CHAPTER 8: INTERNAL and EXTERNAL VALIDITY

INTERNAL VALIDITY

An experiment is internally valid if there are no confounds.... that is, the only reason why the groups are different (with respect to the DV) is "actually and only" because of differences in the IV

EIGHT THREATS TO INTERNAL VALIDITY

all of the following are a potential source of confounds:

1. **History**

   Can be a problem in a repeated measures (within subjects) design where each participant is tested in each group.

   A history effect is present when an event (external to participants) occurs:

   a) Between presentations of the levels of the IV
   e.g. IV = day of the week: between taking a quiz on Tuesday and a quiz on Thursday, the campus "shuts down" on Wed when a student goes on the rampage (must of gotten his stats test back)

   or

   b) From pre-test to post-test with the IV presented in between
   e.g. students take a questionnaire on assertiveness > then receive assertiveness training > then take the assertiveness questionnaire again. What if something happens between the pre-test and post-test during the time the IV is presented (e.g. US goes to war with Canada)

2. **Maturation**

   Systematic, time-related changes in the participants that occur between presentations of the levels of the IV (or while the IV is being presented) in a repeated (within subjects) design (e.g. participants may be growing board, anxious, hungry, tired etc... they are also getting older)

   Beware of a maturation effect especially if testing takes place over a long time, or the task is very difficult, etc...

3. Testing

changes in the DV occur simply because the DV was measured (i.e. not because of the particular level of the IV).

Examples

i. The Hawthorn effect (also called reactance or reactivity effect)

Elton Mayo's Hawthorne Studies

The Hawthorne Studies (or Hawthorne Experiments) were conducted from 1927 to 1932 at the Western Electric Hawthorne Works in Cicero, Illinois (a suburb of Chicago), where professor Elton Mayo examined productivity and work conditions. Elton Mayo started these experiments by examining the physical and environmental influences of the workplace (e.g. brightness of lights, humidity) and later, moved into the psychological aspects (e.g. breaks, group pressure, working hours, managerial leadership).

The Hawthorne Effect

In essence, the Hawthorne Effect can be summarized as "Individual behaviors may be altered because they know they are being studied." Elton Mayo's experiments showed an increase in worker productivity was produced by the psychological stimulus of being singled out, involved, and made to feel important.

Additionally, the act of measurement, itself, impacts the results of the measurement. Just as dipping a thermometer into a vial of liquid can affect the temperature of the liquid being measured, the act of collecting data, where none was collected before creates a situation that didn't exist before, thereby affecting the results. Another example is measuring attitudes toward discrimination. If the survey is not “disguised” well, participants could alter their responses (the DV) to provide only socially acceptable responses

You can avoid this problem by using **non-reactive measures.** For example, measure the DV in such a way that participants do not know what’s being measured, or perhaps even that they are being observed. (One way mirrors, hidden cameras, deception)

ii. Practice effects (or fatigue)

Changes in the DV occur simply because of practice with the task (i.e. has nothing to do with the particular levels of the IV)... note that this effect is only a factor in repeated designs.

E.g. if you take the GRE several times, you can expect your score to increase a little each time…it's not that you know more, you just have more practice and you are more familiar & comfortable with the procedure

4. Instrumentation effects and human error

Values of the DV change because of faulty equipment, the human scorer gets tired etc... That is, changes in the DV which result from changes/errors in the recording device (whether synthetic or human)

Control by testing a few participants from each group all around the same time. NEVER test all of group A, then test all of group B, then all of group C…a HUGE faux pas!

Check your instruments/equipment before testing each day

5. Statistical Regression

Can occur in repeated measure designs when people score either extremely high or extremely low. You could see regression toward the mean. The next time you measure them, there will be a tendency for their scores to move in the direction toward the mean... This is a confound because you will not know if the DV changed because of the IV, or because of regression toward the mean.

6. Selection

This is a problem that could arise when using an IV that is a classification (or subject) variable**.** Examples include: gender, SES, academic major, mental diagnosis… There are certainly *pre-existing* differences between the levels (categories) of each factor. If you do see a difference between the levels, how do you know if the IV produced the difference vs. the *pre-existing differences* which are only tangentally related to the IV

e.g. Does watching American Idol increase singing in the shower? To find out, you record how long American Idol fans sing in the shower vs non-American Idol fans. However, American Idol fans are probably different from non-fans in other respects eg. AI fans are more intelligent!

7. Mortality

refers to attrition due to death and “no shows”
If mortality occurs in one condition more than the other conditions, then you have a problem, specifically, a confound

The “survivors” in the group that was hit particularly hard are probably very different from the subjects in the other groups. If you now see a difference between the groups, you won’t know if it’s because of the IV or something particular to that one group of survivors

Even if mortality is approx the same for each group, you still have a problem. To what extent do the survivors represent the population you had originally targeted? I.e. you have a problem with **external validity**

8. Diffusion or imitation of treatment

Participants in one treatment group become familiar with the treatment of another group. They then either copy that treatment or are just otherwise affected by what they have learned. As such, they are no longer “naive” and this changes their behavior. This will minimize or mask completely the difference between your groups in an experiment.

We try to prevent this from happening by asking people to refrain from talking about the experiment with any other participant until the experiment is over.

Interactions with selection

Occur when one or more of the effects discussed above (e.g. history, maturation, mortality, testing, instrumentation etc) are systematically different between the different levels of the classification IV

E.g. cross-cultural research is prone to a selection x history effect. That is, different cultures differ not only by culture, but also by their historical experiences

PROTECTING INTERNAL VALIDITY

These actions need to be taken before you run the experiment.

First, you must sit down and think about all the potential confounds. Ask yourself, “what could go wrong”.

Second, implement one or more of the control techniques discussed in chapters 6 and 7. (e.g. balance, random assignment to group, hold the EV constant etc…)

Third, carefully follow one of the standardized experimental designs, to be discussed in chapters 10, 11, 12. (e.g. repeated measures t-tests, mixed ANOVAs)

Fourth, have a knowledgeable person(s) review your proposal before you conduct the experiment.

The book says that statistics do not control/eliminate confounds, nor detect them. This is mainly true, but note exceptions:

> Analysis of co-variance can control for potential confounds
>
> Chi squares can help detect a potential confound by seeing if an extraneous variable is evenly distributed across the different levels of the IV

EXTERNAL VALIDITY

The extent to which your results apply to populations/situations/times/environments different from those in your experiment… concept of generalizability

Different types of generalization

Population generalization: the extent to which your results generalize to people/animals beyond just the participants you tested.

Environmental: the extent to which your results generalize to situations or environment beyond those used in the current experiment

Temporal: the extent to which your findings apply at all times, not just the specific time/season your study was conducted.

NOTE: in all cases, the lack of generalization could in and of itself be VERY interesting and could yield vital clues about human/animal behavior

Relationship between internal validity and external validity

Remember this relationship from the previous chapter: as one goes up, the other goes down… as a general rule…

As we implement more and more controls to reduce confounds (i.e. increase internal validity) we are making the experiment more and more artificial and thereby it's generalizability (external validity) suffers.

An exception would be in reference to <u>specific control techniques</u>

> e.g. the balance technique would allow for more generalizability than would the eliminate or hold constant techniques

Relationship between within group variability, power, and external validity

Recall that one way to increase power is to test *homogeneous* groups. But, in the *real* world, people are not homogeneous. So, by testing homogeneous groups, our results may not generalize well to the real world.

FOUR THREATS TO EXTERNAL VALIDITY BASED ON METHODS

Often, the design of our experiment threatens its generalizability

1. **Interaction of testing and treatment**

   In a pre-test, post-test design (also called a before-after design), the pre-test may sensitize people to the treatment yet to come. Since pre-testing does not occur in the real world, our results may fail to generalize well

   You can estimate the effect your pre-test has by adding another group: those who only get the post-test

2. **Interaction of selection and treatment**

   This occurs when the groups of participants you test are so unique, your results do not generalize beyond them.

3. **Reactive arrangements**

When people know they are part of an experiment, they typically change their behavior no matter what the IV is that they are exposed to (the Hawthorne effect). This means that our results may not generalize to the real world where people are not part of an experiment and whose behavior is not thus affected

Demand Characteristics present in reactive treatments

Recall that these are just about anything (other than what the experimenter says or does) that participants use to figure out the hypothesis or how they should behave. You cannot eliminate demand characteristics in an experiment where people know they are part of the study. Since the DC's change participants' behavior, and these DC's are not present in the real world, your results may not generalize well to the real world. The only way to completely remove DC's in a study is if you use *naturalistic observation*

4. **Multiple Treatment Interface**

These occur in repeated designs where the same people are tested in each group or condition. It's possible that the effect you observe is present *only when* people are exposed to this constellation of treatments. That is, in the real world you would not observe the same effect of a specific treatment because it was not accompanied by the other treatments.

FIVE THREATS TO EXTERNAL VALIDITY BASED ON PARTICIPANTS

Limited types of population tested (animal)

Especially in animal research, we tend to test mostly rats (and specific breeds at that). To what extent will the results generalize to other species, including humans?

Limited types of population tested (human)

In human research, we tend to use convenience sampling and test mostly college students. To what extent do college students represent the general population? To the extent that they don't, our external validity suffers.

Gender Bias

For many reasons, some of which persist today, men were studied more than women in psychological and medical research. To what extent can you generalize from research conducted on men to women (or vise versa for that matter)?

Racial Bias

Same idea as above. If your research does not include certain racial groups, you must exercise caution when trying to generalize to them.

Cultural Bias (ethnocentric research)

If you study American culture in America, your results can only be assumed to generalize to American culture in America

Four goals of research that do not stress external validity

1. Finding out if something CAN happen, not when if it usually happens
2. Finding out if a real world phenomenon occurs in the lab
3. Finding out if something occurs in the lab's unnatural settings can strengthen the validity of the phenomenon
4. Studying a phenomenon in the lab that doesn't have a real world counterpart

According to Smith & Davis, internal validity is essential for an experiment, whereas external validity is not.

Replication with extension can be used to establish the validity of the finding and its external validity

Once an effect is demonstrated under one set of circumstances (testing environment, time, participants), you can replicate the finding by changing one or more of these variables. It is advisable to change only one thing at a time. Why?