

Evaluating experiments

Andy J. Wills

Spurious correlations

<http://www.tylervigen.com/spurious-correlations>

Depression and memory

- ▶ Depression is associated with over-general memory.
- ▶ Depression causes memory problems?
- ▶ Memory problems cause depression?
- ▶ Both causal directions?
- ▶ Neither causal direction (e.g. both caused by childhood trauma).
- ▶ It is not possible to distinguish between these accounts on the basis of correlational data.

Longitudinal data does not solve this problem

- ▶ Use of night lights in infancy is correlated with myopia in later life (true).
- ▶ Seems causal? Causes must precede effects. The later myopia cannot cause the earlier use of night lights. So, night lights must be causing myopia?
- ▶ Ban night lights? (genuinely recommended on basis on these data).

Third factor explanations are still possible in longitudinal research

- ▶ A third factor causes both the presence of night lights and myopia.
- ▶ Developing myopia in later life has a genetic component. If your parents are myopic, this increases the chance you will become myopic.
- ▶ Myopic adults, on average, favour higher levels of illumination. This drives their decision to use night lights in their baby's room.
- ▶ The parents' myopia causes both the presence of infant night lights and later myopia.
- ▶ Ban night lights? Clearly, this would be ineffective.

Correlation does not imply causation

- ▶ Correlational research is fundamentally limited.
- ▶ It is extremely unlikely that any two variables are completely unrelated.
- ▶ Many correlations in psychology are very small e.g.
 - ▶ Extroversion explaining 2% of the variation in some other variable.
 - ▶ 2% is detectably different from no correlation
 - ▶ but not meaningful (everything likely to be related to some degree).

Determining causation through the Experimental Method

- ▶ Simplest form
 - ▶ Take two groups of people
 - ▶ Do different things to those two groups.
 - ▶ Measure something
- ▶ Independent variable - Intended difference in what we do to the two groups
- ▶ Dependent variable - The thing we measure

Example: Testing a treatment for depression

- ▶ Group 1 - 6 weeks of the new therapy
- ▶ Group 2 - Nothing.
- ▶ Take measure of depression at end (e.g. Beck Depression Inventory).
- ▶ Group 1 are less depressed than Group 2
- ▶ This has the potential to show that the therapy *causes* a reduction in depression.
- ▶ ...but there are other explanations.

Pre-existing differences

- ▶ Group 1 - 6 weeks of the new therapy
- ▶ Group 2 - Nothing.
- ▶ What if Group 1 were happier to start with?
- ▶ Approaches to this problem
 - ▶ Detection
 - ▶ Prevention

Detection

- ▶ Take pre-treatment measures
- ▶ e.g. Measure BDI of both groups before (and after) treatment period.

	Pre	Post
Therapy	25	5
Control	25	25

Prevention

- ▶ Construct groups such that we eliminate pre-existing differences.
- ▶ 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, minimised).
- ▶ Randomisation - Allocate people to groups randomly.
- ▶ Matching versus Randomisation - pros and cons.

Our therapy experiment

- ▶ Use large, randomised groups.
- ▶ Take pre-treatment measures
- ▶ Treatment caused the reduction in depression?

	Pre	Post
Therapy	25	5
Control	25	25

Attrition

- ▶ Attrition - participants dropping out before the end of the study
- ▶ If attrition rates vary between conditions, you may have a major problem.

Example

- ▶ Pre-treatment BDI scores

						Mean
Therapy	6	8	12	15	30	14.2
Control	6	8	12	15	30	14.2

- ▶ The most-depressed 20% drop out of therapy (perhaps because the therapy is quite demanding).
- ▶ There are no drop-outs in the control condition (there's not much to drop out from).
- ▶ Both therapy and control are inert (no effect) - post-treatment BDI equals pre-treatment BDI.

Example

► Pre-test BDI scores

						Mean
Therapy	6	8	12	15	30	14.2
Control	6	8	12	15	30	14.2

► Post-test BDI scores

						Mean
Therapy	6	8	12	15		10.25
Control	6	8	12	15	30	14.2

- A therapy we know to be ineffective appears to have worked, due to non-random attrition.

Placebo effect

- ▶ Classic example
 - ▶ Someone has a headache
 - ▶ Give them a pill with no active ingredient
 - ▶ Tell them it's a headache tablet
 - ▶ Their headache symptoms reduce
- ▶ Lesson - In order to assess drug effectiveness you need to test drug vs. placebo, NOT drug vs. nothing.

Placebo effect in psychological therapy

- ▶ Perhaps the therapy is inert?
- ▶ The treatment group are happier because they have the expectation that what they are receiving will work.
- ▶ Problem - a placebo pill is known to be inert; what is the equivalent in therapy?
- ▶ There is no agreement - there's someone willing to endorse the effectiveness of almost any therapy.

Placebo effect in psychological therapy

- ▶ Solution - set out to show that your new therapy works better than an existing treatment (or, as well as existing treatment, if yours is better in some practical way e.g. cheaper).
- ▶ Problem - this is seldom done.

Experimenter Effects - Data analysis - Example

- ▶ Diary entries as a measure of happiness.
- ▶ Participants write about their feelings
- ▶ Experimenter rates for level of happiness.
- ▶ If experimenter knows which condition the participant is in, this may bias their assessment of happiness.

Experimenter Effects - Data analysis

- ▶ Machine-recorded measures (e.g. reaction time) immune?
- ▶ No! - Data analysis typically involves many decisions, all open to bias.
- ▶ If the experimenter knows which condition the participants are in, this could bias their decisions.

Blind testing

- ▶ Single-blind testing - participant does not know which condition they are in.
 - ▶ e.g. Drug vs. placebo. Participants do not know which condition they are in.
- ▶ Double-blind testing - single-blind testing plus the experimenters do not know which condition is which until after they have completed their analysis.

Pre-registration

“The first principle is that you must not fool yourself, and you are the easiest person to fool” - Richard Feynman.

- ▶ Record your hypothesis, method, and analysis plan, before you analyse the data.

Difference versus no difference designs

- ▶ The preferred hypothesis is that people differ in the speed with which they react to auditory and visual alarm signals.
- ▶ The alternative theory against which this is compared is that there is no difference (nil hypothesis).
- ▶ Problem - Experimental control is never perfect.
- ▶ Thus - the nil hypothesis is almost certainly wrong, and detectably so if you test enough people.
- ▶ Thus - the result of the study is known before you run it.
- ▶ Thus - There was no point in running it.

Better alternatives 1

- ▶ Directional hypotheses

- ▶ The preferred theory is that auditory is faster.
- ▶ The alternative theory against which this is compared is that there is no difference (nil hypothesis).
- ▶ If you find visual faster, you have disproved your theory.
- ▶ So, whatever the result, there was a point to running this experiment (because the theory was falsifiable).

Better alternatives 2

- ▶ Strong inference
 - ▶ One well-established theory predicts that auditory is faster.
 - ▶ Another well-established theory predicts that visual is faster.
 - ▶ Whatever you find in this study, you've gained information (except in the unlikely case where the nil hypothesis was true).

Evaluating an experiment

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological science*, 25, 1159-1168.

1. Find the full text of this paper on Google Scholar
2. Read from the title up to, but not including, the “Study 2” subheading.
3. Evaluate how good Study 1 is, using the *checklist* to help remind you of what we’ve covered today.
4. Agree on a score, be ready to report your score, and to answer some questions.

Further reading/ watching

The notes for this lecture cover a number of additional relevant topics.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International Licence. This license does not cover the work contained in the documents linked above. Last update: September 23, 2021