Desktop. Use the Open Source Edition (Free).
    - Download will adjust to your OS. The Windows download is 171M, so be aware of bandwidth constraints and speed.
    - Current version is 1.3.
    - MS Windows complains to me that it isn't a Microsoft verified app. However, it is safe to install.
    - Once RStudio has downloaded, install it on your computer.
    - Note: You will not be able to install/run RStudio until R has been installed.

Use the RStudio program to start an analysis session. You will also need to have downloaded the provided data from instructors, usually as a file named something like "Exp1_data.csv".

- Launch RStudio. You should see a screen with 4 panels. We will be primarily working with the left 2 panels.
    - The top left panel will have lines of code, a 'script' for carrying out the steps

required for an analysis.

    - The bottom left panel will have the output results of executing those steps, including error messages if something goes wrong.

- Use File -> Open and navigate to the folder on your computer where you've installed the files and associated data from our experiments
- Open the file provided for data analysis. This is an R script for testing your installation and re-running the t-test analysis from our Experiment 1 data for the in-class experiment.
    - On a fresh install, this will produce a warning that there are required packages that are not installed. The option to install them is provided. You can also install them by working through the script analysis steps.
- Set the **working directory** to where your data are stored on your computer. If you have put the data file in the same folder as the analysis file, navigate to the Session menu, then to Set Working Directory and select the top option **To Source File location**.
- To run a single step of the analysis press the **Run** button that is in the upper right part of the top-left panel. This carries out the step in the script on which the cursor is currently. If you didn't do the installation of the 'psych' and 'ez' packages above, put the cursor on line 2 and Run. Then put the cursor on line 3 and Run.
- The installation process will also download and install a series of other packages needed (called dependencies). The process should only take a few minutes to run.
- Now move down to line 6, "library(psych)" and press Run. This loads a set of routines for data analysis for psychology experiment data that are helpful.
- The cursor moves down to the next line after each Run. Press it again to load the library on line 7, 8, and 9 ('psychTools', 'tidyr', and 'ez').
- With luck you are not getting error messages in the bottom left panel. If you are, something may have gone wrong with the above steps.
- The next step, line 12 will start loading our actual data. If everything is working you should see: "Data from the .csv file Exp1_data.csv has been loaded." In red in bottom left panel.

- Run on line 13 will cause the data table to be printed in another tab. It should look a lot like what the source data file looks like if you open it in Excel or another spreadsheet program.
- Run on line 16 to see the output of the describeBy function, which provides descriptive statistics for our data. You may notice that this needs to be unpacked a bit to find the key numbers, which are the Recognition.score values for each condition. Check that these numbers are identical to the descriptive statistics you calculated in your spreadsheet previously.
- Run on line 19 to carry out the two independent samples t-test for the data.

If everything works up to this point, then congratulations! You have just run your first formal analysis of experimental psychological data.

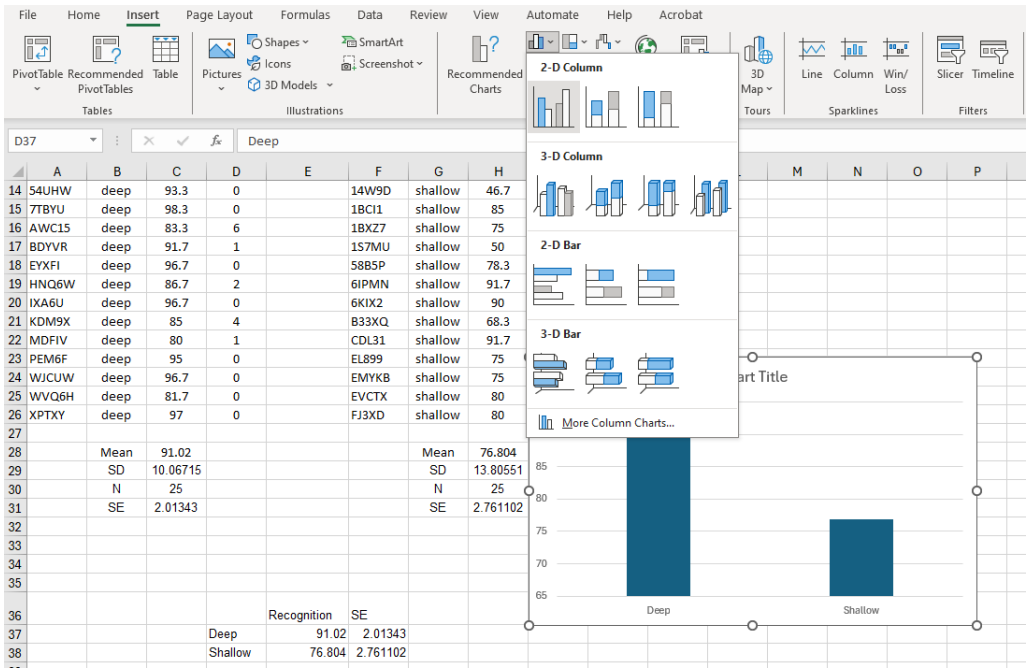
It should look something like:

```
     Two Sample t-test

data:  Recognition by Condition
t = 6.005, df = 46, p-value = 2.846e-07
alternative hypothesis: true difference in means between
group deep and group shallow is not equal to 0
95 percent confidence interval:
 11.52491 23.14705
sample estimates:
   mean in group deep mean in group shallow
             91.66190              74.32593
```

The output lines from the analysis will contain information about the calculated t-value and the associated p-value to assess the reliability of the effect. Refer to Chapter 5 for how to proceed with presenting this information in APA format for inclusion in a manuscript reporting the results of your experiment. The statement of inferential statistics should look like:

$t(46)=6.0, p<0.001$



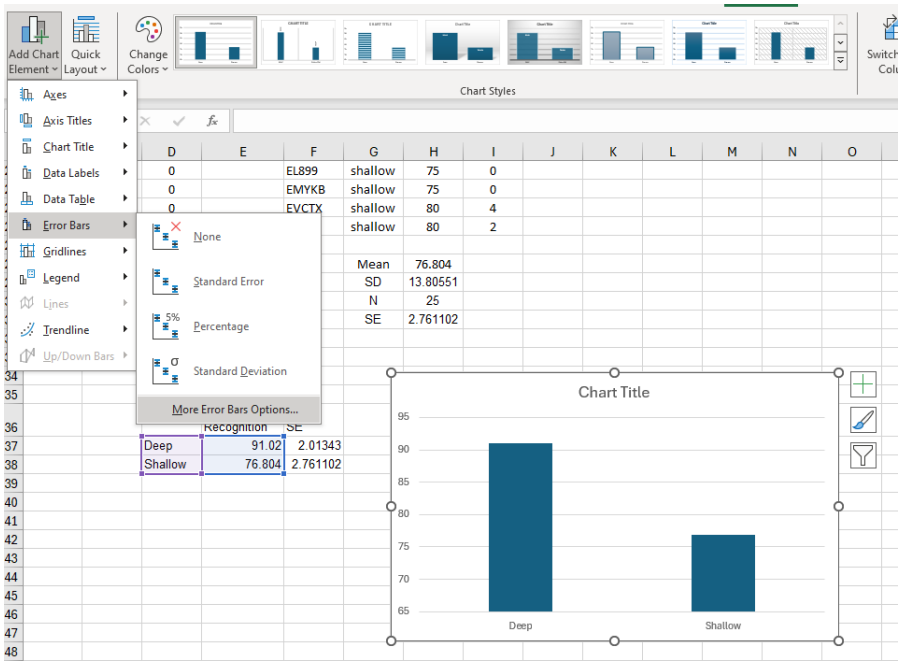
Making a graph in Excel

Here we will illustrate some basic elements to include to accomplish this. For this we will be working with the same data in Excel format where formulas, graphs and othe features can be included with the data. The file used in R, which ends with the .csv suffix ("comma separated values") is internally a simple text file of information separated by commas. An Excel spreadsheet can contain a lot of additional information useful for data review and also provides facility for making data visualizations, specifically Figures to support presentation of experiment data.

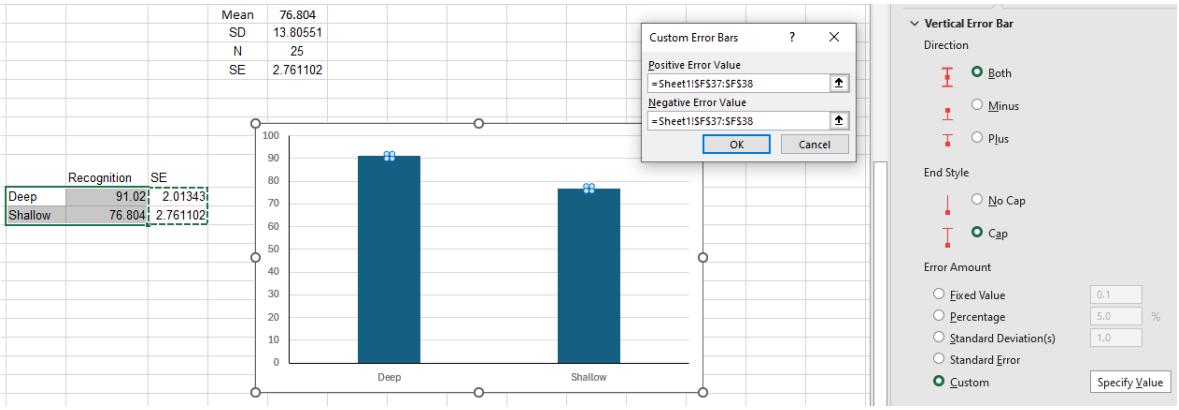
The first step towards creating a figure is generally to create a separate, labeled table of the key numbers that will contribute to the graph. The numbers we need for the graph will be the mean and the SE for each of the two conditions in our study: deep and shallow. For simplicity, we want these organized into a separate data table. In the figure below, you can see a labeled 2x2 table where the mean recognition memory percent correct scores are copied from the descriptive statistics to cells where they are adjacent at the bottom of the image.

In this image, those two cells and their labels are selected with the mouse and then in the Insert tab, the upper left option of Bar Charts is selected and then the upper left option of 2-D Columns is selected. As you do this, Excel already renders an initial image of what the starting bar plot is going to look like overlaid on the spreadsheet.

From this initial draft, we need to do some editing to the layout of this chart to make it effective and in approximately standard format. First, the Chart Title can be cut as we generally do not include titles on figures in manuscripts. Titles are used to help describe data in presentation formats, but APA reporting format requires that Figures be accompanied by a figure caption which is where the description of the illustration should be included. In addition, both the x- and y- axis should be drawn in black to ensure visibility of the axes. You will want to label the y-axis by adding the Chart Element, Axis Title -> Primary Vertical and then change the text label to Recognition Score. You may also optionally choose to remove the horizontal lines (these are chart elements called gridlines accessed through the Chart Design menu) or even change the color of the bars.



Once the basic layout is set, the last element to be added is brackets reflecting the SE of mean. We kept these numbers near our means above, but note that we did not select those numbers when making the graph (if you have 4 bars in your graph, you may have selected them accidentally).



To add error bars correctly the size of the calculated SE, click on the graph and specifically one of the two bars. Then in the Add Chart Element menu, select the Error Bars option and the bottom choice, More Error Bars Options from there. In the Format Error Bars pane, choose Custom for your error bar size (bottom option) and select Specify Value. We will need to specify both the positive and negative sizes of the error bars, above and below the mean. For our data, these are the same sizes. For both the Positive Error Value and Negative Error Value choose the range where the SE values have been copied (F37:F38 above). Because they are next to each other, we can select both values and these will be correctly applied to both bars. The figure below shows roughly what this will look like. When done. select Ok.

This feature of setting the error bars flexibly to specific values for each group allows for correct presentation of both the means of the observed data, shown in the height of the bars, and the variance, shown in the SE bars. This puts many of the key descriptive statistics into the Figure visually. In addition, a useful trick is to look at the range implied by the error bars. For

an independent samples experiment, if the error bars to not overlap (touch), then you most likely have a reliable difference between the groups. That means that the figure is also carrying some implicit information about the inferential statistics. You should always check or carefully include the actual statistical test in the reported text, but a well-made figure acts as a very effective overview of the results.

You can further edit the graph you have made to improve the style and effectiveness of the visualization. Always remember that the goal of a Figure in a research report is to effectively communicate results. For example, Excel currently defaults to drawing the axes in a nearly invisible light gray color. You can select these and change to a more visible black. Excel also defaults to including horizontal grid lines which you will rarely see in published papers and you can choose to remove this element of the graph if you wish.

The graph can be copy/pasted into your manuscript document from here, but note that you should refer to Chapter 6 for correct formatting of a Figure in an APA manuscipt and a Figure Caption must be included with the graph.

Results for Experiment 1

With the output of the R script for analysis of the Experiment 1 data, the descriptive statistics (from either Excel or R) and the graph visualizing the performance on the Recognition Memory dependent variable, you are now ready to write the Results section of your experiment report for the first study. Your report should follow standard APA scientific writing style, which is reviewed in Chapter 7 and is also available through the Publication Manual of the American Psychological Association. Note that a lot of the numbers you have worked through and looked at here do not end up appearing in the final report. The report presents specifically the descriptive (means, sandarad deviations and standard errors) and inferential statistics (t-test, p-value) in standard formatting. The figure is accompanied by a note or caption explaining what the visualziation is intended to communicate.

The data matrix of individual results is not included in a standard experiment report, nor is the full output of the statistical program. These are sometimes included as Supplemental Materials in modern publishing protocols, but are not considered standard parts of the report. Because the quantitative reports are often essential to supporting conclusions drawn from a research study, it is particularly important that the numbers be accurately reported in the manuscript. Typographic errors or mis-formatting the numerical report can lead to catastrophic mistakes in communcating your findings. Proofreading those sections multiple times is highly recommended. Report the results of Experiment 1 using provided data from class the class.

This should include:

- Descriptive statistics for both groups
- Inferential statistics about the difference in performance across groups
- A data visualization of the results. This should be a Figure, which includes a graph results, properly labeled, including standarad error bars and a caption.

Note: Save these for inclusion in the report of Experiment 1 to be described in Chapter 6

Glossary of Research Terms

Collection of terms from first chapters for describing experimental methodology taken from the main textbook chapters. Terms in the Chapters 1-9 should be familiar for the first in-class exam. Terms throughout the textbook are expected to be familiar for the second in-class (midterm) exam.

Chapter 1

- **Experimental research**: The experimenter manipulates an independent variable and measures a dependent variable to test if the manipulation has an effect.
- **Construct**: The high-level concepts we aim to do research about. Typically, these things we that have an intuitive understanding of but need to be translated into specific experiment elements.
- **Operational definition**: Turning an intuitive but imprecise concept into something that can be measured quantitatively, or controlled categorically.
- **Measured operational definition**: A quantitative measure of a construct, essentially turning an idea into something that can be characterized as a number. For example, Experiment 1 operationally defines "memory" as percent correct on the test, a quantitative measure of the amount of memory obtained. A similar process might turn other constructs like anxiety, impulsiveness, attention into numbers that could be used as dependent variables in experimental design.

- **Experimental operational definition**: A controlled method of implementing a specific definition of a construct into levels or categories that can be manipulated by an experimenter in order to create the independent variable(s) for an experiment protocol.
- **Independent variable (IV)**: Often referred to by the acronym IV, this is the element manipulated by the experimenter to see if or how it affects the measure being collected in an experimental design. Controlled manipulation of the IV is the defining feature of experimental research.
- **Dependent variable (DV)**: Frequently referred to by the acronym DV, this is the measurement collected by the experimenter. The core idea in experimental research is to see how the scores on the DV change across the manipulation of the IV. If they do, we can conclude that the IV affected the DV.
- **Experimental Hypothesis**: A statement about the relationship between experimental variables that can be tested and importantly, falsified. If there are no data that would render a statement false, then it is not a falsifiable statement and is typically a description rather than a hypothesis. Typically the hypothesis is that the IV affects the DV, and we use statistics to reject the **null hypothesis** (that the IV does not affect the DV). Note that hypotheses can be stated about the specific IV and DV used in an experiment but also stated separately about the constructs from which the IV and DV were operationally defined. Experimental data gives us confidence to make statements about the specific IV affecting the implemented DV but the goal of research is to draw inferences about the relationship among the constructs.
- **Limitations**: Concerns that conclusions about the underlying constructs might not be true in all cases and conditions other than the specific operational definitions used in the experimental design. Generally these are not issues with the fundamental **validity** of the experiment (Chapter 3), but questions about how widely the results can be applied. Identifying what limitations should be considered often requires some knowledge of the underlying theoretical ideas for a research study and can also indicate directions for future research. Using Experiment 1 as an example, we have data about memory for word lists measured with a recognition test a few minutes later. We might wonder if deeper encoding similarly affects memory for pictures, or if the effect might change with another measure of memory like recall. Studies

examining those questions would reflect different operational definitions of memory, using different DV's and/or different operational definitions of deeper encoding as IV.

- **Statistical reliability**: We will evaluate whether the IV has a robust effect on the DV using standard statistical tools. Our focus here will be selecting the correct tool and reporting the use of the tool accurately. Statistics are often presented as a simple binary outcome: did the IV affect the DV reliably, can we reject the null hypothesis, was the probability of the null less than the criterion of .05 (these three statements are essentially synonymous). However, we will see that Psychological Science is moving towards a model of reporting **effect size** rather than relying on these binary descriptions. The effect size is helpful both with understanding the reliability of the statistics and also communicating the results. For Experiment 1, we might want to be able to say not just that deep encoding improved memory, but how much did this study approach increase our measure of memory?

Chapter 2

- **Reliability of measurement**. The degree to which the measure being used accurately assesses the intended construct. Any measure in psychological science will have some inherent measurement error such that repeated measures of the same participants will generally not produce the exact same value. In addition, measures of constructs that are hypothesized to be stable traits will sometimes be affected by current states (that are not stable over time), for example, measures of mood or emotional stae.
- **Validity of measurement**. The degree to which the measure captures the intended construct accurately. When coming up with a novel operational definition for research purposes, the specific methodology can inadvertently include errors such that the measure no longer accurately assesses the original intended idea.
- **Measurement types: ratio, interval, nominal**. A ratio scale is a measure with a true zero and the spacing between steps on the scale is consistent and interpretable. An interval scale has relative levels, but may not have a true zero or consistent intervales. A nominal scale is a set of potentially non-ordered categories to choose among.

Chapter 3

- **Extraneous variable**. Any variable that can affect the measurement of the dependent variable that is not part of the planned design is considered an extraneous variable. These are not necessarily problems for the reliability or validity of the experiment, but can be. When evaluating the effectiveness of experimental design or planning a study, all the extraneous variables that can be identified should be considered and managed as well as possible.
- **Confound**. An extraneous variable that is synchronized with the independent variable is an experimental confound. These are the most significant problem for experimental design as if a confound exists in the study, no conclusions can be drawn about whether the intended independent variable affected the dependent variable. However, not all extraneous variables are confounds. Contextual variables that affect the dependent variable but are randomly distributed across participants in all levels of the manipulated independent variable are not confounds.
- **Control condition**. For many intuitive hypotheses, we start with an idea that X might cause Y (e.g., social media use causes anxiety). To test these hypotheses with an experimental design, it is necessary to have a control condition that in some cases can be challenging to define (e.g., what is the control condition for "social media"?).
- **Treatment conditoin**. Typically the treatment condition is the condition of interest that is thought to affect the dependent variable that will be tested against the control condition.
- **Counterbalancing**. A major tool for managing the impact of extraneous variables by distributing effects across levels of the independent variable. This keeps the extraneous variable from being confounded (in sync) with the independent variable and allows conclusions to be drawn from reliable results.
- **Constancy** (across conditions in design). Wherever possible, conditions should be kept constant across everything that is not part of the experimental manipulation of the independent variable. This maximizes the reliability and validity of results, although in some cases may raise questions about the broad generalizeability of the findings.

Chapter 4

- **Demand characteristics** refers to the fact that the wording or experimental context itself might prompt participants to respond in a way that they feel is expected. This can bias behavior to undercut the validity of the results from a research study.
- **Blind designs (single, double)**. When participants are blind to the hypothesis and conditions of a research study, as is commonly done, the study is described as a single-blind design. In certain cases, it is also important for some of the research team to be blind to the conditions of the independent variable, referred to as a double-blind design. Whenever there is an element of subjective judgement as part of the scoring of the dependent variable, it is best to have blind raters carry out that scoring process to avoid bias (even inadvertent).
- **Random assignment to conditions**. A critical and basic tool for distributing differences in participants, or participant variables, across conditions. With a sufficiently sized sample, random assignment protects against accidental assignment of systematic differences in participants to conditions. The mathematics of this process are already embedded within the standard statistical models used to assess a reliable effect of an independent variable on the dependent variable.

Chapter 5

- **Convenience sampling**. Most psychological research is carried out with participants who are conveniently available for participantion. While alternatives to this approach can be extremely difficult and expensive to carry out, it is often useful to reflect on whether the conclusiond drawn from a study should be expected to generalize to parts of the population beyond the source of the convenience sample.
- **WEIRD samples**. Most published psychological science comes from convenience samples from university communities. These Western Educated Industrialized Rich Developed countries may not always reflect characteristics of the population of all humans across the planet. Where possible, we consider when the conclusions drawn from an experimental process might be limited to the population sampled.

- **Balanced, stratified sampling**. Detailed sampling methods that drawn from specific characteristics from the population are often used in clinical health studies, epidemiological assessments and attitude polling (including politics). These methods depend on a well-quantified sampling goal that describes the population and then aims to sample in a way that matches the population characteristics.
- **Non-response bias in sampling**. Because all research participation is voluntary, it may be important to consider if the characteristics of behavior of participants who decline to participate might affect the conclusions drawn. This is particularly evident in processes using online surveys where motivation to complete the survey may lead to samples that do not reflect the broader population.

Chapter 6

- **Measurement error** refers to the difference between the actual value of a psychological construct and the value obtained using a measurement instrument. It is the portion of a measurement that does not reflect the true score of the variable being assessed. Measurement error is what causes a person's observed score (e.g., on a test or survey) to be different from their true score.
- **Descriptive statistics**: mean (average), variance, standard deviation, standard error. These are defined quantitative values that are used to describe a sample of data that has been collected. The mean or average describes the central tendency of the data (the middle if it is normally distributed). The variance or standard deviation describe how much the scores vary around the mean. The standard error is a measure that supports statistical inference because it estimates how well the sample measure likely captures the true population statistics.
- **Inferential statistics** are used to calculate whether the effect of the manipulated variable produces a reliable change in the dependent variable. These methods are how we identify whether a research study has appeared to "work."
- **T-test** (two independent samples, paired or dependent samples) variations are used to infer differences for simple designs with one or two ground or conditions.
- The **p-value** is defined as the probability of the data having occured in the observed

distribution under the null hypothesis that there was no effect. Practically, it provides our key criterion for what we use to present or publish scientific findings. The standard formulation is that we have sufficient confidence to present findings if the p-value probability is less than .05 (1 in 20). However, p-values should not be overinterpreted and are just one component of a rigorous approach to scientific inference. Not only should research be replicated beyond an initial finding of $p<.05$, but statistical reliability does not mean that the experiment is internally valid (there can be confounds, errors in operational definitions, etc.).

- **Null hypothesis** is a technical term for the idea that the independent variable has no effect on the dependent variable. Inferential statistics are build around rejecting the null hypothesis at a certain probability value. Sometimes hypotheses are advanced that would require accepting the null hypothesis and it should be noted that these hypotheses do not fit into the standard approaches for statistical inference.
- A **data visualization** is a graphical illustration of a set of data meant to effectively communicate the findings of a study. These support the textual report of descriptive and inferential statistics. Visualizations can take many forms: line graphs, bar graphs, violin plots. Regardless of the form, visualiations are presented as Figures in a written scientific report.

Chapter 7

- **APA format for scientific writing**. The American Psychological Association provides a style guide for formatting scientific reporting of study results. The details of the style guide update periodically with the current version being the seventh edition. Even more important than the details of formatting style and preferred language is the conceptual idea of reporting methodology and results with enough detail for the reader to evaluate limitations in the conclusions. Every detail relevant to the interpretation of the study should be included clearly. The authors may indicate limitations or weaknesses themselves, but there should be enough detail for other scientists to identify additional issues. There should also be enough detail in the description of the methods that another group could undertake a full, exact replication of the study as part of extending rigorous scientific methods.

- **The Abstract** is a short summary of the entire report that precedes the detailed description. This includes the results and conclusions of the study.
- **The Introduction** section reviews the relevant background including references to prior related research, hypotheses, theoretical frameworks, constructs and key operational definitions that were used in the research study.
- **The Methods** section contains a full and exact description of the research that was carried out. It will typically be organized by each experiment within the overall program of research and each experiment will be rdescribed with subsections for the Participants, specific Materials, and detailed Procedure.
- **The Results** section is a listing of the findings from measurement of the dependent variable broken down by the independent variable(s). It includes both description and inferential statistics, but does not generally include interpretation of the findings or conclusions about the hypothesis.
- **The Discussion** section provides the interpretation of the findings described in the Results section. It should generally have few or no numbers (those go in the Results) and focus on the conclusions and meaning of the findings. These should be discussed in the context of the background material reviewed in the Introduction section. This section can also include a consideration of any limitations on the conclusions arising from the methods or sampling. It can also include disucssion of how future research might address those limitations.
- **The Reference** section follows a precise and exact format for the listing of citations to prior research aimed to allow other scientists to find and read all the relevant background material.

Chapter 8

- **Within-participants designs** have all the participants enage with all of the conditions of the experimental design. This provides the advantage of perfectly matching any relevant participant variables and increasing the statistical sensitivity to effects of the independent variable across conditions.
- **History effects** (order effects) are intrinsic to any within-participant design as the

conditions often need to be assessed in order. This can lead to effects across the order of conditions such as fatigue effects, learning effects, or carryover effects.

- Within-participant designs also generally expose the participants to all levels of the independent variable, which can potentially allow them to guess the design and hypothesis behind the study with the potential for response bias effects.

Chapter 9

- **Research ethics** refers to the fact that science is always carried out with attention to the fair treatment of participants. All research must meet standards of justice, beneficence and respect for persons.
- **Informed consent** reflects an agreement from participants acknowledging their awareness that they are engaging in a research study with an understanding of what will be expected of them and their rights as a participant.
- The **Institutional Review Board** reviews all research before it can be carried out to ensure that all researchers follow ethical research practices.
- **Voluntary participation** is a core prinicple of scientific research that participants must choose to participate in research and always retain the option to not participate or stop participation at any time.
- A **Risk/benefit analysis** is carried out by the IRB to ensure that any potential harm that could possibly occur to the participants is justified by the benefit of the scientific gains achieved by carrying out the study.
- **Vulnerable populations** are ones who may not be able to clearly provide voluntary consent of their own accord (e.g., children, prisoners) and any research aimed at these populations is very carefully overseen.
- **2** refer to conditions where participants must be misled for scientific purposes in research, which is at least a mild transgression to treating them with respect. When scientifically justified, deception can be allowed providing that participants are also debriefed afterwards to explain how they were deceived and why.

ANOVA Practice Examples

As we did in Chapter 5, here we will document practical steps required to carry out ANOVA, Analysis of Variance, analysis within R/Rstudio. We will review hands-on examples of three different analysis from three hypothetical experiments.

The first will demonstrate analysis across a single factor with three levels, a one-way ANOVA. This demonstrates the simple extension of the two independent samples t-test to experimental designs with three conditions instead of two.

The second will demonstrate analysis of a 2x2 factorial design with both factors having two levels between participants. This is the simplest factorial design. From the output of the ANOVA analysis, we will extract the key statistical parameters including the F-ratio, the degrees of freedom and the p-value. As with earlier t-test analysis, a simple reporting frame will be provided for reporting the results. However, it should be noted that the simple report of statistics from the output of an ANOVA is particularly uninformative without supporting statements about the descriptive statistics, statements of the direction of the results and ideally, a good data visualization.

In a third example, a mixed-model ANOVA will be demonstrated in which

there is one factor between participants and one factor within-participants. This changes the output information from the analysis as well as requires some reformatting of the input data files. Once the correct information is identified in the table, reporting and visualizing the results is a similar process to other ANOVA analysis.

In our return to hands-on statistical analysis, we will also review how reports of observed **effect sizes** are increasingly a part of modern statistical reporting in psychological science. Several different measures of corrected effect sizes are used to attempt to provide context for conditions where the independent variable as a small, medium or large effect on the dependent variable. These can be used to support the $p<.05$ formalism, but different effect size measures require familiarity with their underlying ranges.

At the end of this chapter, we will touch very briefly on the idea of **Bayesian analysis** as an alternate model for statistical inference. The Bayesian approach has aspects that are very intuitive and reflect a natural way to think about accumulating evidence for a hypothesis. However, the mathematics of employing a Bayesian approach require making assumptions about the experimental hypothesis that have proven difficult to accept broadly.

Learning objectives

- Carrying out an ANOVA in R/RStudio
- Reporting the ANOVA results in APA format, extracting key numbers from the output table
- Understanding how to read and how to make figures for factorial designs to illustrate main effects and interactions.
- Modern reproducibility theory: effect sizes
- Power analysis and sensitivity to observing reliable effects when planning research
- Bayesian analysis as an potential alternate approach to drawing inferences

In this chapter, we will present a series of analysis examples using R/RStudio and the function *ezANOVA* to carry out an ANOVA on simulated factorial data. The data files for these analyses should be available so that you can run these analyses in parallel to become familiar with the general process. The goal of these examples is to review how to extract the information to report from the output of the ANOVA calculation and how to format it for reporting in an APA scientific report.

This is the process we will use to analyze the data from the in-class Experiment 2. The results of this experiment will be reported in the second major writing assignment as an extension of the ideas from Experiment 1. You will also need to be able to carry out your ANOVA analysis for the in-class research projects, which are reported in the final term paper.

Example 1: One-way ANOVA in R

To test a Mozart Effect hypothesis, participants were assigned to listen to one of three kinds of audio while performing a spatial cognition test with 21 challenging problems. The audio sounds were either soothing Ocean noise, Folk dance music or Classical music. The number of problems solved was the dependent variable.

Simulated data are shown as the mean number of problems solved while the different sounds are playing. The standard deviations are shown under the means for each condition.

Music type	Problems solved
Ocean sounds	11.6 (2.72)
Folk music	13.1 (2.38)
Classical music	15.4 (2.27)

The analysis output is shown to the right as it would be printed in RStudio after running the ezANOVA command. The command parameters are included here for your reference. The key part of the output occurs after the *print(anova_result)* command, which reports the statistical output from the analysis. As written, the tabular format of the output is not completely clear.

As a table we can improve the formatting:

	Effect	DFn	DFd	F	p	p<.05	ges
2	Music	2	27	6.04215	.006782334	*	0.3091855

Now we can see the connection from the statistical information to the numbers. For the factor *Music*, the degrees of freedom in the numerator are 2 (DFn) and 27 in the denominator (DFd). The F-ratio value is 6.04. This

R Output

```
> anova_result = ezANOVA(
+   music
+   , dv = .(Problems.Solved)
+   , wid = .(N)
+   , within = NULL # NULL if no within factors
+   , between = .(Music) # NULL if no between factors
+   , observed = NULL
+   , diff = NULL
+   , reverse_diff = FALSE
+   , type = 3
+   , white.adjust = FALSE
+   , detailed = FALSE
+   , return_aov = FALSE # TRUE for showing details
+ )
Warning: Converting "N" to factor for ANOVA.
Warning: Converting "Music" to factor for ANOVA.
Coefficient covariances computed by hccm()
> print(anova_result)
$ANOVA
  Effect DFn DFd       F           p p<.05       ges
2  Music   2  27 6.04215 0.006782334     * 0.3091855

$`Levene's Test for Homogeneity of Variance`
  DFn DFd SSn SSd          F         p p<.05
1   2  27 0.2  65 0.04153846 0.9593736
```

would be written as $F(2,27) = 6.04$. The p-value is just as in our previous analysis and would be written rounded as, $p<.001$ or $p = .0068$ (one or the other, not both).

This is a reliable result where the different audio input types affected the score on the problem solving test. In the R output, the reliability of the results can be accidentally mis-read because of the two rightmost columns. The very rightmost column that is labeled **ges** is reporting a generalized eta-squared effect size to help characterize not just how reliable the effect is but how large it is. We will discuss measures of effect sizes at the end of this

chapter. It is slightly unfortunate that the ges measure is in the range from 0.0 to 1.0, so when there is very little effect of the IV, it can sometimes look initially like a p-value and mislead the reader into thinking an non-reliable effect is reliable. The second column from the right is only an asterisk when the p-value is less than .05 and is designed to help find reliable effects in much larger, more complex analysis with more factors and interactions. It will not usually be very helpful in our simpler designs.

Example 2: 2x2, Anagrams and Ink Color

In the example below, we have simulated data from a hypothetical experiment on stress and eating preferences. In this experiment, participants were given anagrams to solve which were either hard or easy. This difficulty factor was intended to create more stress for the harder problems. The problems were presented in either red or black ink under a theory that red ink presentation implicitly stresses participants more than traditional black ink. After several minutes of solving anagram puzzles, participants were offered candy and the number of pieces of chocolate taken was scored as the dependent variable. As an exercise, you might consider all the potentially questionable operational definitions in this study, but for our simulation we are concerned with interpreting the analysis.

R Output

```
> print(anova_result)
$ANOVA
           Effect DFn DFd          F             p p<.05        ges
2          Color   1  76 31.9657273 2.618672e-07     * 0.29607291
3     Difficulty   1  76  9.0441736 3.571545e-03     * 0.10634678
4 Color:Difficulty   1  76  0.3617669 5.493165e-01       0.00473754

$`Levene's Test for Homogeneity of Variance`
  DFn DFd    SSn   SSd         F         p p<.05
1   3  76 0.1375 17.25 0.2019324 0.8947498
```

The output of analysis using R/Rstudio is shown in the table above which just shows the **ANOVA table** output from the ezANOVA function (not the function call itself or the descriptive statistics). For this analysis, which is a 2x2 design, we have three main possible effects reported. These are the two main effects, of ink color and difficulty, and the interaction between these effects. The interaction term is listed in the row with the Effect, Color:Difficulty.

The first effect reported is the main effect of Color (line following "2"). The F column contains the F-ratio and the two columns to the left indicate the degrees of freedom in the numerator and denominator. This would be written as $F(1,76) = 32.0$. The p-values are all in scientific notation but we should be able to see that for the main effect of Color, this would be .00000026, which we can simply write as $p<.001$. The rightmost two columns are just the asterisk for a reliable result and the ges effect size report.

Similarly, the main effect of Difficulty was found to be reliable as well. Reading on line 3, we can find that $F(1,76) = 9.04$ and $p<.01$ (or $p = .0036$) for this effect.

However the interaction between the two factors is not reliable here. On line 4, the Color:Difficulty interaction produced an $F(1,76) = 0.36$ and translating the scientific notation for the p-value, we see that it is $p = 0.55$ which is greater than .05. Between the scientific notation and the very low ges score, it is possible to mis-read the output for a non-reliable effect like this, so care must be taken when understanding the analysis output.

We might also note at this point that we have no idea what the reliable effects in this study actually are. We have confirmation that the ink color affected the amount of chocolate eaten but ANOVA output itself provides no information about the direction of the effect. Obviously, we need to describe the direction of the effects in order to effectively communicate the results of this kind of analysis to a reader. To do this, we will need to look at the descriptive statistics

	Red Ink	Black Ink
Easy problems	3.45 (1.0)	4.75 (0.91)
Hard problems	2.95 (0.89)	4.00 (0.92)

Above is the means table for the average number of chocolate pieces taken after completing the stressful problem solving exercise. The numbers below the means in parentheses are the standard deviations. Remember to always check the descriptive statistics in both R and in Excel to be sure they are the same values. The output format from R will be somewhat harder to read quickly but may serve as an example of why the above format for means tables is preferred in order to quickly see the data pattern.

Note that in our simulated data, the group who completed the puzzles written in black ink are taking more chocolate pieces than the red ink condition. This was counter to our initial hypothesis. Nothing in the ANOVA report itself would have alerted us to this surprising finding. Careful review of the descriptive statistics is always necessary to accurately explain and interpret our experiment results.

General 2x2 ANOVA Heuristics

For a 2x2 design, the degrees of freedom in the numerator, the first number in parentheses after the F, will be 1 for all three contrasts, both main effects and the interaction term. For each factor, this value is the number of levels minus one, which is 1. For a design with both factors being between-participants, the degrees of freedom in the denominator is the total number of participants minus 4. You can think of this as starting with the sample size and reducing this by 1 for each of the three contrasts plus one more.

In the ANOVA report, there will be 3 lines reporting the reliability of effect results. The first two lines report the main effects, that is, the difference

Reporting the F-ratio

Different statistical programs may format the information describing the evaluation of the statistical analysis in different ways. They should all provide the same core information somewhere in the output. The main statistical parameter resulting from an ANOVA analysis is an F-ratio, typically written as F. The F statistic is reported with two degrees of freedom, for the numerator and the denominator, which are included in parentheses. First is the numerator df (DFn), which is related to the number of levels within the condition being reported on. The second df is the denominator (DFd), which is related to the number of participants in the study across all conditions. There will also be a report of the *p* value, which is the probability of the data occuring by chance under the null hypothesis.

In a written description of the results, the format follows the frame below for each of the main effects and interactions and all should be reported:

$$F(df_n, df_d) = \mathrm{X.xx}, p < 0.\mathrm{yy}$$

between levels of that factor ignoring the other factor. The third line is the interaction term, typically listed as something like *Factor1:Factor2* and will tell you whether there is a reliable influence across factors.

As with all other inferential statistics, we also obtain a p-value which means the probability of having observed the difference occurring in the data under the null hypothesis. We use the same standard criterion for this, p<.05.

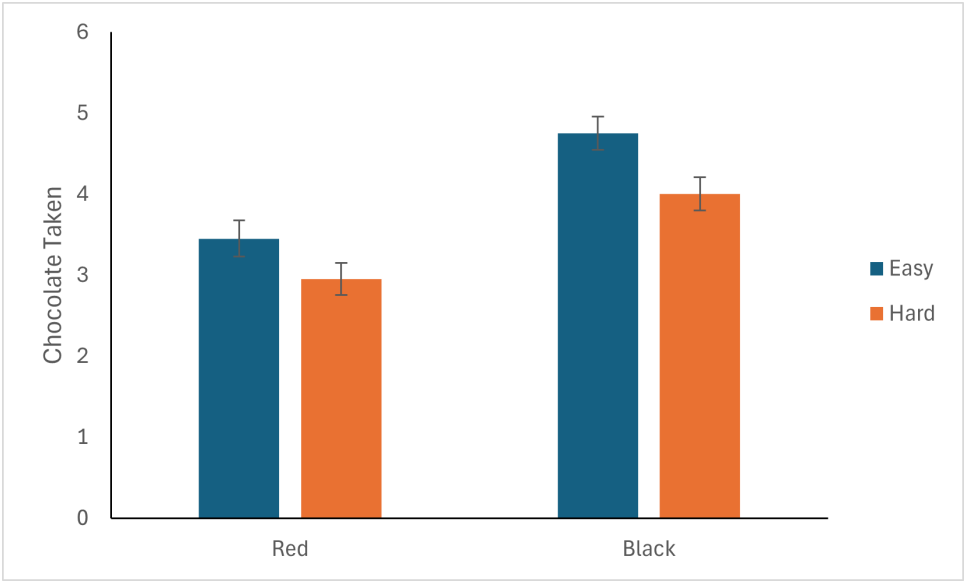
Importantly, the statistical reports of the main effect and the interaction term do not tell you anything about the direction of differences or what kind of interaction might have occurred. A reliable interaction could super-additive,

a 3:1 interaction (or sub-additive) or a cross-over interaction. Just looking at the ANOVA table cannot tell you which. It is necessary to review the descriptive statistics to understand the interaction so that you can report it accurately in the Results.

Making a 2x2 bar plot

To make a figure with 2x2 data in MS Excel, start by creating a labeled means table in a spreadsheet that contains the mean performance in each condition. It will look like the table above, but without the SD information. Start by selecting these 9 cells and Insert a 2-D Column chart.

With a little formatting it should look like this



The formatting applied here was to remove the Chart Title, add a vertical axis label, replace the Legend to the right side, change both axes to be black instead of gray and remove the y-axis gridlines. To add the standard error bars, it is necessary to prepare a separate 2x2 table of just the SE values for each of the cells in the design. Then select the Custom Error Bars option and select values across the row for each of the two series. That will get

accurate error bars for each of the four cells in the design, which each have a slightly different standard error. As a reminder, Google Sheets and the online versions of Excel do not currently have a method for individualized error bars across the conditions within a series of data. As a result, you should not use these programs because your error bars are inaccurate and it is very important not to present your data in a misleading or inaccurate way.

Example 3: 2x2 Mixed-model ANOVA

Consider the adage *the grass is always greener on the other side*. If we were to design an experiment to test whether this adage is true, we would need to come up with operational definitions of the metaphor that is based on viewing somebody else's situation more positively than one's own. For the purposes of this example, we might add an additional element that we hypothesize that this effect interacts with the personality variable optimism/pessimism such that the effect is much larger for optimists than pessimists.

For our hypothetical design, we will suppose that participants are given a description of a moderately lucky event, like winning money in a charity raffle, and asked to rate how happy they would be on a 1 to 7 scale. Participants will also be asked to rate how happy somebody else would be after the same event (order balanced, of course). This is a within-participants factor in this design since every participant answers twice, from the metaphor, once about the *other side* and once about their own side. In addition, we would use a personality scale to measure optimism and split our participants into two groups of 15, optimists and pessimists. As a participant variable, this is necessarily a between-participants factor.

For this 2x2 design, we have one between participants factor and a within-participants factor, which is referred to as a **mixed-model** ANOVA. We will see that with R/RStudio, the ANOVA results for this approach are presented in a very similar way with the only difference being slightly different df in the denominator.

However, this design approach requires some additional work with the spreadsheet tabulations of the data. In a typical data table, data are organized with one row per participant and all data collected from that participant listed across columns. For the within-participant variable, we would simply list the data as two columns. This is a useful format for reviewing data because it is easy to quickly compare scores across conditions within each participant. It is also relatively easy to calculate the descriptive statistics across conditions from this format.

However, the ANOVA analysis within R requires the data input to have a single variable per row and multiple rows for within-participant data. As a result, the sample data provided in the examples Excel (.xlsx) file has the same information organized differently than the file to be used as input for R (.csv). The need to re-organize the data is one of the many reasons why it is always

R Output

```
anova_result = ezANOVA(
+   greener
+   , dv = .(Green)
+   , wid = .(N)
+   , within = .(Side) # NULL if no within factors
+   , between = .(Personality) # NULL if no between factors
+   , observed = NULL
+   , diff = NULL
+   , reverse_diff = FALSE
+   , type = 3
+   , white.adjust = FALSE
+   , detailed = FALSE
+   , return_aov = FALSE # TRUE for showing details
+ )
Warning: Converting "N" to factor for ANOVA.
Warning: Converting "Side" to factor for ANOVA.
Warning: Converting "Personality" to factor for ANOVA.
> print(anova_result)
$ANOVA
            Effect DFn DFd         F            p p<.05        ges
2      Personality   1  28  5.316854 2.873922e-02     * 0.13530825
3             Side   1  28 53.200000 6.101031e-08     * 0.25052047
4 Personality:Side   1  28  7.221053 1.198682e-02     * 0.04340124
```

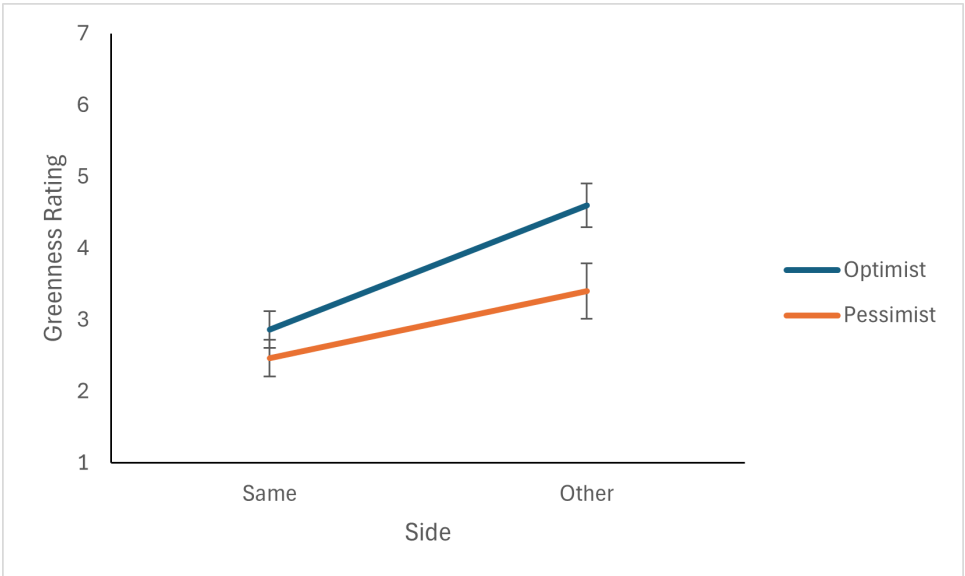
best to redundantly calculate sample descriptive statistics and verify the values across formats.

The output of the ANOVA looks very similar to the prior example. You might note that the df in the denominator (DFd) is the number of participants in the study, 30, minus 2 instead of 4 due to the within-participants factor of same/other side. The rest of the table is read the same way. There are reliable effects of Personality, Side and an interaction between these factors for the simulated data. To see what these effects are, we inspect the means table and a data visualization.

Means table:

	Same Side	Other Side
Optimists	2.87 (0.99)	4.6 (1.18)
Pessimists	2.47 (0.99)	3.4 (1.5)

Here we will use a line graph to visualize the effects the ANOVA indicated are reliable. As in the prior example, the graph is constructed from a means table in Excel with the four cell means used. In this case, we select a 2d Line graph as the starting template to make the following figure.



The additional formatting applied here to the standard Excel template was: remove the Chart Title, add vertical and horizontal axis labels, move the Legend to the right side, change both axes to be black instead of gray and remove the y-axis gridlines. Standard error bars are added to a line graph the same way they are added to a bar graph, using the Custom Error Bars option and selecting the SE values from a table of these in order to show the SE values correctly for each of the cells in the design.

The line graph allows us to see what the reliable effects reported in the ANOVA are. The Optimists are rating higher on the metaphorical greenness measure than the Pessimists, reflecting the main effect of the personality variable. Both groups are rating the Other Side higher than the Same Side, reflecting the main effect of the within-participants variable. These data would be consistent with the hypothesis implied by the adage that a good event happening to somebody else (the other side) would be perceived as having a bigger effect on their happiness. In addition, we have a reliable interaction between optimism and the side variable which the graph shows us is the hypothesized super-additive effect. Optimists see an even stronger effect of good things happening to somebody else than pessimists (in this made up data).

In the results section, these effects would be reported indicating the direction of the effect and supported by the statistical reports:

Optimists rated the benefit of a lucky event as more impactful than the pessimists, $F(1,28)=5.31$, $p<.05$. All participants rated the impact of the lucky event as being larger for somebody else compared with themselves, $F(1,28)=53.2$, $p<.0001$. There was a reliable interaction between the personality variable and the side factor reflecting the fact that the optimistic participants felt the lucky event would have an even larger effect on other people than themselves, $F(1,28)=7.22$, $p<.05$.

Modern Reproducibility Theory

You may have heard that psychology, as well as a variety of other scientific domains, is currently experiencing a *replicability crisis*. This has been inspired by a series of attempts to replicate well-known findings that have not produced reliable differences among conditions that were originally observed as reliable. There are substantial issues with the replication methodology that has been used that likely indicate that the term "crisis" is more extreme than warranted. However, the concern has usefully drawn attention to some aspects of how we carry out statistical inference in psychological science that we can use to improve our overall scientific progress.

The statistical model we have used so far reflects the approach used in the bulk of psychological research aimed at rejecting the null with a criterion of $p<.05$. As a reminder, this mathematically means that there is less than a 5% chance of the data appearing as observed if the null hypothesis were true and randomness produced the apparent effect. This leads to reporting results with a binary outcome: either the effect was reliable or not. There are several difficulties created by trying to make the outcome as simple as yes/no.

Marginal effects. It is not uncommon for research to be carefully carried out, analyzed properly and find that the probability of rejecting the null does not meet the .05 threshold but is instead in the range of .051 to 0.10. This poses some challenges for drawing interpretations of the results. We cannot claim that the results are reliable because they are not. However, the null hypothesis has actually been found to be somewhat improbable so simply saying that the effects are not reliable seems to miss an important aspect of the data. The simple binary model does not provide guidance for how to deal with these kinds of results.

Minuscule effects. It is also possible to have a statistically reliable effect that is actually extremely small. For example, if we found that an alternate studying method led to an reliable increase in memory performance of 1% accuracy, we would have a significant but somewhat uninteresting effect. This problem is fairly uncommon in experimental work as even small effects

can have theoretical implications, but comes up in more applied research or in some large-scale non-experimental studies. Here the simple binary model does not help us explain a reliable but not particularly useful effect.

Null findings. Sometimes our experimental hypothesis depends on providing evidence for a null effect. For example, we might want to show that sugar does not lead to hyperactivity in children. The simple binary model does not provide a method for evaluating this hypothesis since a *non-significant* findings could reflect a marginal effect or a true absence of an effect.

Effect sizes

Increasingly, the way researchers have sought to improve communication of results is to focus more on measures of the effect size. This changes our inference from "did the IV affect the DV?" into, "how much does the IV affect the DV?" In this approach, note that the null hypothesis is now the same as saying the effect size equals zero. Whenever we carry out an analysis, we are estimating the effect size based on our sample, which is a subgroup from a broader population. Unless we measure the entire population, we can never assert that the effect size is exactly one specific value. This is the difficulty of arguing for the null hypothesis. Our estimates can provide evidence that the effect size is not very different from zero, but not that it is exactly zero. When we fail to reject the null, we can only say that we are not sure that our current effect size estimate is different from zero.

As we reviewed earlier, an unstandardized effect size is simply the difference in the DV between conditions of interest across the average (mean) scores. In some cases, this can help communicate the results of an experiment, but it has the weakness of not incorporating any information about the variability of performance that was observed. Standardized effect sizes all incorporate a measure of variance to rescale the difference in means with the intention of providing a common scale for denoting effects across a scale something like 'small,' 'medium,' and 'large.' Unfortunately, the field of psychological science has not yet converged on a standard methodology analogous to the reporting

of p-values. Instead, there are several different forms of standardized effect sizes that are used depending on methodology and analysis type. Here we will briefly review two of these.

One common standard effect size measure is **Cohen's d**, which is often reported with t-tests to help communicate the findings. It is calculated as a ratio of the mean difference to variance and produces a number than can be used to scale the effect size into categories: small, medium, large, very large. Large effect sizes intuitively reflect factors that are particularly important to understand in their relationship to the dependent variable measure.

Another common effect size measure that is reported in the ANOVA results above is **generalized eta-squared** or η2 in the column titled **ges** in the *ezANOVA* output in R. This can be treated as an effect magnitude estimate like Cohen's d, but the scale is different. In the table below, values for both of these effect size measures are shown for the common effect size descriptive categories.

Table of effect size ranges

Effect Magnitude	Cohen's d	Generalized eta-squared
Small	0.2	0.02
Medium	0.5	0.13
Large	0.8	0.26
Very Large	1.2	0.40

While this effect size approach improves on the simple binary categorization based on whether p is less than .05 or not, the effect size statistics require becoming familiar with their relative scale values. It is also obviously very important to know what effect size measure is being provided by the analysis function. The ges values in the analyses reported above are generally very robust, many being medium or large, but if one accidentally compared them to Cohen's d effect sizes, they could be mistaken for small effects.

One of the advantages of the effect size approach is to identify reliable but small effects. Small effect sizes can be reliable but reflect factors that do not have a large effect on the dependent variable. In the third example above, the interaction between optimism and side is a small effect. This could help us accurately communicate the results that the more substantial effects were due to the main effects and while there is a reliable interaction, it has less impact on the scores.

Effect sizes are also very helpful for planning research and understanding conditions where effects are found to not be reliable. In both effect size types, the null hypothesis that the IV does not influence the DV is the same as indicating that the effect size is zero. While our statistical models do not provide a method for establishing confidence in a null finding, consistently observing effect sizes around zero would be a method for eventually supporting a conclusion that an IV has reliably no effect on the DV.

For planning research, if we have an estimate of the effect size we can use that to help plan the sample size for our study. If we think the effect size may be small, we know that we will likely need a large sample and very rigorous procedure to avoid a Type 2 error. When the effect size is expected to be large, smaller sample sizes are likely to be enough to observe a reliable effect. The process of planning the sample size from the effect size is carrying out a **power analysis**.

Power Analysis and Sensitivity in design

When planning a research study, particularly a rigorous Randomized Clinical Trial (RCT), it is important to be able to specify in advance exactly how many participants are expected to be in the research study. This is done by carrying out a power analysis, which is based on an a priori estimate of the effect size to be observed in the study. A power analysis takes a standardized effect size and with a specific number of participants expected to be recruited, provides a probability estimate of the chance of obtaining a reliable statistical difference between conditions. The math of carrying out this analysis is

beyond our scope here, but the underlying idea is that even where there is a real, true difference between conditions, data can still be variable enough that our statistics do not work (we fail to reject the null, a Type 2 error). In many formal research proposals, studies are designed around a power analysis based estimate of 80% or 90% likelihood of success.

In many experimental research studies, the researchers do not start with a strongly held numerical estimate of the expected effect size. In this case, it is impossible to carry out a formal power analysis before starting research. However, if the data indicate no reliable statistical differences, it may lead the researchers to consider that their design lacks sensitivity to the observed effect size. That is, the effect size is smaller than could be detected with the sample size available. This is often the case in results termed "marginal" above. The best practice in this case is to estimate the effect size from the "failed" study and use this to design a better follow-up study with larger n and/or a more powerful manipulation.

A consideration of power and sensitivity points out the difficulty of interpreting findings that "fail to replicate" prior studies that have been commonly reported as inspiring a "replicability crisis." We should actually expect studies to fail to replicate some of the time, even with real effect sizes when the effect is subtle, as many interesting effects are. Power analysis with effect sizes in the 'small' range can indicate that it may take several hundred participants to have a high probability of obtaining a reliable effect. There are certainly publications that have found reliable effects with smaller sample sizes, suggesting the researchers may have been lucky. We will consider the implications of this later in Chapter 19 (Responsible Conduct of Research).

Bayesian analysis

An entirely alternate approach to statistical inference exists based on Bayesian analysis. This approach focuses on the probability that the experimental hypothesis is correct and how this is influenced and updated as data becomes available. The probability of truth of the hypothesis acts like a quantitative effect size measure and follows a very robust mathematical tradition. The core element of this approach is to start with an estimate of the probability that your hypothesis is correct before you begin your research study. This number is referred to shorthand as the experimental **prior,** or **prior odds**.

After a study has been completed, if the data are consistent with the hypothesis, we can say that the probability that the hypothesis is true has increased. The data from the experiment has made us more confident in our hypothesis. The probability that the hypothesis is true including the experimental data is the **posterior probability** or **posterior odds**.

The Bayesian model is very intuitive because it reflects the way scientists think as research is carried out. We generally start planning an experiment with some confidence than the hypothesis is true and then over a series of studies, this confidence increases with additional consistent data. It also provides a mathematical approach to gaining confidence in the null hypothesis when it may be true.

Unfortunately, the mathematical basis of this approach has a major limitation in that calculating the posterior probability depends very heavily on the specific prior probability. That means if two researchers have different priors, for example one of them does not believe the hypothesis, they evaluate the statistics of the study completely differently. Since researchers often do not agree, it is very difficult to objectively quantify the effect of data on everybody's beliefs. As a result, this approach has not replaced our more traditional statistical models in spite of its benefits.

Key Takeaways

- Learn how to carry out an ANOVA analysis of factorial data in R
- Understand how to read the statistical output of this analysis and translate the result into the format you would use in the Results section of a scientific report
- Learn how to make a figure to visualize the results of a 2x2 ANOVA, both as a bar plot and a line graph so that you can choose the most effective presentation for your data
- Understand **effect sizes** in results reporting and how to use these to interpret large or small overall effects
- Understand the meaning of **marginal effects**, which do not meet the reporting requirements for reliability but do not indicate that the IV has no effect on the DV.
- Understand how a **power analysis** is derived from estimates of effect size to help plan sample sizes for proposed research.

Exercises

Analyze the data from Experiment 2. Start by review the data in Excel using the provided .xlsx file.

Calculate descriptive statistics as we did with our earlier analysis of Experiment 1. Note that there are now four conditions to calculate condition means, SD and SE. You should also examine marginal means, where two conditions are combined, so that you can observe the magnitude of the main effects.

To run the analysis, use the RStudio program to start an analysis session. Open the provided Exp2 ANOVA.R file. Set your *working directory* to the location of the source file (if that is where your .csv data file is located). Step through the commands that load the ezANOVA and related packages.

Load the datafile, run the describeBy function and verify that the output matches the descriptive statistics you calculated in Excel.

Run the ezANOVA command. Review the output, locate the key statistical parameters, the F-ratio's, the degrees of freedom and p-values. Prepare your report of the analysis, remembering to include directional statements about the main effects and a thorough description of any observed interaction.

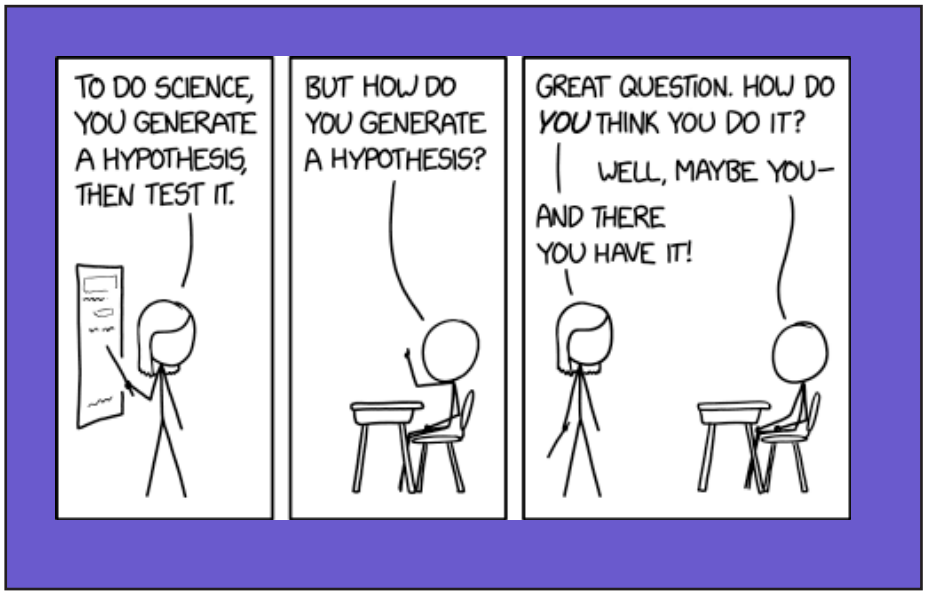
Save your work for when you prepare the next experimental report write-up.

Final Project Proposals

For our hands-on approach to learning experimental research methods, an important part of the learning experience is to develop, propose, carry out and write up a small project using experimental methods. Given the constraints of class, this necessarily is a very constrained process with limited scope. These projects can be carried out with generally good adherence to experimental design principles but cannot be considered formal, publishable research. Should a class project identify something that looks like a novel and interesting research finding, it is recommended that interested students seek out a laboratory group working in that area.

Learning Objectives

- Understand how to prepare and write a research proposal
- Explaining the experimental hypothesis related to prior research
- Describing the novel procedure including details of the independent and dependent variables
- Communicate what is novel about the new study, how it adds to or goes beyond the findings of previous research
- Prepare a protocol description at the level of detail that would be required for IRB review



Developing a research hypothesis: intuition

It's very common to start the research process with a statement that begins "I wonder what would happen if…" These ideas are usually more exploratory in nature and not initially expressed in the form of a testable hypothesis. You can turn these ideas into a testable hypothesis by not stopping with the exploratory statement. Make a prediction about what would happen and then argue for your prediction. Now you have a testable hypothesis.

Keep in mind that your hypothesis may very well be wrong at this point. One reason why it is often more comfortable at this stage of idea development is that you might not be very confident that your hypothesis is correct. A highly effective hypothesis for driving a research idea needs to be specific and testable, but it does not have to be correct. Often the most interesting research ideas are where the outcome is difficult to predict, or if there are two different theoretical approaches that make different predictions. A very useful heuristic for research proposals: it is far better to be wrong than to be vague. Make a specific, testable prediction and motivate your rationale. Even if you are wrong about the outcome, the study will likely be well-designed.

Preliminary background research

The original idea can come from background research, or it might come from a textbook or even more purely from intuition. Developing the idea requires examining the relevant published research around the question in the subdomain of interest. Google Scholar is the currently best recommended search engine for scientific work. It works through a familiar search term interface and covers all available published research across psychology and broader scientific fields. The amount of research included is vast enough that it is important to identify the correct key terminology, the right search terms, to find the related published work.

Many subfields of psychology have specific technical terms or jargon that are precisely defined and identifying these is very important to being able to do thorough background research. Many areas of science have jargon as

shorthand, but this is a particular feature of psychological science because we are often studying concepts that have familiar names. For example, a concept like "depression" is one that most people have an idea of but not everybody's idea may actually be the same. Research in this area will tend to cluster around the term "mood disorder" which is a less common but more precisely defined construct. Similarly, the idea of "altruism" is often done as part of research on "pro-social behavior." Getting started on background research is often a matter of first learning the best keywords for searching.

Once you have found the terms that lead you to the prior work in the area, you can start reviewing the methodologies that are commonly used. As an aside, if you cannot find prior published work, it is much more likely that you have not found the right search terms than nobody has ever thought to do research in this area before. Most research areas will have reported findings using a variety of methodological approaches. Many of these approaches may not be suitable for your proposal. They might require expertise in a complex

Literature Searching with Google Scholar

https://scholar.google.com/

This will provide lists of articles matching search terms. Clicking the article title will take you to the journal where the article was published. However, you may only have access to the title and abstract for that publication without paying a large fee.

Your university likely has a license to access the full text of most published work in reputable peer-reviewed gournals. If you connect through your university, either on the campus internet or via a VPN connection, you should see another option to access the publication. At Northwestern, the link is marked "Find@NU" on the right side of the page. This link will take you past the paywall to give you direct access to the full content of the vast majority of published scientific work.

or demanding technique requiring specific training. They might require access to a special population, or a very large number of participants to be adequately powered.

When a tractable methodological approach has been identified, the next step is to incorporate the original ideas behind an experimental hypothesis into that methodology. That requires constructing the operational definitions of the key constructs, identifying what can be manipulated and how and what the key dependent measure will be.

Operational Definitions

As noted previously in our approach to basic experimental design, we first need to come up with operational definitions of the constructs embedded in our hypothesis. Existing published research is the best place to start with ways to implement complex psychological ideas with tractable methodology. Examine the methodologies used in published work and consider how effective they are with respect to face validity, that is, how obviously the capture the idea. Where they seem imperfect, there may be necessary compromises made to make experimentation possible. Or it can be the case that the idea is so complex that there are many different ways to reconceptualize the idea for a research paradigm. Making adjustments to the methodology can improve the design, especially if the published work might have been constrained by technology of the time in which the research was accomplished.

In general, the process of setting up the operational definitions is the same as described earlier. Identify the key independent variable(s) and the level across each that can be controlled (for experimental designs) or measured (for non-experimental factors such as participant variables). The dependent variable needs to be a measured operational definition that can be collected and exhibits a roughly normal (Gaussian) distribution. The main issues to assess for the dependent variable are that participant scores will not tend to cluster at floor, the lowest possible value, or ceiling (perfect performance).

Operational definitions will often require defining the stimuli that will be used in the research. Any surveys to be used for measurement or words, images, pictures to be shown to the participants should be characterized. The published work may provide exact examples of the stimuli and instruments used or may give a broad overall description. Many published studies are accompanied on the journal's website by Supplementary Materials that may contain the exact stimuli or questionnaires used in the research. In other cases, it is necessary to go and obtain exact stimuli to be used in the research protocol. This should be done early in the research development process to be sure that the stimuli are obtainable and any surveys that are planned to be sued are available. Some research depends on research instruments that are held under copyright and may not be openly available to other researchers. In some cases, authors indicate that the stimuli used are "available upon request" but are not as responsive as would be desired. Before committing to the research plan, the availability of the key research elements needs to be assured. In addition, evaluation of the specific operational definitions used helps guide the analysis of possible extraneous variables to consider.

Open Access Science

There is a movement in the scientific community, often tied in with efforts to improve the reliability and reproducibilty of research, to make more scientific reports more generally accessible. When publishing a research paper, the authors often have an option to designate their report as *Open Access*, which means the publication will never be placed behind a paywall and difficult to reach for non-university researchers. However, many journals currently charge large fees at publication to authors who wish to have this designation and not all research is carried out with funds set aside to cover these costs. As a result, you may only occasionally find recent interesting research to be Open Access and easier to reach. You might note that when you do, the authors have made an explicit effort to make their work available to you.

Extraneous Variables

For planning the experimental procedure, it is necessary to identify as many possible extraneous variables as possible. There is no guaranteed approach to figure out all of these in advance, unfortunately. Looking at the detailed procedure from prior published work will provide a lot of insight into known factors that influence the dependent variable. General knowledge of the research area is the other main source of ideas. Increasing your background knowledge of the theoretical ideas in the area through additional research and reading the published literature is a great benefit.

Once the known extraneous variables are identified, the tools to manage these are the same as we have seen before: constancy, counterbalancing, and random assignment to conditions. Across the manipulated levels of your design, keep as much constant as possible. Anything you cannot keep constant, counterbalance across groups to keep this variable from confounding your research. This process will give you a detailed structure for your research protocol.

For data collection to be carried out in-person with the participants, it is often a good idea to fully script out the research procedure. This helps maintain constancy across multiple participants and especially when research is done by a collaborative team of experimenters who should all try to administer the task exactly the same way.

Data collection that will be carried out online is often very efficient. It relieves the burden of scheduling meetings with individual participants and reduces the need to carefully script participant interactions. However, it does require attention to the method by which the procedure is implemented online. Survey systems such as Qualtrics are very popular for online studies. Some time and effort will need to be invested in learning how to use the system and how to configure the presentation of stimuli or survey questions.

Recruiting Plan

Once the procedure is known, the nest step is to develop a recruiting plan. The two key questions to answer are (1) how many participants will be included in the study? and (2) how will these participants be found? Since all research participation is voluntary, the plan involves outreach to the population of interest with the opportunity to participate. If a specific subpopulation is the focus of research, a plan for finding and recruiting participants is necessary. The number of participants can be technically accurately estimated via the use of a power analysis (from Chapter 12). In many research projects based on convenience sampling, the main constraint is how many people can be recruited making the answer to this question "as many as you can." A good rule of thumb is 15-20 participants per manipulated condition, i.e., 40 for a two-group design and 80 for a fully between-participants 2x2 factorial design.

In formal research, an important aspect of the recruiting plan is developing a fair compensation plan for participants who volunteer. In some cases, this is based on class credit and therefore the experimental protocol is generally highly constrained in length (e.g., 30 m or 1 hr). The length of time needed to carry out the experimental protocol is important for this step as both financial compensation and credit are generally scaled on an hourly basis.

Analysis Plan

Best practice for experimental design is to have a formally written analysis plan for the DV as a function of the IV's before starting data collection. This can be as simple as noting that the analysis will depend on independent samples t-tests or a factorial ANOVA. It can also require more complex analysis approaches planned in advance. However, in a lot of research cases where a novel set of ideas are being tested against each other, unexpected findings will inspire additional analytic ideas in the course of the research process.

As a rule of thumb, if the analysis plan is significantly different than originally planned, the research should most likely be further explored with additional studies. Those studies can be planned with a more accurate understanding of the analytic needs. Using very creative and flexible analytic strategies runs the risk of research being bias by *p-hacking* as will be discussed in Chapter 19 (Responsible Conduct of Research).

IRB approval

Once the entire research plan is complete, the protocol is submitted for review to the Institutional Review Board for approval and/or revision. No systematic data collection from human participants intended for broad distribution should ever be carried out without review. Classroom research by not being intended for broad distribution is typically seen as not under the purview of the IRB. However, it is still important that class projects be carried out under the general principles of ethical research: Respect for Persons, Beneficence and Justice.

Participants should be informed that they are participating in a research study and indicate that they agree to this of their own choice. This can be done by including that information on paper for in-person data collection. For online data collection, the first element should be a notification that they are participating in a research study, what is expected of them and that they can decline to participate. Continuing with the protocol from that point is consenting to participate.

Practical Guidelines for Class Research

The most important first step for planning a psychological science project that can be completed in a classroom is to find a published report in a peer-reviewed journal to work from. You may start from intuition, interesting results you have seen in other classes or elsewhere, but it is extremely

valuable to have a closely related publication for reference. The reason for this is that the operational definition process in psychological science is often a lengthy one with false starts, mistakes and gradual improvements. Most published research implicitly relies on a series of pilot studies that guided the design through a variety of pitfalls. In a new subdomain, the first paper could easily reflect several years of preliminary research developing the methodological tools to test the hypothesis. Those often do not get included in the final publication – making science often look a lot easier than it is – but for classwork there is not time to do this methodological exploration. A published report will contain information on a set of definitions that worked, which is a good place to start.

As noted above, Google Scholar is the tool to use to find this first background publication. Be aware that it may take some exploration to identify the key technical terms used in your area of interest to find the published work. Also be aware that Google Scholar indexes outside of psychology. Pay attention to the journal the work is published in to identify if your search has drifted into related areas that are more physiological in nature (e.g., neuroscience, health) that may be impractical for class. Try to verify that the journal is peer-reviewed if the name is unfamiliar by checking if the publication is cited in recognizable outlets (use the Cited By link) and avoid publications with "Proceedings" in the name as these are conference proceedings which may not robust findings.

Once you have the first paper, you should look for something new to add to their approach. Even for class projects, we should approach research with the idea of extending findings to something novel and not just simply replicating a famous finding. The new idea to add can come from intuition, from the authors discussion of future research in their Discussion section, or from another related publication in the field. Blending two papers together often works well to create a 2x2 design from two publications that each had contrasts between two groups. Note that even if the two published papers used more complex designs, you may be able to take their main effect findings as evidence that a two-group study would work and use this as a

factor in your design. Check the interaction terms in their work, of course, to ensure that these are not indicating critical extraneous variables that you need to plan for.

For classroom work, you will prepare a 2x2 design with at least one manipulated variable. If you are combining published papers, you may come up with a design plan that is more complex. If you find that the design that best captures the previous work is a 3x2 or a 2x2x2 design, you will want to simplify down to 2x2 even if it weakens the scientific impact of your potential findings. Anything more complex than a 2x2 adds too much difficulty to be plausibly carried out in a classroom context. They require too much data, extending the time needed to recruit and test participants. The analysis is also necessarily much more complex and will significantly slow both the analysis and interpretation of the data when writing up the results later.

As discussed previously, there are a variety of ways to design a factorial study with 2x2 complexity. In general, for the manipulated variable, it is best to try to follow a published successful study as much as possible. The second factor can be a participant variable that is measured or recruited for instead of manipulated. However, avoid the temptation to lazily use men versus women as the second variable. This is an area where intuitions are often not at all grounded in a theory that can be articulated to motivate the study. To make the case that this is an important question to ask in your study, you must find research that shows your manipulated variable is explicitly affected by gender. Even so, be aware that modern understanding of gender does not reduce this variable to a simple choice of two options which will make this factor not suitable for a 2x2 unless you restrict recruiting.

With good sources, most of your work establishing the operational definitions can be taken from those publications. Use existing surveys, stimuli, or other materials from those papers as much as possible. If you need to create something new, keep it as simple as possible and maximize face validity, e.g., 1-10 scales asking participants to subjectively rate their current state.

Once you have the basic design and materials, you need a plan for carrying

out the procedure. It is very popular to collect data using online tools such as Qualtrics. Many aspects of experimental control can be implemented within these robust systems. Simpler systems such as Google Forms may also work. Be careful of fees associated with systems not affiliated with the university. Systems with university site licenses often provide access to a great deal of technical documentation to help set up the design and will have local experts to can answer questions (e.g., Northwestern University has a site license with Qualtrics and it is very effective for this purpose).

If you are not doing data collection online, write out a script for how participants will carry out the design procedures. The script helps maintain consistency in interactions with participants through the 40-80 repetitions of the process needed to accumulate the data. It also helps maintain consistency across a collaborative group where 4 people might each be responsible for portions of the data collection.

The recruiting plan should also be specified in advance as part of the research proposal. It may be as simple as social media posts or emails to a locally available convenience sample. If your research plans to recruit from specific populations such as athletes or engineers, be sure to plan how that group will be reached.

Once all the pieces are in place, the entire research protocol is written and provided to class instructors for review. This must include all stimuli that will be used in the planned research. That is, you should not at this point say, "we will collect images of famous celebrities from the internet." You should collect the images you will use and include those in your protocol submission.

Given time constraints, there is generally not time for a formal IRB review of these research plans and the instructor and teaching assistant will act as an informal IRB. As a result, all research should be absolutely minimal risk. All aspects of deception or any issues with privacy should be minimized or eliminated as much as possible. This may render some very interesting and motivated scientific research unable to be carried out in the classroom environment, but this should not be surprising given how important

adherence to ethical research is in science.

Data collection can not start until the entire protocol is reviewed and explicitly approved by classroom instructors. This is necessary for ethical research but has the risk of delaying projects and placing classroom researchers under severe time constraints. Prepare your proposal early and expect feedback about adjustments and revisions to your plan. Make those and resubmit the proposal as quickly as possible. Data collection can take significant time and there is a lot of work still to do after collecting data. The results need to be organized, analyzed and then the writeup of the results needs to be prepared. It is very ambitious to try to carry out an independent project in the scope of a month. It is possible but requires good time management throughout the process.

Grant proposals to funding agencies

The process of preparing a research proposal bears some resemblance to the process of writing grant applications that is an important part of the operation of major research laboratories. This process is somewhat more focused on obtaining funds to support these research projects. Many of the staff in most large research labs are not supported by the institution or university housing the lab but are entirely paid through outside funding to the lab. Research funds also support more expensive methodologies and participant compensation to carry out a series of studies organized around a core theoretical framework.

These proposals often look like research papers to some degree, although written in future tense rather than past tense. They will typically include a fair amount of "preliminary data" that has already been collected but not yet published that indicates that the research plan is feasible. The research plan will detail a series of experiments over a time frame that can vary from less than a year up to five years. These proposals have three major components: the collaborative team (led by a Principal Investigator, PI), a budget (cost/year) and a specific scientific research plan. The format of these sections

varies very widely across funding agencies. Research staff supporting grant applications spend a lot of time reading detailed formatting requirements and necessary levels of detailed information. The research plan is generally reviewed by a committee of scientific peers in a competitive fashion. Grants are reviewed on an annual cycle and depending on availability of funds to the funding agency only the top 5%-15% of proposals may be awarded funding.

Ideas for research proposals to granting agencies virtually never start with intuitive ideas and background research. Agencies tend to award grants to established experts in a field, so most grants build on the prior work of the collaborative team and PI. This does have some known issues in potentially creating a barrier to entry for researchers to become established or to move into a new area. At the same time, much of the money available for research funding comes from governmental sources which have a requirement to obtain some value from those funds. It is very hard to tell in proposal review which projects are going to have the largest scientific impact. Practically speaking, experts with robust track records in an area are most likely to produce scientific advances.

Within the USA, two major institutions that fund psychological science are the National Science Foundation (NSF) and the National Institute of Health (NIH). Within NSF, most psychological research is in the broad category of Social, Behavioral, Economic Sciences (SBE) which is then further subdivided into Behavioral and Cognitive Sciences and Social and Economic Sciences. The NIH is much larger in size and budget than NSF and houses 21 divisions across a very wide range of health-related research areas. Examples of programs that fund psychological science research include National Institute of Mental Health (NIMH), National Eye Institute (NEI), National Institute of Child Health and Human Development (NICHD), National Institute of Aging (NIA), National Institute of Deafness and Other Communication Disorders (NICHD), National Institute of Neurological Disorders and Stroke (NINDS).

There are also research projects funded through scientific divisions within the Department of Defense (DoD). These include a collection of laboratories such as the Air Force Research Laboratory (AFRL) and Army Research Laboratories

(ARL). The Office of Naval Research (ONR) acts as a funding agency similar to NSF but with research aimed at application at military personnel. Most DoD research is aimed at more immediate application of findings rather than long-term scientific understanding. However, it should be noted that these projects can be aimed at psychological questions across the large range of both active and retired (veteran) military personnel, making this sample fairly similar to the overall population. There are also specialty agencies within the DoD such as the Defense Advanced Research Projects Administration (DARPA) which fund very basic science aimed at extremely novel ideas (which has, unfortunately, led historically to support of ideas with little credible scientific support).

There are also private foundations that support psychological science that often have specific areas of interest. Many of these foundations approach scientific support with the same goal of highly rigorous, robust and internally valid research. However, there are some foundations that look for work that advances an agenda regardless of the robustness of science. Most universities or large research institutions have a Development office that provides guidance on private funding sources that support high quality psychological and other science.

Most of this information is not immediately relevant to undergraduate researchers but if you have the opportunity to work in a university laboratory, you may encounter some work aimed at seeking external funding. Some universities have some internal funds set aside to support undergraduate research and if you have the opportunity to apply for these, you will find yourself working through the same process as the lab PI. For example, Northwestern University has undergraduate research funding available for projects done over the summer as well as during the academic year. These can be a great opportunity to do formal, high quality research within a professional laboratory context.

Key Takeaways

- Preparing a research proposal is similar to writing a research report, only in the future tense.
- Providing a planned research protocol in enough detail for IRB evaluation includes at least as much detail as a the Methods section of a report, usually also including all the stimuli to be used in the study.
- Recruiting and sample size planning are done with both experimental rigor and ethical considerations in mind.
- New research builds on prior research for robustness and guidance in design and tools for experimental control.

Exercises

Prepare a research proposal outline for a project to be carried out as a short class project for a final paper.

The outline should contain all of the following information:

- Name of the researchers carrying out the project, including all group members
- Tentative project title
- Identify a first main background source and provide the APA-style reference to this peer-reviewed, published research. This source experiment will provide some theoretical background and starting ideas for the operational definitions.
- Describe the design of the main inspirational experiment in this paper including the IV(s), the DV, the number of participants and the outcome.
- For the proposed research, clearly indicate what new element you are planning to add to this design to expand on this published work. Describe your experimental hypotheses driving your proposed study.
- Diagram your 2x2 design, describing both factors and both levels of each factor. Identify how many participants you think you will need to test your hypothesis.

Final Project Guidelines

Final projects are short studies carried out by small groups of students working together in the Research Methods class (typically 3-4 students). The research project culminates in a final term paper due at the end of the quarter that is a substantial portion of the grade and meant to demonstrate mastery of the class materials.

The Final Project process has the following steps:

- Prepare a research proposal. This starts with idea-based brainstorming with feedback from the instructors and teaching assistants to provide guidance. Typical feedback about project development focuses on feasibility of the project in the limited time available, rigor of the design towards testing a hypothesis and practicality (and occasionally ethics) of carrying out the procedure.
- Develop the materials necessary to carry out the design. Often run online as a survey using a system like Qualtrics (which can also randomize across groups for your manipulated independent variable). Once the survey is completely finalized, it will be reviewed by the instructor and TA for accuracy. The instructor and TA are acting as an Institutional Review Board at this point, verifying the appropriateness of the procedure.
- Collect data, typically from a local, convenience sample of participants (other students, friends, family).

- Carry out the analysis of data using ANOVA tools used in Experiment 2. If your design does not easily fit within these analysis tools, plan to meet with the instructor or TA to review the analysis process together (typically a short zoom meeting during reading or finals week).
- Write up your project as a full APA-style scientific report (note that it is will still typically be fairly short by page count since scientific writing tends to be precise but also concise).

For guidance in coming up with research topic ideas, the recommendation is to read recent publications on a topic of interest to your group and pick a core research paper to build on. You will add something novel to their design, but you can use a lot of the reported methodology as a starting point. That helps identify effective operational definitions for potentially tricky constructs and maximizes the likelihood that the project will "work." It is not required that the project produce a statistically robust effect. However, it is frequently the case that final project papers written that have interpretable results tend to be better papers. It is possible to write a highly effective paper based on a null finding, but it takes more work and effort.

The proposed design should have a 2x2 structure with at least one of the independent variables manipulated (randomly assigned) by the experimenters. Designs more complex than this are strongly discouraged. However, if your initial idea development comes up with a 3x3 or 3x2 design, it is often very possible to simplify to a more focused 2x2 design.

Common elements for Online Surveys

Each survey should start with a clear statement that participation is voluntary and confidentiality will be maintained. Use the following unless your project requires additional specifics:

You are being asked to participate in a research study for a class at Northwestern University on research methods. Your participation is completely voluntary and you may stop participating or withdraw at any time during the study with no penalty to you. Any answers or data you provide will be kept confidential or anonymous and any descriptions of the experiment results will be done using averaged data across groups of participants from which you cannot be identified. If you have any concerns about this study or your participation, contact the experimenters or Professor Paul J. Reber, preber@northwestern.edu.

If you need basic demographics information from your participants, use these questions in the form provided here:

Please provide your age (do not participate in this study if you are under the age of 18 as we do not have permission to collect data from minors):

[Use a textbox for entry]

What is your gender?

[Use a drop-down multiple choice with 5 options: "Male", "Female", "Non-binary", "Gender not listed here", "Prefer not to say"]

Note that recruiting should be "word of mouth" since we don't have IRB approved recruiting materials. That is, you can ask anybody you know or email any list you are part of but not post broad requests on open, public sources. Remember to avoid keeping any identifying information about your participants with your data to maintain privacy.

Qualtrics tips

Northwestern University has a site license with Qualtrics to make it available at no cost to all students. This system has been very effective for student projects and is recommended as a research tool for student research projects.

NUIT page explaining how to register

https://services.northwestern.edu/TDClient/30/Portal/Requests/ServiceDet?ID=133

Qualtrics FAQ

https://www.qualtrics.com/support/

Additional tips:

- How to display one item per page: Looks and Feel -> General -> Questions Per Page
- How to randomly assign conditions: Use a tool called randomizer. https://www.qualtrics.com/support/survey-platform/survey-module/survey-flow/standard-elements/randomizer/
- How to randomize order: Click on block, Question Randomization -> Randomize order of all questions.
- Question type: For scales, use matrix table or slider; otherwise multiple choice is usually preferred. For each question, response requirements -> force response to make sure data is collected for all the questions.
- Preview the survey: Make sure the layout of your content is as designed.
- Generate test response: Tools -> generate test response. This is useful for looking at the format of your dataset before you actually distribute the survey.
- Distribute the survey: Publish your survey. Then go to Distribution -> Anonymous link.
- How to export data: Data & Analysis -> Export data. If you need the condition as a column, click on More options -> Export viewing order data. The name of the column can be found from Survey flow -> show flow ID.

Project Presentation Guidelines

At the end of the quarter, we will set aside some class time for all groups to present their project to the class. This helps gain some experience with presenting research (or proposed research) and also allows all the groups to see what your colleagues are working on.

Plan your presentation to be about 10 minutes, splitting the time up among your research team. Use PowerPoint to present your project roughly following the structure of the eventual report: Introduction, Methods, Results & Discussion. The Results section can be hypothetical explaining what you expect to see and how you will interpret that with respect to your hypotheses.

The presentation is a progress report and will almost never be a presentation of a completed project with all data collected and analyzed. Present progress at whatever point the project is currently at. Some feedback will be provided to help guide design, interpretation or explanation of findings. That feedback is aimed to help strengthen the final project term paper due at the end of the quarter.