11 Participant Sampling

In every psychological research study, we have a group of participants who are recruited to voluntarily participate in the research project. These participants are our sample, and they are drawn from a larger population. This is part of our statistical model as we take our descriptive statistics about the mean performance and variance from our sample and try to evaluate how accurate these numbers are with respect to the population. For statistics and our inferences about difference, the number of participants is a critical aspect of this calculation with more participants, or larger *n*, always being better.

In drawing inferences from our data, it is also necessary to consider where this sample came from, how it was recruited and how it might relate to or differ from the broader population. When we draw a conclusion from our data, who do we expect (hope) the conclusions will apply to? The answer to that question will vary across subdisciplines within psychology. Some areas are ambitious, such as cognitive psychology, which hopes to draw inferences that apply to all humans about memory, perception, or other basic human cognitive processes. Some areas are much more specific, such as developmental psychology, which may aim to draw inferences about

Psychological science is a science of people, by people and for people.

behavior in a very specific age range (e.g., 6-month old infants, or 12-16yo adolescents). Clinical psychology and neuropsychology often aim to draw inferences about people with specific psychological challenges, sometimes aiming to both understand these challenges and what they might apply to the broader population (healthy controls). Some areas are more complex, such as social psychology which looks for behavior that may be general to all humans, but may also be strongly culturally or socially influenced (e.g., stereotype bias). Biologically-oriented psychology disciplines such as health and neuroscience look for generalities but acknowledge that these can be influenced by biological differences such as genetics.

In some of these domains, the research hypothesis provides a very specific approach to identifying participants who can be invited to participate in the research study. However, for a wide range of research topics, participants are recruited broadly from the available population and in these cases, some attention needs to be paid to the sampling procedure to identify if that affects the conclusions of the study. Thus far, most of our research design has been aimed at good operational definitions and control of extraneous variables to avoid confounds to maximize our internal validity. The separate question of whether the conclusions from an internally valid study apply broadly to the whole human population is a question of external validity.

External validity describes the degree to which the conclusions of the research study can be applied to the rest of the population outside the specific sample who participated. Good external validity means the results can be **generalized** to the whole population. External validity can be limited if some aspect of the procedure to recruit participants into the sample accidentally introduced some bias such that the sample is no longer representative of the population. Examples of this are studies run exclusively on undergraduate populations, which typically have a very restricted age range (among other characteristics). Inferences drawn about the operation of memory or perception from this range may not apply to all other ages. Some of these limitations are obvious and implicit. Nobody expects studies on memory for reading prose passages to apply to pre-verbal infants. Others are subtler,

such as the finding that some visual illusions are not perceived universally across cultures but may reflect the experience of the participants with stimuli related to the illusions.

In some cases, the recruiting (sampling) method used to carry out the research can introduce bias into the results. For example, research on high-risk behavior has to carefully consider how to find participants, e.g., an advertisement for a study on sexual attitudes or behavior may not recruit a sample of participants that is representative of the population. This can be a difficult issue to resolve since we ethically require participants to voluntarily participate in a research study. People who are reluctant to talk about this topic even when their privacy is guaranteed may be persistently underrepresented in those studies and it can be difficult even to assess the size of the problem (which in this case, likely also varies substantially across different social and cultural groups).

The question of sampling methodology is most often considered in the context of non-experimental research, especially survey research and the related area of polling. Here we will introduce the main underlying ideas and methodologies as they can be applied to sampling and generalizability in experimental research. In Chapter 17 we will return to this topic in the context of non-experimental research. Many of the sophisticated sampling methods described here, such as stratified random sampling are critical for survey/polling methodology but less commonly used explicitly in experimental research.

Learning Objectives

1. Understand the differences in different kinds of **sampling** and their implications for drawing conclusions from the research.
2. Define **sampling bias** in general and **non-response bias** in particular.
3. Understand how to explain the limitations of the recruiting process using in a research study and how this might affect the **external validity** of the conclusions.

Sampling and Measurement

In addition to identifying which variables to manipulate and measure, and operationally defining those variables, researchers need to identify the **population** of interest. Researchers in psychology are usually interested in drawing conclusions about some very large group of people. This is called the population. It could be all American teenagers, children with autism, professional athletes, or even just human beings—depending on the interests and goals of the researcher. But they usually study only a small subset or sample of the population. For example, a researcher might measure the talkativeness of a few hundred university students with the intention of drawing conclusions about the talkativeness of men and women in general. It is important, therefore, for researchers to use a representative sample—one that is similar to the population in important respects.

One method of obtaining a sample is **simple random sampling**, in which every member of the population has an equal chance of being selected for the sample. For example, a pollster could start with a list of all the registered voters in a city (the population), randomly select 100 of them from the list (the sample) and ask those 100 whom they intend to vote for. Unfortunately, random sampling is difficult or impossible in most psychological research because the populations are less clearly defined than the registered voters in

a city. How could a researcher give all American teenagers or all children with autism an equal chance of being selected for a sample? The most common alternative to random sampling is convenience sampling, in which the sample consists of individuals who happen to be nearby and willing to participate (such as introductory psychology students). Of course, the obvious problem with convenience sampling is that the sample might not be representative of the population and therefore it may be less appropriate to generalize the results from the sample to that population.

Essentially all psychological research involves sampling—selecting a sample to study from the population of interest. Sampling falls into two broad categories. The first category, **probability sampling**, occurs when the researcher can specify the probability that each member of the population will be selected for the sample. The second is **non-probability sampling**, which occurs when the researcher cannot specify these probabilities. Most psychological research involves non-probability sampling. For example, **convenience sampling**—studying individuals who happen to be nearby and willing to participate—is a very common form of non-probability sampling used in psychological research. Other forms of non-probability sampling include **snowball sampling** in which existing research participants help recruit additional participants for the study, **quota sampling** in which subgroups in the sample are recruited to be proportional to those subgroups in the population, and **self-selection sampling** in which individuals choose to take part in the research on their own accord, without being approached by the researcher directly.

Compared with non-probability sampling, probability sampling requires a very clear specification of the population, which of course depends on the research questions to be answered. The population might be all registered voters in Washington State, all American consumers who have purchased a car in the past year, women in the Seattle over 40 years old who have received a mammogram in the past decade, or all the alumni of a particular university. Once the population has been specified, probability sampling requires a sampling frame. This sampling frame is essentially a list of all the members

of the population from which to select the respondents. Sampling frames can come from a variety of sources, including telephone directories, lists of registered voters, and hospital or insurance records. In some cases, a map can serve as a sampling frame, allowing for the selection of cities, streets, or households.

There are a variety of different probability sampling methods. Simple random sampling is done in such a way that each individual in the population has an equal probability of being selected for the sample. This type of sampling could involve putting the names of all individuals in the sampling frame into a hat, mixing them up, and then drawing out the number needed for the sample. Given that most sampling frames take the form of computer files, random sampling is more likely to involve computerized sorting or selection of respondents. A common approach in telephone surveys is random-digit dialing, in which a computer randomly generates phone numbers from among the possible phone numbers within a given geographic area.

A common alternative to simple random sampling is stratified random sampling, in which the population is divided into different subgroups or *strata* (usually based on demographic characteristics) and then a random sample is taken from each *stratum*. **Proportionate stratified random sampling** can be used to select a sample in which the proportion of respondents in each of various subgroups matches the proportion in the population. For example, because about 12.6% of the American population is African American, stratified random sampling can be used to ensure that a survey of 1,000 American adults includes about 126 African-American respondents. Disproportionate stratified random sampling can also be used to sample extra respondents from particularly small subgroups—allowing valid conclusions to be drawn about those subgroups. For example, because Asian Americans make up a relatively small percentage of the American population (about 5.6%), a simple random sample of 1,000 American adults might include too few Asian Americans to draw any conclusions about them as distinct from any other subgroup. If representation is important to the research question, however, then disproportionate stratified random sampling could be used to

ensure that enough Asian-American respondents are included in the sample to draw valid conclusions about Asian Americans a whole.

Yet another type of probability sampling is **cluster sampling**, in which larger clusters of individuals are randomly sampled and then individuals within each cluster are randomly sampled. This is the only probability sampling method that does not require a sampling frame. For example, to select a sample of small-town residents in Washington, a researcher might randomly select several small towns and then randomly select several individuals within each town. Cluster sampling is especially useful for surveys that involve face-to-face interviewing because it minimizes the amount of traveling that the interviewers must do. For example, instead of traveling to 200 small towns to interview 200 residents, a research team could travel to 10 small towns and interview 20 residents of each. The National Comorbidity Survey was done using a form of cluster sampling.

Sampling Bias

Probability sampling was developed in large part to address the issue of sampling bias. Sampling bias occurs when a sample is selected in such a way that it is not representative of the entire population and therefore produces inaccurate results. This bias was the reason that the Literary Digest straw poll was so far off in its prediction of the 1936 presidential election. The mailing lists used came largely from telephone directories and lists of registered automobile owners, which over-represented wealthier people, who were more likely to vote for Landon. Gallup was successful because he knew about this bias and found ways to sample less wealthy people as well.

There is one form of sampling bias that even careful random sampling is subject to. It is almost never the case that everyone selected for the sample actually responds to the survey. Some may have died or moved away, and others may decline to participate because they are too busy, are not interested in the survey topic, or do not participate in surveys on principle. If these survey non-responders differ from survey responders in systematic

ways, then this difference can produce **non-response bias**. For example, in a mail survey on alcohol consumption, researcher Vivienne Lahaut and colleagues found that only about half the sample responded after the initial contact and two follow-up reminders (Lahaut, Jansen, van de Mheen, Garretsen, 2002).The danger here is that the half who responded might have different patterns of alcohol consumption than the half who did not, which could lead to inaccurate conclusions on the part of the researchers. So to test for non-response bias, the researchers later made unannounced visits to the homes of a subset of the non-responders—coming back up to five times if they did not find them at home. They found that the original non-responders included an especially high proportion of abstainers (nondrinkers), which meant that their estimates of alcohol consumption based only on the original responders were too high.

Although there are methods for statistically correcting for non-response bias, they are based on assumptions about the non-responders—for example, that they are more similar to late responders than to early responders—which may not be correct. For this reason, the best approach to minimizing non-response bias is to minimize the number of non-responders—that is, to maximize the response rate. There is a large research literature on the factors that affect survey response rates (Groves et al., 2004). In general, in-person interviews have the highest response rates, followed by telephone surveys, and then mail and Internet surveys. Among the other factors that increase response rates are sending potential respondents a short pre-notification message informing them that they will be asked to participate in a survey in the near future and sending simple follow-up reminders to non-responders after a few weeks. The perceived length and complexity of the survey can also make a difference, which is why it is important to keep survey questionnaires as short, simple, and on topic as possible. Finally, offering an incentive—especially cash—is a reliable way to increase response rates. However, ethically, there are limits to offering incentives that may be so large as to be considered coercive.

Online Data Collection

An increasingly popular methodology for psychological research is based on using web sites that provide access to research participants as a service. One popular option as been Amazon's Mechanical Turk service, often referred to by the shortened **mTurk**. The mTurk service was not originally designed for human participants for research but has been applied to this purpose by many researchers who were able to carry out data collection with online methodologies.

Initially, several concerns were raised about online data collection related to data collection methodologies that did not rely on ever directly interacting with experiment participants. For example, if participants were effectively anonymous, how could we ensure that research was carried out ethically with respect to vulnerable populations such as children. Researchers need to trust and rely on the company running the online marketplace to be rigorous with respect to knowing their customers so that they can certify participants meet standard ethical practice. In the years that psychological science has been carried out with mTurk, no evidence of ethical compliance problems have yet been identified so this type of concern has abated somewhat.

There were also several preconceptions about samples recruited through the internet potentially not being representative of the broader population. Concern was initially raised that internet-based samples might not be demographically diverse. However, studies of online participant demographics have shown that these samples tend to be more diverse than traditional samples that have depended mainly on undegraduate students. There was an early stereotype that heavy internet users might be more likely to be maladjusted, socially isolated, or depressed, which has not been found to be true in practice. In general, there was also concern that internet-based findings might differ from those obtained with other methods but this has also not appeared to be the case whenever methodologies could be compared directly.

Some of the concerns initially raised reflected the fact tha the first availabile

History of mTurk

The name Mechanical Turk refers to a 19th century fraud where a machine was purported to be a chess-playing automaton, an early robot. The device was shown for a fee and wealthy patrons could pay to play against it. In reality, the machine was simply a device that allowed a very short human hiding within the machine to manipulate pieces and play chess. The name reference was likely chosen by Amazon to reflect the fact that the original goal of the mTurk service was to hire humans inexpensively to do cognitive tasks that could not be completed accurately by artificial intelligence programs some years ago. For example, evaluating the accuracy of key words or search terms being related to online postings or determining if photos provided on a site selling cars were actually usable pictures of automobiles. Their model was to create a marketplace where *requesters* could post *human intelligence tasks* that could be completed by *workers* for pay. Since psychological research can easily be thought of as a human intelligence task, this online marketplace presented an interesting opportunity for psychological science for paradigms that could be carried out entirely online.

marketplace for recruiting participants online, mTurk, was not designed for the purpose of systematic, scientific data collection. One consequence of this is that it is possible to collect data entirely anonymously and not even be aware of the demographics of the participants sampled. While this seems to protect participant privacy very effectively, there is no way to know if there was some unexpected sampling bias in data collection that might be important for understanding the validity of conclusions.

Another newer marketplace company, **Prolific**, has recently become available for recruiting human participants for data collection. Their system was built explicitly around the idea of access to research participants. As a result, they can provide averaged demographic information for participant samples without the researcher needing to directly collect identifiable information together with research data.

Data collection online still faces the general issue that the context in which the research protocol is being carried out is under far less control than is possible in laboratory conditions. Participants might be engaged with the research task on mobile devices, in loud or distracting environments or not able to provide their full attention to the research task. A tendency to not comply with more complex tasks online might inadvertently embed a new kind of non-response bias. To date, the fact that online methodologies appear to result in similar patterns of data as in-person protocols suggests this is not a major problem for the kinds of tasks currently run online.

Deploying a protocol within an online environment can also place an implicit burden on the researchers to be familiar with tools for online protocol development. For many paradigms, the wide range of online survey tools makes this process very straightfoward. For some forms of online studies where precise control is desired, researchers need to become familiar with online programming tools.

Overall, the ability to access a very large population quickly and easily online appears to provide an opportunity to accelerate research processes in psychological science. Participants can be recruited and complete protocols onlie at much more rapid rates than can be accompished with in-person procedures. Further technological advances will likely continue to influence how these processes work in practice and how these affect the process of recruiting human participants into psychological studies. Attention will have to be paid to challenges that arise, such as the potential for more effective AI to simulate human behavior, while capitalizing on the potential to carry out effective science at a much greater pace.

WEIRD samples

As we noted in Chapter 8, psychological research is increasingly developing a sensitivity to the fact that a great many well-known findings about psychology depend largely or entirely on undergraduate participants who are at universities with substantial research programs. University students are already a very restricted demographic based on age and education. In addition, research universities tend to be ones that are more competitive with respect to admissions and therefore reflect populations that have succeeded in that competition. That may bias samples with respect to both individual difference variables and also factors like socio-economic status, which have substantial impacts on student preparation, success, and ability to attend competitive universities.

The acronym WEIRD, from Western Educated Industrialized Rich Democratic, has been used to describe the potential sampling issue involved in depending on undergraduate participants. Note that the Western and democratic elements of the acronym reflect the fact that to date, the overwhelming bulk of published psychological research that has been done with populations drawn from the United States, Canada and Western Europe. Acknowledgment of this issue has mainly been used to be more explicit about the demographic characteristics of the participants in research in publication. Some effort has also been made to increase outreach to broader and more diverse communities.

This is not a simple problem to solve as research at universities where research tends to depend on convenience samples of undergraduates. Because this kind of recruiting is far easier that investing time and energy into community outreach, the scientific research is therefore less costly to carry out. Making research more difficult or expensive will lead to less science being accomplished, which is not necessarily the goal of broadening our sampling procedures.

Online/internet based recruiting holds some promise for improving this, but will still tend to over-represent aspects of the WEIRD demographics.

Within research on internet use, there was documented a **digital divide** that reflected less access to the internet among poorer communities. This effect may be attenuating with greater accessibility through increasingly sophisticated mobile devices. It is still likely the case that online-based recruiting procedures are not reaching a fully diverse and representative population, although they are likely similar or better than standard WEIRD dependent approaches.

Considerations of the broader population the research sample is drawn from correctly brings attention to the question of: who is the conclusion drawn from research aimed at? This is essentially a question of **external validity**. Reliable data is known to reflect an effect seen in the participants sampled. The question is then how broad is the population this sample was drawn from to which the conclusion can be generalized?

Limitations in External Validity

Identifying that a research study is based on a WEIRD sample does not necessarily imply that the results are importantly limited by this fact. To identify a limitation, we need to be able to communicate an alternate hypothesis. Specifically, we should be able to identify a different sample that might plausibly not be expected to show the same behavior as the participants in a research study being reported. For example, in our Experiment 1 study examining the effect of encoding "depth" on word memory, the data were collected from undergraduates in a research methods class. We might note this and worry that it limits our conclusions. However, we would need to identify what different samples might not show the depth effect. There is no existing work that suggests that non-college-attending participants, or older/younger participants do not show the effect of memory enhancement following a study process that connects items to be remembered to existing semantic knowledge. Obviously, participant who cannot read (young children, illiterate) would probably not show the depth encoding effect for word list stimuli. We might also see this as a limitation of

the stimuli that could potentially be addressed in future studies.

In many cases, the limitation arising from sampling is completely clear. Research on stereotype bias based on race that are run in the USA are likely to show different patterns of behavior than bias studies carried out in other countries. There may be important commonalities that provide insight into human behavior for all humans on the planet, but the sample context is an important part of understanding the result of a research study run in one location. Psychological research based on attitudes, identity, or moral values are all examples of research that is very likely to be related to the population from which the participants were sampled.

Identifying external validity limitations is typically done by brainstorming as many conditions as you can think of where the effects of the study might not apply, then decide if any of these are important limitations to include in the discussion of your results. Hypotheses about effectiveness of limitations in generalizability usually must be done based on general knowledge of people's behavior. Our intuitions are often useful here, but expertise within the specific subdiscipline of the research is also very helpful. Obviously, the more experience you have in psychological research, the better your intuition about what sampling issues may be relevant.

External validity judgments can virtually never be made perfectly or with absolute confidence. They may look accurate, but then a new idea about differences across people advances our scientific understanding and modifies previous broad statements. Interval validity of studies, when established, rarely changes when new evidence about the phenomenon at study becomes available. However, the external validity of findings may change as science progresses and new factors and context elements are discovered in subsequent research. Often these advances further refine our understanding of the groups of people to whom the results apply, demonstrating the need to be complete and accurate about the samples participating in each research study.

Key Takeaways

- The method by which participants are recruited into the study to be part of the sample can affect the **external validity** and **generalizeability** of the scientific findings to the broader population.
- Participant samples drawn from undergraduate classes may over-represent WEIRD populations and reflect sampling bias.
- Methods to sample participants from broad populations may use either probability or non-probability sampling approaches.
- Population-based surveys and polling techniques can use complex, balanced **stratified sampling** approaches to avoid bias.
- **Non-response bias** reflects the possibility of a shift in the overall patterns of responses based on participants who elect not to participate due to the content of the research.

Exercises

TBA

12 Statistics 2: ANOVA

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.

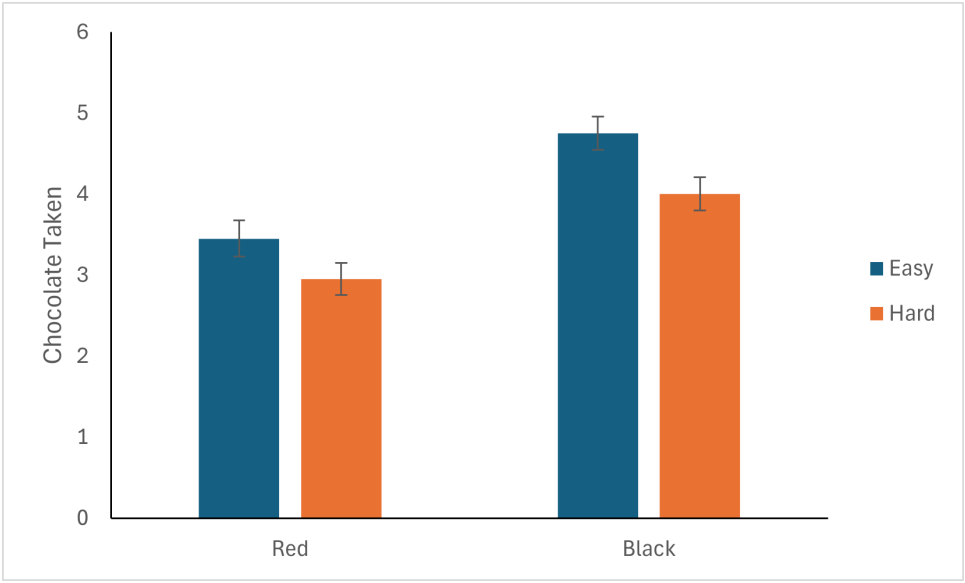
Importantly, the statistical reports of the main effect and the interaction term do not tell you anything about the direction of differences or what kind of interaction might have occurred. A reliable interaction could super-additive,

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
```

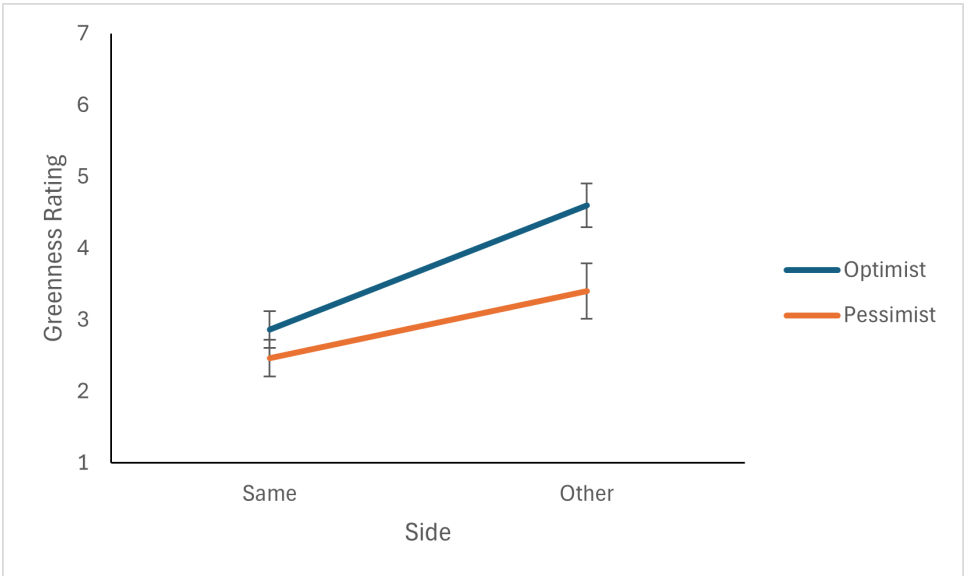
best to redundantly calculate sample descriptive statistics and verify the values across formats.

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.

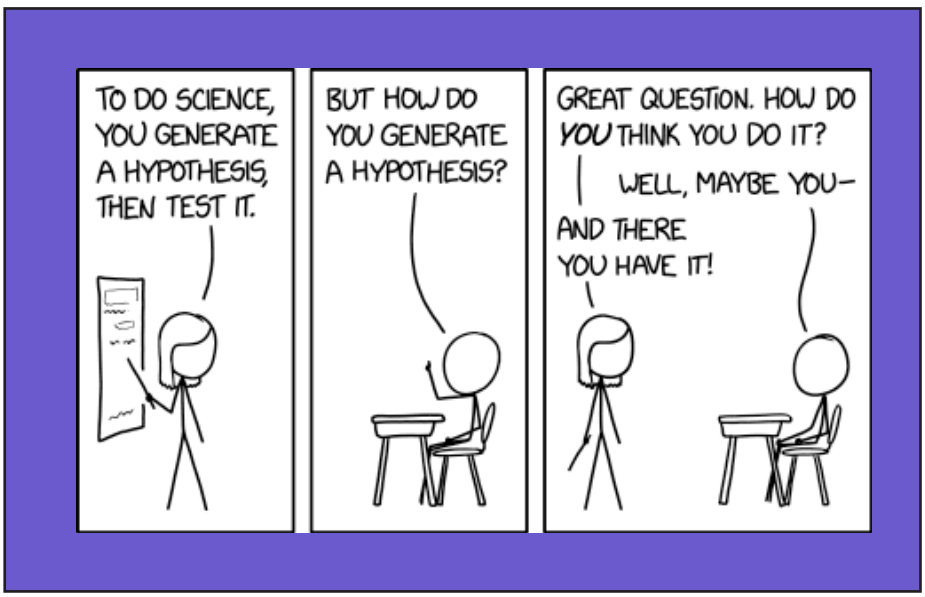
Load the datafile, run the describeBy function and verify that the output matches the descriptive statistics you calculated in Excel.

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.

13 Proposing Research

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

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.

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 or demanding technique requiring specific training. They might require

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.

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.

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.