Mike Gruzynski

W241 – 2

Essay 2 - Final Draft

Mechanical Turks were developed by Amazon.com originally for in-house use to detect duplicate product postings on Amazon’s web site. It has since evolved from its original purpose, and now is a vast social network platform where questions can be asked and answered. One can think of it as an inverted artificial intelligence where instead of a human calling on a computer for analytics, the computer calls on the human for input. A requester (or researcher) can post a human intelligence task (or HIT) out for the public to be completed for money by a worker (or test subject). The workers can be from anywhere in the world and be associated with a variety of socio-economical groups. *The question proposed for an experimental research proposal is at what incentive package and testing constraints will the examination answered yield the most precise response*. Typically, payments for HITs range depending on the task at hand and how long it takes to complete the task. One of the big concerns with (unsupervised) Mechanical Turk experiments is that the workers can just be completing the current HIT quickly (or just clicking answers to get through the exam) in order to get paid quicker and move to the next HIT task to get paid once again. The question of at what incentive level does the worker respond with valuable attention and answers is an important question to ask, if this type of experiments are used.

Many universities and researchers use Mechanical Turk in order to get high sample size experiments quickly and at a low cost. The use of this network raises the concern that the results from these experiments might be imprecise in the terms that the workers speed through, with possibly randomly answers in order to finish the assignment quicker. Research in this subject has been performed previously [1], but additional experimentation can be developed in order to answer additional questions. The experimental result is of interest because if researchers or experiment creators are using Amazon Mechanical Turk for either an experiment or a pretest experiment, there is some benefit to figure out if the answers from the experiment have any variation of results based on incentives and types of questions asked. Another concern with this experimental setup is if the Mechanical Turk process provide value (on both capturing heterogeneity differences in population and testing results) above other options of college, or other internet experiments. The anonymous worker subjects have an unknown motivation and unknown level of focus at examination time and this could produce results that are not ideal for experimental results. An optimal incentive package is important to study in order to get a better understanding of what you need to offer to workers to get any value back from the experiment. An arbitrary value for payment is not ideal because you will either pay too much money for the experiment, or pay too little and the workers will not answer in a way that will yield useful results.

There will be a variety of treatments that will be exposed to the subjects. For the control group, the constraints will be the standard 20 question exam that will have no time limits and upfront compensation will be calculated from the average monetary amount for a 20 question exam on the network and no monetary reward (bonus). The workers will receive no payment until the exam is completed and if the exam is not completed, the exam results will have a special code-in value that signals an unfinished exam. The control amount will be gathered by data mining the Amazon mechanical Turk website and calculating the average payment of this type of experimental task. There will also be no reward based system in the control group, so the payment upfront is the total payment for the HIT. The control will give a baseline of expected means and variances in order to compare to various treatment groups. The treatment cases will need to vary incentive packages in order to get main effects and to discover what conditions in Amazon Mechanical Turk experiments give more optimal and precise results (highest likelihood of correct responses range). Incentive packages will need to be varied in the total upfront payment for completing the task. This allows us to figure out a key question of measuring causal effects of upfront payment on optimal experimental results. In addition to upfront payments, the assigned amount of reward money will be given to a worker who gets a certain percentage of the 20 questions correct. The information of the percentage correct needed to earn the reward will be given before the exam is administered and the worker will only know the percent correct on the test once the exam is finished. In other words, there will be no indicator if the previous question was answered correct or not during the test. This condition will be used to study the incentive mechanism term on an exam to see how focused and dedicated the subjects become with a varying amount of promised and rewarded money. The bonus (variable by treatment case amount) will be awarded to the worker if they get an 80% or higher on the exam.

The test will be set up to run on the Mechanical Turk web page and the test will be a set of 20 logic and illusion based questions evenly split. The even split will also allow for an observation on which type of question is answered better and could be used as data to incite an additional experiment on what types of questions are best suited for an unsupervised web-based exam. More research is needed, to explore the types of questions asked on a typical Machine Turk experiment. The research will help define if the types of questions in the experiment are a representative subset of the types of questions other Mechanical Turk experiments contain. This questions will be similar to ones you see in undergraduate psychology classes or early school logic tests and will not be impossible to get the correct answer if the workers actually read the prompt and think about the answer. This is in order to trap people that quickly move through the exam in order to receive payment faster so they can move on to another task to get paid once more. The answers to the exam will be skewed and not equally distributed over the options. For example there are 20 questions in this exam with options of A, B, C, and D for answers. In order to reduce the likelihood of workers randomly selecting correct answers, there will be an uneven amount of answers on two of the four choices. The non-equally distributed characteristics will lower the probability of answering the questions correct at random.

The testing output data will contain percent correct of each exam question, time taken on each exam question, and amount of changed answer on each exam question (or the amount of different choices clicked before answer submittal). The expected value of the exam is unknown until we collect enough samples to get the overall mean and variance of the data. The difficulty level of the test will be designed so that anyone eligible to take the exam will be able to score a higher score. In other words, there will not be advanced differential equations on the exam, but instead logic questions that can be answered accurately if the worker has a basic high school education. Exam question difficulty level will be explored during the pretest stage of the experiment.

More research will need to be done on the Mechanical Turk social network and to see how experiments are actually run. For now, the randomization will be done on an individual basis, so that every worker’s incentives will be varied. The actual incentive package (upfront and reward) ranges in value will be investigated during the pretest phase of the experiment. Independence will be managed by not allowing a worker to take the exam more than once. That way we will not get a subject taking the same exam multiple times, skewing the results of the exam. One thing to note is that we cannot control if user makes additional accounts and signs up for the study again, under another alias.

The outcome measurement that matters are the amount of exam questions answered correctly for each incentive plan along with more consistent time spent on each question so that each question has similar amount of attention. The results will need to be studied with variables of upfront money, and reward money and will be analytically combined together to get main effects of the input variables on the output objective along with other statistical analysis. Some question to consider are:

* Is there an optimal amount for incentive package vs success of exam questions?
* Is there a law of diminishing returns on amount given to worker and their performance?
* Are different mechanism inside the test such as constraints on how long the minimum time taken on a full experiment related to overall performance of the exam?

This experiment has the capability of using blocking and/or clustering mechanisms in the experimental design. We could group on country that the survey was taken in, age groups, ethnicity, level of education, etc. However, this experiment's purpose is get a causal inference on amount of payment to the workers with interaction terms of reward system and testing constraints on the quality of performance of the worker and therefore no clustering will be used in this experiment. This is not a social-economical examination of varying group type and their expected value and variance of the test. This test is open to anyone registered with the Amazon Mechanical Turk and participants will be randomly assigned to control or treatment cases. Blocking may be considered in order to get even amount of variation treatments in each group in order to help rate comparable test subjects, and will be explored in the pretest phase as well.

From the output data, discussed above, we can do statistical analysis on the data in order to get a clearer picture of the experiment. We can perform a pooled regression on the dataset in order to gather slope coefficients characteristics and how efficient and unbiased our results appear. In addition to a pooled OLS, we can bin data into groups based on ranges of incentive package to see the variation in slope among testing bins, which will help gather a trend on what incentive package and figure out fixed effects of each category in order to figure out best testing scenarios. In addition to regression, we can perform t-test comparison on the groups to see if the bins are statistically or practically significantly different. All this data will help figure out expected mean of each binned testing situation along with the standard error of variables. Combining the expected mean, the standard error, and regression can help uncover the causal inference of the experiments and ranges of confidence in each treatment group. This will help determine a more ideal amount of money to pay workers and help discover if incentive packages cause more precise results.

With online unsupervised tests, there are a lot of risks that comes with the task. The first thing to consider is selection bias. The question of: does the online community sample registered to the Amazon Mechanical Turk social network represent the population as a whole? [2] The answer is most likely no, because we are only offering this to English speaking workers who happen to be online and willing to take this experiment. This subset of the population is not randomly selected and may appear to represent a small sample of the population rather than the population itself. Data may be skewed in responses during this test and may have to come up with some normalization factors. In addition, we have no idea if these are bots created to take the test or actual humans. Bots may be programmed to randomly select answers based on types of words used inside the sentence. Also the experiment is occurring at one time of the year, for a fairly short amount of testing window (due to length of the class). There may be some reason why these people are taking the exam and may have gotten other types of users if it was given at another time during the year.

Researchers can also perform pretest simulations in order to get experimental fallouts figured out before the real (more expensive) experiment is performed. If a pilot test is ran before the larger experiment, we can test to see if we can figure out if a majority of the workers are bots, and how to filter them out of the final data. Solutions like using images of questions vs actual text, so that the bot can't parse the text and then answer the question, or the use advanced logic flips in the English language to confuse the bot. In addition to the bot check, the pilot experiment can help make some predictions on how many subjects will actually sign up for this type of exam and to see if we can actually secure a minimum of 100 observations for the final experiment with at least 30 observations in each treatment group. Also, we can tune the level of difficulty of the test questions. We may create an exam that is too difficult or is too easy, so that a majority of the test data results are consistently low or high. This will hurt the regression and statistics of the exam and we will have unclear causal inference. Thus the pilot test will help us fine tune the experiment and maybe figure out a distribution of easy, medium and hard types of questions (along with question types) in order to get a smoother distribution of worker examination scores for the final statistical analysis. One of the biggest concerns of the experiment is if we have enough statistical power in the experiment with the various characteristics of the experiment. This will be examined with the pretest will help determine if we have enough statistical power with the varying amount incentive packages. The concern is having too many variation in the design of experiment, which will weaken the statistical power vs a smaller design variation with larger range of inputs. In addition to the statistical power investigation, there we will also see if there are certain incentive packages that obtain huge variation of responses or even a huge lack of sign up interest or large drop rates. All of the information from the pretest investigation will be used to update the experimental design in order to get a more ideal experiment setting for the final experiment.

References:

1: <http://journals.sagepub.com/doi/abs/10.1177/1745691610393980>

2: https://scholar.harvard.edu/files/dtingley/files/whoarethesepeople.pdf