Exam Field Experiments

Spring 2021

Please count the words of each part and write the number behind your answers.

1. **Interpreting Results and Policy Recommendations - 25% of grade** In the paper, Al-Ubaydli et al. 2020 called *"What Can We Learn From Experiments? Understanding the Threats to the Scalability of Experimental Results"*, the authors talk about different threats to external validity and scalability. They write in section II: "One common source of scaling bias is adverse heterogeneity, whereby the participants' attributes make them systematically predisposed to exhibiting a stronger relationship than in the population at large."

   (a) Explain in your own words what "adverse heterogeneity" means and why this can lead to a scaling bias. (maximum 150 words) (5% of the grade)

   Solution: This question tests the following learning outcomes:

   - Argue for optimal choice of outcome variables, sample size, clustering and randomization balancing.

   Adverse heterogeneity means that some people have characteristics that make them more likely show a stronger effect when given a treatment than other people. Often this means that they can benefit the most from the treatment. The bias results from these people being either more likely to self-select themselves into the experiment (for example when there is informed consent) or when researchers purposely select these individuals to take part in the experiment.

   (b) Choose TWO papers from the course for which you think "adverse heterogeneity" could be a problem. You can choose papers presented in the lectures, the homework videos or the student presentations. If there are multiple treatments and outcomes, you can choose one treatment and one type of outcome to make your argument. For each of the chosen papers:

   1. Name the paper (Authors and Title)
   2. Explain what the experiment was about and what the main effect was (maximum 100 words) (5% of the grade)
   3. Explain why you think there could be adverse heterogeneity among the sample (maximum 150 words) (7,5% of the grade)
   4. Should we expect a larger or smaller effect if the intervention was scaled to the full (relevant) population? (maximum 100 words) (7,5% of the grade)

   Solution: This question tests the following learning outcomes:

   - Discuss state-of-the-art empirical research in experimental economics conducted in the field.

   This is one example. Depending on the chosen papers the solution might look slightly different.

1. Belot, Michele, Noemi Berlin, Jonathan James, and Valeria Skafida. "The formation and malleability of dietary habits: A field experiment with low income families." (2018).

2. The field experiment measures to which extent dietary habits are malleable during early childhood. They conduct an experiment on 285 low income families with young children. There was a control group and two treatment groups. In Treatment 1 families received free groceries for 12 weeks and were asked to prepare meals at home. In Treatment 2, families were asked to reduce snacking and eat at regular times. They collected food preferences, dietary intake and biomarkers. They find significant effect on the children, but no effect on the parents.

3. The families were recruited via flyers, posters and advertisements and had to consent to be part of the experiment. Even though, they were not informed about the treatments, they did know that this was an experiment on health and dietary choices that would run for three years and that it involved their children between the age of 2 and 6. It is very likely that families who had an interest in improving the dietary choices of their children would be more likely to notice the advertisement and consent to taking part in the study. If they are motivated to change, they can benefit the most from the intervention.

4. We should expect the treatment effects on the general, eligible, low-income population of Edinburgh and Colchester with small children to be smaller if this intervention is rolled out to this population. On average, it is likely that they are less motivated to change and to adhere to the experimental protocol. The effects they find are most likely only representative for a motivated sub-group.

2. **Discussing limitations of experimental research - 35% of grade** Read the attached paper “Indian Water Use". Tip to save time! Focus on the parts that are relevant for the questions below.

   (a) Write a summary of the research question, experimental design and main results. You can also make a graph or flowchart. Focus on what is important for discussing the experimental design. (maximum 400 words - 10% of grade)

Solution: In this paper the authors intend to determine whether the introduction of a new practice - Alternate Wetting and Drying - when combined or not with social messaging, induces a reduction in water consumption on a sample of Indian farmers. The experiment took place in two districts in India. The area was chosen because in these regions farmers use ground water to water their crops and ground water levels are starting to be depleted. The experiment was conducted in collaboration with an irrigation research and management institute. The experiment ran over a period of 9 months. 30 villages were chosen to participate with 10 households per village. That results in 300 households. There were the following treatments:

1. Control: Eight weekly phone surveys on water use

2. Treatment 1: Eight weekly phone surveys on water use + Farmers were trained in alternate wetting and drying, a new technique. They also received weekly reminders about this technique

3. Treatment 2: Same as treatment 1, but in addition the farmers received social comparison messages that told them about the pumping hours of the other farmers

Before the treatment, there was an in-person baseline survey asking the farmers about their irrigation practices. And at the end there was another in-person survey. In addition to the main randomization into treatments, there was an additional cross-randomization into whether their pump was being monitored or not. 200 households were randomized into the monitoring treatment. 100 did not receive a monitoring device. Results: The study finds suggestive effects that the training program had a positive effect on water reduction. The effects of the social norms on behavior change are inconclusive.

(b) Discuss the experimental design based on what you have learned in class. You should answer the following questions (maximum 1000 words for all sub-questions together - 25% of the grade):

1. What were the main outcome variables? Are they well chosen? Why, why not?
2. How were the participants selected? Was there self-selection?
3. Was there stratification before randomization? If yes, on what variables? If no, what would you suggest?
4. Is there attrition from the experiment? What type of attrition? If yes, how could it affect the results?
5. Is the experiment well-powered? What evidence do you have for that?
6. Is it likely that there was a Hawthorne effect? How could it affect the results?
7. Is there evidence of spillover effects? If yes, how could they be mitigated?
8. What is a experimental design problem that has not been covered by the questions above? Why is it a problem?

Solution: This question tests the following learning outcomes:

- Argue for optimal choice of outcome variables, sample size, clustering and randomization balancing.
- Discuss state-of-the-art empirical research in experimental economics conducted in the field.
- Conduct power calculations for determening the correct sample size of an experiment.
- Analyze experimental data using econometric tools.
- The main outcome variables are a) self-reported pumping hours and b) the meter value on the farmer's pump. The farmers were asked to self-report how many hours they had pumped the past week on each day. For the meter, they were asked to report the number on the meter. A problem with the self-reported outcome variables is that there might be re-call bias. Since farmers only report once per week, they might not remember what they did every day. They might also round up the time the pumped because that is easier to remember. The metered value is more reliable, but could also have errors because the farmers were not told what the numbers represented. A

problem that appears when comparing metered and non-metered outcomes is that because of the intentional or unintentional bias, the self-reported values are likely to be incorrect. However, this assumption cannot be verified. When comparing the values for farmers who are monitored, they seem similar, but not identical. Overall, having metered outcomes is better than the non-metered ones despite potential drawbacks from the metering.

- The authors selected two regions that fit with their requirements. More than 50% of the households needed to grow rice and be within 30km of each other. They randomly selected 30 villages from the list and 10 rice growing households from each village. It is not clear how they knew who was rice growing before the households were selected. There is no possibility to self-select into the experiment, but selected households could decline to participate.

- There is no information on stratification in the experiment. This would have most likely been a good idea and also feasible given that there was a baseline survey conducted before the treatments were assigned. The balance check shows that the randomization was not successful on phone ownership. This is a big problem, when the data is collected via phone. They also find imbalances in age and marital status. Another dimension to stratify by could have been size of farm because the outcome will vary a lot based on size.

- The attrition bias is potentially high in the results since only 75% of the farmers picked up their phones when called. All the interventions aim at saving water. Knowing this, the farmer that keeps using water might be reluctant to pick up the phones and admit that they keep using water. This would threaten the validity of the results. It is likely that there is selective attrition, even if there is no differential attrition.

- The authors conducted a power analysis (footnote 7) and decided on enough households to identify a 2.26 hour decline in water use. They decided they needed 300 households. However, in the final analysis they only have 116 households that are metered. They did not take into consideration to do sub-sample analysis, even though their outcomes depend on it. The main effect is larger than 2.26 hours (3-4 hours) so the power seems ok given their Minimum Detectable effect size.

- The Hawthorne effect is very likely in this experiment. The farmers know that they are being monitored on their water use and they are reminded of this every week. They have an incentive to use less water or at least report using less water. However, all three groups had to report water use (also the control group), so the treatment effects can still be valid. The results are less likely to have external validity.

- There is no direct evidence of spillover effects, but they are likely, especially in Treatment 2 were farmers are told about the water use of their neighbors. This could lead them to talk to the neighbors and find out who is in the experiment. Some farmers had to use the phone of a neighbor for the survey, which makes talking about the experiment even more likely. It is not clear in which direction the spillovers would go. Farmers in treatment 2 might talk to farmers in treatment 1 and let them know about the social comparisons, which would then reduce the treatment effect between the treatments. The

paper says that because the farmers are richer and have their own irrigation they will communicate less. But 9 months is a long time and they might talk eventually.

- Different answer are possible here. For example one can talk about the experimental design not being able to disentangle the mechanisms. One could also say something about the external validity of the results. It is a special group of farmers that were used. Or one could suggest how some surprise in-person measurements could have verified the reported results.

3. **Designing your own experiment - 40% of the grade** The pandemic has strongly affected teaching and learning at the university. The study board at KU is interested in the answer to the following research question: *Are students' learning outcomes higher when TA sessions are live (but virtual) compared to TA sessions being pre-recorded with a one-hour per week office hour to ask questions?* There are no differences in content between TA sessions (the problem sets are the same for all students).

   You are tasked to find an answer for the study board. They give you 10,000kr and 6 months maximum for the task. The dean of the Faculty of Social Science has agreed that you can use students from his faculty for the experiment. You do not need to use all of the money or the time. It is important that the experiment is doable in the given time, with the money and amount of students available. He wants to see a detailed proposal/ pre-analysis plan from you how you will tackle the question. It is important for him that the experiment is ethical and it does not place an unnecessary burden on the students in the experiment (maximum 2000 words for all sub-questions together).

   Remember! You are supposed to answer the question posed by the dean - this is a program evaluation. You are not supposed to come up with alternative interventions.

   (a) Write up the experimental design - Elements to consider are: Outcomes, level and process of randomization, number of observations, length of the experiment...

   (b) For each element of the design you should argue why you have made this design choice based on what we have covered in class.

   (c) Discuss what problems could arise and how your design intends to control for those.

   (d) Explain briefly what analysis you will conduct with the data you collected to test for the effect.

   Solution: This question tests the following learning outcomes:

   - Argue for optimal choice of outcome variables, sample size, clustering and randomization balancing.
   - Plan and develop a field experiment design, both conceptually for testing economic theory and practically.
   - Independently identify challenges that could arise when conducting an experiment and argue how to overcome them.
   - Analyze experimental data using econometric tools.
   - Initiate and participate in discussions of the implications of experimental results for policy makers or managements in the public and private sector.

There are different solutions possible. It is important that the answer addresses the following elements in a convincing way. It is also important the the experiment is realistic to carry out given the constraints.

- How are the students selected? For example only courses with several TA sessions are eligible here so exam result are comparable.
- How many students do you need? How would you do a power analysis? You probably need to allow for clustering in your power analysis if you have TA groups. You might also need to cluster by teaching assistant.
- What are the outcomes? Just exam grades at the end of the semester or also evaluation? Perhaps your design needs intermediate tests?
- On what level will the randomization take place? Individual? Group level? On pre-existing study groups.
- Should the students be stratified before randomization? On what variables? This needs to be coherent with the level of randomization.
- How will you avoid spillover effects? Students in different TA sessions most likely all know each other and will talk to each other.
- How long will the experiment run and why?
- Should both groups receive the same treatment eventually? So for example group A first recorded and group B live and then switch after half of the semester? This would make it more ethical, but outcomes would need to be measured in a mid-term exam.
- How do you avoid attrition? Is it a problem here? Students will probably not drop out of a course, but might try to attend the other sessions if they like those better. Is it possible that students who feel badly prepared will not go to the exam causing bias? Can you exclude students from TA sessions? Can you collect data to make sure that if they do, you know about it?
- For the analysis you probably want to analyze the intention to treat effect and you want to cluster students at the TA group level.
- How representative are the results from your sample for the rest of the KU students, Danish students, all students in the world?