Chapter 4: Experimental Control

The fairly well-known phenomenon of “stereotype threat” reflects a situation where exposure to some commonly-held cultural expectations about poor performance actually causes poorer performance. Methodologies for studying stereotype threat are often excellent design examples, but I have sometimes hesitated to rely on them in classroom situations because of the effects they cause. If simply repeating the expectation that some subgroup is, for example, bad at math causes poorer math performance, then even discussing the research imposes a cost on students in the class. As a general rule, stereotypes are examples of a general misunderstanding of correlation and causation as they describe current relationships that might exist in the world that lead to misleading conclusions about why (e.g., the existing relationship is thought to be intrinsic instead of reflecting cultural bias).

In spite of being a sensitive topic, research on stereotype threat provides an excellent example of unexpected extraneous variables influencing conclusions drawn from research. Stereotype threat research can be carried out with a simple design with two levels of an independent variable. In one condition, participants are exposed to the stereotype threat content and in the other condition, participants are exposed to control content that does not mention the stereotype. The dependent variable is measured performance on a related test. If performance is reliably lower after being exposed to the stereotype, then we see that the IV affected the DV and can draw the inference that stereotype threat affects performance.

However, not every study of stereotype threat produces a reliable effect of the IV on the DV, leading to questions about the robustness of the phenomena. In Aronson et al. (1999) some insight into this variability was provided in a study that examined stereotype threat on math performance but further asked participants how important math was to them. For students who self-reported that math was extremely important to their identity, stereotype threat was found to impair performance. Their study was also notable in that they used an unusual threat stimulus where white males were exposed to the stereotype that “Asians perform better on math tests,” showing that this effect also applies to non-minority participants. But for students who reported that math was not important to them, no effect of stereotype threat was found. In fact, the stereotype threat led to better performance in the low math identification group.

This study both advances the understanding of the phenomenon but also illustrates some of the challenges of extraneous variables. How important math is to the participants is a variable that might have been overlooked in previous studies. If there were many participants in a research study who did not care about math, that study would likely not have observed a reliable effect of the threat on their performance. This uncontrolled extraneous variable would lead to increased variance on the math performance measure (some participants showing the effect and some not), leading to statistical results that are not reliable. Note that the absence of a statistical effect here does not allow us to conclude that stereotype threat does not exist. This is an example of an important idea to be discussed in Chapter 5 that we cannot draw inferences from non-reliable results.

This example does illustrate the most difficult aspect of coping with extraneous variables in experimental design in that they are often not known in advance of the research. It can take a lot of experience and expertise in the specific research domain to learn where design problems might emerge from. And in cases like this, the new variable ends up significantly extending the known theory about the main constructs for the study.

## Learning Objectives

1. Managing the effect of extraneous variables on experimental measures
2. Constancy: keep as many factors as possible consistent across levels of the independent variable
3. Counterbalancing: if factors cannot be kept constant, distribute them evenly across the independent variable to avoid confounds
4. Practical concerns: defining the procedure, instructions to participant, pilot testing and manipulation checks

# Non-confounding Extraneous Variables

In Chapter 3, the problem of experimental confounds was used to illustrate the importance of planned experimental control. Once all potential confounds are eliminated from the design of a study, the next challenge is to manage the extraneous variables to reduce variance on our dependent measure. Any measure derived from human participants is going to have variance in performance associated with it. We will see the term variance used to evaluate the DV statistically. Conceptually, this variance results in part from **measurement error**, which reflects the important idea that no quantitative operational definition is ever perfect. We can also think of this variance as “noise” in that it reflects aspects of our data that we are not directly concerned with as part of our experimental hypothesis.

As we have seen previously in Chapter 3, an extraneous variable is anything that varies in the context of a study other than the independent and dependent variables. In an experiment on the effect of expressive writing on health, for example, extraneous variables would include participant variables (individual differences) such as their writing ability, their diet, and their gender. They would also include situational or task variables such as the time of day when participants write, whether they write by hand or on a computer, and the weather. Extraneous variables pose a problem because many of them are likely to have some effect on the dependent variable. For example, participants’ health will be affected by many things other than whether or not they engage in expressive writing. This influencing factor can make it difficult to separate the effect of the independent variable from the effects of the extraneous variables, which is why it is important to control extraneous variables by holding them constant.

Extraneous variables make it difficult to detect the effect of the independent variable by adding variability or “noise” to the data. Imagine a simple experiment on the effect of mood (happy vs. sad) on the number of happy childhood events people are able to recall. Participants are put into a negative or positive mood (by showing them a happy or sad video clip) and then asked to recall as many happy childhood events as they can. Even in the happy mood condition, some participants would recall fewer happy memories because they have fewer to draw on, use less effective recall strategies, or are less motivated. And even in the sad mood condition, some participants would recall more happy childhood memories because they have more happy memories to draw on, they use more effective recall strategies, or they are more motivated. If the effect of these extraneous variables was large, then the added variance in performance can make even a real effect of the manipulation difficult to detect (a Type 2 error).

# Control of Extraneous Variables

The principles for implementing best practices for reducing the effect of extraneous variables are simple in theory. Once the variables have been identified, keep as many as possible **constant** across conditions. Anything that cannot be kept constant but can be controlled, **counterbalance** across conditions so that it occurs equally often across levels of the independent variable. These two basic techniques remove the possibility of extraneous variables being confounds and maintain the internal validity of the study.

Practically speaking, implementing these aspects of experimental control can be difficult. As seen in the example with stereotype threat, it can be difficult to identify all possible extraneous variables in advance of running a study. In addition, differences in performance arising from individual differences in the participants cannot be externally controlled and can only be handled by random assignment. However, there are generally a set of variables related to the stimuli used in the experiment and testing conditions that can be managed in order to both avoid confounds and minimize noise in the DV.

## Constancy

As much as possible in any experimental design, keep things constant across the levels of the independent variable. Use a meticulously written procedure for carrying out the procedure and follow the guidelines to the letter. Keep the stimuli the same across conditions as much as possible. Avoid changing anything about the recruiting process, task instructions or context in which data collection is carried out.

It can be surprising to students who get to participant in psychological science research how explicitly detailed data collection procedures typically are. Many studies have carefully- written scripts for interacting with participants in research. This is done to keep interactions as constant as possible across conditions and also across experimenters. The importance of this level of experimental control is also seen when research is reported through the Methods section of an APA-formatted research report (Chapter 6). Many of those specific script details are included with the presentation of the experimental results so that the reader can identify key aspects of experimental control in carrying out the study.

## Counterbalancing

For any factors that cannot be kept constant, distribute how these are implemented equally across conditions. For example, if participants are being run throughout the day, collect data from both of the experimental conditions equally early and late in the day to avoid confounds due to circadian (time of day) effects. If it is necessary to have multiple experimenters, make sure they each contribute to data collection in each condition. If the stimuli are presented in different orders to participants, make sure the orders are distributed properly across the conditions of the study.

Note that counterbalancing is focused on making sure the extraneous variables do not confound the study but does not address the issue that these variables may contribute to measurement noise. That means that these factors that affect the DV importantly may increase the variance of that measure and creates the risk of a Type 2 error (where we fail to obtain reliable results even though the hypothesis was correct). We prefer to take risk of a Type 2 error over the risk of a Type 1 error where we incorrectly claim the IV affected the DV but our inference is incorrect due to a confound embedded in the design.

In some cases, we may not be able to fully control variables like the time of day the participants complete the study (e.g., online) and therefore cannot formally counterbalance to guarantee the same number of participants complete the study in the morning or evening. In those cases, we might rely on an approach more similar in spirit to random assignment to keep these variables from being confounds for the study.

## Constancy by Restricting Recruiting?

Keeping extraneous variables constant can also be applied to holding participant variables constant. For example, many studies of language limit participants to right-handed people, who generally have their language areas isolated in their left cerebral hemispheres. Left-handed people are more likely to have their language areas isolated in their right cerebral hemispheres or distributed across both hemispheres, which can change the way they process language and thereby add noise to the data.

In principle, researchers can control extraneous variables by limiting participants to one very specific category of person, such as 20-year-old, heterosexual, female, right-handed psychology majors. The obvious downside to this approach is that it would lower the external validity of the study—in particular, the extent to which the results can be generalized beyond the people actually studied. For example, it might be unclear whether results obtained with a sample of younger lesbian women would apply to older gay men. In many situations, the advantages of a diverse sample (increased external validity) outweigh the reduction in noise achieved by a homogeneous one.

Historically, a great deal of early health-based research was done with insufficient attention to maintaining appropriate diversity in participant recruiting (e.g., all participants were white males). The attempt to justify this at the time was that this reduced variability in participants, increasing the power to detect whether a health improving intervention was clinically effective. Technically this approach increased internal validity of the design while reducing the external validity of the conclusions (to be further discussed in Chapter 13). However, it should be clear that this also raises significant ethical concerns that these research studies were not being designed to provide benefit widely across the population. The tension between improving the scientific process with homogeneous recruiting samples and the ethical goal of benefiting all people will be one of our examples of “Where Ethics Gets Interesting” (Chapter 8/19).

Modern approaches to psychological science have reinforced the idea that we should avoid restricted sampling as much as possible in research. Concerns have been raised about the reliance on “WEIRD” participant samples: Western, educated, industrialized, rich and democratic. These are kinds of participant groups that are included in research that depends on university undergraduates in the USA, Canada, and Western European countries. Commonalities in social or cultural expectations in these participants may be implicitly embedded in many psychological research reports. The main implication of this idea is that there may be unknown extraneous variables that vary across social and cultural groups that affect behavior in ways we have yet to explore in research. That does not invalidate research that depends on WEIRD populations, but may affect applications of the findings to broader, more diverse populations.

One technique for increasing potential diversity of research is to use methodologies for collecting data online. Research on how online methodologies affect recruiting diversity is ongoing. Collecting data online likely improves diversity compared to WEIRD samples, but may still restricts sampling to relatively higher socio-economic status due to the need to have technological access depending on a device and internet connectivity.

## Design of Experiment 1

Our Experiment 1 reflects a handful of design decisions aimed to keep extraneous variables constant across the two conditions in the study: deep and shallow encoding. All participants rated the exact same set of 30 words, although the instructions for the rating varied as the independent variable. The words themselves were selected to be between 5 and 8 letters in length and to have a “written frequency” occurrence of 30-80 times per million. The characteristics of the words were kept similar to reduce variance in memory for the words chosen for the experiment.

Unless you have some experience in memory research using word lists, you might not have anticipated that the length or frequency of the stimulus would be important for the design. Knowing what potential extraneous variables are relevant to a specific study often requires some prior knowledge of research in that domain. Once the variables are identified, the technique for controlling them is straightforward: select words in a restricted range from a database of word frequency information.

In addition to the stimuli, note that the two scales used for rating the stimuli were also constructed to have 5 levels. Although it is unlikely that the specific number of levels on the scale will affect memory, it is good practice to keep as many design elements the same as possible across conditions.

In cases where the data collection for Experiment 1 are done in the classroom, we also gain the benefit of all the participants complete the study in the same conditions in terms of surrounding and time of day. When this experiment is completed by participants outside the classroom, there may be influences of outside distractions and attention that are outside of experimental control. Note that these would be examples of extraneous variables that increase variance, but do not confound the study because we have no reason to believe that either of the conditions of the independent variable would be more affected by distraction.

The design of Experiment 1 also includes 3 minutes of irrelevant trivia questions to be completed after performing the word rating and before the surprise recognition test. The time of the trivia task is kept constant across participants, but the number of questions answered and the content of the questions is not. The number and content of the questions experienced is allowed to vary randomly across all the participants in the study, potentially contributing to variance in the memory measure but not in a way that is confounded with the study conditions.

## Practical considerations

Best practices for controlling extraneous variable in carrying out psychological research can lead to fairly elaborate and precise procedures for research personnel. As a consequence of this, it is very common for research procedures to be evaluated with a short period of **pilot testing** before staring formal data collection. Sometimes this can mean simply practicing carrying out the research procedures under observation of other researchers to ensure it is working as intended by the planned operational definitions. It can also mean running a small preliminary sample of participants to evaluate the procedure and scripts. It should be very clear in the overall research plan when pilot testing is underway and when that process is complete and formal data collection for the planned study starts. Pilot testing data is not intended for inclusion in published research and may often depend on knowledgeable members of the research team (or collaborating teams). This can affect demand characteristics of those participants making their behavior or performance importantly different from the main intended recruited sample.

A common feature of pilot testing of procedures is to include a measure referred to as a **manipulation check**. This is a measure that will often look like a dependent variable but is not part of the research hypothesis. For example, in a mood manipulation study using music to create positive/negative moods, participants might be asked after listening to the music to rate their mood. If mood ratings were not consistent with the independent variable (music type), we would have concern about the operational definition being used. In some research publications, manipulation check data may be included and even analyzed statistically but note that no real hypothesis is being tested. A statistically reliable effect that the music manipulation affected self-rated mood only validates the operational definition of the IV and does not lead to any general conclusion.

Pilot and preliminary testing can also be used to examine the distributional characteristics of the dependent variable. As we will see in the next chapter, our ability to draw inferences from our data will depend on observing statistically reliable effects of the IV on the DV. Poorly controlled extraneous variables may lead to high levels of variability in performance, which will show up as high variance and may indicate a need to improve experimental control in design. Accurate estimates of variance often require large participant samples, though, so this cannot always be anticipated.

Pilot testing is often very useful to identify potential statistical problems with **floor effects** or **ceiling effects** in the DV. Ceiling and floor effects occur when the dependent variable measurement range is not properly anticipated in the experimental design. For example, a floor effect will occur when a task is too difficult for participants. If participants are given a problem-solving task with the intention of the measure being the number of problems solved but nobody is able to solve any of the problems, everybody will score zero regardless of the IV manipulation (no reliable difference can be detected). Similarly, if all participants get all the answers correct, performance is at ceiling for all groups and again there is no possibility of observing a statistically reliable effect. Pilot testing is often used to verify that scores on the dependent variable will be within a range that allows for detectable influence from the independent variable so that we have some chance that our statistics will be effective.

## Key Takeaways

* Extraneous variables that do not confound the study increase variance in performance (measurement error, noise).
* Constancy: as much as possible, keep things the same across levels of the independent variable
* Counterbalancing: for anything that cannot be kept constant, keep this factor from being confounded with the independent variable
* Restricted participant sampling may reduce variance (increasing constancy of participant variables) but should be used carefully due to effects on generalizability of findings
* Identifying all the possible extraneous variables is harder than controlling for ones that are known
* Rigorous systematic procedures for data collection are important and contribute to research success
* Pilot testing is the process of working out details for research procedures and often precedes formal data collection

## Exercises

TBA