Chapter 19: Quasi-Experimental Design

Quasi-experimental design are designs that blend elements of experimental and non-experimental research designs. The goal in this approach to research is to try to capture aspects of the higher external validity of non-experimental designs, but still try to support strong causal statements about the relationship between the independent variable(s) and dependent variable. We will consider two broad classes of quasi-experimental designs here: Field Research and Intervention Research.

# Field Research

The next major distinction between research methods is between laboratory and field studies. A laboratory study is a study that is conducted in the laboratory environment. In contrast, a field study is a study that is conducted in the real-world, in a natural environment. This increases external validity but at some cost to internal validity. These can be real experimental designs with manipulated independent variables. In theory this allows for stronger causal inferences than non-experimental designs. However, the lack of full control over extraneous variables makes these less reliable.

Internal validity challenges in field research come from the same sources as all other experimental research, mainly arising from accidentally confounded variables or participant selection bias. In laboratory research, extraneous variables can be identified, and an experimental procedure constructed to minimize their effects (constancy) or at least de-confound these (counter-balance). In field research, there is often a very limited ability to employ these techniques, so it is more likely that conclusions will have to be considered in the context of alternate explanations (like in non-experimental research).

Laboratory experiments typically have high internal validity. Internal validity refers to the degree to which we can confidently infer a causal relationship between variables. When we conduct an experimental study in a laboratory environment we have very high internal validity because we manipulate one variable while controlling all other outside extraneous variables. When we manipulate an independent variable and observe an effect on a dependent variable and we control for everything else so that the only difference between our experimental groups or conditions is the one manipulated variable then we can be quite confident that it is the independent variable that is causing the change in the dependent variable. In contrast, because field studies are conducted in the real-world, the experimenter typically has less control over the environment and potential extraneous variables, and this decreases internal validity, making it less appropriate to arrive at causal conclusions.

But there is typically a trade-off between internal and external validity. External validity simply refers to the degree to which we can generalize the findings to other circumstances or settings, like the real-world environment. When internal validity is high, external validity tends to be low; and when internal validity is low, external validity tends to be high. So laboratory studies are typically low in external validity, while field studies are typically high in external validity. Since field studies are conducted in the real-world environment it is far more appropriate to generalize the findings to that real-world environment than when the research is conducted in the more artificial sterile laboratory.

Finally, there are field studies which are non-experimental in nature because nothing is manipulated. But there are also field experiments where an independent variable is manipulated in a natural setting and extraneous variables are controlled. Depending on their overall quality and the level of control of extraneous variables, such field experiments can have high external and high internal validity.

## Extraneous variables in Field Research

Procedure is designed to be carried out in the real world. How are participants going to be identified to “recruit” into the study? In many cases they might not even know they are in the study, raising questions about ethics that must be addressed. Beyond ethics, the implications of the recruiting method may have important implications for the generalizability of results if there are any concerns that the recruiting method accidentally biases the sample.

Many field research procedures effectively require experimenters to become actors to carry out the procedures of the study as if it were a script. The validity of the research can end up depending on the acting ability of the experimenters and/or confederates. Field research procedures can also put a lot of pressure on observational techniques that require scoring aspects of behavior that can have a significant subjective component. Where possible, standard techniques should be used such as multiple raters and assessing consistency, or to structure the procedure so that the raters are blind to the experimental condition. These are not always possible and can weaken the internal validity of the results.

The process of “recruiting” participants can also be highly sensitive to the place, time and date of planned experimental data collection in field research. These can inadvertently incorporate bias into the recruiting procedures that needs to be considered when interpreting the results. Recruiting bias typically affects the external validity of the research with the idea participants found at a different time or place might have behaved differently. While field research is generally higher in external validity, recruiting processes can pose a challenge.

## Ethics in Field Research

Field research has several specific characteristics of ethical concerns that occur frequently. The most obvious is that data is being collected from participants without their awareness that they are in a research study. That is an elevated level of deception that is almost always inherent in this type of research that needs to be evaluated by the IRB before data collection can occur.

Field research also often involves areas of psychology that are sensitive since these are the kinds of topics for which behavior can be markedly different in the laboratory. In a laboratory setting, when participants are asked about attitudes related to stereotypes or high-risk behavior, their responses may incorporate expectations about what they believe the experimenter things is most socially appropriate. Collecting data on these ideas in the field may provide a much more direct insight into what people actually do rather than allowing them to report what they think is commonly considered acceptable. This exacerbates the issues related to deception and lack of clarity about informed consent. However, since these are often question of great scientific interest, they can certainly be judged to be ethical in that the benefit of the research outweighs the costs to the participant in risk, or embarrassment.

# Intervention research

A great deal of psychological science research aims to further our understanding of human behavior and the underlying cognitive processes. Studies that have a robust theoretical foundation and are designed to add new facts to what is known are termed “basic science” research. However, many psychological science research areas of study are aimed more directly at identifying ideas that might immediately and positively impact the world which are described as “applied research.” These ideas are not necessarily mutually exclusive, but in practice, basic science tends to be carried out within the laboratory with maximum control and the best internal validity. Applied research is typically done as field research where there is a controlled independent variable that involves the intervention.

In medical research, studies that aim to establish the effectiveness of a new drug are done as intervention studies. These follow many of the same methodological issues, but often do not involve psychological measures. In addition, many of these studies are done in the context of very specific methodological practices related to ethics and rigor (e.g., double-blind methods) that not always available in psychological studies. Research on improvements in educational practice captures more of the challenges of psychological research as there are important questions about efficacy of interventions that we want to observe in the context they would be used, e.g., a classroom. But in this work, it is challenging to implement all the preferred aspects of experimental control that we would want in laboratory research.

## Basic Intervention Designs

In a one-group posttest only design, a treatment is implemented (or an independent variable is manipulated) and then a dependent variable is measured once after the treatment is implemented. Imagine, for example, a researcher who is interested in the effectiveness of an anti-drug education program on elementary school students’ attitudes toward illegal drugs. The researcher could implement the anti-drug program, and then immediately after the program ends, the researcher could measure students’ attitudes toward illegal drugs.

This is the weakest type of quasi-experimental intervention design. A major limitation to this design is the lack of a control or comparison group. There is no way to determine what the attitudes of these students would have been if they hadn’t completed the anti-drug program. Despite this major limitation, results from this design are frequently reported in the media and are often misinterpreted by the general population. For instance, advertisers might claim that 80% of women noticed their skin looked bright after using Brand X cleanser for a month. If there is no comparison group, then this statistic means little to nothing.

In a one-group **pretest-posttest design**, the dependent variable is measured once before the treatment is implemented and once after it is implemented. Let’s return to the example of a researcher who is interested in the effectiveness of an anti-drug education program on elementary school students’ attitudes toward illegal drugs. The researcher could measure the attitudes of students at a particular elementary school during one week, implement the anti-drug program during the next week, and finally, measure their attitudes again the following week. The pretest-posttest design is much like a within-subjects experiment in which each participant is tested first under the control condition and then under the treatment condition. It is unlike a within-subjects experiment, however, in that the order of conditions is not counterbalanced because it typically is not possible for a participant to be tested in the treatment condition first and then in an “untreated” control condition. This is basically a non-experimental approach but gets used fairly often because of the ease of implementation.

A common approach to ruling out the threats to internal validity in one-group designs is by designing the research protocol to include a control group, one that does not receive the treatment effect. A control group would be subject to the same threats from history, maturation, testing, instrumentation, regression to the mean, and spontaneous remission and so would allow the researcher to measure the actual effect of the treatment (if any). Of course, including a control group would mean that this is no longer a one-group design.

**Control intervention designs**. Can be combined with a pre/post design. Intervention is compared with a control intervention. It can be difficulty to identify an appropriate control task. The best control tasks manage demand characteristics for the participants. For example, a study on improving educational outcomes using intense extra training with personal tutors is going to have the issue that participants receiving tutoring are aware that something is being done to help them. If the control condition is simply additionally assigned, optional reading, the control condition may not respond to the post-test with the same expectations. However, a really well-matched control condition can be difficult to construct in many cases. As a result, it is not that uncommon to use the simplest alternative and simply not intervene. This is sometimes called a ‘wait-list’ control condition to reflect the fact that recruiting is done among people who are seeking the benefit of the intervention, who are then randomly assigned to receive the intervention or not. Those not receiving the intervention are designated the waitlist control condition.

**Phased designs**. You might notice that there is an ethical issue with recruiting people seeking an intervention and then randomly assigning them to the waitlist control condition. A method for mitigating this problem is to use a phased design such as an AB or ABA design. In these designs, interventions alternate with baseline control conditions. The advantage of this approach is that the comparison group can alternate with the opposite conditions. The simplest design of this type is to use a waitlist control condition in the first phase and then in the second phase, apply the intervention to the control group. Contrasts on the DV at the end of the first phase are used as the core statistical test, but to manage ethical concerns the control group then obtains the potential benefit of the intervention in the second phase.

**Crossover Interaction Design**. A more complex variation of the phased designs is to combine pre/post testing and multiple intervention phases. In this design, all participants receive a pre-test at recruitment. They are then randomly assigned to intervention or control condition in the first phase. Then everybody receives a post-test. Then the intervention/control conditions are swapped in the second phase, which is then followed by a second post-test. While this is potentially complex and time-consuming to implement, a series of potentially interesting, related hypothesis are tested.

At the first post-test, differences in performance or differences in improvement from the pre-test across groups provide a measure of the intervention effectiveness. This is usually the primary goal of the research study. In addition, performance from the first post-test to the second for the intervention group provides a measure of whether the benefit of the intervention fades over time. This is very commonly an important question in interventions, not only whether they help but whether the benefits persist. Performance of the group that gets the intervention on the second phase can be compared as a pre/post design that effectively replicates the pre/post design of the early intervention group. This embedded replication, if successful, helps increase confidence in the efficacy of the intervention. And since both groups eventually get exposure to the intervention, ethical issues caused by random assignment to conditions are mitigated.

Some of the complexities of intervention design are best realized by considering a specific example of practical research questions and how they might be addressed.

## Example 1: Laptops in classrooms

Many teachers have been concerned that the use of laptop computers in classroom environments are not conducive to effective learning. In theory, laptops allow for more effective notetaking, which should theoretically enhance learning. In practice, concerns have been raised that other processes associated with laptops in class are overwhelming the additional value of better notes. Two core hypotheses about why laptops might reduce classroom learning have been raised. The first is that enhancing the easy of notetaking is counterintuitively less helpful for learning. Laboratory studies based on the idea of “desirable difficulty” have suggested that extra attention and effort during learning might lead to better memory representation formation reflecting better retention of class-presented information. Under this hypothesis, facile notetaking leads to less attention and engagement with class material, and it is more easily and quickly forgotten.

A second hypothesis is that laptops carry a high risk of distraction from class activities by allowing for other type of information to be engaged with that reduces attention to class-presented material. Students may engage with non-class activities via social media or other content via laptops in class, leading to poorer learning. In laboratory studies, the phenomenon of “dual tasking” where participants are asked to try to pay attention to two differing tasks simultaneously unsurprisingly leads to poorer performance.

The laboratory, or basic science, studies of learning and memory provide several hypotheses about what the effect of notetaking on laptop might be in class, but it is clear that the actual impact would require field studies. In addition, there are many extraneous variables to consider in these kinds of field studies. Students in high school may exhibit different effects of laptops than college students. These effects may also be very different in workshop-based training environments for corporate training events. The quality and style of the teacher in the classroom may have large effects or interaction with other characteristics of the environment to modify the effect of using a laptop during class. Any non-experimental field research also encounters a series of validity challenges related to variety in existing conditions about when, why and how laptops might be currently used in the classroom.

An intervention research study would be aimed at drawing a causal conclusion about the laptop effect by random assignment of laptop-use or not during class. This is basically a simple two-group design, but there are a lot of challenging decisions required to come up with the operational decisions and implement a procedure. First, the definition of what is laptop-use in class needs to be precisely described. With resources, identical laptops could be provided to all students in one of the conditions. This would provide for better control over the extraneous variable of differences in laptop type or quality, but at the same time would potentially reduce some of the external validity of the study. One goal of field research is to support a claim about a practical effect under typical conditions, so allowing students to use their own might be a better measure of real consequences.

Operationally defining the non-laptop condition is also challenging. If students to be disallowed from laptop use in class, can this be enforced for all students? If there was, for example, a grade penalty for violating the rule, this would itself potentially lead to grade outcome differences across the class, reducing the internal validity of the study. Inability to use laptops could also influence students’ enrollment in the class, which could distort results due to recruiting bias. Since research participation is ethically required to be voluntary, differential recruiting across conditions is a major risk for an intervention study. In addition, exceptions for cases for students with special needs could impact the ability to randomly assign conditions.

There is also a crucial ethical issue related to using class performance as the dependent variable. Randomly assigning participants to a condition thought to lead to worse performance is a risk of harm that we would normally design research to avoid. In this particular case, if the standard expectation of students is to use laptops, we might be able to motivate the research ethically by the need to demonstrate that constraining laptop use is of value to students. Without testing this intervention, it would not be possible to determine if a policy recommending not using laptops in class is helpful. If it turned out that avoiding laptop use impaired student performance, it would be unethical to impose that constraint broadly. Differing expectations of students or teachers about the impact of laptop use in class might also influence whether it was practically possible to mix conditions within a single class or be necessary to assign entire classes to use/non-use.

Because of these ethical issues, it might be preferable to design this kind of intervention using a small number of students or classes. This, however, then highlights the potential problem of differing content across classes, or differences in educational style or presentation of teachers. A single instructor with multiple class sections would be most effective, but this then weakens the external validity of the results by potentially limiting them to that instructor and that class material. The question of whether the instruction can proceed in a completely unbiased way across these conditions is also important.

The main point of this example is to consider what seems like a fairly simple and straightforward two-group experimental design and realize that as an intervention study, it is exceptionally difficulty to carry out a study with a very high level of internal validity. Intervention-based research is a kind of field research that generally has to attempt to work around a large number of challenges that affect what inferences can be drawn. The elements that can not be approached with the best experimental design principles do not mean that valid inferences can not be drawn, but that our critical evaluation skills for identifying alternate explanations need to be applied to these studies. At the same time, a well-constructed and well-executed intervention design should be appreciated for the methodological care and effort that needs to be invested to make this type of research effective.

## Longitudinal Research

Intervention research is often done over a more extended time period that our simpler experimental designs run in an hour or two in a laboratory. This feature requires consideration of some history effects like those that apply to within-participant designs. One such effect is the phenomenon of **regression to the mean**, which is a natural consequence of variability in human performance. A well-known example was reported in discussions with instructors who provided feedback to pilots who felt that it was most important to harshly criticize poor performance and ineffective to compliment good performance. If some of the daily variation in performance is simply due to random variance, today’s poor performers are likely to improve to perform near the mean on the next session. At the same time, better performers will fall back to their average performance. The tendency for performance to center around a mean level plus variability leads to the phenomenon of regression to the mean. In this case, today’s top performers are most likely to do nominally worse next time, and the bottom performers ought to do better even without a robust learning effect. This effect can look a lot like a response to feedback, as in this example, but with the data described here, it is impossible to tell which account for the finding.

A related idea in clinical or medical intervention research is the idea of **spontaneous remission**, which is the rate of improvement that is not influenced by an intervention. This idea is another rationale for placebo control groups. Participants in the control group may improve due to demand characteristics (knowing they are in a study) or may simply improve due to other effects related to the passage of time. Intervention research aimed at treatment very rarely relies on one-group designs due to the risk of improvement being due to effects unrelated to the intervention.

## Mortality Effects

The term **mortality effect** refers to a challenge to the internal validity of a study based on participants dropping out of the study. In this context mortality often does not refer to death of participants but can simply reflect a decision not to continue with participation in the research project. Since all participation in research should be voluntary, if one of the conditions is seen as unpleasantly onerous and not worth the time required or compensation offered, participants have the right to opt out of the study. If one of the experimental conditions experiences a higher rate of dropout, then this can bias the subsequent comparison between groups.

To illustrate this problem, consider a simple research design comparing the effect of reward and punishment on learning. Participants in the reward condition are asked to do a moderately challenging perceptual task and are rewarded with 50 cents for each correct response. A separate group of participants attempts to learn the same task but are punished with a mild electric shock for each incorrect response. If the study runs over several sessions and several days, a very plausible outcome is that the participants receiving shocks will exhibit higher levels of performance. However, without detailed information about participant dropout, we can draw no conclusion about the relative value of reward versus punishment. It is very likely in this simple design that participants in the punishment group who were poor at learning the task would drop out of the study. If more of the poor learners drop out of the punishment than the reward group, any differences at the end of the study could be entirely due to **differential mortality** across groups.

This effect does not only happen in intervention research but can occur with any longitudinal design where data is collected across many days, sessions or even weeks or years. Research using this approach will generally need to report detailed information on participant retention across the study and compare this across all independent variables to try to rule out this problem. Intervention research with a challenging intervention can be weakened by participants opting out of the study because the intervention is unpleasant (e.g., diet or exercise interventions). Even participants in the control condition can opt out of participation if the recruitment sought participants with a condition to be treated but they are temporarily in a control or placebo condition. In some designs, recruiting and compensation practices are design explicitly with reducing participant dropout by including compensation bonuses for completing the study. These techniques are always considered carefully by the IRB for any accidentally coercion to complete the study that they might imply.

Mortality effects in research are closely relate to the idea of **survivorship bias** in which inferences about data available at the end of a study or other ongoing process can be distorted by the missing data unavailable at the end. The most famous example of this was an analysis of military airplane damage done in World War II on planes that returned from combat. The key inference was that the undamaged parts of the planes that returned were likely the most vulnerable and critical regions based on the inference that the planes that had been hit in those regions failed to return. This idea has been raised as an issue in highly competitive fields (business, athletics, acting) where studying only people who have succeeded may not reveal critical differences that actually predict success for which information from less successful people needs to be included, essentially a sampling issue in the language of experimental research.

## Maintaining Good Scientific Practice

Intervention research is also an area where challenges can arise related to **conflict of interest** and **experimenter bias** and can make maintenance of best rigor practices difficult. Because intervention research is challenging to implement, it is generally done with a strong hypothesis and expectation about the outcome by the research staff. Research procedures should be designed with attention to the possibility of bias in scoring the dependent variable or any aspect of implementing the conditions manipulated for the independent variable(s). In addition, large scale intervention research is resource intensive to do effectively, often meaning that a funding agency is supporting the research, usually with the idea of wanting to find effective interventions to address social problems. This can create the appearance of conflict of interest in that successful research can have fairly immediate direct benefits for researchers.

The existence of a potential conflict of interest does not mean that research cannot be carried out with a high degree of rigor and integrity. Standard practice is to disclose all information related to possible conflicts of interest so that reports of results can be interpreted in that context. Results in this context may be considered with a heightened degree of skepticism, but should not be treated as if they cannot be accurate. As with all research, confidence in conclusions accrues over time, replication and extension to novel operational definitions, samples and research contexts.

## 

## Key Takeaways

* Quasi-experimental research involves the manipulation of an independent variable without the random assignment of participants to conditions or counterbalancing of orders of conditions.
* There are three types of quasi-experimental designs that are within-subjects in nature. These are the one-group posttest only design, the one-group pretest-posttest design, and the interrupted time-series design.
* There are five types of quasi-experimental designs that are between-subjects in nature. These are the posttest only design with nonequivalent groups, the pretest-posttest design with nonequivalent groups, the interrupted time-series design with nonequivalent groups, the pretest-posttest design with switching replication, and the switching replication with treatment removal design.
* Quasi-experimental research eliminates the directionality problem because it involves the manipulation of the independent variable. However, it does not eliminate the problem of confounding variables, because it does not involve random assignment to conditions or counterbalancing. For these reasons, quasi-experimental research is generally higher in internal validity than non-experimental studies but lower than true experiments.
* Of all of the quasi-experimental designs, those that include a switching replication are highest in internal validity.

## Exercises

* Practice: Imagine that two professors decide to test the effect of giving daily quizzes on student performance in a statistics course. They decide that Professor A will give quizzes but Professor B will not. They will then compare the performance of students in their two sections on a common final exam. List five other variables that might differ between the two sections that could affect the results.
* Discussion: Imagine that a group of obese children is recruited for a study in which their weight is measured, then they participate for 3 months in a program that encourages them to be more active, and finally their weight is measured again. Explain how each of the following might affect the results:
  + regression to the mean
  + spontaneous remission
  + history
  + maturation