Chapter 3: Experimental Control

We have described the process of setting up an experimental design as starting with the high-level constructs and then implementing operational definitions that allow us to create an experimental procedure that assesses the effect of an independent variable on a dependent variable. The previous chapter discussed some aspects of creating a measured operational definition that can be used as the dependent variable. In Chapters 3 and 4, we will review issues and methods for creating effective independent variables that will allow us to draw strong conclusions from our research studies.

As an example, consider starting with the hypothesis that “listening to music while studying is helpful.” To evaluate this idea as a research question, we would need to choose operational definitions for music and also define what we mean by “helpful.” We could potentially define what we mean by helpful as a measured operational definition of some academic performance, e.g., grade on an upcoming exam. That would give us our dependent variable. For the independent variable, a surprisingly common design error is to come up with a way of implementing a way of listening to music during studying – and then stop there. As experimental psychologists, we should get in the habit of asking a follow-up question after stating the hypothesis: “compared to what?”

Our hypothesis implicitly implies that music should help with studying more than some condition that does not involve music. In this example, the choice of the contrasting condition is not completely straightforward and whatever is chosen will affect the strength of the conclusions drawn. We could make a bad choice of comparison condition such as playing obnoxiously loud, distracting noise. In that study, we would almost certainly observe an effect of our independent variable on the dependent variable but not be able to infer whether music helps studying or loud noise hinders studying.

In this chapter, we will start with consideration of designs based on treatment/control and then extend this idea to using an IV with two levels. Most of the designs we consider in this class are based on two levels. While the extension to three or more levels is fairly simple conceptually, this change can dramatically increase the complexity of the data analysis tools required to draw inferences about statistical reliability.

## Learning Objectives

1. Understanding the independent variable in experimental design
2. Define what a control condition is, explain its purpose in research on treatment effectiveness, and describe some alternative types of control conditions.
3. How to construct two treatment conditions as the IV for a study and how this might be extended to more complex designs
4. Recognize examples of confounding variables and explain how they affect the internal validity of a study.
5. Bias and demand characteristics as potential confounding variables and the importance of random assignment
6. Situations requiring the use of single-blind and double-blind methodologies

# **Treatment and Control Conditions**

In psychological research, a treatment is any intervention meant to change people’s behavior for the better. Interventions includes psychotherapies and medical treatments for psychological disorders but also interventions designed to improve learning, promote conservation, reduce prejudice, and so on. We will discuss methodologies specific to intervention-based research in detail in Chapter 14. Here, we will start with the simplest kind of intervention research as an example of a very basic design. To determine whether a treatment works, participants would be randomly assigned to either a treatment condition, in which they receive the treatment, or a control condition, in which they do not receive the treatment. After the intervention, the dependent variable is assessed and if participants in the treatment condition score better, we can infer that the treatment led to the change in condition. In the earlier example, we could compare exam scores after studying with music to studying in silence with music acting as an intervention and silence as the control.

Choosing an effective control condition is not always a straightforward process. One challenge is that participants might be aware that they are in the treatment condition, which creates an expectation on their part about the research study. Participants who are aware they are receiving a treatment that is hypothesized to improve their performance might exhibit an influence of **demand characteristics**, where their scores on the dependent variable incorporate this expectation. Because of this, while minimally engaging control conditions, sometimes termed a no-treatment control condition where participants receive no treatment whatsoever, can create an interpretation problem due to the existence of placebo effects. A placebo is a simulated treatment that lacks any active ingredient or element that should make it effective, and a placebo effect is a positive effect of such a treatment. Many folk remedies that seem to work—such as eating chicken soup for a cold or placing soap under the bed sheets to stop nighttime leg cramps—are probably nothing more than placebos. Although placebo effects are not well understood, they are probably driven primarily by people’s expectations that they will improve. Having the expectation to improve can result in reduced stress, anxiety, and depression, which can alter perceptions and even improve immune system functioning (Price, Finniss, & Benedetti, 2008). Placebo effects are interesting in their own right, but they also pose a serious problem for researchers who want to determine whether a treatment works.

## The Powerful Placebo

Many people are not surprised that placebos can have a positive effect on disorders that seem fundamentally psychological, including depression, anxiety, and insomnia. However, placebos can also have a positive effect on disorders that most people think of as fundamentally physiological. These include asthma, ulcers, and warts (Shapiro & Shapiro, 1999). There is even evidence that placebo surgery—also called “sham surgery”—can be as effective as actual surgery.

Medical researcher J. Bruce Moseley and his colleagues conducted a study on the effectiveness of two arthroscopic surgery procedures for osteoarthritis of the knee (Moseley et al., 2002). The control participants in this study were prepped for surgery, received a tranquilizer, and even received three small incisions in their knees. But they did not receive the actual arthroscopic surgical procedure. Note that the IRB would have carefully considered the use of deception in this case and judged that the benefits of using it outweighed the risks and that there was no other way to answer the research question (about the effectiveness of a placebo procedure) without it. The surprising result was that all participants improved in terms of both knee pain and function, and the sham surgery group improved just as much as the treatment groups. According to the researchers, “This study provides strong evidence that arthroscopic lavage with or without débridement [the surgical procedures used] is not better than and appears to be equivalent to a placebo procedure in improving knee pain and self-reported function” (p. 85).

# Independent Variable with Two Levels

The challenges of interpreting a placebo condition can be avoided by using an independent variable with two levels that vary across the underlying construct. In our Experiment 1, the two conditions were operational definitions of varying levels of depth of processing. Viewing the independent variable this way highlights how the operational definition of this part of the design is similar to the measured operational definition for the dependent variable. However, the independent variable needs to be conceptualized in a way that allows for implementation of two conditions that vary in the amount that the underlying construct is brought to bear.

Independent variables that can be manipulated to create two conditions and experiments involving a single independent variable with two conditions are often referred to as a single factor two-level design.  This approach is also referred to as a **two groups independent samples** design, which will connect to a common statistical approach for this design (two independent samples t-test; see Chapter 5).

This approach is straightforward conceptually, but not always simple to put into practice. In our music and studying example, it is not necessarily immediately clear how we might construct two levels of “music” to compare. This uncertainty is often very helpful in experimental design to focus the experimenter’s attention on the underlying question of “why” one might hypothesize that music helps with studying. The question of the mechanism by which the independent variable is hypothesized to affect the dependent variable, the “why?” question is often the most difficult and interesting part of psychological science. No hypothesized mechanism was given in the above example, but we might conjecture that music helps with studying by inducing a state of calm while blocking out ambient noise. In that case we could compare music to an audio recording of ambient noise that participants found not be calming. In Chapter 4, we will discuss additional considerations needed to implement novel comparison condition to maintain the validity of the research study.

## Extending this design approach

We will discuss a series of examples with two-group designs as these are the simplest designs that still illustrate the critical issues of **extraneous variables** and potential **experimental confounds**. However, sometimes greater insights can be gained by adding more conditions to an experiment. When an experiment has one independent variable that is manipulated to produce more than two conditions it is referred to as a single factor multi level design. So rather than comparing a condition in which there was one witness to a condition in which there were five witnesses (which would represent a single-factor two-level design), Darley and Latané’s experiment (Chapter 1) used a single factor multi-level design, by manipulating the independent variable to produce three conditions (a one witness, a two witnesses, and a five witnesses condition).

In addition, we will discuss creating designs with two conditions that all participants get to experience, which are termed **within-participant designs** (Chapter 7). This approach can be very effective as long as the unique challenges of condition order can be managed. In addition, we will extend our consideration of experimental design two cases with two independent variables, typically called **factors** and **factorial design** (Chapter 9). Even the simplest factorial designs add a lot of complexity to the experimental design, procedure and data analysis. They allow for a much broader and more interesting range of hypotheses to be tested. Most modern published psychological research uses multi-level factorial design. Although we will keep to simpler designs for examples here, the extension from a two-factor design to arbitrarily complex designs is conceptually straightforward.

# Internal Validity

The term **internal validity** is used to characterize an experimental design that will be able to test the underlying hypothesis. Any major problem that impairs the ability to draw a conclusion from the experimental data is a problem with the internal validity of the study. One way this can happen is if there is a mistake in the operational definitions. If they do not accurately reflect the underlying construct, the main inference about the constructs cannot be drawn from the data.

This core idea is distinct from **external validity**, which reflects the degree to which the conclusions can be applied to participants outside the research lab, e.g., in the real world. External validity generally depends on the methods of sampling participants, that is, how they are found and recruited into the study. This issue will be discussed in depth in Chapter 13, but as a preview, you can consider the concern being raised about the general dependence of psychological research on behavior measured from undergraduate students at major American universities. The question is whether the results obtained from university participants correctly predict the behavior of the broader population and whether we need to consider broader sampling or limiting the expected breadth of our conclusions.

The question of internal validity is not whether the results apply outside the population engaged in the study, but whether the experimental design itself is robust enough to draw any conclusions at all.

# **Confounding Variables**

The most important consideration to consider first when evaluating whether an experiment has a high degree of internal validity is whether there is a **confounding variable** embedded in the design. In a two-group design, an **experimental confound** is an **extraneous variable** that varies with the planned independent variable that leads to different conclusions about the results of the study.

## Extraneous variables

**Extraneous variables** are any variables that affect scores on the dependent variable that are not part of the experimenter’s design. There are always a large number of these implicit in any experiment. For example, in the Experiment 1 memory study, the words themselves affect how well they will be remembered later. Uncommon words and longer words are more memorable than short, frequently encountered words. Details of the experimental context such as what time of day, where, when and with what external distractions will affect performance. Individual differences in memory, or familiarity with the words will affect scores on the recognition test. None of these variables are confounds for this experiment. Instead, these reflect factors that affect the recognition score dependent variable that will mostly show up as variance in the observed data.

Most of the time, extraneous variables do not affect the internal validity of a study. That is, they create noise in measures that can lead to a failure to reject the null hypothesis statistically. They create Type 2 errors, but this is the less consequential problem of the two main errors in experimentation.

For an extraneous variable to be a **confound**, it has to vary with the independent variable. As an example, while some words are more easily remembered than others, for this to be a confound problem for our Experiment 1, participants would have to see different words in the two conditions (rating liking or counting vowels). If the participants in the rating liking condition also saw words that were more memorable, e.g., were longer, we would have an inference problem with our data and be at risk for a Type 1 error, the type we strongly try to avoid. The inference problem should be clear: if the participants in the liking condition also had longer words, we would be able to say (a) rating liking was associated with better recognition score and (b) longer words were associated with better scores. We would have no way to tell which effect actually produced our data and this is the general problem when a confound has been detected. The real problem is that when there is an experimental confound, we learn nothing from the study. We cannot say later that the IV affected the DV, nor do we have any evidence that the IV did not affect the DV.

A properly designed study is planned around anticipated extraneous variables and carefully designed to avoid confounds. Basic approaches for this are the subject of Chapter 4. Confounds are more likely to occur in non-experimental designs where the independent variable is not under the experimenter’s control.

In experimental design, the risk of a confound often comes from practical questions associated with implementing the procedure in ways that are unexpected in planning the experimental design. As an example, consider the hypothetical scenario where after planning Experiment 1, we discovered that there was a mistake in the online protocol so that all the participants in the class received the shallow encoding instructions (count vowels). But we found another group of students unable to attend the first class who could participate in the deep encoding condition, but these were non-native English speakers who had been unable to travel to attend the first class. In this scenario, we would have accidentally created a confounded study where all the non-native English speakers were in the same condition and if their memory for English words was different, we would not be able to draw any conclusions from our data.

When planning a research study, or readying about a completed study, the standard method to try to identify potential confounds is to try to think of as many extraneous variables as possible that might affect the DV. There will generally be quite a few, but most or all of these will not vary with the IV so we do not have to worry about them reducing the internal validity of the study by creating a confound.

One very common aspect to consider as a potential extraneous variable is **individual differences** in performance. On a memory test, maybe some of the participants are just better at memorizing lists of words than others. For almost any study, individual differences will almost always be an extraneous variable, but very rarely will this be a confound for the results. The reason this is rarely a confound is the common use of a simple, but important procedure called **random assignment** to conditions.

## Random Assignment to Conditions

As long as participants are randomly assigned to conditions, individual differences should never confound the final result. It is tempting to worry that it is possible to get unlucky in our randomization and assign all the better/worse participants to the same condition. However, this is exactly what our statistical tools are designed to test. For all our statistical tools for deriving inferential statistics, the final “p value” is formally the probability that we accidentally observed the difference we did due to this random chance (under the null hypothesis that there was no effect of the IV). When we reject the null, we explicitly consider and reject the possibility that individual differences, or any other non-confounding extraneous variable accounted for our results.

It is important to note that for random assignment to work, it has to be carried out correctly and there needs to be an adequately large sample of participants recruited for the study. We will discuss sample size in the context of statistics (Chapter 5), sampling (Chapter 13) and designing research (Chapter 15). A good, simple rule-of-thumb is to try to have at least 30 participants in each of your experimental conditions, if possible. It isn’t always possible to obtain that many volunteers, however, and 15-20 per condition also often works.

Smaller sample sizes weaken the effectiveness of random assignment. In some specialized cases with restricted populations (e.g., neuropsychological studies) it is not possible to recruit large samples. In these cases, it may be necessary to use designed based on **matched participants**, where participant-based extraneous variables are assessed and explicitly balanced across the IV. This and related techniques were used in some older psychological science studies that pre-date the modern recommendations to use larger sample sizes. The challenge of matching procedures is the need to identify all possible participant-based extraneous variables and then have reliable, independent measures of all of these prior to assigning conditions. It is generally much simpler just to randomly assign a large group of participants to conditions and trust that the statistical model will account for assignment luck.

## Demand Characteristics and Bias

Random assignment, properly carried out, will prevent individual differences from confounding an experiment. However, incorrectly following the randomization procedure can lead to embedding **bias** in a study. Bias in a study creates a problem similar to a confound but which is generally smaller in effect, but often much harder to detect. An example of where sampling bias can creep into a research study is when a novel experimental procedure is being developed with a complex IV and to test the procedure, the experimenter runs the first group of volunteers all in the treatment condition. This might be done ostensibly to test the procedure to make sure it is working as intended. However, this can create an accidental bias in that the first participants to sign up for your experiment are often the most engaged and motivated participants who really want to do well on your DV measure. Their data is now disproportionately in the treatment condition. Later when you compare treatment to control, there has accidentally been a bias included where treatment is correlated with motivation, weakening the internal validity of the study in the same manner as a confounded variable.

Because of the risk of this somewhat subtle kind of bias, we strongly prefer procedures that cannot be influenced by experimenters’ expectations or desires. Virtually all researchers want their experiments to succeed, so avoid the possibility of implicitly embedding bias by removing the experimenter’s opinion when assigning participants to conditions. Standard experimental methods will have meticulously detailed protocol instructions to be followed to the letter to avoid weakening the conclusions. Cases where this was not done will be discussed in the context of Research Ethics (Chapter 8) and specifically the Responsible Conduct of Research (Chapter 19).

The expectations of the experimenter are not the only source of concern for implicit bias in carrying out our research procedures. Participants in research are often very sensitive to the **demand characteristics** of the protocol. These effects are similar in spirit to placebo effects in that these expectations affect performance on the DV. In general, if the participants in a research study are aware of the underlying hypothesis, this may influence their performance on the task that measures the dependent variable. This will confound the study and make the conclusions inaccurate.

As an example, suppose participants in our Experiment 1 knew about both conditions being studied and that we expected that rating liking would produce better memory than counting vowels. They might then try harder to remember the words if they were asked to rate liking and score better on the recognition test based on their motivation. This would have the effect of creating a confound between motivation and depth, potentially leading to a Type 1 error in conclusions.

The simplest way to avoid this problem in a two-group independent samples design is to not inform the participants about the hypothesis or the other condition of the study that they are not participating in. This is termed a **single-blind** procedure and was the way we implemented our Experiment 1 here. This is an extremely common method for designing psychological research that strengthens the internal validity of the experiment by eliminating concerns about demand characteristics. Later in Chapter 8, we will touch on the subtle ethical implications of this common approach (we prefer participants to know what they are engaging in when participating in research, yet we usually cannot explain everything in advance).

More complex procedures need to be used in experimental design when there is concern about the possibility of **experimenter bias** affecting the measurement of the dependent variable. This effect needs to be considered whenever there is a subjective element to the quantification of the DV. While many measures and scales are scored objectively (e.g., our recognition memory measure or the Self Esteem scale from Chapter 2), there are many areas of psychological study that are not as externally objective. For example, to evaluate a treatment aimed at reducing stage fright and improving stage performance, it would be necessary to quantitatively evaluate performance. Or we might need to measure an aspect of emotional expression such as laughter or quality of partner interactions in a study of relationships. For any subjective judgment, we assume that experimenters who are aware of the design and are invested in the outcome of the study are at risk for experimenter bias and should not be the source of the DV measure.

One common method for when DV requires a subjective evaluation is to use **independent raters** who provide the scores of the judgment without knowing the condition the participant was in. These raters are blind to the experimental condition, so that their rating cannot be influenced by the experimental hypothesis. The raters must generally be trained with detailed instructions on how the scoring of the DV is to be carried out. It is also common to have multiple raters and compare scores for overall consistency to establish the reliability of the procedure.

In cases where independent raters cannot be used, a **double-blind** methodology may be employed to remove experimenter bias. This approach is most commonly seen in medical research, such as pharmaceutical intervention designs aimed to test whether a new drug is effective at treating a disease. In medical research such as this, it is difficult to implement an external scoring system for a complex DV like improved health outcomes because the participants are patients under care of a physician who is often also the experimenter.

The double-blind procedure involves administering the IV in a way such that the researcher does not know which participants are in each condition. In a drug study, this is done by a pharmacy providing numerically labeled doses that are half treatment drug and half placebo. The research staff administers the drug without knowing whether the participant is receiving treatment or control so that all subsequent health measure assessments are done blind to experimental condition. At a specific later planned date, the conditions are revealed so that data analysis can be done to identify the efficacy of the drug.

A double-blind procedure is an extremely rigorous and robust procedure for assessing efficacy of interventions. It is, however, difficult to implement properly, which makes research depending on this approach slower and more expensive to carry out. Because of this, the approach is not in common use in psychological science. In most psychological research, we can identify bias-free measures for our dependent variables or implement independent-rater procedures which are much easier and simpler to deploy in practice.

## Key Takeaways

* An experiment is a type of empirical study that features the manipulation of an independent variable, the measurement of a dependent variable, and control of extraneous variables.
* Experimental research on the effectiveness of a treatment requires both a treatment condition and a control condition, which can be a no-treatment control condition, a placebo control condition, or a wait-list control condition. Experimental treatments can also be compared with the best available alternative.
* An extraneous variable is any variable other than the independent and dependent variables. A confound is an extraneous variable that varies systematically with the independent variable.
* Experiments can be conducted using either between-participants or within-subjects designs. Deciding which to use in a particular situation requires careful consideration of the pros and cons of each approach.
* Random assignment to conditions in between-participants experiments is a fundamental element of experimental research. The purpose of this technique is to control extraneous variables so that they do not become confounding variables.
* Studies are high in internal validity to the extent that the way they are conducted supports the conclusion that the independent variable caused any observed differences in the dependent variable. Experiments are generally high in internal validity because of the manipulation of the independent variable and control of extraneous variables.

## Exercises

* Practice: List five variables that can be manipulated by the researcher in an experiment. List five variables that cannot be manipulated by the researcher in an experiment.
* Practice: For each of the following topics, decide whether that topic could be studied using an experimental research design and explain why or why not.
  + Effect of parietal lobe damage on people’s ability to do basic arithmetic.
  + Effect of being clinically depressed on the number of close friendships people have.
  + Effect of group training on the social skills of teenagers with Asperger’s syndrome.
  + Effect of paying people to take an IQ test on their performance on that test.
* Discussion: Imagine that an experiment shows that participants who receive psychodynamic therapy for a dog phobia improve more than participants in a no-treatment control group. Explain a fundamental problem with this research design and at least two ways that it might be corrected.