

# MACS30000 Assignment 5

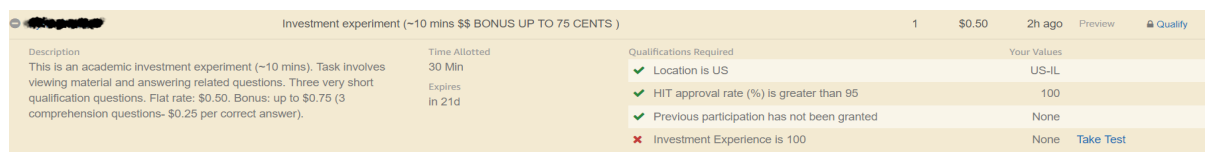
Dr. Richard Evans

Submitted by: Junho Choi

Due November 12, 2018 (11:30 AM)

## Problem 1

Figure 1: Amazon Mechanical Turk Investment Experiment Detail



The screenshot shows the details of an Amazon Mechanical Turk investment experiment. The title is 'Investment experiment (~10 mins \$\$ BONUS UP TO 75 CENTS)'. It includes a description of the task, time allotted (30 Min), and expiration (21d). A table lists four qualifications required: Location is US (US-IL), HIT approval rate (100), Previous participation (None), and Investment Experience (None). A 'Take Test' link is provided for the last qualification.

Investment experiment (~10 mins \$\$ BONUS UP TO 75 CENTS)		1	\$0.50	2h ago	Preview	Quality
<b>Description</b> This is an academic investment experiment (~10 mins). Task involves viewing material and answering related questions. Three very short qualification questions. Flat rate: \$0.50. Bonus: up to \$0.75 (3 comprehension questions- \$0.25 per correct answer).	<b>Time Allotted</b> 30 Min <b>Expires</b> in 21d	<b>Qualifications Required</b>		<b>Your Values</b>		
		✓ Location is US		US-IL		
		✓ HIT approval rate (%) is greater than 95		100		
		✓ Previous participation has not been granted		None		
		✗ Investment Experience is 100		None <a href="#">Take Test</a>		

- (a) For this problem, I have chosen an “investment experiment” from Amazon Mechanical Turk (MTurk). The details of this experiment are captured in Figure 1. I note that I have crossed out the name of the HIT requester just in case it may cause any problems.
- (b) While the up-front reward is given as \$0.50, the description actually elaborates that there are three “comprehension questions” that may provide experiment participants with further rewards. Since it says that for each correct answer to a comprehension question, the reward is \$0.25, the total reward can be written as:

$$\text{total reward (in dollars)} = 0.5 + 0.25x$$

$$\text{where } x := \text{number of correct answers, } x \in \{0, 1, 2, 3\}$$

and this would describe the full payment structure of this particular MTurk experiment.

- (c) There are four qualifications for this experiment. The first is that an experiment participant needs to be located in the U.S. The second qualification is that one’s HIT approval rate needs to be higher than or equal to 95. The third is that one was not granted participation in the experiment before. Finally, one needs to have an “Investment Experience” of 100, which can be measured by taking a very short questionnaire composed of three qualification questions. The details to this “Investment Experience” preliminary questionnaire are shown in Figure 2.
- (d) Referring back to the description of this HIT, it says that while the time allotted for this HIT is 30 minutes, the experiment itself is expected to take up to only **10**

Figure 2: The Details of “Investment Experience” Questionnaire

Please answer the following questions to qualify for this study.

Thank you for your willingness to participate in our study. The following questions are important for the validity of our study and I ask that you answer them honestly and to the best of your knowledge. If you do not qualify, it will not affect your MTurk rating.

Are you 18 years or older?

- ☒ Yes  
☐ No

Do you have previously bought or sold an individual company's common stock or debt securities?

- ☒ Yes  
☐ No

Do you have, at least once, evaluated a company's performance by analyzing its financial statements?

- ☒ Yes  
☐ No

**minutes.** With the assumption that on average, participants answer 1.5 out of the 3 questions correctly (50%), the implied hourly rate will be \$6.75 per hour.<sup>1</sup>

- (e) This particular job expires in 21 days (3 weeks) from when I accessed it, which was on November 11, 2018. Therefore its expiration date is December 3, 2018.
- (f) According to the Amazon Mechanical Turk website for pricing (for requesters), it says that there is an additional “20% fee on the reward and bonus amount (if any) you pay Workers” (Amazon Mechanical Turk n.d.). There are also additional clauses for those HITs with more than 10 assignments (Amazon Mechanical Turk n.d.). However, this seems irrelevant for this particular experiment that I have chosen. From the description, it is observable that there are at most six questions (3 comprehension questions which give bonuses, and 3 qualification questions).

The above information, combined with the fact that the most an experiment participant can earn is \$1.25 ( $= 0.5 + 0.25 \times 3$ ), will allow one to figure out the maximum cost of this particular experiment. If there are 1 million people participating in the experiment, the maximum cost for this HIT requester is **\$1,500,000**, where \$1,250,000 ( $= 1000000 \times 1.25$ ) is for paying the experiment participants and \$250,000 ( $= 1250000 \times 0.2$ ) is the additional MTurk fee.

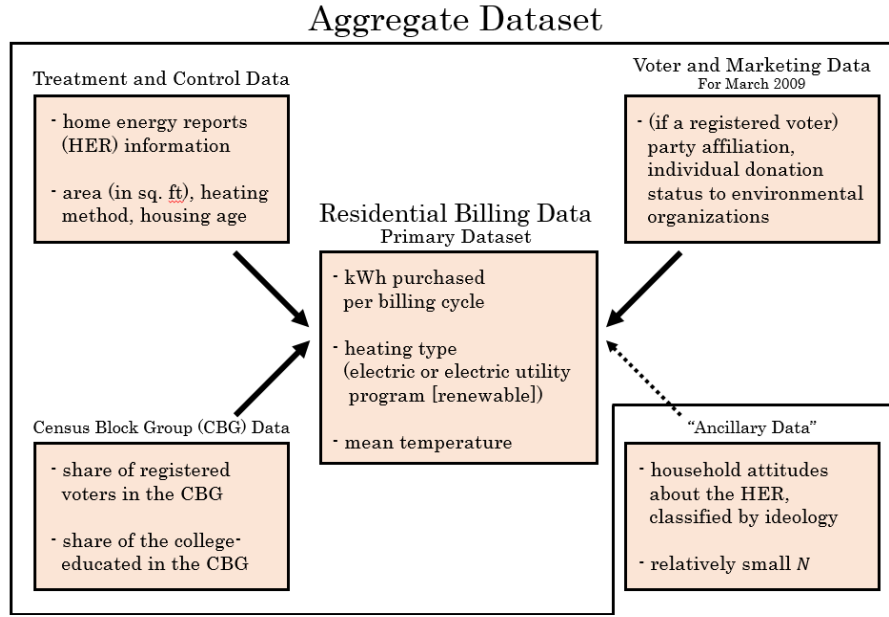
## Problem 2<sup>2</sup>

In their 2013 paper, the authors Dora Costa and Matthew Kahn seek to find answers to an important and relevant research question: “do liberals respond more effectively to electricity conservation ‘nudges,’ which compare one’s energy use to those of neighbors?” (CK 2013, p. 680). Costa and Kahn elaborate that the United States has “low taxes and little political will to sacrifice for the sake of conservation” despite the potential severity of environmental issues such as the global warming (CK 2013, p. 680). Therefore, the

<sup>1</sup>This is calculated as  $\frac{0.5 + 0.25 \times 1.5(\text{dollars})}{1/6(\text{hours})} = 6.75(\text{dollars}/\text{hour})$

<sup>2</sup>Note that for the sake of convenience, I refer to the paper by Costa and Kahn (2013) simply as CK 2013 throughout this problem.

Figure 3: Composition of the Aggregate Dataset in Costa and Kahn (2013)



NOTE: Solid arrows indicate that they have been matched with the primary dataset; dotted arrow indicates that it is available to be matched with the primary dataset.

authors draw from research in psychology to find a potential alternative that can work effectively (namely, the “nudge”) and “posit that liberal/environmentalists are more likely to respond to energy conservation nudges” (CK 2013, pp. 680-681).

In order to test the aforementioned research question, the authors design the study so that there is a “treatment” to divide the observations into treatment and control groups. The said treatment is in the form of home energy reports (HERs), which can be understood as “nudges” or small incentives to reduce the usage of electricity. A HER contains one’s monthly usage of electricity in absolute amount and provides information about how this “compares with that of 100 neighbors living in similar-sized homes” (CK 2013, p. 683). Additionally, the HERs have some advice on conserving energy written on them. Such a treatment is randomly assigned across the sample, in which each observation is a household satisfying the following conditions. Firstly, it is from one of the “85 census tracts with a high density of single-family homes.” Secondly, it has to have a “current account with electric utility” with at least one year of activity. Furthermore, an observation cannot be that of an apartment unit. Finally, it has to be between 250 to 99,998 square feet in housing area (CK 2013, p. 683).

The data used for this study (which will be dubbed as the “aggregate dataset”) was constructed from four different sources, and if one includes what the authors describe as “ancillary dataset” it would be from five different sources (CK 2013, p. 686). For a more concise representation, I refer to Figure 3. The main or “primary” dataset is the data on residential billing for electricity use, containing information about the amount of electricity (in kWh, or kilowatt hour, units) bought during each billing cycle, heating type of a residence, and mean temperature associated with it (CK 2013, p. 685). This is matched

initially with the data on treatment and control groups, which contains information about the treatment of the study – home energy reports (HER) –, size or area of housing unit (in square feet), heating method, and age of housing (CK 2013, p. 685).

There are two more data sources that the authors add in the construction of the aggregate dataset. One of them is “voter registration and marketing data,” which contains information about “party affiliation, and whether the individual donates to environmental organizations” (CK 2013, p. 686). The other is data by census block group, which contains the information on “the share of registered voters who are liberal... in 2000 and the share of the college-educated” (CK 2013, p. 686). In addition to the above datasets, another dataset, which is not used in the main analysis of the study, is also acquired by the authors. This is the previously-mentioned “ancillary dataset” that has to do with the “household attitudes about HER by ideology” (CK 2013, p. 686).

While adding different steps of analysis onto the work by Schultz et al. (2007), the authors also introduce extra layers of heterogeneity to participant characteristics. Beginning with Schultz et al. (2007), it is observable that the only heterogeneity that they introduce is the division between “[households] with energy consumption above average for the community and those with energy consumption below average for the community” (p. 430). However, in order to tackle the above-mentioned research question, Costa and Kahn consider, not only energy usage, but also a participant’s political leanings, status of donating to environmental causes, and the block group’s political leanings. This is done in collaboration with allowing for other control variables, such as home square footage, type of heating, and so forth (CK 2013 pp. 685-688).

The main finding of the authors confirms the speculation from research question – that politically liberal households are more likely to be affected by the said nudges, to a degree of two-to-four times (CK 2013, p. 680). In addition to this main finding, the authors also find that conservatives are “more likely to opt out of receiving [HER] and to report disliking the report” (CK 2013, p. 680).

### Problem 3<sup>3</sup>

- (a) One very noticeable trade-off for this problem is that between the number of observations (i.e. eligible patients) and the number of clinics to cover. Therefore, by focusing on only a small number of clinics, it would be possible to gather a large number of observations. This would indeed enable us to have a better approximation of the average treatment effect (ATE), which is can be understood as a way of averaging individual causal effects (Salganik 2018). For instance, an experiment designer may choose to focus on only 2 clinics and get 800 observations, which is the maximum number of observations given the current budget of \$1,000.

However, the above argument forgoes the possibility that there may be viola-

---

<sup>3</sup>I note that due to using the electronic version of *Bit by Bit: Social Research in the Digital Age* by Matthew Salganik (2018), I was unable to pin down and reference the exact page numbers.

tions with the *no spillovers* condition.<sup>4</sup> Why might this particular experiment be prone to spillovers? Consider the case in which for a given hospital, the news of this experiment spreads like some contagion. As more are treated with receiving text messages, there is a chance that even the control group is affected by this news. One way to avoid this predicament, then, is to lower the number of observations per hospital (so that spillover can be minimized) and, instead, increase the number of hospitals to be observed. Of course, this is giving up the number of observations (which could lead to problems with unbiasedness of our estimator for ATE), and assuming that there are no inter-hospital spillovers.

Another condition that one must consider covering a number of hospitals is when there are heterogeneous qualities between hospitals. In such a case, covering only a few number of hospitals will not measure an estimate to ATE, but rather that to conditional average treatment effect (CATE) which is said to be a “treatment effect for a subset of people” (Salganik 2018). Unless there is some assurance to assume that patients across different hospitals are somewhat homogeneous, it is difficult to argue that a difference-in-means estimator will pick up a treatment effect that approximates to the ATE of the entire population. Of course, there may be ways to incorporate differences by hospitals adding fixed effects to the regression, but it would still not be able to escape from the argument about approximating to CATE.

In summary, it would be a wiser choice to spread the experiment onto a number of schools when there are worries that the no spillover condition is violated and/or there seems to be heterogeneous qualities among hospitals and their respective patients. If not, it would be better to focus on a small number of hospitals so that the number of observations could be maximized.

- (b) While there can be a multitude of factors that may determine the effect size of study that is reliably detectable, I will consider three of such. The first has to do with the “balance” between the numbers of observations for treatment and control groups. According to Salganik (2018), there needs to be equal amounts of those in treatment groups and in control groups in order to minimize the standard error of the ATE estimator (given that the variances of potential outcomes for treatment and control are similar). However, for this particular example experiment, because an experimenter seems to have the ability to assign treatment and control, this issue does not seem to be very problematic.<sup>5</sup>

Another has to do with the type of estimator one uses. It is elaborated by Salganik (2018) that a “difference-in-differences estimator... can lead to a smaller

---

<sup>4</sup>No spillovers condition tells that for any observation, one’s potential outcome is affected, not by how others are treated or controlled for, but by one’s own status of being treated or not (Salganik 2018). This is one of the key assumptions in using the potential outcomes framework, which is described by Salganik as the “best way to understand experiments” (Salganik 2018).

<sup>5</sup>This is given that one can reasonably control for no spillover condition that was mentioned in the previous question.

variance than a difference-in-means estimator.” In order to implement a difference-in-differences analysis, then, one would have to have at least one measure each for the periods before and after the treatment. Depending on the assumptions about the experiment design, it may be the case in which additional budget is required to capture information from both of the said time periods. In that case, there would be some trade-off between number of observations and the magnitude of variance (or standard error). Also mentioned is Frison and Pocock (1992)’s work on comparing difference-in-means, difference-in-differences, and analysis of covariance (or ANCOVA), in which the ANCOVA method is recommended when there are “multiple measurements pre-treatment and post-treatment” (Salganik 2018). As there are even more data to be accrued for the ANCOVA method, it may be even harder for one to get a decent number of observations by the previous logic of limited budget.

Finally, there can be the violation of no hidden treatment effects that may interfere with the correct measurement of the impact of treatment. No hidden treatment is noted as the second part of SUTVA, and if there is any suspicion that this is violated, the reliability of the estimate will be compromised (Salganik 2018). Avoiding no hidden treatment effect could be done with a careful design of the study, but given that the treatment suggested is simply sending text messages, it may not be very robust. Considering the idea that hidden treatment effects can be associated with omitted variable bias and other forms of endogeneity, it may be worthwhile to use methods such as the previously-mentioned difference-in-differences or instrumental variables approach which may be able to account for the problem (Meyer 1995; Angrist and Pischke 2009, pp. 115-116).

## References

- Amazon Mechanical Turk. n.d. “Amazon Mechanical Turk Pricing.”  
<https://requester.mturk.com/pricing>.
- Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princeton: Princeton University Press.
- Costa, Dora L., and Matthew E. Kahn. “Energy Conservation Nudges and Environmental Ideology: Evidence from a Randomized Residential Electricity Field Experiment.” *Journal of the European Economic Association* 11 (3): 680-702.
- Frison, Lars, and Stuart J. Pocock. 1992. “Repeated Measures in Clinical Trials: Analysis Using Mean Summary Statistics and Its Implications for Design.” *Statistics in Medicine* 11 (13): 1685–1704.
- Meyer, Bruce D. 1995. “Natural and Quasi-Experiments in Economics.” *Journal Business & Economic Statistics* 13 (2): 151-161.

Salganik, Matthew J. 2018. *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.

Schultz, P. Wesley, Jessica M. Nolan, Robert B. Cialdini, Noah J. Goldstein, and Vidas Griskevicius. 2007. "The Constructive, Destructive, and Reconstructive Power of Social Norms." *Psychological Science* 18 (5): 429-434.