# Basic Experimental Design

## I. The Basic *Between Subjects* (Between Groups) Experiment

Start with a hypothesis about something that can cause a difference in behavior. This potential cause will become the independent variable (IV). The behavior being observed will be the dependent variable (DV).

Form two equivalent groups of participants. The only factor that determines which group a person is assigned to is chance (*random assignment* to groups).

Introduce the IV manipulation then measure behavior on the DV.

Since the two groups were equivalent before the IV, any difference in participant behavior must have been caused by the IV.

This simple, *between subjects design* has internal validity. That means that we can determine a cause and effect relation between the IV and the participants' behavior.

For example, if we want to evaluate the efficacy of a nicotine patch in helping people quit smoking, we would take a set of smokers and randomly assign them to one of two groups. This would give us two equivalent groups with differences between participants randomly distributed across the two groups. The only difference in the treatment of the two groups would be the nicotine patch. (How would you treat the other/control group?)

Following treatment, we would assess the degree to which participants smoke. Any difference between the two groups would be due to the patch.

# II. Some Non-Experimental Designs

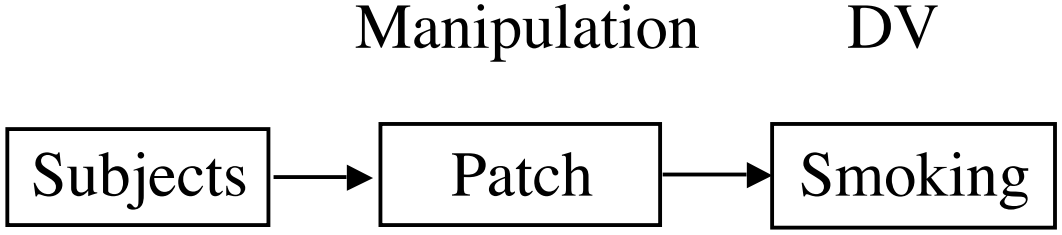
For comparison, we'll consider some non-experimental designs. This will help to illustrate why only the true experiment has internal validity.

They are: One Group Designs
Non-equivalent Control Group

## A) One Group Designs

### 1. One Group Posttest Only

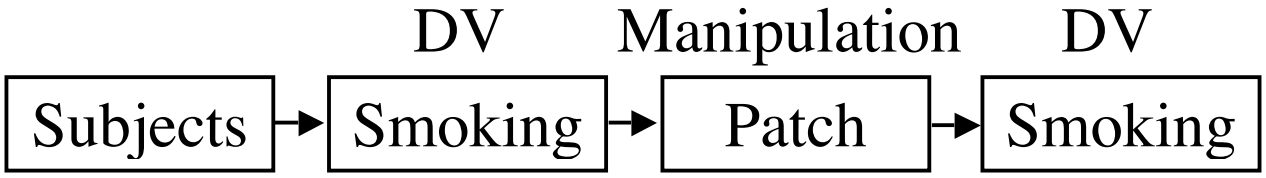
In this design, after subject selection, a manipulation is introduced and then the participants' behavior is measured (DV).



This design is problematic because there is no comparison (alternative treatment or baseline or control condition) for what the participants would have done without the manipulation (patch in this example). Would they have quit smoking anyway, or reduced smoking without the patch? Was there some other event that co-occurred with the manipulation that could have been the real cause?

2. One Group, Pretest and Posttest

Here, we try to increase our control by measuring all participants on the DV before the manipulation and then again afterward. We will look at the difference in their behavior on the DV.



Even though we can evaluate the degree to which each participant's behavior has changed, we still *cannot determine why*. Was it some other event that occurred while the manipulation was taking place? Was it that the participants realized they were getting older? Did someone famous who was a smoker die of lung cancer? Were they trying to quit anyway? Was it simply that they expected the patch to work?

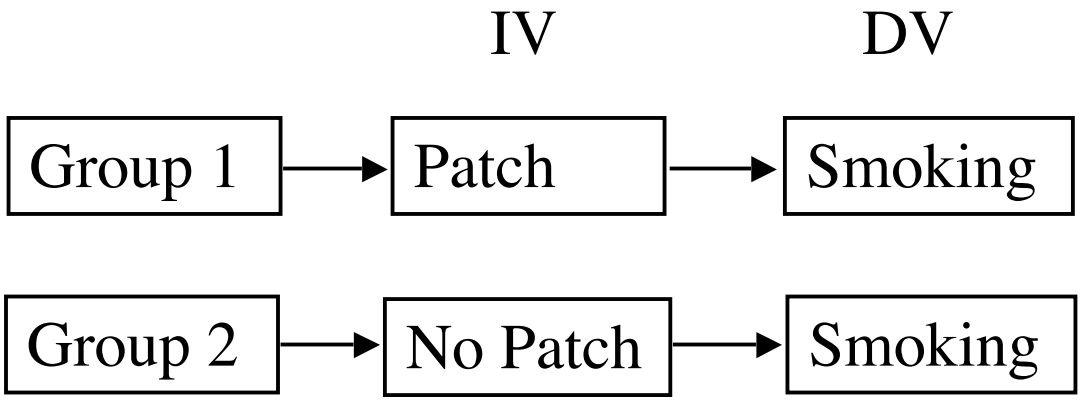
The lack of an appropriate control or comparison makes it impossible to answer these questions.

Both of these two one-group designs lack an independent variable. Since there is no independent variable, there is no basis for determining cause (independent variable) and effect (dependent variable).

# B) Non-equivalent Control Group

In this design, we have two groups with one getting each level of our *independent variable*. However, the two groups are NOT formed by random assignment. Rather, we take advantage of naturally occurring groups.

For example, to study the effect of the nicotine patch, we take a group of smokers who have decided to quit smoking and they get the patch (Group 1). For comparison, we take another group of smokers who have not decided to quit smoking (Group 2) and they do not get the patch.



This is still *not* a true experiment and does not allow a cause and effect conclusion. It could be that once a smoker is ready to quit, the patch has little or no additional effect. Conversely, it may be ineffective on someone who does not want to quit (they simply get a higher dose of nicotine).

There were (are) pre-existing differences between our two groups. Some of these pre-existing differences could reasonably explain any difference we observe in behavior (on our DV) in the study. We *can not conclude* that the manipulation produced (caused) any difference on the DV because of *alternative explanations* based on the *pre-existing differences* between the two groups.

# C) Summary

All three of these non-experimental designs have less internal validity than a true experiment.

The one group designs do not even have an independent variable, since everyone received the same, single treatment.

All three suffer from one or more of the problems that will be described in the next section.

# III. Threats to Internal Validity

A) Each of these is a *potential cause* for changes in the dependent measure (behavior) in some types of studies.

1. Maturation - The biological processes of aging and growth that normally occur *while* the study was conducted.

2. History - External events that occurred (that may have influenced the participants' behavior) *while* the study was conducted.

3. Testing - Changes (or lack of any change) in the participant's score that result from previous testing (practice, fatigue, reactivity, learning, memory of previous performance).

4. Instrumentation - Changes in the accuracy/calibration of the instrument over time. This can be a problem with human observers.

5. Regression to the mean - On repeated testing, extreme scores tend to be less extreme. If subjects are originally selected because of extreme scores, then it is possible that they will change on repeated testing because of error in the original measurement.

6. Selection - Are the groups of subjects equivalent at the start of the study? How do we know?

7. Attrition - Do participants leave the study so that the conditions (groups) are no longer equivalent? For example, if one condition is difficult and participants drop out more than in other conditions, this destroys any equivalence of the groups.

# B) Comparing the Designs

Some of these problems can happen with *any* study. Most are just a problem with the non-experimental designs. To illustrate, we'll look at each of our three non-experimental designs for each potential problem.

| Problem | One-Shot | Pre/Post | Nonequiv |
|---|---|---|---|
| Maturation | Yes | Yes | No? |
| History | Yes | Yes | No? |
| Testing | No | Yes | Maybe |
| Instrum. | Maybe | Yes | Maybe |
| Regression | Maybe | Yes | Maybe |
| Selection | | | Yes |
| Attrition | Maybe | Maybe | Maybe |
| Baseline | Yes | Yes/No | Yes/No |

Note that even though the one-shot (post-test only) design seems to have fewer problems than the pre/posttest, the one-shot is the only design that lacks any baseline. The lack of any baseline means that there are *no* comparisons that can be made.

# IV. Between Subjects Designs

When we form two or more groups of participants using random assignment to conditions and each participant/group goes through only one condition or treatment, this is a *between subjects* (between groups) design. The manipulation of the IV is "between" participants since each participant sees only one condition (one level of the IV).

There are three basic versions of this design:

Posttest only
Pretest-Posttest
Soloman four group

A) Posttest only

Randomly assign subjects to conditions. Introduce IV manipulation. Measure effects on DV (post-test). This is the type of design described in Section I above.

## B) Pretest-Posttest

Randomly assign subjects to conditions. Measure DV (pretest). Introduce IV manipulation. Measure DV (posttest).

## C) Comparison

1. The advantage of the *pre-post* design is that we can compare each participant's score after the IV to that before. This allows us to eliminate some of the differences between participants as a source of error. The design is *more sensitive* than the posttest only because it has a *baseline for each participant*.

2. The advantage of the *posttest only* design is that there is no possible contamination of the posttest DV by the subjects' familiarity with the DV from the pretest. The *pre-post* design has the *potential confound of testing effects*.

D) Soloman Four-Group

1. Posttest only, experimental (level 1 of IV)
2. Posttest only, control (level 2 of IV)
3. Pre & Posttest, experimental (level 1 of IV)
4. Pre & Posttest, control (level 2 of IV)

*This design has two independent variables*. One is the experimental versus control conditions (like the posttest only and the pretest-posttest designs). The other is testing (pre-post vs. post only). Forming all possible combinations of the 2 levels on each variable gives us 4 conditions (2 x 2).

If there is an effect of *testing*, this design can measure it. Future work would then use the post-test only design. This design can also detect if the manipulation interacts with repeated testing.

The disadvantage of this design is that it is twice as much work as the other two designs and takes twice as many participants.

# Summary so far – Between Subjects Design

As long as the two (or more) groups of participants are equivalent at the start of the study and the only difference between the groups during the study is the IV, the logic is sound and we can determine cause and effect.

But, how do we "know" that the participant groups are equivalent at the start? Random assignment only says that, on average, the groups will be equivalent. *Random assignment does not guarantee equivalent groups*.

# E) Using Matching in Subject Assignment

1. *Matching prior to assignment to groups*. Subject qualities are measured in advance and sets of equivalent subjects (subjects with the same measured characteristics) are formed. Then, subjects from these sets are randomly assigned to conditions such that each condition has the same number of subjects from each of these sets.

If we had 10 males and 10 females, we could randomly assign half of each sex to a control group and half to an experimental group. This would ensure that the groups were *equivalent* on the variable of sex.

In a study of how to improve memory using various mnemonics, we might match on age because it has a large influence on memory performance.

*Why do this?* Matching **increases the odds** that the groups in a between subject design are equivalent. However, since you can not match on everything, it can not guarantee equivalent groups. Matching is generally just done for variables known to pose a problem if they are not equally distributed. It is used *with random assignment* to form equivalent groups.

2. Post-hoc Matching – Analysis of Covariance. Another technique is to form our groups using random assignment only. During the experiment, we also collect information from each participant that might be related to the DV (e.g. age in our memory example). A statistical technique (Analysis of Covariance) is then used to eliminate the influence of this subject characteristic from the data.

3. The problems with matching:

a) It is a lot of work, and since some subject qualities occur in clusters, you can not necessarily match on everything: matching on some things can produce a mismatch on others. In the most extreme case, if you measured everything about your subject, the subject would be unique and no match could be found.

b) If you get subject attrition, it destroys the equivalence of the groups. If a subject drops out of the experiment or performs so poorly that their data are unusable, the groups are no longer matched. Unless you have additional subjects whose equivalence has been established ahead of time to substitute, the groups are now *guaranteed NOT to be equivalent*. Subject attrition is a problem in most cases, and a bigger problem when matching is used.

# V. Within Subject (Within Group or Repeated Measures) Experimental Design

Each subject is run through *every* level of the independent variable(s). Put another way, each participant goes thru all of the conditions. This is the *Within Subject* design.

## A) The Design

Since each participant goes through all of the conditions, each participant serves as their own control. This makes the within subject design *extremely sensitive*.

Because the DV is used repeatedly (with each IV condition), this is also called a *repeated measures* design.

# B) Counterbalancing.

Because each participant goes through multiple conditions, it is possible for their experience in one condition to alter their performance in the next condition. This is a *carry-over* effect.

Two examples of this are effects of practice (better with repetition via learning) and fatigue (worse with repetition via tiring, boredom, distraction).

The *key idea* here is to have different participants go through the IV conditions in different orders. This is called *counterbalancing*.

Counterbalancing the order of conditions distributes carry-over effects (it does not eliminate them). That is, we want all levels of the IV(s) (all of our conditions) to have equivalent influences of carry-over. Thus, carry-over is minimized as a confounding variable by making sure that *each condition contains the same carry-over effects*.

There are three basic approaches to counterbalancing: *complete*, *random* and *Latin-square*.

## 1. Complete counterbalancing

If your independent variable had only two levels, then there are only two possible orders for presenting the conditions. Together, these orders are a complete counterbalancing.

Complete counterbalancing requires that all possible orders of conditions be used equally often. For n conditions, there are n! (read as n-factorial) orders for a complete counterbalancing. n! is:

n x n-1 x n-2 x n-3 x ... x 2 x 1

so 2! is 2
3! is 6
4! is 24
5! is 120

If your experiment has lots of conditions, this would require lots of subjects. For more than 3 or 4 conditions, this is usually impractical. However, whenever it can be used, complete counterbalancing is preferred.

## 2. Random counterbalancing

The order of conditions for each subject is determined by chance. This is analogous to random assignment in the between subject design. Averaged over many orders, the carry-over effects will be equally distributed.

## 3. Latin-square

In the Latin-square, the goal is to distribute carry-over effects equally by having each condition occur in each ordinal position (1st, 2nd, 3rd, ..., last) equally often. In addition, we would like each condition to be preceded and followed by each other condition equally often. Here, we are balancing carry-over effects that may reflect particular combinations of conditions. This is a *balanced Latin-square*.

There is a simple formula for the orders to be used. We'll illustrate with the orders for a set of 6 conditions.

The formula for ordering the conditions for the first subject (row) is:

1, 2, n, 3, n-1, 4, n-2, ...
(n is the number of conditions)

Then, for each succeeding subject (row), add one to each condition. If the condition was number n, make it number 1. For 6 conditions (labeled A, B, C, D, E, F), we have 6 orders:

| | Order of Conditions | | | | | |
|---|---|---|---|---|---|---|
| Subject | 1st | 2nd | 3rd | 4th | 5th | 6th |
| 1 | A | B | F | C | E | D |
| 2 | B | C | A | D | F | E |
| 3 | C | D | B | E | A | F |
| 4 | D | E | C | F | B | A |
| 5 | E | F | D | A | C | B |
| 6 | F | A | E | B | D | C |

The number of subjects must be some multiple of the number of orders. If you have 6 orders, then there should be 6, 12, 18, ... subjects with each order run equally often.

## C) The Time Interval Between Conditions

In a within subject design, we need to choose the time interval between conditions carefully.

1. To minimize fatigue. Here, we want to reduce or eliminate any fatigue the participant may experience from repeated testing.

2. To reduce carry-over effects. Some conditions produce longer lasting effects. In drug trials where different dosage levels are tried within subjects, time is needed for one dose to clear the patient's system before starting the next dose.

If a design manipulates mood (e.g. anger, sadness, joy), then we need time for the participant to return to baseline before the next condition.

# VI. Comparing the Experimental Designs

A) *Within* designs are more efficient.

Since each subject is exposed to all conditions, fewer subjects are needed. However, more time per subject is required.

If the time commitment is too large, participants will drop out. Consequently, a design is sometimes chosen based on this pragmatic consideration.

B) *Within* designs are more sensitive.

Since each subject participates in all conditions, each subject serves as her/his own control. Thus, this design is *insensitive* to individual differences and more likely to reveal differences caused by the independent variable. That is, the within design *guarantees equivalent groups* before the experiment is started.

C) *Between* designs do not suffer from carryover effects.

Since each subject participates in only one condition, there are *no carryover effects*.

This is critical when a condition produces *permanent* carryover effects.

For example, to investigate the influence of Headstart on young children's performance in school, we have to use a between subjects design. You can not return the child, once they have gone through the program (or not) to their earlier age to start again. The effects of maturation, which co-occur, are permanent.

Similarly, a study of two methods of teaching algebra must use a between design. Participants can't "unlearn" the material that they have been taught.

# D) Which Design Mimics the "Real World"

This question requires us to know something about how the factors we investigate occur in the world.

A study to investigate perception of colors under different lighting would probably use a within design since people are exposed to colors under different lighting conditions. The within design would mimic our experience in the real world.

To study the influence of a defendant's appearance on jury behavior (verdicts), we would use a between design. In real court cases, jurors participate in one trial rather than the same trial done repeatedly.

# Answers to Chapter 7 Sample Questions

1) – b;  2) – a; 3) – a; 4) – d

# Sample Multiple Choice for Chapter 8

1) Matching is used in forming equivalent groups in an experiment because: a) it reduces the possibility that the groups differ (are not equivalent) b) it is a convenient, easy substitute for random assignment c) it guarantees that the groups are equivalent on all relevant variables d) all of the above

2) Subjects in a non-equivalent control group: a) are similar to the group that received the treatment, but do not receive the treatment b) are controls for maturation and history effects c) are randomly assigned to the control group d) a & b above

3) Differences between a within-subjects and a between-subjects design include: a) there is no problem of forming equivalent groups in a within-subjects design b) confounding cannot occur in a between subject design c) each subject serves as her or his own control in a within subject design d) a & c above

For the next two questions, refer to the following paragraph.

A researcher administers a mood scale to a sample of anxious adults. Then, the subjects' television set is removed and after two months of no TV, the subjects' anxiousness is measured again.

4) This design is an example of: a) a natural control group, pretest-posttest b) a no control group experimental design c) a single group, pretest-posttest design d) b & c above

5) If the study found that the subjects' anxiety decreased, this would imply: a) that TV viewing leads to higher levels of anxiety b) that TV viewing and anxiety are related for most individuals c) that reducing TV viewing might be an effective treatment for anxiety d) all of the above

# Conceptual Review

Below is a description of an experiment that has two independent variables. Please read it and answer the following:

1. What is the dependent variable?
2. What are the two independent variables?
3. Which of the independent variables was done as a between subjects (participants) manipulation?
4. Are there any subject variables and if so, what are they?

In a memory experiment, all participants initially answered a series of questions about words. For half of the questions, the question was designed to get the participants to think about the sounds in the words (how they are pronounced). The other half of the questions were designed to get participants to think about the meaning of the words (semantics). After answering the questions, participants' memory was tested one of two ways. For half of the participants, a standard recognition test was given in which a single word was presented at a time and the participant indicated that the word was one of the ones from the set of sentences or not. In the standard recognition test, participants were more accurate at recognizing the words from the meaning based questions (82 percent correct) than from the sounds based questions (60 percent correct). The other half of the participants were given a rhyme recognition test. A word was presented and they were asked what word from the original sentences rhymed with it. For the rhyme recognition test, Words from sentences that focused on the sound of the word were better recognized (48 percent correct) than words from sentences that focused on meaning (31 percent correct).