Quasi-Experimental Designs

An *experimental* design is one in which participants are *randomly assigned* to levels of the independent variable. As we saw in our discussion of random assignment, experimental designs are preferred when the goal is to make cause-and-effect conclusions because they reduce the risk that the results could be due to a confounding variable. However, there are some circumstances in which it is not practical or ethical to randomly assign participants. We cannot randomly assign participants to be male or female, and we should not randomly assign pregnant women to consume different quantities of alcohol. Random assignment is not an option for these variables, but we still want to make cause-and-effect inferences about gender and about alcohol. Is there anything we can do? Yes. Quasi-experiments are designed to maximize internal validity (confidence in cause-and-effect conclusions) despite being unable to randomly assign. At the outset, it is important to know that quasi-experiments tend to have lower internal validity than true experiments. In this reading, we'll discuss five quasi-experimental approaches: 1) matching, 2) mixed designs, 3) single-subject designs, and 4) developmental designs.

The major threats to quasi-experimental designs are confounding variables: variables other than the independent variable that (a) tend to co-vary with the independent variable and (b) are plausible causes of the dependent variable. Quasi-experiments are designed to reduce confounding variables as much as possible, given that random assignment is not available. For example, say that you are investigating the effects of bedtime stories on children's reading performance. Your hypothesis is that children whose parents read to them will score higher on a reading test than children whose parents do not read to them. Since many parents feel strongly about reading to their children and are unlikely to give it up, random assignment will not be appropriate for this study. What confounding variables are likely in this situation? In other words, how might children who receive bedtime stories be different from children who do not receive bedtime stories? Perhaps their parents have a greater love of reading themselves, and either through genetics or environment the parents are influencing their children to read. Perhaps children who do not receive bedtime stories tend to have parents who work longer hours and have less time to spend with their children. Either of these factors – parents' love of reading or total contact time with parents – could be a confounding variable in the relation between bedtime stories and reading performance. How do you rule those out? The first approach we will look at is *matching*.

Matching

Let's say that our two biggest concerns with the bedtime story study mentioned in the last paragraph are parent reading and parent contact time. Matching begins by measuring these potential confounding variables, along with the independent and dependent variables. Thus, all the participants in our study would be measured on four variables: 1) whether they receive bedtime stories (our independent variable), 2) how much their parents read (one of the potential confounding variables), 3) average number of parent contact hours per day with the child (the other potential confounding variable), and 4) reading performance (the dependent variable). Perhaps the most common matching procedure is **pairwise matching**, in which you identify a pair of participants – one who gets a bedtime story and one who does not – who are as similar as possible on the two potential confounding variables. Ideally, each pair would be exactly the same on each matching variable, but as the number of matched variables increases, this becomes less likely[1]. Once you identify your matched pairs, you conduct a paired-samples

1 It turns out that pairwise matching is relatively inefficient because it throws out a lot of perfectly good data. Much better is a method called *full matching* that makes it possible for a person in one condition to be matched to several similar people in the other condition. See Hansen and Klopfer (2006) and the *optmatch* package in R.

t-test (or a within-subjects ANOVA) on the reading performance scores. If the reading scores differ significantly, you can be confident that the difference is not due to either parent reading or parent contact hours because each pair was matched on those variables. Those two potential confounding variables have been "controlled" through matching.

Matching is superior to random assignment in one way: it guarantees that the participants in different conditions are not different on the matched variables. The problem with matching is that *it does nothing to equalize the conditions on other variables*. For example, the availability of children's books in the home may be an important cause of children's reading performance, but because it was not measured or matched, it will remain a potential confounding variable.

Mixed Designs

Mixed designs are designs with at least one between-subjects variable (such as "treatment condition") and at least one within-subjects variable (such as "time: before and after"). We've already discussed mixed designs as a way to combine the advantages of between- and within-subjects variables, but it turns out that they are especially useful when random assignment is not available. Consider a study on the effect of television exposure on children's aggressiveness. Randomly assigning children to watch different amounts of television is problematic both because it is intrusive into family life and because it may be unethical if you suspect that television makes children aggressive. For these reasons, most studies on the relation between television exposure and aggression are correlational: you measure children's level of television exposure as it is and do not manipulate it. But these studies suffer from a problem with inferring cause and effect. If you find that exposure to television is strongly correlated with aggressiveness, it could be because 1) television makes children aggressive, 2) aggressive children have a greater preference for watching television, or 3) some third factor, like amount of parent supervision, is actually causing both television exposure and child aggressiveness. To solve this problem, Joy, Kimball, and Zabrack (1986) took advantage of an unusual circumstance: television coming to a small Canadian town. Joy and colleagues used a mixed design in which the within-subjects variable was time (before and then two years after television arrived) and the main between-subjects variable was town (the town that had no television – nicknamed Notel – as well as two other towns that were chosen to be very similar to Notel). Thus, their design was a 2 (time: before and after) by 3 (town: Notel and the two other towns) mixed design with repeated measures on the first factor. The researchers measured aggressive acts per minute during free play on the playground in each town, and the results are plotted in Figure 1 below.

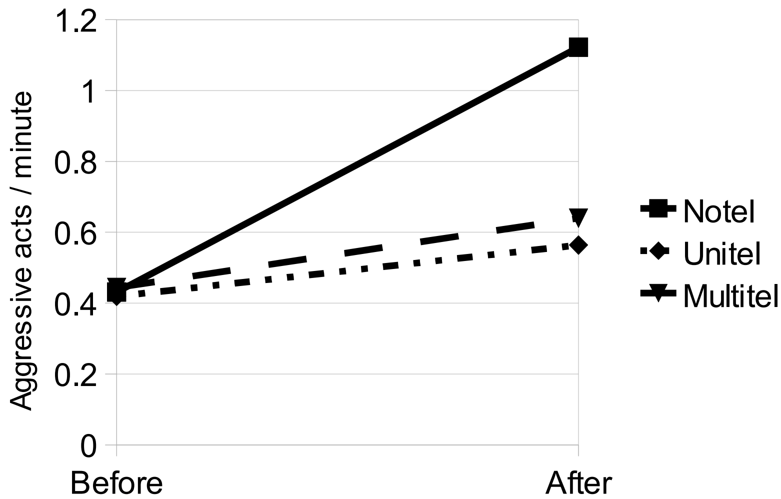

Figure 1. Number of aggressive acts per minute in three towns both before and two years after Notel received television. (adapted from Joy, Kimball, and Zabrack (1986) Table 7.A2 on page 341).

Figure 1 shows that before Notel received television, its rates of children's aggression were no different from the rates in towns that had television[2]. But two years later, children's aggression scores were significantly higher in the town that received television than in the other two towns.

To understand how a mixed design contributed to this study, first consider the problems that the researchers would have faced if they had not included any comparison towns. The results would show that the single town Notel experienced an increase in rates of children's aggression from before television to after, but that could have been due to the children getting older (a *maturation* effect) or the culture shifting to become more aggressive (perhaps because of a terrorist attack, a *history* effect). The fact that the two comparison towns did not increase as much as Notel enables the researchers to rule out history and maturation effects. If those effects had been operating, aggression rates in all three towns would have increased.

Now consider the problems the researchers would have faced if they had only measured aggression before or after television came to Notel. If they had measured aggression before, they would have concluded that television has no effects: the three towns had equal levels of aggression at the first time point. If they had only measured aggression after television came, it would be unclear whether the higher levels of aggression in Notel were due to its recent change in television or to some pre-existing difference between it and the other two towns (a selection effect). By showing that aggression increased only in the town that received television and not in the other towns, the researchers have more confidence that the change in Notel was due to the advent of television and not to selection, maturation, or history effects.

Single-Subject Designs

If you are a therapist or coach and are working with a single individual to improve their condition, how might you test whether your interventions are having an effect? If you observe a

2 It is a problem for the researchers that the three towns showed no differences at the beginning of the study. If television exposure causes aggression, Notel should have had lower aggression levels at the start. Still, the increase in aggression levels only for Notel is evidence for an effect of television on aggression.

change from time 1 to time 2, it might not be due to your intervention. It could be due to any of the potential threats to validity from a within-subjects design: history effects, maturation effects, testing effects, instrument decay, regression to the mean, etc. Perhaps the person would have improved even if you had done nothing. How could you know?

A popular solution to this problem is to use an "ABA" design, sometimes called a **reversal design**. The "ABA" refers to a sequence in which the participant is measured without treatment (A), then measured again after treatment (B), and measured a third time when treatment has been withdrawn (A). For example, let's say that an elementary school teacher has a student who is acting out and disrupting class. She formulates a plan to change his behavior by rewarding him when he works on an assignment and ignoring him when he yells out answers without being called on. To test whether the plan is working, she should begin by measuring his behavior every day for 2 weeks before she begins the plan (A). The next two weeks, she will measure his behavior every day while she implements her feedback plan (B). For the final two weeks, she will stop using her plan and go back to her old way of interacting with the student, measuring his behavior every day (A). If the feedback plan was effective, then we should see the student's behavior improve from the first segment (A) to the second (B) and then deteriorate again when the plan is withdrawn (A). This pattern would increase our confidence that changes in the student's behavior were due to the feedback plan and not to maturation, history effects, or any of the other threats to validity for within-subjects designs.

Of course, the ABA design is not suited to many situations. A depressed client receiving antidepressants could not very well be taken off his medication without considerable risk to his safety, and recipients of artificial joints could hardly be asked to have the old joints put back so that some threats to validity could be ruled out. Still, when it can be used, the ABA design can help to improve confidence in cause-and-effect conclusions when you are working with a single subject.

Developmental Designs

In developmental research, the main independent variable is typically age. Developmental researchers are interested in how some behavior of interest (e.g., friendship, memory, reasoning) changes across different ages. There are four major research designs used to study developmental changes, and I present them below in increasing order of validity.

Retrospective design. In a retrospective design, participants are asked to retrospect (literally, to "look back") and try to remember what they were like at an earlier time point. For example, Hazan and Shaver (1987) asked participants to choose which one of three descriptions best captured the style of attachment they had with their parents while they were growing up. They also asked participants to rate their most important romantic relationship on several dimensions. The researchers found that people who reported secure relationships with parents when they were children tended to rate their romantic relationship as involving more trust and less fear than participants who classified their relationship with their parents as insecure. The advantage of retrospective designs is that, compared to other developmental designs, they are much easier to do. The problem with retrospective designs is that our memories are not perfect. In fact, evidence suggests that people are motivated to reconstruct the past in a way that makes it consistent with the present (Bem & McConnell, 1970). For that reason, evidence from retrospective studies should be followed up with one of the designs described below.

Cross-sectional design. In a cross-sectional design, researchers collect data at a single point in time from participants of different ages. For example, researchers might hypothesize that people become more traditional in their attitudes and more resistant to social change as they get older. To study this, they might get participants in their 20s, 40s, and 60s to complete a measure of traditionalism and then test whether there is a positive correlation between age and traditionalism. Although cross-sectional designs do eliminate the problem of biases in memory that is intrinsic to retrospective designs, they face another problem: cohort effects. A **cohort** is a group of people with a common history. For example, first-year college students are a cohort because, in general, they will face similar challenges as they move through their college career. A **cohort effect** occurs when the difference between people of different ages is due not to age itself (i.e., the process of getting older), but is instead due to differences in the histories of the groups. For example, if you found that people in their 60s were more traditional than people in their 20s, it could be because the 60-year-olds grew up in a culture that valued tradition more that the culture that 20-year-olds grew up in. One way to think about cohort effects is that they are a type of **history effect**, in which a change over time is due to some outside event. In this case, the outside event would be the distinctive historical events experienced by each age group in your study (e.g., World War II versus the Vietnam War).

Longitudinal design. In a longitudinal design, the same people are measured at different ages. For example, Lewis Terman identified a sample of gifted children in the 1920s and measured them periodically for several decades. By following one group of same-aged people over time, you eliminate the possibility that the changes you observe across age are due to cohort effects. However, there is still a possibility that the changes you observe are due not to age itself but to changes in the culture. For example, if you measured the conservatism of people when they were 20 years old in 1960 and found that they had become more conservative by 1980, when they were 40, it would be unclear whether the change in conservatism was due to aging 20 years or due to an increase in conservatism from 1960 to 1980. To rule out this possibility, a cross-sequential design is often used.

Cross-sequential design. A cross-sequential design is a combination of cross-sectional and longitudinal designs. At the first time point, groups of people from several different ages are measured. If the design were to stop there, it would be a simple cross-sectional design, but these groups are then followed over time. Hagenaars and Cobben (1978, as cited in Rosenthal & Rosnow, 2005) used a cross-sequential design to examine patterns of religious non-affiliation across age. Table 1 presents the percentage of a sample of women in the Netherlands who reported no religious affiliation.

Table 1.

Percentages of Women in the Netherlands with No Religious Affiliation According to Age and Time Period

	Period 1 (1909)	Period 2 (1929)	Period 3 (1949)	Period 4 (1969)
Age 20-30	Cohort 4 4.8%	Cohort 5 13.9%	Cohort 6 17.4%	Cohort 7 23.9%
Age 40-50	Cohort 3 3.1%	Cohort 4 11.9%	Cohort 5 17.2%	Cohort 6 22.0%
Age 60-70	Cohort 2 1.9%	Cohort 3 6.7%	Cohort 4 11.9%	Cohort 5 19.4%
Age 80+	Cohort 1 1.2%	Cohort 2 3.8%	Cohort 3 6.6%	Cohort 4 12.2%

The first data collection was conducted in 1909, when religious affiliation was measured in four age groups. The researchers collected data every 20 years through 1969. The researchers were interested in whether women tend to become more or less religiously affiliated as they get older. If they had conducted a cross-sectional study in 1969 (see the Period 4 column in Table 1), they would have found that the highest rates of non-affiliation were in their 20-year-old sample and the lowest rates were in their 80-year-old sample, suggesting that non-affiliation declines with age. But by following the cohorts over time (best illustrated by Cohort 4 in Table 1, but a similar pattern emerges for the other cohorts), we see that a more appropriate conclusion is that women in the Netherlands tend to become *less* religiously affiliated as they get older. The reason a cross-sectional approach in 1969 would lead to inaccurate conclusions is that there are cohort effects: each cohort has a different starting level of non-affiliation. Compare 20-year-olds in 1909 to 20-year-olds in 1929, 1949, and 1969 to see how the culture has changed to favor non-affiliation.

References

Bem, D., & McConnell, H. (1970). Testing the self-perception explanation of dissonance phenomena: On the salience of premanipulation attitudes. *Journal of Personality and Social Psychology, 14*(1), 23-31. doi:10.1037/h0020916.

Hazan, C., & Shaver, P. (1987). Romantic love conceptualized as an attachment process. *Journal of Personality and Social Psychology, 52*(3), 511-524. doi:10.1037/0022-3514.52.3.511.

Hansen, B., & Klopfer, S. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics, 15*(3), 609-627.