Threats to Internal Validity for Within-subjects Designs

In a within-subjects design, each participant is in more than one (and usually all) of the levels of an independent variable. Within-subjects designs have more statistical power than between-subjects designs, but there are a number of potential threats to their internal validity. Many threats come from the fact that subjects cannot be in every condition at exactly the same time. Instead, they must proceed through the conditions of a within-subjects study in a particular order. Any factor that is confounded with that order is a potential confounding variable.

1. **Maturation effects**

Imagine a study on the effectiveness of a new cold remedy that recruited participants with colds, measured their symptoms, gave them the remedy, and then measured their symptoms again. If the symptoms showed a significant decrease, could you conclude that the cold remedy worked? No, because they might have gotten over the colds at exactly the same rate without the treatment because their immune systems responded. The cold remedy example belongs to the first category of threats, maturation effects.

A maturation effect occurs when changes in a score over time are due to naturally-occurring internal processes (e.g., immune response, cognitive development) rather than to the independent variable in a study. Maturation effects may be relatively fast, as in an immune response to colds, or relatively slow, such as the cognitive changes from age 6 to age 7. In both cases, the independent variable is time, and a potential confounding variable is maturation. Imagine a teacher of first-graders who wants to test the effectiveness of a new math curriculum. The teacher gives a test at the beginning of the year, uses the curriculum, and gives the test again at the end of the year. The teacher wants to attribute improvements to the curriculum, but they might simply be due to intellectual maturation that would have occurred even if the children had been kept in boxes.

One procedure to detect maturation effects is to add a **control group** to the study. The teacher in the above example could give the test to two classes, one that uses the new curriculum and a control group that uses the old curriculum. The researcher investigating a cold remedy could measure the symptoms of two groups: one that gets the remedy and a control group that doesn't. In both situations, if the treatment group (cold remedy, new curriculum) improves the same amount as the control group (no cold remedy, old curriculum), then the researcher should *not* conclude that the independent variable is causing the change. Instead, the changes may be due to maturation. However, if the experimental group improves *more* than the control group, then the researcher can be more confident that the independent variable is having an effect. For example, if both classrooms of students improve on a math test, but the control class improves 10 points while the new curriculum class improves 20 points, then the new curriculum is probably doing something.

2. **History effects**

Whereas maturation effects involve an internal process, history effects involve an external event that occurs between the two measurements. For example, consider a researcher testing whether a particular medication increases anxiety. The researcher measures New York City residents' anxiety on September 5, 2001, gives them the medication, and measures them again 20 days later. Scores are likely to have increased because of the terrorist attacks on September 11th – an external event. In addition to well-publicized national events, history effects can include subtle factors such as changes in the weather (for example, improving mood

because people are outside more) or changes in public policy (for example, increasing stress because of changes to bankruptcy laws).

As with maturation effects, one way to identify (and hopefully rule out) history effects is to add a control group. If you see the same change in both the control and experimental groups, then the change is not due to the independent variable but rather to history or maturation effects. Only when the change is different in the control and experimental groups can you conclude that the independent variable is having an effect.

3. **Testing effects**

A testing effect occurs when being tested (or measured) in one condition influences responses in later conditions. The most typical example of testing effects is a **practice effect**, where performance at post-test is higher than at pre-test simply because the participant is more experienced with the test. Practice effects can be reduced by using a different form of a test at post-test, but some improvement may occur anyway simply because participants have become more familiar with the testing procedure. A more subtle effect can occur when the pretest sensitizes participants to a topic, leading them to change their beliefs or behavior. For example, a survey on health behaviors could lead participants to reflect on their eating and exercising habits, which could lead to healthier behaviors simply because they filled out the pre-test and not because of any exercise-related intervention scheduled between pre-test and post-test.

In addition to using alternate forms of the same test, another way to reduce testing effects is **counter-balancing**, in which an equal number of participants receives each possible order of conditions. For example, if you are testing whether people have better recall when listening to classical music than when listening to rock music, and you give them similar memory tasks in each condition, you may be concerned that their scores will improve from the first to the second condition because they have practiced the memory task. If rock music always came first and classical music always came second, practice effects would be a confounding variable, but if music condition was counterbalanced, practice effects are no longer a confounding variable. After counter-balancing, you may still find higher scores in the second condition, but because that second condition is rock half the time and classical half the time, the increase will no longer be a confounding variable.

4. **Instrument decay**

Instrument decay occurs when the standards of measurement change over time. The term “instrument decay” calls to mind a failing mechanical measurement device. For example, imagine that you are using a spring scale to measure sacks of flour. In the morning, the spring scale indicates that each sack weighs 20 pounds, but by late afternoon, each sack appears to weigh 22 pounds. One reason this might happen is that the spring in the spring scale has become stretched out, leading to heavier estimates for the same sack of flour. The instrument (the spring scale) has “decayed”, leading to an apparent change of weight over time that is misleading. Although instrument decay does apply to changes in mechanical instruments, it also applies to human judges who are making measurements. For example, imagine a within-subjects experiment in which a teacher is testing whether oral presentations improve with practice. Each student gives a presentation, which she grades, and then gives a second presentation, which she also grades. The teacher is the human judge making the measurement, and the independent variable in this study is time: first or second. Her hypothesis is that the second presentations will receive higher scores than the first presentations. Assuming that there is a significant increase in grades from first to second

presentation, how might instrument decay have contributed? One way would be for the teacher to have used more lenient standards when grading the second presentations.

To reduce the possibility of instrument decay, the researcher could **counterbalance** the order of presentations during scoring. She cannot counterbalance the order of the actual presentations, because the first presentation is first by definition, but she could counterbalance the order in which she grades the presentations if she recorded all the presentations on video. She would then counterbalance by grading them out of sequence, sometimes viewing the second presentation before the first presentation. By doing this, she has made herself blind to condition (time: first or second).

Look for instrument decay any time that a human judge or measurement device could show a consistent shift in standards over time, such as becoming more strict or more lenient.

5. **Fatigue effects**

Fatigue effects can occur when the conditions in a within-subjects study go on for so long that subjects begin to feel tired and perform more poorly. For example, in a study on the effects of font on readability, subjects may each view 10 different reading passages, each one presented in a different font. By the 9th or 10th passage, the subjects may respond more slowly not because of the font but because they are tired. If the passages are always presented in the same order (e.g., Arial, Courier, … Times Roman), then the researcher may mistakenly conclude that Times Roman is less readable because subjects performed more poorly. This is an example of a **bias** in the study because the design of the study is leading to a consistent change in scores: scores in later conditions are lower.

To remove this bias, the researcher can **counterbalance** the order of conditions. If every subjects gets a different order of fonts, then there can be no consistent effect of fatigue on particular conditions. Fatigue may still lower responses in later conditions, but because those later conditions don't always correspond to particular fonts, fatigue will contribute to error rather than bias.

6. **Statistical regression toward the mean**

Sometimes described simply as “statistical regression” or “regression toward the mean,” this refers to a phenomenon that only occurs when participants are selected based on extremely high or low scores, such as scoring very high or very low on an intelligence test. The phenomenon is that when tested again, the group's scores will tend to be closer to the mean. For example, let's say a school puts students with I.Q. scores above 130 into a gifted class, which has a mean I.Q. score of 135. Regression toward the mean will predict that one month later, when the group takes the test again, their mean will be lower than 135. Likewise, if a group of students who score below 80 are placed into a special education class with a mean of 75 and are then tested one month later, their mean will be higher than 75. In each case, the group's scores have “regressed toward the mean” of the population. Why does this happen?

Regression toward the mean occurs because of two factors: 1) a measurement is always a combination of true score and chance events, and 2) in a distribution of scores, there are always more scores toward the mean than there are on the other side of an extreme score. A score on an I.Q. test is mostly a function of cognitive ability but it is also a function of sleeping well the night before, just having a fight with a friend, or a hundred other chance events. An I.Q. score of 130 is at the 98th percentile, meaning that it occurs just 2% of the time. There are many more scores lower than 130 than above 130. Given a score of 130, there are three possible explanations:

A. the true score is below 130 (and chance events boosted it to 130)

B. the true score is exactly 130

C. the true score is above 130 (and chance events lowered it to 130)

Which of these three is most likely? Given that only 2% of people score 130 or above, A is much more likely. That means that the average person who gets 130 will score *lower* the second time they take the test, not higher. A group of people who score 130 at time 1 will thus tend to score below 130 at time 2.

Two other contributing factors to regression toward the mean can be **ceiling** or **floor effects**. A ceiling effect occurs when scores pile up at the high end of the scale, such as when sixth-graders are given 2nd-grade spelling tests. If you selected all the people who got perfect scores and then gave them another 2nd-grade spelling test, they could not do better but they could do worse, leading to an apparent decline in scores. A floor effect is the same phenomenon but with scores piling up at the low end, such as when 2nd-graders are given 6th-grade spelling tests. If you put all the people who got 0 correct into a group and then gave them another test, none of them could do worse and some of them would probably do better, leading to an apparent increase.

Regression toward the mean is a problem any time a sample is selected because of extreme scores. For example, if the Department of Education identifies a group of schools as "low-performing," implements some corrective measures, and then checks their performance later, they will very likely find that the scores have gone up *simply because of regression toward the mean and not because of any real improvement.*

The solution to regression toward the mean is to add a control group that does not receive any treatment. If the control group shows the same change as the experimental group, you know the change is not due to the treatment.

Summary

Within-subjects designs can be very statistically powerful, but without a control group or counterbalancing, they are vulnerable to the threats listed above. Any time you see results that show significant change over time, consider whether the change could be due to the factors discussed above.