# CPP 524: Final Exam Practice Questions

## Estimators

- In the formulas for estimators (e.g. post-test only = T2-C2) what do the T's and C's represent?

  ANSWER: they represent a group mean. For example, *T2 is the average Y for the treatment group in the post-treatment time period.* The models estimate "average treatment effects" because they capture differences in group means.

- We have two studies, one where half of the participants receive 200 mg of caffeine and half receive none, and another study where participants are randomly assigned a dosage between 0 and 500mg of caffeine. The outcome Y is heart rate. In both models below the coefficient b1 will represent the effect of the treatment (expected increase in heartrate). Explain the difference in interpretation of the two b1 coefficients:

  Y = b0 + b1( caffeine_dummy ) Study 1
  Y = b0 + b1( caffeine _level ) Study 2

  ANSWER:

  In the first case the treatment is discrete (you received caffeine or you didn't) so the coefficient is a test of group difference:

  Study 1: b1 as a t-test
  b0 = heart rate of control group (consume 0mg of caffeine)
  b0+b1 = heart rate of treatment group
  NULL: (b0+b1) = b0 → b1=0, significance tells us groups that receive the treatment perform differently

  Study 2: b1 as an input-output machine
  b0 = heart rate of someone consuming 0mg of caffeine
  b1 = slope representing heart rate increase for 1mg of caffeine
  NULL: b0 = 0, significance tells us the slope is different from zero

  Main difference is both tell us whether caffeine has an impact, in the first case we can say caffeine matters, in the second if we know the amount of caffeine a person consumes we can guess their heart rate (the input-output interpretation of a slope).

  Note that you can estimate similar effects from Study 1 by multiplying the slope of Study 2 by 200 since the slope represents the rise in heart rate associated with a one-unit change, one-unit in this case representing one milligram of coffee.

- Explain the different between the intention to treat (ITT) estimate of program impact vs the treatment on treated (TOT) estimate of program impact? What does each represent? Which is more accurate?
- What is the difference between refusing treatment (household that refuses to take daily vitamins in a study about anemia) and attrition (someone moves in the middle of the study)?

- Explain difference between a control group and a comparison group.
    - Control group is the counterfactual. Comparison group provides useful information such as trend (expected gains independent of the treatment).
    - Comparison groups are often used to construct the counterfactual, but are not directly used as the counterfactual.
- Which study design uses a "control group" and which uses a "comparison group"?
- What is the Achilles heel of each study design?
    - Group equivalence prior to treatment (T1=C1)
    - Parallel trends assumption (C2-C1 captures expected gains for T1)
    - No expected gains independent of the treatment over the study period
- Match counterfactuals and estimators
    - Reflexive: T2-T1
    - Post-test only: T2-C2
    - DID: (T2-T1) – (C2-C1)

**Dummy Variables and Hypothesis Tests**

- Match counterfactuals to regression models
    - Data has ID, Y, group (T and C), time (2 periods)
    - $Y = b0 + b1(D_{group})$
    - $Y = b0 + b1(D_{time})$
    - $Y = b0 + b1(D) + b2(T) + b3(DxT)$
    - What does b0 represent in each case?
    - Which coefficient or coefficients represent the counterfactual in each case?
    - Which coefficient represents program impact?
    - Explain how the test b1 = 0 is a test of program impact.
- Consider case from slides: teach for America vs regular teaching training, suburban vs urban schools. Dataset with four dummies, one for each group and Y represents student performance on standardized exams. Which regression would you run to test the following questions:
    - Do teach for America instructors perform overall better than regular?
      Y = b0 + b1(D_tfa)

- Do suburban school students perform better?
  Y = b0 + b1(D_suburban)
- Who performs best in suburban schools?
  Y = b0 + b1(D_suburb) + b2(D_tfa) + b3(D_suburban x D_tfa)
- Can you answer the following with the last regression: Who performs best in urban environments?

  No, b3 is a test of (b0+b1+b2) = (b0+b1+b2+b3)

  b0=regular teachers in urban schools
  b0+b1 = regular teachers in suburban schools
  b0+b1+b2 = tfa teachers in suburban schools ***counterfactual*** (if no difference)
  b3 = test of counterfactual to observed

  For TFA in urban schools need to include urban dummy. We can recover each group mean from both models, but we can't test all hypotheses with each model.

- Which research design / estimator do these models represent? Will it capture performance accurately?

**Tests for Group Equivalency**

- Explain how a table of contrasts can be used for a test of "happy" randomization
- What is the NULL hypothesis in this case?
- Is statistical significance a good thing or a bad thing?
- In a normal context, what would a p-value of 0.07 tell us?

  Given the observed effect size (group difference or slope) and variance (width of CI), we would only expect the observed outcome 7 times out of 100 if the two groups were equivalent.

- Why do we need to apply a Bonferroni correct to perform an omnibus test (several contrasts at once)?

**Attrition Tests**

- Consider study with a covariate X which describes the study sample at the start of the study prior to any treatment. For example, X could be.
- It is reflexive design, and dataset contains dummy D indicating attrition.
- Write down the regression for the random attrition test.
- Is a statistically significant coefficient a good thing or bad thing?
- Interpret the sign on the coefficient – what would it tell us about attrition in the study? And how would it bias our findings (assume higher IQ is associated with higher outcome Y in the study).

| Y | IQ | Attrit? |
|---|---|---|
| 242 | 95 | 0 |
| 234 | 115 | 0 |
| 432 | 87 | 0 |
| 342 | 98 | 0 |
| - | 102 | 1 |
| - | 117 | 1 |

**Measurement**

- Define "latent construct"
- What is an "instrument" in the social sciences?
- How can we tell if an instrument is reliable?
- What is predictive validity?

  It is common to create categories of things that have high reliability but low validity. For example, "red-headed students" is easy enough to define and observe, but it would not predict much in the classroom. Meyers Briggs tests have decent reliability but very low predictive power (employers should never give the test to determine candidate fit with the organization!). The Big Five personality test, on the other hand, does predict performance in many contexts.

- Define measurement error.
- TRUE or FALSE: Latent constructs will ALWAYS be measured with error.

**Campbell Scores**

- Which items have the guilty until innocent criteria, and why?
- What are the effects of measurement error in the DV?
- What are the effects of measurement error in the IV?
- How do we determine an appropriate time frame for a study (the length of time needed for effects to manifest if the program is effective)?
- If a hurricane impacts the treatment group and the study group is it an intervening event?

  Hurricane reduces performance by Z.
  Post-Test Estimator: (T2-Z) – (C2-Z) → (T2-C2) + Z – Z → T2-C2 *(unbiased)*
  DID Estimator: (T2-T1-Z) – (C2-C1-Z) → (T2-T1) – (C2-C1) + Z – Z → (T2-T1) – (C2-C1) *(unbiased)*
  Reflexive Estimator: (T2-Z) – T1 → (T2 – T1) – Z *(biased)*

Intervening event only reduces bias when it impacts one group and not the other, unless it is a reflexive design.

**Study sample**

- True or false, we always want the study sample to reflect the general population. For example, our study of charter school effectiveness should aim for a study sample that represents the average or typical student in the city.

  FALSE, if programs are voluntary we want to study sample to represent the type of person that would elect to participate in the program.

  It would be odd to study weather a smoking cessation or addiction recovery program works for the average citizen since most are not smokers or addicts. The **study population** should focus on success rates of those likely to participate in the program.

  The one exception is when a program is being scaled. Is Medicare an effective insurance program for low-income households is a different question than would a Medicare for all program make the population healthier. In the former it is a question of having healthcare or not, in the latter it is a matter of replacing current insurance with Medicare.

  The program scale fallacy occurs when you take inferences from a voluntary program and tabulate the benefits if the program is scaled to the entire population. Those that select into a program are rarely the same as those that choose not to participate. Is a program effective (does it do what it was designed to do – help addicts recover?) is different than the question, would a program benefit everyone?

  Make sure you differentiate population → program eligible / target audience → program participants.

- TRUE or FALSE: non-participants of voluntary programs often make a good control group. FALSE, they are often very different than study participants.

- TRUE or FALSE: selection is a type of omitted variable bias. TRUE

- TRUE or FALSE: selection is the biggest source of bias in impact studies in evaluation. TRUE

**Example essay question:**

**CRUDE OIL CHEMICAL LINKED TO HEART DEFECT IN BABIES**

Babies who are exposed before birth to ethyl benzene, a toxic component in crude oil, may have a higher risk of developing congenital heart disease, U.S. researchers said Saturday.

Another chemical used as an industrial metal degreasing agent, trichloroethylene (TCE), also boosted heart risks, said the research to be presented at the Pediatric Academic Societies annual meeting in Denver, Colorado.

Congenital heart disease occurs when the heart is malformed before birth, and is the most common of all birth defects. Previous studies have suggested it could be caused by chemicals in the environment.

"Congenital heart disease is a major cause of childhood death and life-long health problems," said D. Gail McCarver, lead author of the study and professor of pediatrics at the Medical College of Wisconsin.

"Thus, identifying risk factors contributing to CHD is important to public health."

Researchers collected stool samples from 135 newborn babies with the heart condition and 432 infants without it. A full 82 percent of all the infants showed exposure to at least one of the 17 solvents measured in the study.

White, but not black, infants who showed exposure to ethyl benzene had four times the risk of CHD.

Black infants exposed to trichloroethylene showed an eight-fold risk for the heart condition, and white infants with the traces in their stool had a two-fold higher risk, said the findings.

**Question 1: Which estimator is implied by the story?**

They are comparing two groups at the same point in time, and after one group was exposed to high levels of ethyl benzene. So this is a post-test only comparison (T2-C2).

**Question 2: What is the problem with this research design? Specifically, is the identifying assumption of the estimator met?**

They collected stool samples after birth (so the post-treatment period), and correlate chemicals in stool samples to rates of heart disease. The information we would need to feel comfortable with this estimator is that the two groups – those with ethyl benzene present and those without – were otherwise the same prior to exposure. For example, was there an equal proportion of male and female babies in each group? What about the proportion of white and black children? The article commits the mortal sin of the Campbell Scores – failure to establish group equivalency. **It appears that the treatment and control groups are likely different (different SES, race, geography, etc.). Thus, the post-test only estimator loses it's internal validity.** We cannot conclude that ethyl benzene is causing the cancer. That does not mean the study is not useful in the public health context – early work is often descriptive – just identifying relationships between variables so that they can be further examined in subsequent research.

**Question 3: Explain one or two competing hypothesis it fails to eliminate.**

The researcher are relying on correlations to try to identify causality. For example, babies that have congenital heart disease also have high levels of ethyl benzene exposure prior to birth. Does the benzene cause the congenital heart disease?

The authors are not able to eliminate the competing hypothesis that some other factor is both correlated with levels of benzene and heart disease (the classic omitted variable bias). For example, perhaps the kids with heart disease come from the same neighborhood. In this neighborhood there are many factories. The factories may produce ethyl benzene, but also other toxins that were not measured in the study. Perhaps the other toxin is the actual cause of the heart disease, not ethyl benzene.

The control group comes from a different neighborhood so they show low levels of ethyl benzene and low rates of heart disease. As a result, the researcher might assume that it is benzene that causes the heart disease, when it is actually the other unmeasured toxins or other characteristics of the neighborhood.

**Question 4: Why is it impossible to do this type of study using a rigorous Randomized Control Trial?**

In general we deal with omitted variables by creating a group of willing participants and then randomly assigning the treatment to some of them. The fact that assignment is random ensures that any omitted variable will be evenly distributed across both groups, and as a result effect the treatment and control groups equally. The independent effect of the treatment can then be discerned through the difference in outcomes between the treatment and control group.

In this case, however, the treatment is exposure to a toxic substance. There is no way to "treat" a baby in this kind of research without serious ethical violations. As a result, correlation analysis is all that the researcher can use.

**Question 5: Explain why a reflexive design would be impossible here.**

Reflexive design requires a pre-treatment measure. It's impossible to determine if a baby has a congenital heart problem before birth, so no pre-treatment outcome exists.

**BONUS: Describe what a difference-in-difference design could look like in this case.**

We would need a group that has no treatment in the first period and has received the treatment in the second. And we need another group that has no treatment in either period.

The hard part of this question is determining how to operationalize pre-treatment period and post-treatment periods. Since we need outcomes in both periods, however, we would need to observe two births to have two study periods.

Thus, a study might find a sample of women with multiple children that have had stool samples in their first few months of life. Isolate the cases where the first child did not have ethyl benzene present and the second child did. Compare to cases where neither child had ethyl benzene present, but other chemicals were found. See if rates of heart disease increase after exposure.

With this difference-in-difference design you could also include levels of all other chemicals as control variables in the study to further control for other chemicals in the environment.