Pre-lecture Notes II.5 – Between- vs. Within-Subject Designs

In previous discussions of internal validity, we focused on making sure that, when you manipulate one thing – namely, the potential cause of interest, such as room-lighting – in order to test whether it has an effect on something else – such as later recall – you need to make sure that you create one and only one difference between the conditions. But the choice of manipulation is only one place where a failure to be careful can reduce the internal validity of an experiment. A second place where you need to watch out for confounds occurs when you choose the type of experimental design.

Expressed in another way, after coming up with a manipulation that produces one and only one difference between the conditions in the experiment, the second-most important decision to be made in setting up an experiment is whether to have separate groups of subjects in each of the conditions – which is called a "between-subjects design" – or to have all of the subjects participate in all of the conditions – which is called a "within-subjects design."

There is no simple answer to the question of which type of design is better. This is true because the two options have opposite strengths and weaknesses in terms of the three kinds of validity that we have seen so far in the course. One type of design usually has higher levels of both internal and construct validity and the other type of design usually has higher levels of statistical conclusion validity. Understanding the details of why the two designs differ in terms of these three kinds of validity should help you to make the "best" decision for your particular research question. But please keep in mind that there's never a perfect option because there's always a trade-off.

Let's start by thinking about statistical conclusion validity. In general, within-subject designs have more power (in the technical/statistical sense of the word *power*) than between-subject designs. In other words, within-subject designs are better at finding (small) differences between conditions, assuming that a difference exists, given a fixed number of subjects; within-subject designs also give you more precise answers to questions, since they produce much smaller standard errors and, therefore, much smaller confidence intervals. This is true because, under a within-subjects design, all of the pre-existing differences between the individual subjects "cancel out" because all of the subjects are in both of the conditions. Thus, for example, if you have one subject who consistently scores higher than anyone else on your dependent variable (for reasons that have nothing to do with the experiment), then this won't have much if any effect on the statistics from a within-subjects design because this subject will contribute a high score to both of the conditions. In contrast, if you have one subject who scores higher than anyone else in an experiment that is using a between-subjects design, then whatever condition was "lucky" enough to get this subject will end up with a higher-than-expected mean (and a lot more variability).

On the flip side, between-subject designs usually have an advantage over within-subject designs when it comes to both internal and construct validity. To get an idea of how this works, imagine that you are a subject in a within-subjects version of the lighting/memory experiment and you have already done the bright-room condition and you are now about to do the dim-room condition. In other words, until a moment ago, the room was well-lit and you did a memory test. Now the room is really dark and you're about to do the memory test again.

With regard to internal validity, which always comes down to some sort of confounding, there are actually two differences between a moment ago and right now. First, it is now dark when it used to be

bright. This is the intended difference between the conditions. Second – and this might not seem very different from the first – it is now darker than before; in other words, the lighting has changed and it went down in particular. This is also a difference between the two conditions and it isn't the one that we wanted to study. The experiment wasn't intended to study the effect of ***changing*** the lighting; the experiment was supposed to only concern the effect of the current level of lighting.

The second problem with within-subject designs is a bit more straight-forward, but forces us to think about construct validity again. As above, imagine that you're a subject in the lighting/memory experiment and you've done the bright-lights condition and are about to do the dim-lights conditions. If you're like most people, you'll probably not only notice that the room is a lot darker now, but you could well suspect that this is a key aspect of the experiment. In fact, you might now guess (correctly) that the researchers are studying the effects of lighting on memory. This brings us to another definition:

> Demand Characteristic – *<u>any</u> aspect of the experiment that provides the subjects with a clue as to what is being studied*

Now, demand characteristics, on their own, are not a problem. The reason why we worry about them is that subjects often change their behavior – i.e., they "react" (in the technical sense of *reactivity*) – when they have a guess as to what is being studied. Most often they behave like "good subjects" in that they alter their behavior to provide the researcher with the data that they think that the researcher wants. Thus, for example, when you enter the dim-room condition after already having done the bright-room condition, you might say to yourself: "self, I think these people believe I'll do less well on their stupid memory test in the dark than I did in bright lights." And, then, because you're a nice person, you might also say to yourself: "OK, self. Fine. I'll play along. I'll make a couple of extra mistakes this time (on purpose) so that I actually do get fewer correct in the dark."

Before being clear why demand characteristics plus good-subject behavior is a serious problem for researchers, let's go back and imagine, instead, that the experiment was conducted using a between-subjects design. In this case, each subject only does the memory test in one condition. They should, therefore, have no idea what the other condition is, so they are much less likely to figure out that the experiment concerns lighting. Because of the lower level of demand characteristics under a between-subjects design, researchers are much less likely to have the kind of problem that we're discussing. That's why between-subject designs are said to suffer less from the problems that are based on demand characteristics.

OK, so which types of validity do demand characteristics plus good-subject behavior threaten? Let's do the easy one first. Good-subject behavior is a threat to construct validity because, when this happens, you are no longer getting a pure measure of what you are interested in, which, in this case, was memory. You are also getting a measure of what the subject thinks the experiment is about and/or whether the subject is a helpful person. Therefore, any amount of good-subject behavior lowers the discriminant validity of the measure being used. In the worst case, where the subject completely stops doing the task as requested and behaves in a way that is completely determined by what they think the researcher wants, then you have lost your convergent validity, as well. This is the major problem caused by having lots of strong demand characteristics.

But, when looked at in a slightly different way, you can also think of all this is terms of a threat to internal validity, too. Start with the general rule for internal validity: all threats to internal validity are due to some form of confounding. So what becomes confounded with condition in this case? Answer: the

presence of demand characteristics (which are objective, observable things) with condition. The demand characteristics are not equally strong during both conditions of a within-subject design. While the subject is in their first condition, the demand characteristics are the same as they are for a between-subject design, which is relatively low. It's when the subject enters the second condition – which only occurs under a within-subjects design – that the demand characteristics become strong, because that's when the subject gets the very strong clue as to what's being studied. So the demand characteristics don't influence both conditions equally. There's a huge correlation between the amount of demand and whether the condition was run first or second.

We can reduce this last problem by using a trick that I'll demonstrate in lecture, but the general problems that are caused by letting subjects see both of the conditions will never go away completely. This is why we have the general statement that within-subject designs have more threats to internal and construct validity than between-subject designs. So why do we ever use within-subject designs? Because they have much better statistical conclusion validity. A two-condition, within-subjects design usually needs only one-third as many subjects as a between-subjects design to produce statistically significant results. And that – in a nutshell – is the huge decision that you have to make when setting up an experiment: do I want lots of statistical conclusion validity (and therefore should use a within-subjects design) or do I want lots of internal and construct validity (and therefore should use a between-subjects design).