Testing Scientific Explanations

Most people are curious about the causes for certain things. For example, people wonder whether exercise improves memory – and, if so, why? Or, we wonder whether using cell phones causes cancer – and, if so, how? Science allows us to systematically test competing hypotheses of such relations and the explanations for them. For example, exercise may affect neurotransmitters or increase blood flow to the brain, either of which could increase memory. Scientists conduct experiments to test their hypotheses. Any conclusions are only as good as the experiments that were used to support the conclusion. In this module, we will teach you to detect problems with experiments that lead to conclusions that seem suspect.

A hypothesis specifies a testable relationship between variables. Thus, a hypothesis has three important elements: the independent variable, the dependent variable, and the type of relationship between them. The independent variable (IV from here on) is the condition that we vary to see if it has an effect on the other variable. In the examples above, the independent variables are (A) exercise and (B) cell phone usage, because we think they will causally affect another variable. The dependent variable (DV from here on) is the variable that we think will be affected. It is also called an outcome variable. The dependent variables in the examples above are (A) memory and (B) cancer.

Finally, both example hypotheses assume a causal relationship, specifically, causes an increase. You should know that hypotheses are not always about causal relationships. One could also have correlational hypotheses, which merely predict that two variables change in relation to each other, but they do not assume that a change in one causes the change in the other. For example, we could change the verb in our example hypotheses to get these correlational hypotheses: (A) more frequent exercise is associated with better memory or (B) cell phone usage is related to the incidence of cancer.

Figure 1. General statement.

<table>
<thead>
<tr>
<th>Conclusion: IV</th>
<th>( \rightarrow ) causes</th>
<th>( \rightarrow ) DV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conclusion: exercise (IV)</td>
<td>( \rightarrow ) causes increase</td>
<td>( \rightarrow ) memory (DV)</td>
</tr>
<tr>
<td>Conclusion: cell phone (IV)</td>
<td>( \rightarrow ) causes increase</td>
<td>( \rightarrow ) cancer (DV)</td>
</tr>
</tbody>
</table>

When evaluating research, the first things you should do are to identify the independent and dependent variables and determine if a causal or correlational prediction is being made. There are more specific issues related to these topics that you should consider when evaluating research. In these pages, you will read about some important flaws that may occur when research is conducted.

A. Possible problems with IVs and DVs

1. **Use valid measure of IVs and DVs**

   Causal hypotheses predict that a change in the IV will cause a change in the DV. To test such a hypothesis, one therefore has to manipulate or vary an appropriate IV of interest. For the cell phone hypothesis, for example, you could compare cancer rates (DV) for participants who are grouped based on minutes of cell phone usage per day (e.g., more than 2 hours a day, .5 - 2 hours a day, less than 30 mins a day). This is a valid definition of cell phone usage (IV) because you are varying the amount of usage. An invalid definition of the IV would be the amount of money spent on a cell phone per month (e.g., more than $100 dollars, $40-99 dollars, less than $40 dollars a month). This is invalid because it doesn’t really indicate how much time one spends on the phone. You could have a very cheap plan, pay for much more time than you actually use, or call a lot but only on off hours. It is also important to have a valid measure of your DV. For example, if you want to test whether exercise (IV) improves memory (DV), you need to measure memory in a way that accurately reflects the aspect you are interested in. A standard memory test would be a valid measure; weighing someone’s head would not.

   **Construct validity.** We are often interested in variables that cannot be directly measured or observed, such as memory, intelligence, or depression. We refer to such variables as constructs. When our variables are constructs, we have to rely on scientifically developed measures as indicators of those constructs. For example, we may use a score on an intelligence test to indicate that a person is intelligent. Because these are not direct measures, they always depend on the developer’s ideas of what the construct is. Because these can vary, an experimenter has to be very careful in the choice of measure. When a measure really reflects the construct well, we say it has construct validity. Construct
validity of the IV in our example is whether working out on a treadmill (6 hours per week) is actually a “good” sense of exercising, whereas construct validity of the DV is whether our test of verbal and spatial memory is a “good” measure of what people mean by memory. For example, giving a person a gym membership would not be a valid manipulation of exercise, and self-reported judgment of one’s own memory is an invalid measure of memory. If your experiment does not have good construct validity, then you are not actually manipulating or measuring what you are claiming to — and thus, you will not test your actual hypothesis.

**Content validity.** A second type of validity is content validity. Most constructs have many aspects, and the more of them you include in your measure or manipulation, the better your content validity becomes. For example, if a professor gives students a study guide for the test but all the questions cover only one or two of the areas, students would complain that it is not a fair measure of what they were supposed to learn. They are really arguing that the test does not have content validity. For the exercise example, to measure the number of items recalled from a list of words may have construct validity, but content validity could be improved by including other types of memory, such as recognition memory or spatial memory (e.g., maps and images). We could improve the content validity of our IV by also including 3 hours a week of anaerobic exercise (e.g., weight lifting).

So, in summary, you have to manipulate and measure the variables you claim you are, not some other related variables, in order to have construct validity. If you don’t, you cannot draw valid conclusions because you are testing the relation between variables that are different from what you stated. Furthermore, you have to vary the most important elements of a variable, not just any one aspect of it, in order to have content validity. Valid measures are required to be able to appropriately interpret the results from an experiment and draw appropriate conclusions.

(2) **Must have comparison group(s)**

When testing the effect of an IV (e.g., exercise) on a DV (e.g., memory), you have to vary the IV. To test a causal hypothesis (e.g., exercise improves memory), you need to manipulate the IV, exercise. To test a correlational hypothesis (e.g., exercise is associated with better memory), you just compare already existing groups that differ on their level of exercise. Most of the time, scientists want to test causal hypotheses. To do so, you need one or more comparison groups that are not exposed to the treatment. For example, your treatment group would exercise 6 times a week while the comparison group would not exercise at all. Your comparison group, however, does not have to be a no-treatment group. For instance, you could have several groups that vary in exercise frequency (e.g., 6 times a week, 3 times a week, 1 time per week, and no exercise). These additional comparison groups allow you to determine how much each unit of exercise adds to improving memory. Although there are many things to consider in designing your comparison groups, the two most important flaws are having no comparison group and having a poor comparison group.

**No comparison group.** Although it may seem good enough to simply show that a DV changes from before to after our IV manipulation (thus, a difference between “pretest” and “posttest” performance), this is not the case. It may be that everyone would improve on the second test even without the treatment. For example, people may perform better on our memory test the second time, even without exercise in between. To be able to attribute any change to the IV, we need a baseline for comparing change. This baseline is a comparison group not given the treatment.

**Poor comparison group.** A more complicated problem is when the experimenters have two or more groups, but they do not vary on the critical IV that they claim they do, or something besides the IV is varied between groups. For instance, if people in the exercise condition work out in a fun group class, then any improvement may be due to the social interaction. To make the groups as similar as possible, you could have the no exercise group also attend a sedentary group class like watching others exercise. Notice how much care has to go into deciding what your comparison group should be.

Comparison groups allow researchers to rule out alternative explanations for any effects, but only if the comparison group is equivalent in all regards except for the IV. If the researcher uses chance to determine each participant’s group assignment, such as flipping a coin (called random assignment), then this randomly assigned comparison group allows the researcher to make causal conclusions about the effect of the IV on the DV. In contrast, the researcher could put people into groups based on their natural, pre-existing characteristics, such as how much they usually exercise per week (0 hours, 3 hours, or 6 hours). But because many other things vary with how much someone chooses to exercise per week (e.g., diet, fitness level, motivation level, weight), it could be that one of these other reasons could explain
the improved memories. Random assignment means that everyone in a study has an equal chance of being in any condition, and this allows us to distribute such differences across conditions so that hopefully, the only thing that varies is our manipulation (IV). Thus, even the best comparison group will not allow the researcher to make a causal statement about the relationship between the IV and DV if people are not randomly assigned to condition.

(3) Use appropriately sensitive DVs

To test a hypothesis, we vary one or more IVs to see whether they affect performance on the DV. We have to make sure that the test or measure we are using for the DV is sensitive enough to detect any possible effect. There are two main flaws that can affect DV sensitivity.

Ceiling or floor effects. Ceiling effects happen when a test is so easy that everyone achieves a near perfect score (at the ceiling or top of the scale). So even if you have a perfect treatment, you would never know it because you cannot make people do better than the control condition, which is already performing at the top of the test. Floor effects occur when the test is so hard that participants are not getting any items right. In this case, your manipulation, even if effective, may not be powerful enough to improve the scores even a little.

Imprecise measures. The second concern with DV sensitivity is the use of an imprecise measure. This requires the researcher to determine how fine a scale should be used. For example, suppose you want to measure how much someone exercises. You could use a three-point scale (none, some, a lot). This is easy to answer, but if people start to exercise a little more, it is unlikely that this would be reflected by this broad 3-option scale. If, however, you measured exercise on a 100% scale, you would be able to see numbers change representing even slight increases in exercising.

B. Problems with eliminating unwanted variables

When you test the effect of an IV on some DV, it is important that you control the rest of the environment. There are always other variables that also vary between groups that could affect the outcome, but researchers try to isolate the effect of the IV by minimizing the influence of these other variables. You want to hold as many things constant between groups as possible especially if you think they may affect the DV. Otherwise, you cannot be sure that your varying of the IV(s) was really the reason for any change in performance (DV). For our exercise example, there are several possible variables that we might worry about (such as, motivation, starting health level, previous level of exercise, attitude and experience with exercise, people working out beyond what is required for the study). Two standard practices to help the groups be comparable before a treatment are to randomly assign participants to your groups, and to manipulate the IV rather than using naturally existing groups (which may differ along other variables). While there are many serious flaws that concern controlling the environment, we will only mention two: attrition and experimenter bias.

(4) Limit attrition

Groups can become different even after random assignment because sometimes participants “drop out” or fail to complete the experiment. We call this attrition.Attrition can be a serious problem, especially if the number of participants failing to complete the experiment is different for the various conditions. This can lead to differences between groups that affect the DV, which makes it impossible to conclude that the IV was responsible for any DV differences between groups. Attrition is most problematic if it is just one group that looses participants, or if participants with specific characteristics that may affect the DV drop out. Attrition is less of a concern when it is random and not due to the participants, such as equipment failure. However, researchers should always report the level of attrition in all conditions.

Drop out. Attrition due to drop out can be a serious flaw. This happens when, after randomly assigning participants to condition or groups, a participant leaves without finishing, or stops showing up for sessions. When this happens, the groups may then be different for a reason other than the manipulation. In the case of the exercise experiment, we might find that unmotivated people may be more likely to drop out of the exercise group than out of the comparison group. This would leave the experimental group with more highly-motivated individuals than the control group. Because highly-motivated individuals will probably try harder on the memory test (DV), we would likely find memory differences between groups even if the exercise has no effect, simply because of the characteristics of those who dropped out.
**Missing information.** Attrition due to missing information is usually less of a threat but is still a concern. This happens when a participant completes the experiment but fails to answer every question. In our exercise-memory experiment, we could find missing information if participants do not indicate their gender or age on the pre-interview survey, or more important, they may skip one or more trials in the memory test. Missing information could lead to biased results, for example if people in one group systematically skip specific problems.

(5) **Limit experimenter bias**

One of the easiest factors of the environment to control is the effect of the researcher or experimenter. We know from several experiments, that the behavior or expectations of a researcher can have a significant impact on the results of an experiment. When this happens, it is not possible to determine whether the results were due to the IV or to actions of the experimenter. There are two important flaws to consider here. The first type of flaw focuses on our motivation (conscious or unconscious) to bias the results; the second type of flaw focuses on our ability to bias the results.

**Conflict of interest.** Conflict of interest occurs when a researcher has a strong investment (fame or fortune) in a particular outcome of the experiment. This is a problem because the researcher may intentionally or unintentionally bias the experiment. For our exercise example, if we were paid by an exercise equipment company to show that using their equipment leads to better memory, we could have a conflict of interest. You should know that a conflict of interest doesn't guarantee that a researcher will bias the results, but it creates a situation in which a researcher is more motivated to influence the outcome.

**Opportunity for bias.** Opportunity for bias is a flaw in which the experimenter does not take important precautions to reduce his or her chances of affecting the behavior of the participants. The best precaution is using a double-blind technique in which the researcher is unaware of which condition the participant is in. A second method is to reduce contact with the participants. Most experiments are automated so that a computer presents the instructions rather than the experimenter doing so. A third method is to use automated or objective scoring of data. If that is not possible, then two raters should score all data independently (inter-rater reliability) without knowing the participant’s condition (double-blind). For our memory experiment, if we use a recall rather then multiple choice test of memory, then the experimenter will have to determine which recall responses are correct and which are incorrect. This is an opportunity for bias. Therefore, two raters who are unaware of conditions should be used. We need to reduce the opportunity and motivation for experimenters to bias the results of an experiment, or we will not be able to rely on the results.

C. Problems with the sample

(6) **Use appropriate sample size**

Of course, we can only test a hypothesis on a sample from the larger population. There are several possible flaws that can occur in an experiment that could be due to the characteristics of such a sample. We will only focus on those that relate to the size of the sample. In experiments, we use statistics to compare performance on the DV(s) to see if there is a difference between conditions. The size of the sample that we need to get reliable results is determined by how variable the population is. **Variability** refers to the extent to which the population’s scores on a DV tend to naturally vary, regardless of any sort of manipulation.

Let’s say we have two conditions (exercise and no exercise) in the memory experiment above. Imagine using a highly variable population such as 5 year-olds worldwide to study the relationship between exercise and memory. It is likely that there is automatically a great deal of difference in the memory scores of all 5 year-olds even prior to any sort of exercise manipulation. When scores vary a lot in a population, and we draw small samples, our groups are likely to be quite different at the outset of the study, because people with extreme scores (high or low) have a big impact on the group average. As we increase sample size, however, individual extreme scores get increasingly balanced out by other scores, leading to more comparable groups. Typically, we have little information about how much a variable varies in a population; so, it is always a good idea to have large samples. There are primarily two flaws associated with an inadequate sample size: lack of power and low generalizability.

**Small sample size and lack of power.** **Power,** or the likelihood of finding a difference that really does exist between groups, depends on the size of the sample. Power increases with larger samples.
because, as we have seen, there is less extreme variation in the groups. Therefore, a null effect found in a study that uses small samples could simply be due to the lower power that comes with smaller sample sizes. In other words, there could really be a difference between groups, but it is not evident in the study because the sample sizes are so small.

**Small sample size and limited generalizability.** Researchers are always interested in results that we can have some confidence are most likely true for all members of a population. The second flaw associated with small sample sizes is the limited extent to which we can generalize a result obtained with a small sample to the larger population of interest. So, you may find a significant difference between two groups, but if the samples are small it is difficult to generalize these results to other individuals. Remember that scores vary in a population, and the chances of having unusual (or unreplicable) results are higher in small samples where extreme scores have a large impact on our statistical measures.

One thing to keep in mind is that there is no perfect size of sample to use – not even 30 subjects. It all depends on the variability of the population. If the population is highly variable (e.g., adult humans), then 20 subjects may not be enough to detect a true difference or to generalize to the population. If the population is relatively homogeneous (e.g., potatoes from a certified lab supplier), then 4 subjects may be large enough.

**D. Problems with Conclusions**

(7) **Make tentative conclusions**

When researchers finish collecting data for a study, they analyze the data and prepare a manuscript to report the results of the study to others in the scientific community. The results of the study should always be reported using tentative conclusions. That is, instead of talking of “proving” a hypothesis true, you will always word conclusions in a way that acknowledges the possibility of future revisions of the hypothesis. This is because findings should be considered in light of the limitations that studies typically possess. Two flaws can occur when a researcher is making a conclusion.

**Conclusions too strong.** The first flaw involves drawing too strong of a conclusion. This occurs when a researcher fails to acknowledge the possibility of alternative (and at times, better) explanations for the results. A researcher’s conclusion must always reflect the limitations of an experiment, and the potential for error. Even in the absence of obvious problems with an experiment, there is always the possibility that some effect was due to chance (i.e., any factors other than the IV) and not to the manipulation.

**Need for replication.** The second flaw related to the tentative nature of the conclusions that are reported has to do with the need for replication in research. Replication involves repeating a study with the same (or very similar) methods but with different participants and sometimes with different researchers. When many different experimenters obtain similar results when conducting a similar study with different subjects, we can be more confident that the results generalize to a larger population. To the extent that participants are varied in replications, we gain information about how well an effect holds across age groups, countries, cultures, species, and so on. Furthermore, a consistent result allows researchers to make stronger conclusions.

(8) **Requirements for causal conclusions**

Recall that the goal of an experiment is to try to find support for your hypothesis and if possible rule out alternative explanations at the same time. To find support for a hypothesized causal relationship, you need to consider several things. First, consider your IVs and DVs. Your DV must be valid and sensitive. Your IV should also be valid and manipulated by the experimenter. This means that participants are randomly assigned to conditions with at least one condition being a good comparison group. Second, you want to eliminate unwanted variables, such as reducing attrition (especially unequal drop out among conditions) and limiting the effects of the experimenter. Finally, you need to make sure to have a large enough sample which depends on the variability in the population.

This reading provides only a subset of the things you will want to consider about IVs, DVs, and controls. But, we think these are a good start to helping you think about testing hypotheses through research and the type of conclusions you can make from those studies. We hope this chapter increases your interest in finding out more about evaluating research.
Using the score card below, determine the flaws, if any, with each experiment description. If there are none, write “good experiment”. No credit will be given for questions such as “how many subjects were there?”. It is either clearly a flaw (So do not put in comments like “it doesn’t say how many subjects there were”).

A. Possible problems with IVs and DVS

1. **Use valid measure of IVs and DVs** – manipulate (IV) and measure (DV) the variables you claim you are, especially the most important aspects of the variable. This is important because valid measures are required to be able to appropriately interpret the results from an experiment and draw appropriate conclusions. Flaw: (1) **Validity of DV**

2. **Must have comparison group(s) – randomly assign** participants to one or more groups that are not exposed to the treatment or are exposed to less of the treatment so you can rule out alternative explanations for any effects. Flaws: (1) **Poor or missing comparison group**, (2) **No random assignment**

3. **Use appropriately sensitive DVs** – use tests or measures that people do not score so well or so poorly on that you fail to detect a true treatment effect just because you have insensitive measures. Flaws: (1) **Sensitivity of DV**, (2) **DV is not scored objectively**

B. Problems with unwanted, extraneous or confounding variables

4. **Limit attrition** – try to limit the number of participants who failed to complete the experiment or who skipped information because this can lead to differences between groups that affect the DV, which makes it impossible to conclude that the IV was responsible for any DV differences between groups. Flaw: (1) **Mortality or attrition**

5. **Limit experimenter bias** – try to limit the impact of the behaviors or expectations of the researcher on the results of an experiment. When this happens, it is not possible to determine whether the results were due to the IV or to actions of the experimenter. Flaws: (1) **Experimenter bias**, (2) **Conflict of interest**

C. Problems with the sample

6. **Use appropriate sample size** – make sure that you have a large enough sample to detect a true effect of the treatment and a large enough sample to generalize to the entire population. Flaws: (1) **Small sample size**; (2) **Poor sample selection**

D. Problems with Conclusions

7. **Make tentative conclusions** – experiments always have limits, and conclusions must reflect the possibility that the conclusions can be shown to be incorrect. **Replicating** the experiment and getting the same result helps increase our confidence in the results but will never prove the hypothesis true. Flaws: (1) **Not a tentative conclusion/ Premature generalization of results**

8. **Requirements for causal conclusions** – to make a causal conclusion, one has to randomly assign participants to a valid IV which has at least one comparison condition, measure a DV that is sensitive and valid, and then control unwanted variables (e.g., limit attrition and experimenter bias). Flaw: (1) **Inappropriate causal statement**