# 6: Threats to Validity

## Construct Validity

* Report to Board of Directors that students in the program had better grades, so the program must have worked.
* BUT, don’t actually know. And it could be somewhat of a stretch to argue that commitment to school can be reasonably proxied by test scores.
* Streetlight Effect – person looking where they can see rather than in the dark where the thing actually is, but they can’t see.
* Measuring abstract concepts is difficult because you’re looking in the dark zone.
* Are you really measuring what you want to measure???

## Statistical Conclusion Validity

* Want to ensure the statistics in your analysis are correct.

### Statistical power

* In order to find/detect an effect, you need to have sufficient data.
* Can see more accurately what the ‘null’ world/world with no data looks like i.e. the interval around 0 shrunk considerably.
* Sample size really matters.
* Need to have a big enough sample size to shrink the null world down to something reasonable.

### Test assumptions

* Ensure that all assumptions are reasonably met.

### Fishing and p-hacking

* Causal inference isn’t the same as machine learning. In any situation where you care about the values of the coefficients, you can’t ensemble model/iterate over different combinations of features until you find the results you want.

### Spurious statistical significance

* You can’t truly know if the p-value you find is spurious or not – no magical statistical test.
  + A 5% threshold means 95% of the time it’ll probably be right, but 5% of the time it could be in the opposite (‘wrong’?) direction just due to the nature of probability.
* You can measure the thing you’re interested in in different ways or add more coefficients to check if the conclusion changes.
* You could also recheck the DAG.

## Internal Validity

* Hidden threats or threats to the study itself.
* Selection and attrition are omitted-variable biases.

### Selection

* Subjects signing up for a program will be fundamentally different to those who do not.
* Often this is because these subjects need/want the treatment, and this ‘wanting’ is confounding.
* No easy way to fix the problem with observational data.
  + Specific statistical models like DiD or regression discontinuity.
  + DAGs.
* If time is a factor, ‘waiting’ subjects could be fundamentally different from subjects never intending to enrol in the program.
  + Change from absolute to relative time – doesn’t completely eliminate time-based selection bias but is better than nothing.

### Attrition

* Subjects who drop out prior to a study ending could be fundamentally different to those who don’t e.g. if something in the study is making them drop out.
* Need to figure out how to ensure attrition is as low as possible to minimise bias.

### Maturation

* Subjects in your treatment group will naturally develop.
* Need to isolate or remove that trend to observe the real impact of a program/intervention.

### Secular trends

* Periodic issues or other issues out in the world will distort results across the board.

### Seasonal trends

* Need to use statistical methods to remove the seasonality or compare similar periods.

### Testing

* Repeated exposure to the same test can lead to better performance just due to familiarity.
* Change tests – run the risk of not measuring the same thing.
* Not offering pre-tests – then don’t have a baseline result for comparison purposes.
* No good solution to fixing this issue.

### Regression to the mean

* Extreme values are rarer and often there’s a fair amount of luck involved.
* Being able to separate this ‘return to the mean’ from the true effect of the intervention is difficult.
  + Subjects with extreme values in the treatment group could be systematically different to others and hence may be a good idea to remove them.
  + No systematic way of performing this – justification required. (Or could run the model with and without these subjects and observe the effect on the results).

### Measurement error

* Related to construct validity.

### Time frame

* Need a goldilocks timeframe, which can only be chosen using subject area knowledge.

### Hawthorne effect

* Need to be aware of this problem, but may not be able to fix.

### John Henry effect

* Often occurs if the control group knows there’s an experiment being conducted and they know there could be some positive outcome (and they know they’re the control?).
* The control group is no longer the control, they’re also the treatment group and hence the results are distorted.
* Ensure that the two groups can’t interact.

### Spillover effect

* Similar to the John Henry Effect.
* Sometimes the treatment will have positive effects on the surrounding subjects/environment.
* If the system improves as a result of that intervention, it can make it hard to separate the impact of the program from this general improvement.
* Similar to the John Henry Effect, try to keep the groups very separate to avoid contamination e.g. different geographical locations.

### Intervening events

* Something happens to one group and not the other that’s out of the researcher’s control.
* No easy way to fix this. Impossible? Typically means the study needs to be thrown out.

## External Validity

### Generalisability

* Ensuring the study is generalisable to broader populations rather than anything to do with the internal functioning of the study.

### Lab conditions vs the real world

* Typical study participants are university undergraduates i.e. different from the rest of the world.
* Not everyone takes surveys.
* Need to justify why the results of the study are relative to a broader population.

### Different settings and circumstances

* Could do everything right in terms of study design but then need to justify why a study in one geographical location can apply to another geographical location.
* Could be systemic differences in the two locations.
* Could just run another trial in that location, but this is expensive.