Field, A., & Hole, G. (2002). How to design and report experiments. Sage.
pages 58 – 63

**Threats to internal validity**

Most of the factors that may reduce internal validity can be avoided by sound experimental design. Here are some of the most common threats to internal validity.

• Group threats: If our experimental and control groups were different to start with, we might merely be measuring these differences rather than measuring any differences that were solely attributable to what we did to the participants. Selection differences can produce these kinds of effects - for example, using volunteers for one group and non-volunteers in another, or comparing a group of undergraduates to a group of mental patients. 'Group threats' of this kind can largely be eliminated by ensuring participants are allocated to groups randomly. However, if you are looking at sex- or age- differences on some variable, group threats of some kind are largely unavoidable (more on this below).

• Regression to the mean: If participants produce extreme scores on a pre-test (either very high or very low), by chance they are likely to score closer to the mean on a subsequent test - regardless of anything the experimenter does to them. This is called regression to the mean, and it is particularly a problem for any real-world study that investigates the effects of some policy or measure that has been introduced in response to a perceived problem. Suppose, for example, the police had a crackdown on speeding as a consequence of particularly high accident rates in 2001: if accident rates decreased in following years, it would be tempting to conclude that this was a consequence of the police's actions. This might be true - but the decrease might equally well have been due to regression to the mean. Because accident rates were very high in 2001, they were more likely to go down in subsequent years than up - hey presto, you have an apparently effective traffic policy. The same kind of argument applies to interventions to help poor readers, depressives, 'alternative' medical treatments, etc.

• Time threats: With the passage of time, events may occur which produce changes in our participants' behaviour; we have to be careful to design our study so that these changes are not mistakenly regarded as consequences of our experimental manipulations.

• History: Events in the participants' lives which are entirely unrelated to our manipulations of the independent variable, may have fortuitously given rise to changes similar to those we were expecting. Suppose we were running an experiment on anxiety in New Zealand, a country known for its propensity to earthquakes. We test participants on Monday, to establish baseline anxiety levels, administer some anxiety-producing treatment on Wednesday, and test the participants' anxiety levels on Friday. Unknown to us, there is an earthquake on Thursday. Anxiety levels are much higher on Friday due to the earthquake, but we mistakenly attribute this increase to our experimental manipulations on Wednesday. (Don't worry, there are ways round this problem, coming shortly - and they don't involve avoiding doing research in New Zealand . . .)

• Maturation: Participants - especially young ones - may change simply as a consequence of development. These changes may be confused with changes due to manipulations of the

independent variable the experimenter is interested in. For example, suppose we were interested in evaluating the effectiveness of a method of teaching children to read. If we measure their reading ability at age four, and then again at seven after they have been involved in the program, we can't necessarily attribute any improvement in reading ability to the program: the children's reading might have improved anyway, perhaps due to practice at reading in other contexts, etc. In this case, it's pretty obvious that maturation needs to be taken into account. However, these kinds of effects can occur in more subtle ways as well. For example, in a pre-test/post-test design in adults, any observed change in the dependent variable might be due to a reaction to the pre-test. The pre-test might cause fatigue, provide practice, or even alert the participant to the purpose of the study. This may then affect their performance on the post-test.

• Instrument change: Good physicists frequently calibrate their equipment. This guards against obtaining apparent changes in what they are measuring merely because their measuring device has changed. Imagine working in a nuclear power station and concluding that it was safe to go into the reactor core, unaware that the 'negligible radiation' reading on your Geiger counter was due to the fact that the batteries had run down. Similar (if somewhat less dramatic) effects are less obvious in psychology, but may happen nevertheless: for example, interviewers may become more practised, or more bored, with experience. An experimenter may get slicker at presenting the stimuli in an experiment. Factors such as these may change the measurements being taken, and these changes may be mistaken for changes in the participant rather than in the measuring tool.

• Difrential Mortality: This sounds a bit dramatic! If your research involves testing the same individuals repeatedly, participants may sometimes drop out of the study for various rea- sons. This can make the results of the study difficult to interpret. For example, if all of the unsuccessful cases on a drug treatment program drop out, leaving us only with the successful cases, then a pre-test on the whole group is not com- parable to a post-test on what remains of the group. There might be systematic differences between the people who remain and those that dropped out, and these differences might be wholly unrelated to your experimental manipulations. In the current example, it might be that those who remained in the drug treatment program had higher levels of willpower than those who left.

• Reactivity and Experimenter Eflects: Measuring a person's behaviour may affect their behaviour, for a variety of reasons. People's reaction to having their behaviour measured may cause them to change their behaviour. I was once asked to take part in a long-term study on the relationship between diet and health: completing the dietary questionnaire made me realise that my diet consisted almost solely of pizzas, and so I changed my behaviour (well, for a while, at least). Perhaps I'll now live to a hundred as a consequence of my new healthy lifestyle, whilst the organizers of the study end up with the mistaken impression that living solely on pizzas leads to a long and healthy life. Merely measuring my behaviour caused it to change. There's a huge social psychological literature on 'experimenter effects': the experimenter's age, race, sex and other characteristics may affect the results they obtain (Rosenthal, 1966; Rosenthal and Rosnow, 1969). Experimenters can subtly and unconsciously bias the results they obtain, by virtue of the way in which they interact with their participants. Participants often respond to the 'demand characteristics' of an experiment (Orne, 1962, 1969) - that is, they try to behave in a way that they think will please (or, occasionally, annoy!)

the experimenter, for example by attempting to make the experiment 'work' by giving the 'right' data. Ideally, you could minimise these effects by using a 'double- blind' technique. This involves both the experimenter and the participant being unaware of the experimental hypothesis and which condition the participant is in. If the experimenter is as ignorant as the participant about what's going on, there's little opportunity for the experimenter to bias the results. Unfortunately, as a student, you are probably unlikely to have the resources to employ someone to run your experiment for you, and it has to be said that most psychologists don't bother with double-blind techniques either. Related to demand characteristics is the possibility that participants may show 'evaluation apprehension' (Rosenhan, 1969), a posh term for anxiety about being tested. Many non-psychologists seem to fail to appreciate that the experimenter is usually interested only in average performance, and isn't at all interested in the data of them as an individual. I've run experiments in which participants have treated the experiment as a test of their abilities, and have been so concerned with not looking stupid in front of me that they have failed to supply me with decent data! Finally, questionnaires may give rise to 'social desirability' effects, with respondents telling porkies about their income or sexual practices to look good to the experimenter. ('How many times have you had sex with a horse?' is unlikely to elicit many accurate replies!) Reactivity is especially a problem when obtrusive measures which are under the participant's control (e.g. verbal reports) are used. Participants may show practice or fatigue effects, or become increasingly aware of what the experiment is about. However even something as apparently 'low-level' as reaction times can be affected by these kinds of effects. Some studies have shown that the elderly have slower reaction times than young undergraduates. Some of this difference may be due to age-related cognitive decline, but it may also occur as a result of different age-groups adopting different strategies within the experimental situation. There is some evidence that the elderly are more cautious in novel situations and are more concerned to help the experimenter by making fewer errors. Both of these factors would conspire to increase reaction times in a way which could be mistaken for age-related physiological deterioration rather than an increased desire to please the experimenter.

So, there are lots of extraneous factors that can lead to changes in behaviour, changes that can be confused with the effects of our intended manipulations. Good experimental designs guard against ('control' for) all of these competing explanations for the changes in our dependent variable, and thus enable us to be reasonably confident that those changes have occurred because of what we did to the participants - that is, they are a direct consequence of our experimental manipulations.

**Threats to external validity**

• Over-use of special participant groups: McNemar (1946) pointed out that psychology was largely the study of undergraduate behaviour. Rosenthal and Rosnow (\975) found that, 30 years later, 70-90% of participants were still undergraduates. Research suggests that students have higher self-esteem, take drugs and alcohol less (hah!), and are less likely to be married than are other young people. Young people in general are lonelier, more bored and more unhappy than older people (try telling that to your granny). Using volunteers may also cause problems: Rosenthal and Rosnow (\975) found that the participants recruited as volunteers via adverts were more intelligent, better educated, had higher social status and were more sociable than non-volunteers. On the downside, volunteers for research into psycho-

pathology, drugs and hypnosis are more likely to have mental health problems. Volunteers generally have a higher opinion of, and respect for, science and scientists. That's nice, but a bit of a nuisance if it makes them respond differently than non-volunteers. As mentioned in the section on 'generality', the extent to which this is a problem depends on the kind of research being done: it's not automatically the case that it's invalid to use volunteer student participants. A student's visual system may be pretty much like that of any other human, even if their social behaviour is a bit strange!

• Restricted numbers of participants: This is more a threat to reliability, but it also affects one's ability to generalize to the population as a whole. Cohen (1988) has pointed out that most psychology experiments use too few participants for them to have a reasonable chance of attaining statistical significance. (See page 154 on 'power' so that you don't make the same mistake).

**Maximizing Your Measurement's Generality**

Closely related to external validity is the issue of whether our findings will generalize to other groups of participants in other times and places. This is usually taken for granted by psychologists. The best measure of generality is by empirical testing - replications of the experiment by other people in other circumstances. If food additives make children hyperactive in Chippenham, then they should also do so in downtown Kuala Lumpur. If they don't, that might be interesting in itself, but it would mean that we can't make sweeping statements about the effects of additives on humanity. Generality can be enhanced at the outset by representative sampling - by making sure that you have indeed used participants who are typical of the population that you want to make statements about. Sampling methods include random sampling, and stratified sampling (where the sample is deliberately constructed to mirror the characteristics of the parent population. So, for example, if the population consists mostly of 90% poor people and 10% rich people, so too does your sample). Threats to generality may come from using volunteers and undergraduate students. Generalization needs to be confirmed not only across participants, but also across experimental designs, methods, apparatus, situations, etc. The generality of findings will depend to a large extent on the kind of research that's being done. All other things being equal, the results from a study on basic cognitive processing are more likely to be generalizable than the findings from a study on the social interactions of city office workers, because there is probably greater scope for social and cultural influences to affect the results of the latter.