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 without a control group, there are a number of potential threats to their internal validity. Fortunately, most of those threats can be ruled out with the simple addition of a control group (which would make them *mixed* designs because they now have a mix of both within-subjects and between-subjects independent variables).

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. If a control group had been included who did not receive the remedy and they showed the same improvement as the remedy group, we would conclude that the "improvement" in the remedy group was not due to the remedy. The cold remedy example belongs to the first category of threats, maturation effects.

1. **Maturation effects**

A maturation effect occurs when changes in a score over time are due to naturally-occurring internal processes. These 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. 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. The only way to know whether the results are due to the curriculum or to maturation is to give the two tests not only to the children with the new curriculum but also to children who get the old curriculum. If the new curriculum group improves *more* than the old curriculum group, the teacher can be more confident in the effectiveness of the new curriculum.

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 chemical increases anxiety. The researcher measures New York City residents' anxiety on September 5, 2001, gives them the drug, and measures them again 20 days later. Scores are likely to have increased because of the terrorist attacks – 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).

3. **Testing effects**

A testing effect occurs when the pretest itself influences the post-test. The most typical example of testing effects is a **practice effect**, where performance at post-test is higher than at pretest 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 pretest and not because of any exercise-related intervention scheduled between pretest and post-test. Another common testing effect is **fatigue**, where participants perform worse after repeated testing. If performance is being measured multiple times, it is possible that performance will decrease in later conditions simply because the participant is getting tired. A simple way to eliminate testing effects is **counterbalancing**, in which an equal number of participants receives each possible order of conditions. With just two conditions, half the participants will complete condition A first and half will complete condition B first.

4. **Instrument decay**

Instrument decay occurs when the standards of a measurement device change over time. An example might be a spring scale used at a loading dock. In the morning, the spring scale registers sacks of flour at approximately 50 pounds. By evening, the spring scale is registering the same sacks of flour at approximately 55 pounds. The standards of the spring scale have changed because the spring has become stretched out. Although the term *instrument decay* suggests mechanical instruments, it also applies to measurement by human judges. Judges may get better at measuring (akin to practice effects), they may get worse (akin to fatigue), or their standards may simply change because they attend to different aspects of performance (paying more attention to a diver's feet, for example). In every case, the measurements at different times may be due to changes in the *measurer* rather than actual changes in behavior, making them a threat to internal validity. The terms “practice effects” and “fatigue effects” are reserved for participants, so when they appear in judges they are generally classified under the more general heading of “instrument decay.”

5. **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 who score 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, they are vulnerable to the threats listed above. Any time you see results that show significant change over time without a control group, consider whether the change could be due to the factors discussed above.