Evaluating experiments (part 1)

Andy J. Wills

2016

1 Correlational research

Correlational research in psychology is fundamentally limited. By correlational research, I mean that approach to psychology where one discovers that two variables co-vary, and conclude from this evidence that one may cause the other. As the website "Spurious Correlations" illustrates, this method of research can lead to some daft conclusions. For example:

- Increased suicide rates in the US may be caused by the increase in science spending.
- Nicholas Cage's appearance in films may increase deaths by drowning.
- Eating cheese may increase the risk you will be strangled by your bedsheets.

1.1 Piracy and global warming

It is demonstrably the case that average global temperatures have been rising over the last few hundred years. It is also demonstrably true that the number of cases of piracy on the high seas has dropped over the same period. There is clearly an association — a correlation — between these two variables.

Does this mean that global warming is caused by the absence of sea pirates? Put another way, would an effective way of reducing global warming be to encourage piracy? Or perhaps it means that the absence of sea pirates is caused by global warming? Put another way, would an effective way of reducing piracy be to increase our carbon emissions?

1.2 Depression and memory

The piracy example was deliberately ridiculous, but the same problematic inferences are made in much of psychology. Any study that is purely correlational is not demonstrating cause. For example, it's well documented that depression is associated with over-general memory. Specifically, those with a history of depression seem to be worse at recalling the specifics of episodes (what, where, when) than those without such a history.

Does this mean that depression causes memory problems? Or perhaps that memory problems cause depression? Or perhaps both causal directions are in force? Depression causes memory problems which makes it difficult to recall information about the past which in turn leads to becoming more depressed? Or perhaps depression and memory problems are both caused by some third factor? (e.g. childhood trauma). All of these theories are to be found in the literature, but they cannot be distinguished on the basis of correlational data.

1.3 Longitudinal data (2 slides)

Using longitudinal data does not solve the problem. For example, there's a famous (and accurate) result that the use of night lights in a baby's room is associated with myopia in their later life. Babies with night lights are more likely to be short-sighted in later life than babies without night lights.

This feels causal, because the assumed cause (night lights) occurs earlier in time that the assumed effect (myopia). Causes must precede effects, so one can eliminate the alternative explanation that myopia in later life causes the presence of night lights in infancy. It feels like there's only one option left — night lights cause myopia. We should ban night lights, because this would reduce myopia. Indeed, such recommendations were made on the basis of this research.

What longitudinal research does not rule out is the presence of a third factor that causes both the night lights and myopia. For example, myopia has a genetic component. Thus, if your parents are myopic, there's a good chance you will become so, too. In addition, myopic people tend to prefer more brightly-lit environments (perhaps because this compensates for the myopia somewhat). So, having myopic parents could be the cause of both the presence of night lights and the later myopia.

1.4 Correlation does not imply causation

In summary, correlational research — including longitudinal correlation — cannot establish causation. It is thus fundamentally limited. The limitation is much bigger than is sometimes realised, because it is extremely unlikely that any two variables are completely unrelated. It's a complex, inter-related world, and we're complex, inter-related beings. Many of the correlations we observe in psychology are extremely small. For example, in the study of personality, it is not unusual for a measure such as extroversion to explain as little as 5% of the variation in another variable. This is detectably different from no association (with methods you'll be taught about later in your degree), but it's frankly silly to have 'no association' as your competing theory because more-or-less everything is associated to more-or-less everything else to some degree.

2 The Experimental Method

The best way we have of establishing causation is through the Experimental Method. In its simplest form, we take two groups of people. We do different things to those two groups, and measure something. The difference in what we do is the called the independent variable. What we measure is called the dependent variable.

2.1 Depression therapy example

For example, let's say we're testing a new cognitive therapy for depression. We take two groups of people. One group gets six weeks of the new therapy. The other group get nothing. At the end of six weeks, we take some measure of depression (e.g. the Beck Depression Inventory). The group that get the therapy are less depressed than those who do not.

Such an approach has the POTENTIAL to show that the therapy CAUSES a reduction in depression. But there are a number of ways in which a conclusion of causation might be unsound.

2.2 Pre-existing differences (3 slides)

The first is possibility of pre-existing differences. What if the group who received therapy were happier from the outset than those who were not? We can address this problem in two ways - detection, or prevention.

Detection — We could (and should) have taken BDI ratings before the therapy started. If the two groups were comparable on BDI before the treatment, this rules out a pre-existing difference in happiness.

Prevention — We construct our groups such that we eliminate the possibility of pre-existing differences. There are two main approached to prevention — matching, or randomization.

Matching — Take BDI measures for everyone. Allocate people to groups in such a way that the average BDI for the two groups is identical (or at least, minimally different).

Randomization — Allocate people to groups randomly.

Matching is technically superior, if you are reasonably confident you know what the relevant variables are, and if there are relatively small number of them. Here, we just used BDI, but what about age? Number of previous episodes of depression? All these things might affect how much BDI varies over six weeks (with or without treatment). In practice, we often randomize. Random allocation has the advantage that all variables (measured or otherwise) will be well-matched if your groups are large enough. The problem is that your groups have to be very large for this to be likely to be true.

2.3 Quasi-experimental designs

Prevention of pre-existing differences is not always possible. Sometimes, we have the groups we are given and have to make the best of it. This situation is described as a quasi-experimental design. Don Campbell famously referred to them as "queasy experimental designs", because they made experimentalists feel sick. A more neutral term is nonequivalent groups design (although there are other types of quasi-experimental design, not covered in this course).

Detection of pre-existing differences (pre-post testing) is particularly important for quasi-experimental designs.

2.4 Back to our therapy experiment

So, if we have very large randomized groups, no pre-treatment difference in BDI, but a post-treatment difference — can we then conclude that the treatment CAUSED the reduction in depresssion (as measured by BDI) ?

Not quite yet. Next, we have to look long and hard for possible CONFOUNDING VARIABLES.

3 Confounding variables

A confounding variable is any variable, other than the one you are attempting to study, that varies between conditions. There is no magic formula for detecting confounding variables, because they depend on what you are trying to find out. What follows is a series of well-known confounds psychologists need to be aware of.

3.1 Attrition (3 slides)

Let's start with attrition. Attrition is where some participants drop out before the end of the study. This is not in itself a critical problem, but it becomes one if the attrition rates are different between your conditions.

Returning to our depression therapy example, let's say that 20% of people receiving the therapy decide it is not for them, and drop out of the study. We therefore do not have post-treatment BDI for these people. Our control group — who do not receive treatment — are likely to have a different attrition rate. Let's assume in this example that the attrition rate in the control group is 0%.

Let's further assume, and this is critical, that attrition is not random. For example, let's assume that the more depressed you are to start with, the more likely you are to not complete the therapy.

In the example, the 20% most depressed of the therapy group drop out, and both the therapy and the control have no effect on BDI. Yet the therapy group still looks like it's happier post-treatment than the control group, but it's an illusion caused by differential, non-random, attrition.

3.2 The Hawthorne Effect

Another major issue to look out for is specific to the fact you are testing living beings. The classic, and earliest, demonstration of the issue gives it one of its names — the Hawthorne effect. This refers to a 5-year study in the Hawthorne Works near Chicago. One of their products was electrical relays — these were assembled by hand. Productivity (relays per day) was recorded secretly for two weeks. Then a group of six workers were chosen for a study on how various factors affected productivity. The procedure involved discussing changes with the workers and at times using their suggestions.

One series of investigations started straight forwardly enough. After discussion, they shortened the working day by 30 minutes. This led to an increase in productivity. After further discussion, they shortened the day even more — productivity (at least in terms of relays/hour) went up again. So far, this looks quite straight forward — evidence that over-long working days can hit productivity (perhaps due to fatigue, boredom, etc).

Then they had further discussions, which resulted in going back to the original length of working day. Productivity went up again!

Although there are many interpretations of the Hawthorne Works experiments, and much criticism, the central point is undeniable. Humans are complex systems, and what the experimenter thinks she is changing is not always the only thing (or even the most important) thing that is changing from the participant's perspective.

One popular interpretation of the results of the Hawthorne Works experiments is that the increase in productivity is caused by the increased motivation, sense of worth, importance, and so forth, that comes from the experimenters selecting a group of workers, being interested in them, monitoring their behaviour, and discussing options with them.

3.3 Placebo effect (3 slides)

A related and ubitiqious phenomenon is the placebo effect. The classic example of a placebo effect is that a pill with no active ingredients, that a participant believes to be a headache remedy, reduces headache symptoms. The participant's expectation that the treatment will be effective is sufficient to reduce symptoms.

The lesson, often not heeded in drug testing, is that in order to assess whether your drug is effective, you need to compare against a placebo control. So, taking the drug versus taking the placebo. Rather than taking the drug versus not taking it. It is now widely believed that the effects of anti-depressant medication is almost entirely placebo. The question is harder to address, but no less important, for psychological treatments.

Returning to our therapy example, one possibility is that the therapy itself is basically ineffective. The treatment group are happier than the control group because the treatment group have the expectation they are receiving something that will work, whilst the control group — having received nothing — will not have this expectation.

In psychological therapy, there is no equivalent to the sugar pill — no treatment that will consensually be endorsed as entirely inert. However, one can (and should) attempt to show the new treatment works better than existing treatment (or, as well as existing treatment, if the new treatment is better in some other way e.g cheaper).

Unfortunately, this is seldom examined. Where it has been examined, the results are that pretty much anything works better than no treatment, but treatments do not differ from each other. For example, posting people a short DIY pamphlet on cognitive behavioral therapy appears to work as well as six weeks of one-to-one sessions with highly trained therapists.

3.4 Demand characteristics

Another concept related to the placebo effect is that of demand characteristics. This is just a name for the idea that participants' responses are affected by the desire to comply with what they think the experimenters want to see.

For example, there's an effect called evaluative conditioning. The idea is that pairing something neutral with something people already like increases their liking of the neutral item. Belief in evaluative conditioning underlies much advertising. For example, I show you a picture of a soft drink can, and follow it by pictures of beautiful smiling people. Assuming you like beautiful smiling people, the idea is that this makes you like the drink more. This can be shown experimentally — get liking ratings of the drink, pair it repeatedly with something positive, take liking ratings again. Liking ratings go up, and do not go up in a control condition where drink and smiles are presented, but unpaired.

The claim often made for this kind of experiment is that evaluative conditioning increases the liking of neutral stimuli. An alternative explanation is that participant thinks - "What's going on here? The experimenter is showing me this coke can and then smiley faces. I think they expect me to like coke more as a result. I wouldn't want to disappoint them so sure, let's give it a higher rating than I did last time".