CPP 524: PE2 – Research Design

Prof. Jesse Lecy

Campbell Scores

Instructions:

You will be evaluating the internal validity of a variety of case studies using a list of criteria that we are calling a "Campbell Score." Each item on the list represents an alternative hypothesis to the primary hypothesis, which is "the program created the effect that we see in the data."

For the Campbell Score, an item gets a +1 if the hypothesis can be eliminated – i.e. if the study has adequately controlled for that particular competing hypothesis. The item gets a +0 if the study does not include adequate controls and the item is outstanding at the end of the study. The control in each case will be different. See below for instructions for each item.

In some cases if the item is not relevant to the study because of the particular study questions and design. Testing, for example, is only an issue if (1) there is a pre-test and (2) it is a skill that could feasibly be enhanced by practicing the exams. So, for example, IQ tests are likely to create testing problems because there is an internal logic to the questions that becomes clearer as you answer more of them. Measuring blood sugar levels to see if a diabetes treatment is working, however, could not feasibly be subject to testing bias. If the specific item does not apply to the domain in the chapter then you just need to state that and explain why (the measured outcome cannot be changed because of testing), then the item should receive a +1 because we are not worried about bias resulting from the failure to address the item.

Note that you have here general guidelines, but that each study will have its own particular idiosyncrasies, both in terms of the research questions as well as in terms of how the authors chose to present data and results. In many cases you will have to use your best judgment to determine whether the item receives a zero or one.

To simplify the assignment, we will talk about the outcomes in binary terms – the program had a detectable impact, or the program did not have a detectable impact. In the real world, you would want to go a step further and establish whether the impact has social or economic meaning above and beyond statistical significance.

Please also think through the following questions as you read the studies:

(1) What is the group or scenario that represents the counter-factual?

(2) If so, is it a 'control' group or 'comparison' group?
(3) What is (are) the dependent variable(s) in the study?
(4) How is the program effect calculated (T2-C2, T2-T1, diff-in-diff, etc.)?
(5) Does it measure intention to treat, or treatment on the treated?

You do not need to report the answers to these, but they will be helpful in reviewing the study.

The Campbell Score Items

Omitted Variable Bias

(1) Selection / Omitted Variable Bias
(2) Non-Random Attrition

Trends in the Data

(3) Maturation
(4) Secular Trends
(5) Seasonality
(6) Testing & Hawthorne Effects
(7) Regression to the Mean

Study Calibration and Measurement

(8) Time-Frame of Study
(9) Measurement Error
(10) Intervening Events

Item #1: Selection / Omitted Variable Bias (+ / –)

Problem: If people have a choice whether or not they participate in the program then those that participate will be different than those that chose not to participate. Remember that selection and omitted variable bias are two sides of the same coin.

Solution: Randomization into treatment and control groups.

Note that the presence of randomization is not enough to ensure equitability of the study groups, especially when sample sizes are small. The researcher must present evidence that randomization was successful or "happy." This is done my comparing measured characteristics of T1 and C1 (treatment and control groups).

For the purpose of this assignment questions about "happy" randomization and non-random attrition, use a level of α =0.05 and apply the Bonferroni correction of α/n where n pertains to the number of contrasts performed (the number of comparisons made in the table).

So, for example, in the table that we examined in class we saw that one case out of seven had an alpha of 0.01. Should we reject the null, that attrition was random? According to the criteria, we would need to see an item with a significance level below 0.05/7, or 0.0071. So in this case, we would not reject the null and we would assume that attrition was random. Therefore, the study gets a +1 for the Campbell Item #2.

The Bonferroni correction only applies when the author has reported the actual p-values of the contracts in a table. In many cases (like CH5) the author will simply report, there were no significant differences between the treatment and comparison groups. As long as they report this, we will give them the benefit of the doubt and grant a +1.

Criteria: Guilty until proven innocent. Identify non-equivalent treatment and control groups in the first time period.

Item #2: Nonrandom Attrition (+ / –)

Problem: We worry that the people who leave the study are different than the people who stay in the study. NOTE: failure to participate is different than leaving the study (this relates to the concept of treatment on the treated versus intention to treat), and attrition means specifically that you cannot measure someone's outcomes in the post-treatment time period. There are many cases where individuals or organizations stop participating, but you can still measure their outcomes (for example, if a parolee stops coming to a support group for a recidivism program, you can still tell whether they end up back in jail).

Solution: Comparison of the groups that stayed versus the groups that left.

This comparison can be made in a variety of ways. One can compare measured characteristics of T2 and C2 to see if they differ significantly (similar to selection, except at time period two).

One could also compare the characteristics of the individuals that stayed in the study to the characteristics of those that left. This would amount to a comparison of two subgroups, $T1_1$ and $T1_2$. The former is more common, though.

You will again apply the Bonferroni correction.

Criteria: Guilty until proven innocent. Identify non-equivalent treatment and control groups in the second time period or evidence that those who left are different than those who stayed.

Item #3: Maturation (+) or Aging (–)

Problem: As the subjects age the dependent variable will change over time. For example, in studies of verbal ability children will naturally develop higher capability independent of the treatment. In studies involving aging groups they will naturally lose ability over time. Subjects can be individuals, but they can also be things like organizations, machines, etc. – anything that will experience a change with age in the dependent variable.

Solution: Use of a valid comparison group.

If there is data on the expected change over the study period (C2-C1), this can be subtracted out from the total change observed over the study period (T2-T1).

Maturation will only apply to a subset of studies that have observations that can be expected to mature over the study period. If the study is only a few days long, for example, then maturation would not be considered a problem for something that occurs over several months or years like muscular degeneration.

Criteria: Innocent until proven guilty. In order for this item to get a zero, you must first argue that it exists and then show that a lack of adequate controls will lead to incorrect results.

Item #4: Secular Trends (+ / –)

Problem: The dependent variable will change over time as a result of broad trends in the environment. For example, if there is high inflation then wages will increase as a result of the economy, not as a result of the program. If there is a program to reduce population growth, the program might be implemented in a community where birth rates are already declining, but the program is meant to reduce them even further.

Solution: Use of a valid comparison group.

If there is data on the expected change over the study period (C2-C1), this can be subtracted out from the total change observed over the study period (T2-T1).

Secular trends will only apply to a small subset of studies and only when trend can be expected and the study time frame is long enough to be affected by the trend.

Criteria: Innocent until proven guilty. In order for this item to get a zero, you must first argue that it exists and then show that a lack of adequate controls will lead to incorrect results.

Item #5: Seasonality (+ / –)

Problem: There are seasonal trends in the data that cause the dependent variable to be high during some periods of the study and low during others. A comparison of results across seasons can lead to Type I and Type II errors.

Solution: Use a valid comparison group to subtract out seasonal trends. Alternatively you can compare points across similar seasons.

For example, you would need to account for seasonal trends if you were looking at ice cream sales in August versus January. You could, however, compare ice cream sales in August of 2011 with sales in August 2012.

Seasonality will only apply to a small subset of studies.

Criteria: Innocent until proven guilty. In order for this item to get a zero, you must first argue that it exists and then show that a lack of adequate controls will lead to incorrect results.

Item #6: Testing Effects or Hawthorne Effects (+)

Problem: Because of multiple exposures to the same test, the subject learns how to better answer questions and scores higher over time, independent of program effects. Alternatively, the subjects behave differently when they know they are being observed – a problem known as the Hawthorne Effect or survey response bias. Both cases make the results look better than they are.

Solutions: For testing effects, you can do one of the following: (1) Use a comparison group that does not receive testing. (2) Use a post-test only design. (3) Use a different testing instrument at the pre-test and at the post-test.

For Hawthorne effects, you would want to use a non-intrusive study design, ideally where the subjects do not know they are being watched.

Criteria: In order to get a zero, you must argue that testing effects would be likely, and then show that there is no comparison group that removes testing effects from the estimation.

Item #7: Regression to the Mean (+)

Problem: If the treatment group was all drawn from a lower tail or upper tail of their own distribution (an individual, over-time distribution), then in the next observation period they will regress to their own mean, and thus it will look like a remedial program was effective at improving the outcome when it is just a statistical artifact.

Solutions: Do not draw a study group from the high or low end of their own distributions. If you must (in remedial programs) then use a control group that is also drawn from its high or low end of an individual and over-time distribution.

Criteria: You must argue that the sample is drawn from the low end of each individual over-time distribution, and that the comparison group does not have these characteristics.

Item #8: Measurement Error in the DV (–)

Problem: There is measurement error in the dependent variable, thus increasing the chance of a Type II error.

Solution: Increase the sample size to improve statistical power.

Use better measurements for the DV. This often involves using test instruments that have been developed by professionals in order to measure important latent constructs like intelligence, happiness, etc.

Criteria: In order for this item to get a zero, you must argue that the dependent variable is not or cannot be measured precisely. For the purpose of the homework, we will apply a strict test – if the dependent variable has ANY measurement error, give this item a zero.

Item #9: Study Time-Frame (–)

Problem: The program may take some time to have an effect. As a result, if the study time-frame is too short it makes the program look ineffective, even when it actually is.

Solution: Use previous research to establish an appropriate time-frame for the study.

The time-frame problem will usually lead to Type II errors (failing to reject the null when in fact the program has an impact). So watch for this problem when the study concludes that the program has no impact.

Criteria: In order for this item to get a zero, you must argue that the time-frame is too short, and as a result it could have potential to bias the conclusions through a Type II error (the program looks ineffective when in fact it was).

Item #10: Intervening Events (+ / –)

Problem: There is an event that effects one of the study groups (treatment or control), and as a result makes the effect larger or smaller than it should be. For example, if there is a study about fertilizer and there is rain in the treatment village and drought in the control village, this will affect the estimates. Note, secular trends affect both groups. Intervening events affect only one group. They are also events (drought), not trends (climate change).

Solutions: There is no good solution after an intervening event has happened, unless you can estimate the size of the change caused by the event (which is hard if the treatment and control group experience the event differently).

Criteria: In order to score a zero, you must identify a plausible intervening event in the study, or a good argument why there must have been intervening events given the time frame of the study and unit of analysis.