Experiment Procedure and Research Ethics in the Era of Replication Crisis

Duk Gyoo Kim

Sungkyunkwan University

May 28, 2024
KIEP

Experiment Procedure

- Research question
- Experimental design
- IRB approval and pre-registration
- Grant applications
- Software development
- (Pilot) Experiments
- Payments

Procedure: Research question

- Every research project must start with a question.
- Isn't it too obvious? Not really.
- Many studies HARK (Hypothesize After the Results are Known). The validity of your research question must not depend on the results.
- Dufwenberg (2011) even suggested, "don't mention the results at all in the submitted paper. Put the results in a sealed envelope! Ask the editor and the referees to make their call before opening the envelope. Only once they have decided whether or not to publish the paper, they may open the envelope, study the data, and read your summary."

Procedure: Experimental design

- It may take several days to overview the typical experimental designs for specific topics. Some keywords you can search for:
  - Multiple Price List (MPL)
  - Global preference survey (adaptive questions; dynamically optimized sequential experimentation)
  - Convex Time Budget
  - Real-effort tasks: adding numbers, counting zeros, clerical work, decoding, etc.
  - Ultimatum Game, Dictator Game, Trust Game, Public Goods Game, Coordination Game, etc.
  - Higher-order rationality, higher-order risk preferences, level-k, cognitive hierarchy
- The key is to set experimental conditions and treatments optimal for supporting the empirical relevance of research idea.
- The experiment needs to be simple; much simpler than whatever you think.

Procedure: Experimental design

Example from Kim (2018), skip if running out of time

- Two-by-two between-subject design: linear vs. nonlinear // group size certainty vs. uncertainty
- Random rematch
- A baseline (linear & certain) treatment becomes a replication of a typical public goods game: 6 subjects simultaneously decide how much to put the money into the public account (or how much to keep into the private account).
- Examine the treatment effects: whether people contribute more when the group size is uncertain.

Procedure: IRB approval

- You need to get an Internal Review Board (IRB) approval before conducting experiments with human participants. (unsure whether KIEP has it.)
- Before applying for an IRB approval, I strongly suggest you to take online courses in research ethics. Some institutions explicitly require it.
  - National Institute of Health (NIH) and Collaborative Institutional Training Initiative (CITI) Program offers training courses for protecting human research participants. (SNU Link)
  - Korea Institute of Human Resources Development (KIRD) offers free online courses. Take "Research ethics for graduate students" and "Research ethics for research manager."
- In the course of getting an IRB approval, you may consider potential harms of your research. Experiments with children, fake information, and social stigma may have a long-lasting negative impact.

Procedure: pre-registration

Pre-register your study whenever possible.

- AEA RCT registry: For all AEA journals, "As of January 2018, registration in the RCT registry is mandatory for all applicable submissions. This applies to field experiments. Laboratory experiments do not need to be registered at this time."
- For lab experiments: self preregistration at AsPredicted.org. [an example]
- Other platforms, such as Open Science Framework, offer different ways of making transparent research conduct.
- This pre-registration becomes more important as people aware of research misconduct.

Procedure: Grant applications

- This step is less usual for other research fields in economics, but you need money.
- "I (or my advisor) have some funds to be spent by the end of this year. I gotta do some experiments." is not a good idea.
- If possible, look for seed grants first.
- Writing a grant proposal by itself is helpful in formalizing your research question and the goal of the research.

Procedure: Software development

I can't spend too much time on this.

- For single-person decision-making experiments: Use a survey platform. Qualtrics is recommended as it can add JavaScript for each survey item (Kim and Moon, 2024).
- For interactive (game-like) experiments (1): If you are not tech savvy at the moment, convert a game into a single-person decision-theoretic situation. There are some caveats of doing so, but you can still live on with the survey tools.
- For interactive (game-like) experiments (2): Consider learning z-Tree, oTree, or LIONESS. The first two are Python based, and the last one is JavaScript based. There are much more.
- Basic ideas are identical: A server collects inputs from many clients, computes some intermediate results, and sends outputs to the clients.

Procedure: (Pilot) Experiments

- Before a full-fledged experiment, consider running a small-scale pilot experiment.
- Not for manipulating the research design as the pilot results tell, but for fine-tuning parameters and checking technical issues.
- In a laboratory (a classroom with several separately-located computers is sufficient in many cases), or online (Amazon Mechanical Turk and Prolific.) If your software works online, Zoom-administrated real-time online experiments are also possible.

Procedure: Payments

- Sending money to participants and getting reimbursed are easier said than done.
- Consult your admins.

[IMPORTANT] Research ethics

- Researchers conducting experiments with human participants must pursue research ethics at the highest standard.
- Research misconduct can happen in every field, but those who deal with observations from human participants can, technically, manipulate the data directly.
- As we are living in the era of replication crisis, we collectively need to be more careful.

Credibility of economics research

- Read: Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." Journal of Economic Literature.
- Credibility issue is not only residing in the experimental studies: JMCB project (Dewald, Thursby, and Andwerson, 1986) illustrates how hard the computation codes are replicated. Leamer (1983) points out the specification search problem in empirical economics in general.
- Reproducibility: examining if results of original studies can be reproduced based on the **same data** as used in the original studies. computational/recreate/robustness reproducibility
- Replicability: examining if results of original studies can be repeated using **new data**, direct/conceptual replicability

Replication can be failed. It is by itself not wrong.

- Suppose you conduct an experiment. The mean difference between the control and treatment groups is significant at the 5% level (p-value=0.048). You publish the result at a journal.
- It means that your result can be a false-positive one in twenty replications. In other words, if your study is a subject of replication, 19 out of 20 replications will fail.

False-positive results: Power pose

- Amy Cuddy's power pose lecture is the second-most viewed video in TEDtalk.
- She, back then at HBS, claimed that power pose can significantly increase the testosterone level of the subject.
- It turns out to be false positive.
- Worth reading NYTimes covers: When the Revolution Came for Amy Cuddy.
- Sad truth: Boring corrections are much less popular than tabloid results, even in academia.

It is wrong if p-hacked.

- Specification search to look for the "stars" (p-values lower than 0.05) = Hypothesizing After the Results are Known.
- Isn't this what all empirical economists do? Maybe, when the data provider and the researchers are independent. HARKing is not acceptable when you are *both* data generator *and* analyzer.
- Brian Wansink's downfall: "He just kept analyzing those datasets over and over and over again, and he instructed others to do so as well, until he found something." He'd been proud of his p-hacking misconduct in his entire career.

It is wrong if p-hacked.

HARKing is one way of p-hacking. P-hacking can be also done in many other ways:

- Stop collecting data (although it is under-powered) once statistically significant results are obtained
- Omit some data by calling them "outliers"
- Cherry-pick some "meaningful" sessions
- Group or disarrange data arbitrarily

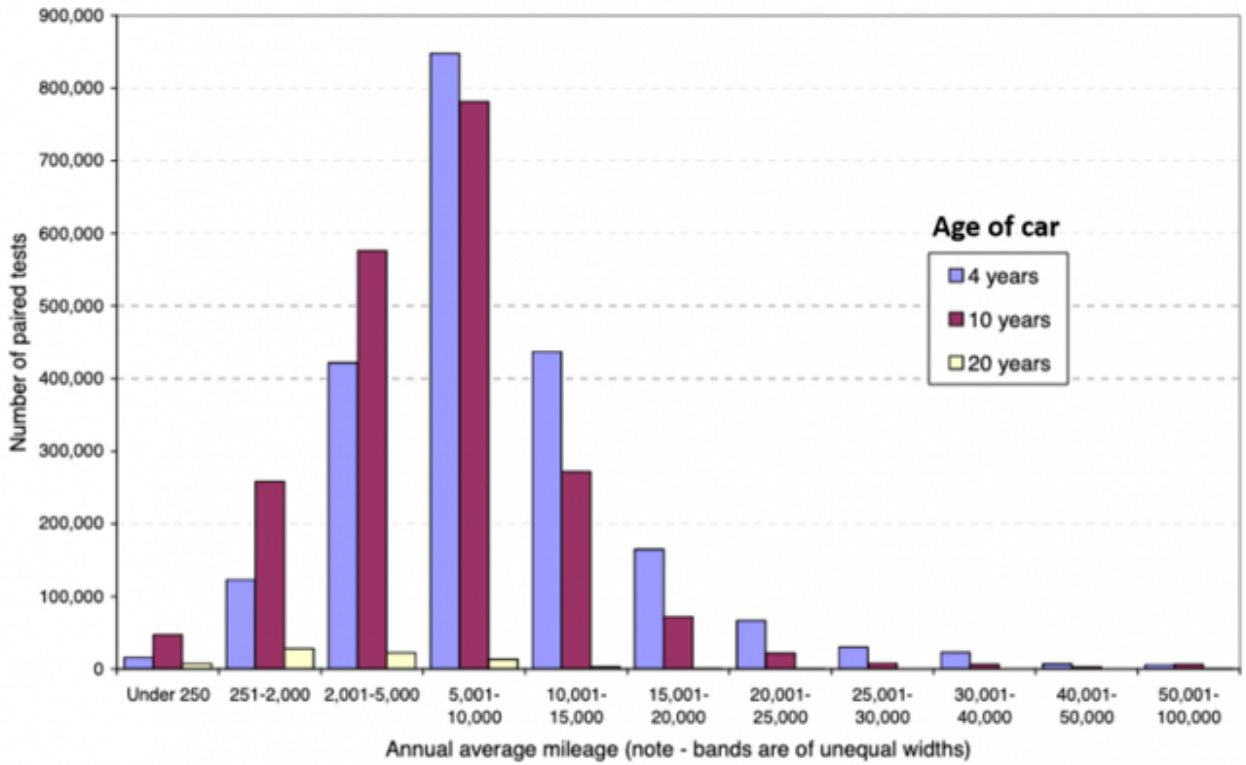
It is wrong with intentionally manipulated data.

- Michael LaCour's paper "When contact changes minds: An experiment on transmission of support for gay equality" published in Science [Wikipedia link]
- The paper placed him, UCLA graduate, an assistant professor position at Princeton.
- David Broockman, Joshua Kalla, and Peter Aronow investigated the paper and concluded that the data had been falsified and no data had been collected. The paper is retracted, and no one knows where LaCour is.

It is wrong with intentionally manipulated data.

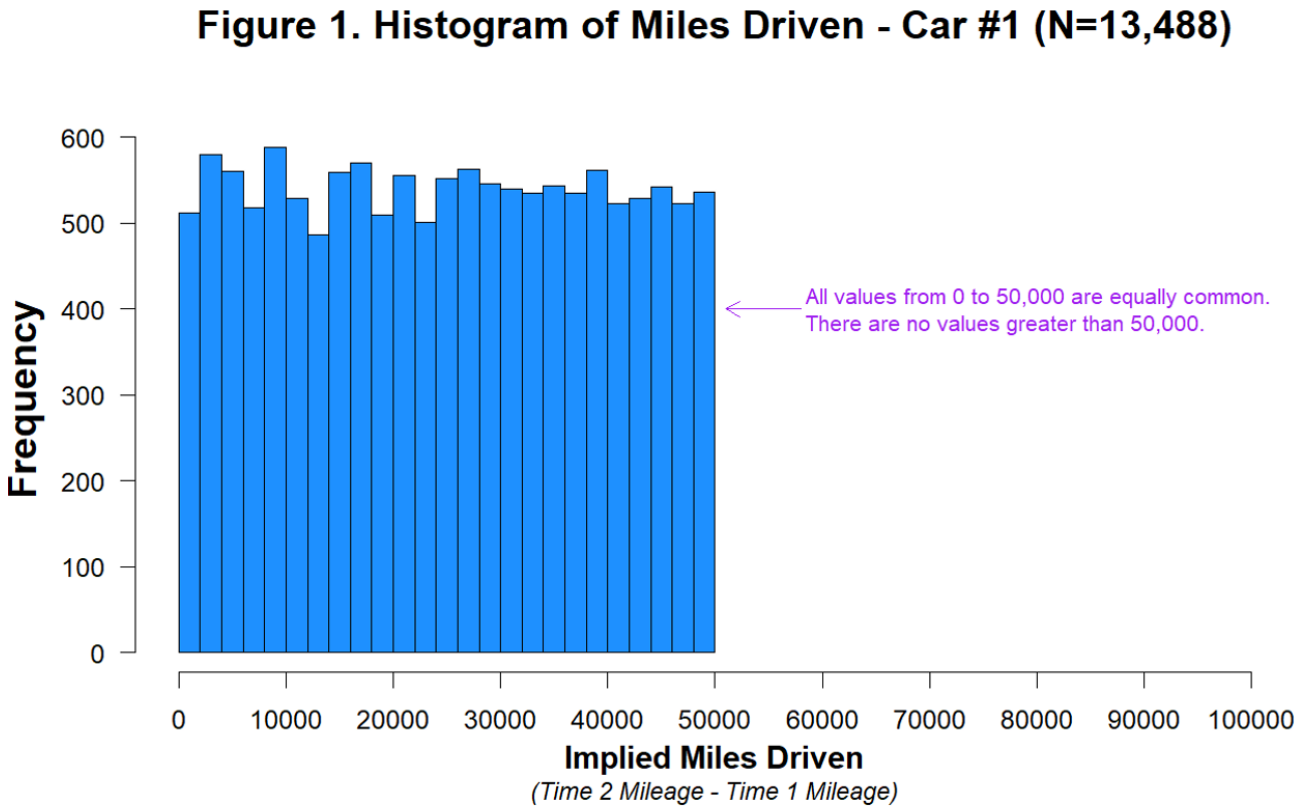
Francesca Gino, Dan Ariely, and coauthors' paper about honesty is retracted. It is ironic that all retracted papers of Gino are about honesty. [Datacolada]

Figure from a UK Department of Transportation Report on Distribution of Yearly Miles Driven in 2010
Source: https://bit.ly/3jwSP2N



It is wrong with intentionally manipulated data.

Francesca Gino, Dan Ariely, and coauthors' paper about honesty is retracted. It is ironic that all retracted papers of Gino are about honesty. [Datacolada]



Why does it happen?

- Publication bias: Studies with no significant results are harder to publish. NSF keeps the record of funded research projects in Time-Sharing Experiments in Social Sciences (TESS). Almost all projects with significant results are written, and more than 60% of them are published. Meanwhile, more than 60% of the projects with insignificant results are unwritten.
- "Lucrative business": The probability of fraud detection is low, compared to the marginal benefit of research output.

Both significantly undermine credibility of economics research.

How to tell it is p-hacked or merely false positive?

- Unfortunately, it is hard to tell with only one study (unless data are fabricated.)
- Fortunately, meta analysis of the similarly replicated studies can tell if the literature is exposed to the publication bias, by checking the p-curve.

P-curve

P-curve is a function that maps a p-value to its likelihood.

- Suppose two sets of data are randomly drawn from the same distribution. The p-value of the two-sample t-statistic (mean difference of the two samples) must be uniformly distributed from 0 to 1. [p-curve1]
- Now suppose two sets of data are randomly drawn from two *different* distributions. (That is, there is a population treatment effect.) The p-value of the two-sample t-statistic is much more likely to close to 0, yielding a monotone-decreasing curve, picking near 0 and phasing out near 1. [p-curve2]

This means, if one study is replicated several times and the treatment effect is real, we should expect to see more p-values close to 0, rather than close to 0.05.

P-curve of lemonade effect studies

The Bitter Truth About Sugar and Willpower: The Limited Evidential Value of the Glucose Model of Ego Depletion (Vadillo, Gold, and Osman, 2016)

- So called the "lemonade effect"
- Willpower, or ego, is believed to be a limited resource: After conducting cognitively burdensome tasks we tend to indulge in easygoing things.
- Several published studies repeatedly report that intake of glucose (i.e., drinking a cup of lemonade) can recover the depleted willpower.
- The p-curve is upward sloping, implying that this literature is publication-biased.

Academia is being looked down...

because of academic misconduct and replication failure.

- It is not just some individual researchers' ethical problems.
- The internal validity of whole research in the discipline is jeopardized.
- Legitimate, boring corrections are less popular than tabloid results, even in academia. We need to intentionally pay attention.

Less incentives for p-hacking / publication bias

- Dufwenberg (2019): Let's seal the results.
- Journal of Development Economics' 'Registered Reports' publish a pre-approved project based on the question and design before the results are known.
- Coffman, Niederle, and Wilson (2017): Enforcing "replication report" and its citation.

Encouraging replications

- Camerer et al. (2016 Science): replication project of economics experiments
- Camerer et al. (2018 Nature): (extended) replication project of social science experiments
- Journal of the Economic Science Association regularly publishes replication studies.
- All respectful journals have data-sharing policies. For experimental studies, not only the dataset but also the experiment instructions and detailed procedure need to be shared. Askarov et al. (2023) report data-sharing policies reduced reported statistical significance, and hence decreases publication bias.

Who conducts replications?

- For now, a small number (< 100) of dedicated researchers including Colin Camerer, Anna Dreber, Felix Holzmeister, Taisuke Imai, to name a few, work on replication studies.
- It may not be a big plus for one's academic career: risk of (unnecessary) fights with the original authors, reputational concern, publishability...
- However, it may be good for grad students, early-career researchers, and **researchers at policy-relevant research institutions**.
  - Get to know the literature
  - Learn how to run experiments and analyze the experimental data
  - Check **validity of policy implications** from experimental evidence (For example, Ayres et al., 2013).