Evaluating experiments (part 2)

Andy J. Wills

2016

1 Confounding variables (2 slides)

These slides summarize the end of the previous lecture; they contain no new information.

2 Allocation of markers example (2 slides)

This is an example of using the Experimental Method in a practical way to solve a real-world problem. In the imaginary HE establishment "Rustling University" class sizes are very large so, in order to get coursework back to students in a reasonable amount of time, multiple people are involved in the marking of each assessment. In this particular case, there are 300 students and two markers. The marking is split equally between the two markers. Specifically, the first marker gets the first 150 scripts that are handed in, the second marker gets the rest. After marking, we find that Marker 1 gave an average mark of B+, whilst Marker 2 gave an average mark of B.

We considered what might explain the difference between markers, and compared two responses the university might make. On the one hand, it might use a correlation analysis to see if there is a significant relationship between date of submission and grade. On the other hand, it might decide to, in future, randomize allocation of scripts to markers. Randomization is better.

3 Therapy example (3 slides)

This is an example of experimenter effects. We considered an experiment that compared meditation-based therapy with relaxation training, and looked at effects on the BDI. The effect is that the meditation-based therapy improves happiness more than relaxation training. Pre-existing differences have been controlled for. There's no differential attrition. Sounds compelling?

However, in this study, the meditation-based therapy is delivered by the people who developed it. The relaxation therapy is delivered by people who have no particular investment in relaxation therapy, but have been on a one-week training course in relaxation therapy.

An alternative hypothesis is that is it not the type of therapy that matters but some combination of how much the therapist believes in the treatment, how much experience they have in delivering it, and possibly their level of generalized expertise at making people feel happier.

You could control for this (or at least get someway towards controlling for it) by ensuring the levels of expertise, ideological commitment, and years of therapeutic experience were matched between conditions.

This is seldom done. What we do know, however, is that the effectiveness of a therapy is often revised steadily downwards as more and more studies are performed. The first studies are typically performed by those with a strong belief in the effectiveness of the treatment, the later studies by researchers who are more agnostic.

4 Experimenter Effects — Data analysis (3 slides)

Another form of experimenter effect is bias in data analysis. This is perhaps most acute when the dependent measure is subjective. For example, say you use diary entries as a measure of happiness. Participants write about their feelings and the experimenter rates these entries for level of happiness. If the experimenter knows which condition the participant is in, this may bias their assessment of happiness.

Although not so often talked about, the problem can still occur with entirely objective measurements (e.g. reaction time). The problem occurs because (as you will increasingly see throughout your degree), data analysis involves a number of steps, all of which have the potential for bias. Should I exclude outliers? What should be the cut-off? Should I use a parametric or non-parametric test? Should I correct for multiple comparisons? If the experimenter knows what condition the participants are in, this knowledge could bias their decisions in a way that favours the experimenter's hypothesis.

For example, say my theory predicts RT is higher in condition A than condition B. I find this result if I exclude people with reaction times > 3 seconds, but not if I keep everyone in, and not if I exclude everyone with reaction times of less than 100 ms. I go with the >3 second exclusion. Am I sure that decision was unaffected by the fact it's the only decision that results in a conclusion favourable to my theory?

5 Blind testing

There's a relatively straight forward answer to experimenter-effect problems in data analysis — blind testing. When analyzing data, make sure you do not know which group is which. This is easy to achieve — get someone else to replace the meaningful labels in your data set with meaningless ones, and ask them to withold the mapping from you until the analysis is complete.

The central concept here is called "blind testing". Single-blind testing refers to experiments where the participant does not know which condition they are in. For example, the experiment has one condition with an active drug, another with a placebo. Participants are not told which condition they are in. Double-blind testing is single-blind testing plus the experimenters do not know which condition is which until after they have completed their analysis.

6 Order effects (3 slides)

Everything we've discuss so far has made the assumption that the two groups (treatment and control) contain different participants. This is a between-subjects design. Experiments can also be within-subject designs. Where one employs within-subject designs, it's critical to consider order effects.

For example, imagine an experiment to assesses whether people react more quickly to visual or auditory alarm signals. For a within-subjects design, you might do something like

study shown on the slides. From the results on the slide, it looks like people can react to auditory signals faster. But auditory signals also come later in the experiment. So perhaps people have had more practice of the general experimental set up, and are getting faster for that reason.

So, we now try the experiment the other way around. Looked at in isolation, this second set of results could be a fatigue effect. It's a long, boring experiment and people slow down later on because they are bored.

However, if we control for order effects properly, we randomly allocate half the participants to each condition and (in this example) find auditory signals are faster, irrespective of order.

7 Difference versus no-difference designs (4 slides)

These are common in psychology, but they are still weak ways to run experiments. Avoid them if at all possible. The term refers to expeirments where the alternative against which your preferred theory is tested is a null (no difference). For example, your preferred theory is that people differ in how quickly they react to auditory and visual alarm signals. Your alternative is that there is no difference. Often called a "null" hypothesis, but perhaps more properly a "nil" hypothesis — a hypothesis of no difference.

Difference versus no-difference designs are problematic because experimental control is never perfect. For example, it is never going to be the case that modality (auditory versus visual) is literally the only thing differing between the conditions of your experiment. So, the null/nil hypothesis is almost certainly wrong, and detectably so if you test enough people. Thus the result of the experiment is known before you run it, and thus there was no point running it at all.

Having a "one-tailed" test is a bit better. Your preferred theory here is that e.g. auditory is faster than visual. Your alternative is that there is no difference. It's possible your experiment would disprove your preferred theory — if visual turns out to be faster than auditory. So, there was a point to running this study. Even better, compare theories that make opposite predictions. For example, one well-established theory predicts auditory > visual whilst the other predicts visual > auditory.

Or, define effect size (how large an effect you expect to see). This is often done in experiments whose results have immediate practical applications (e.g. drug trials).

8 Statistical control myth

Another common problem is the assumption you can "control for" group differences statistically by something called Analysis of Covariance. You'll be taught this statistical method later in your degree, and it has its uses, but "controlling for" group differences is not one of them.

For example, say you have two groups. Group 1 are depressed — you selected them on the basis of BDI score. Group 2 are nondepressed, selected on the same basis. These are clearly neither randomly allocated nor matched groups. You find that the depressed group do worse on a memory test than the nondepressed. Depression causes poor memory? Poor memory causes depression? Not only can you not decide causal direction, it turns out you can't even say depression is the relevant variable. You look at your two groups and notice that the depressed group are, on average, older than the non-depressed group. Perhaps it is old age, rather than depression, that is associated with poor memory?

Some people hold the belief that you can "control for" the age difference by entering age as a covariate into your analysis. You can't. As you haven't been taught the method yet, I won't go into details, but I strongly suggest looking at the Cohen paper cited on my slides you ever are encouraged to do this. It's a really good explanation of why ANCOVA does not allow you to "control for" pre-existing differences. Control for pre-existing differences using matching, or randomization, as discussed in the previous lecture.

9 Evaluating a research paper

In this section, I discussed how to evaluate an entire research report. Evaluating a research report is the bread-and-butter of scientific investigation. It also is central to writing essays and lab reports. As this is my last lecture, some of this material is as much for the future as for now. Hang on to this list, and refer back to it throughout your degree. You should be able to get the gist now, but there are parts of it that may not make sense until towards the end of your degree.

As with evaluating arguments, we have an Infalliable List — this time a checklist. If you are able to answer all these questions about the paper you are reading, you will have done a thorough and insightful evaluation. It also provides a good way of condensing your notes on what you read.

1. **Central topic** (What is it about?)
2. **Central research question** (What do they wish to test?)
3. **Central rationale** (Why bother?)
4. **Essence of Method** (How did they test their question?)
5. **Key Result**
6. **Central Valid Conclusion**

To the extent these questions can be answered clearly for a paper, it is a good scientific paper. This is of course a different question to whether it is correct — but that's science. Similarly, good scientific papers have confounds — the process of science includes the investigation of those confounds.

In the lecture, I illustrated the evaluation of a research paper with a concrete example. I used a fairly recent paper in my area of research, and showed how an evaluation of this paper led directly to us running a follow up study which overturned the conclusions of the original paper.

9.1 Central topic

Your first question about a paper should always be - what is this about? In other words, what is the central topic. Papers can often be complex, attempting to address a number of different, overlapping, questions. But there normally is, or should be, a central key theme — the thing that, if the paper is remembered at all, it will be remembered for.

The title often gives you a big clue. In this case, the main title is eye-catching rather than helpful — the researchers are not really removing the frontal lobes, as a quick look at the methods will confirm. The clue is in the sub-title.

The sub-title tells us it's something to do with learning. Category learning, to be specific. The authors don't do a particularly good job of defining category learning — this is often the case for terms that a psychology graduate would be expected to know. Of course, this means that sometimes you will not know what they mean. Textbooks and wikipedia are good sources here. Category learning is the act of learning how to divide the world into groups of things.

The sub-title also tells us the paper is something to do with executive function. Again, this is a term the authors assume the reader knows. A quick wikipedia or textbook search will tell you that executive function is a broad term, referring to the management of cognitive processes, and is known to involve the frontal lobes.

9.2 Central research question

So...we're getting somewhere, but it's a pretty broad definition of what is being studied. What we need next is the central research question. This can be distinguished from a topic in that a research question is something you can test empirically, while a topic cannot be tested directly.

The central research question is typically found in the Introduction. You should read the whole introduction, but often you will find the critical details of what the authors want to find out in the last paragraph of the Introduction. That is the case for the example paper, and sometimes it is easier to start there and work backwards.

Inevitably, skipping to the end like this means we have some terms we don't understand, but we can go with the flow and look up those terms in a minute. In essence, the idea is that some types of learning ("information integration") are helped by a secondary task, while other types of learning ("conjuctive learning") are hindered by that secondary task.

"Secondary task" won't be fully defined in the paper because it's a standard term, but you should be able to infer it from the intro and/or Method — it's getting the participants to do something else cognitively demanding at the same time as the learning task.

The terms "information integration" and "conjunctive" are specific to this research area, so we look to the Introduction for a definition. We find a definition in the form of Figure 1. So, now we know what their question is, but we don't know why they would bother to ask it. This is our next step.

9.3 Central rationale

Frankly, authors are often a bit vague about this, and it can be hard to pin them down. We are often left looking for clues rather than following a clear argument. It's often best at this point to read the whole Intro and Discussion and see what you can reasonably infer. In this case, the COVIS theory is prominently and repeatedly mentioned, so it's likely this question is related to the predictions of this theory.

Reading the paper in its entirety, it's clear that the COVIS theory predicts the phenomenon they wish to test. The paper is not actually that great on why COVIS predicts this. This is not uncommon. Your best best under such circumstances is to look for a review paper on the theory. For example, Ashby & Maddox (2005). For current purposes, the explanation is basically that II is hard to verbalize and hence learned unconsciously. Concurrent load stops people attempting to use verbal rules that won't work and rely instead on their unconscious system. That's why concurrent load helps. In contrast, CJ is verbalizable, and concurrent load will get in the way of discovering the rule.

9.4 Essence of Method

The next step in evaluating a paper is to look at how they set about to test their question of interest. This paper employs the Experimental Method, so the key questions are:

- What is the dependent variable (DV - thing that is measured)?
- Is the DV appropriate for the question?
- What are the independent variables (IV - things that are manipulated)?
- Are the IVs appropriate for the question?
- Are the IVs confounded with any other variables?

It's often easy to determine the DV by looking in the Results section. From this, we can see that the DV is accuracy. This seems entirely appropriate to the question. In this case we can also get the main IVs from the Results section. This really should be the case if the Results are described clearly. One IV is the type of category structure — conjunctive or II. This seems appropriate given the question. Ignoring the "2D" bars, which concern a secondary part of this paper, the other IV seems, from the graph, to be the presence or absence of "WM". Reading the Method makes it clear this is the presence or absence of a working memory load, specifically whether or not you have to remember some digits at the same time as learning the category structure. Given the hypothesis this seems an entirely reasonable choice.

The final, much harder question, is whether the IVs are confounded with any other variables. You can probably start from the assumption that they are - no experiment is perfect. So try and find one. Sometimes, authors will admit in print that their experiment is confounded or in some other way limited. You should find these statements towards the end of the Discussion.

Failing that, you have to get into the procedural aspects in some detail. Often drawing out what a single trial would look like is a good place to start. This is what my colleagues and I did, as shown on the slides.

Having done this, a central confound became apparent — those with a digit load also had a longer inter-trial interval — more time to think between trials. As it turned out, it was this difference, rather than the difference in digit load, that was underlying their result. This undermined both their result and their theory — but that's a story for another time — take a look at the paper if you like, it's on my website.

9.5 Key result

Now we move on to what they found — what was their key result. Often we'll already know this from finding out the Method, or the prediction. That is the case here, and you'll often find there is a single Figure or Table that contains the key result.

As your expertise in statistics grows, you should also ask whether the difference in the means is statistically significant, and whether the difference of differences (the "interaction") is significant. Here the interaction is critical, and is significant, as is the striking prediction that concurrent load increases accuracy for the II structure. However, we should note that — as is often the case — the data are not quite as clear as the authors' summary of them might suggest. Concurrent load in fact has no detectable effect on the CJ structure. Although a null result, this does not seem particularly consistent with the theory.

9.6 Central valid conclusion

Now, to round this all off, what is the Central Valid Conclusion of the study? By now, you should be focussing on integrating what the author thinks with any problems you have come across. So, this is a statement — not of what the authors claim to have done — but what you believe can be reasonably inferred from the study.

9.7 Last two slides

The last two slides contain my summary of the example paper. This is an illustration that most short papers can be summarized rather briefly.