**Experiment Overview**

To test the hypothesis and the research question, we plan to test each of the three extrinsic motivations one at a time. We also plan to combine two or three of these motivations to improve the treatment power after the learning about each motivation is acquired. We therefore design the experiment around testing one motivation and then repeat this design with only changing the motivation type. Below (Table 1) is a summary of the treatments we plan to test.

Table 1: Mechanical Turk Worker Motivation Experimental Design



**The In-Lab Experimental Design**

Blocking Variables and Clusters

We plan to collect the following information as the blocking variables (or covariates) through a pre-treatment survey to improve the test power.

1. Workers’ demographic information
2. Tools they use to do bounding boxes

At first, we also plan to include current bounding-box accuracy and current motivations as the blocking variables; however, due to the concern that these variables may leak information about the future treatment, we have decided to drop them. Please refer to the Appendix for a copy of the presurvey.

Because bounding box tasks generally don’t require particular skill sets, we decide not to use “location”, “HIT approval rate”, and “number of HITs taken” to limit the worker universe. We do require each worker only taking the task once to isolate each treatment effect.

Since workers are spread all over the world, most workers don’t form a cluster in the study. The only exception is that there is a low chance where more than two workers live in the same house/apartment. Overall, we think there is no need to consider clustering in this project.

Randomization

After workers finish the pre-survey, we acquire their worker IDs, and then we are able to email and/or assign tasks to them. We randomly split the worker IDs into the treatment and the control group by blocking on the blocking variables mentioned above.

Task

We ask each worker to bound 20 images for this task. We hope to see as they go through the images to bound, their accuracy drops due to the mechanical nature of the HIT.

Measurement

We plan to measure the Euclidian distance between the smallest bounding box (ground truth) and those that each worker draws as the success metric. (?may change to ratios)

**The Pilot Study**

We launched the pilot study with a pre-treatment survey on 50 workers. The survey results showed, as expected, that workers were a homogeneous group of people, most of whom are young people between age 20-30, college education, earning annual income between 30K and 60K.

We then randomly created the treatment group and the control group through blocking on the pre-treatment survey result. Either group had 25 workers. Both groups were contacted via emails about being assigned a 20-image-bounding-box task, with the control receiving no treatment message but the test receiving the message, “Your work will be used to develop a government surveillance system.”

Around 50% of the workers (12 workers) in the treatment group didn’t take the bounding box task, and 60% in the control group didn’t do so as well. The reasons for the high attrition could be that there was a time gap, a couple of days, between the pre-treatment survey and the bounding box task. Additionally, in our survey, we didn’t specify that there would be a follow-up task after the survey. Because these workers may not be online every day, the lack of noticing about the follow-up task may be a reason for the high attrition.

From the learnings in the pilot study, we decided to give up using blocking variables, which required collecting data in two steps; rather, we would collect data in one task and use covariates instead to improve the treatment power. Additionally, we also decided to task workers with one image because the current platform design allowed multiple workers to grab any image in the twenty-image task, which disabled one worker to finish the entire twenty images.

**The Implemented Experimental Design**

Table 2: Mechanical Turk Worker Motivation Implemented Experiments



As summarized in table 2, in Experiment I, we varied both payment amount and payment messages. According to current studies (need to cite!), the primary motivation of Mturk workers is for immediate rewards. By raising the payment amount and stating a warning about declining their work, we hoped to see that bounding boxes in the treatment group were more accurate than those in the control group, in which, workers were paid less than a half of the amount and were not given the warning message.

Experiment II’s messaging was inspired by feedback from the workers. A few workers emailed us their feedback about our tasks, expressing their willingness to do more similar tasks for us. Maintaining a good reputation to acquire stable task opportunities fell in the “future payoff” category, and we then decided to try this messaging, “We use the quality of this bounding-box task to help decide if we would assign our future tasks to you for assistance.”

Although the accuracy of Mturk HITs sometimes is indeed important, because they often do not require particularly trained skills to perform, they are often considered low-paid, low-importance, and mechanical tasks. Humans tend to try to do good jobs if they are aware that their work matters. Therefore, we decided to use this messaging to motivate the workers, “Your work will be used to develop a self driving car system.” We also tried a negative impact in this category, “Your work will be used to develop a government surveillance system,” hoping to see the accuracy being worse than the control.

**Appendix**

The pre-treatment survey

1. Have you done any HITs that require putting bounding box on images before? Below is an example: the left side is the original picture ( not shown below but in the actual survey), and you are asked to draw a box on the main object on the picture as shown in the right picture.

A. Yes

B. No

C. I don't remember/I don't know.

2. How large is your monitor that you use for Mturk HITs?

A. Cellphone size (6.5" or less)

B. Tablet size (7" to 10")

C. Small laptop size (11" to 13")

D. Mid-size screen (14" to 17")

E. Large screen (18" or larger)

F. I don’t know

3. To move the cursor do you primarily use which of the following?

A. Touch Screen

B. Mouse

C. Trackpad

D. Other

4. What best describes your age group?

A. Under 21

B. 21-30

C. 31-40

D. 41-50

E. Over 50 years old

5. What best describes your annual household income in US dollars?

A. Equal to or below $10,000

B. $10,001 to $30,000

C. $30,001 to $60,000

D. $60,001 to $90,000

E. $90,001 or higher

6. What best describes your education?

A. Less than high school

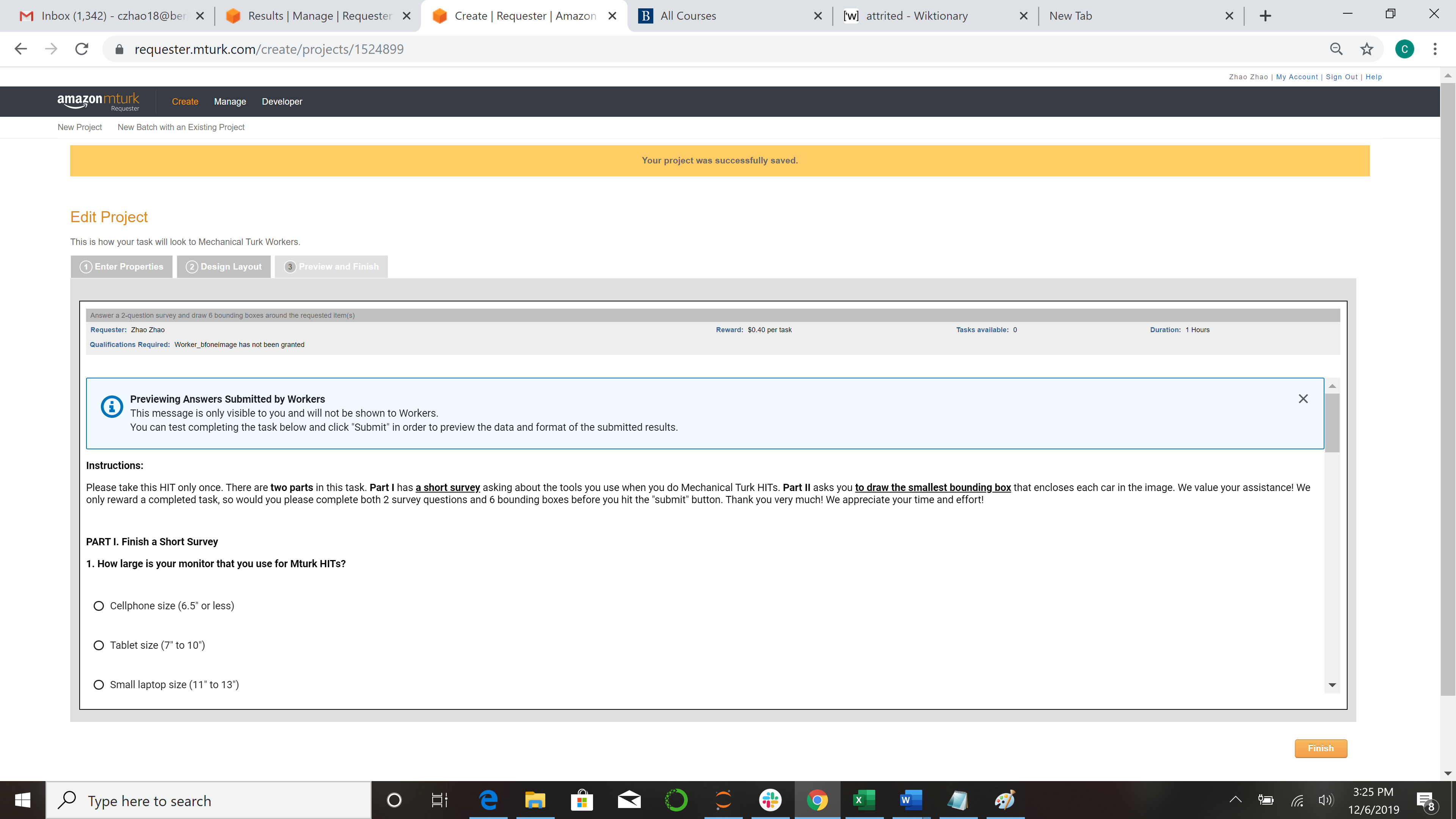
B. High school graduate

C. Some college

D. 4-year college degree

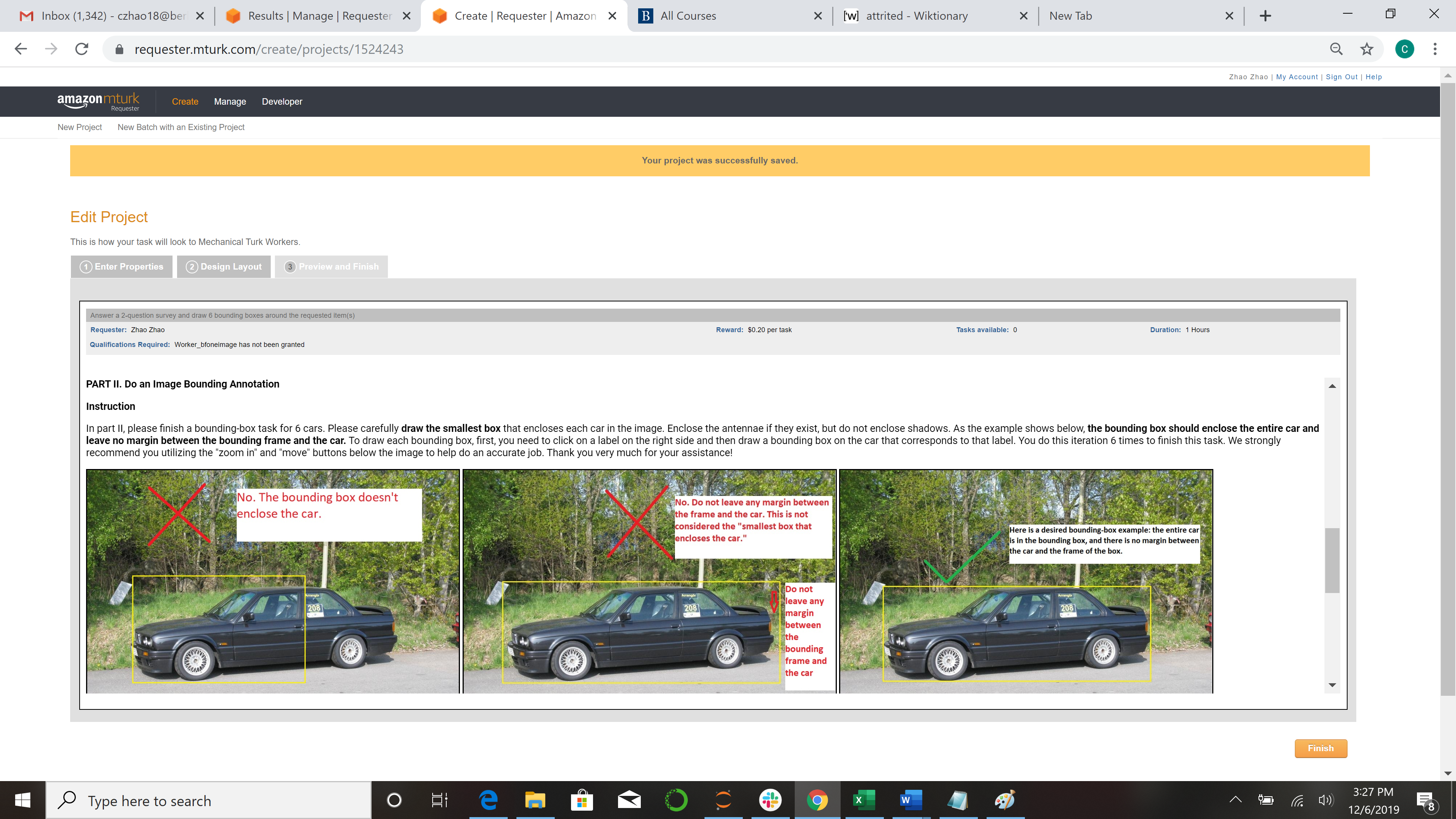
E. Master of professional degree or higher

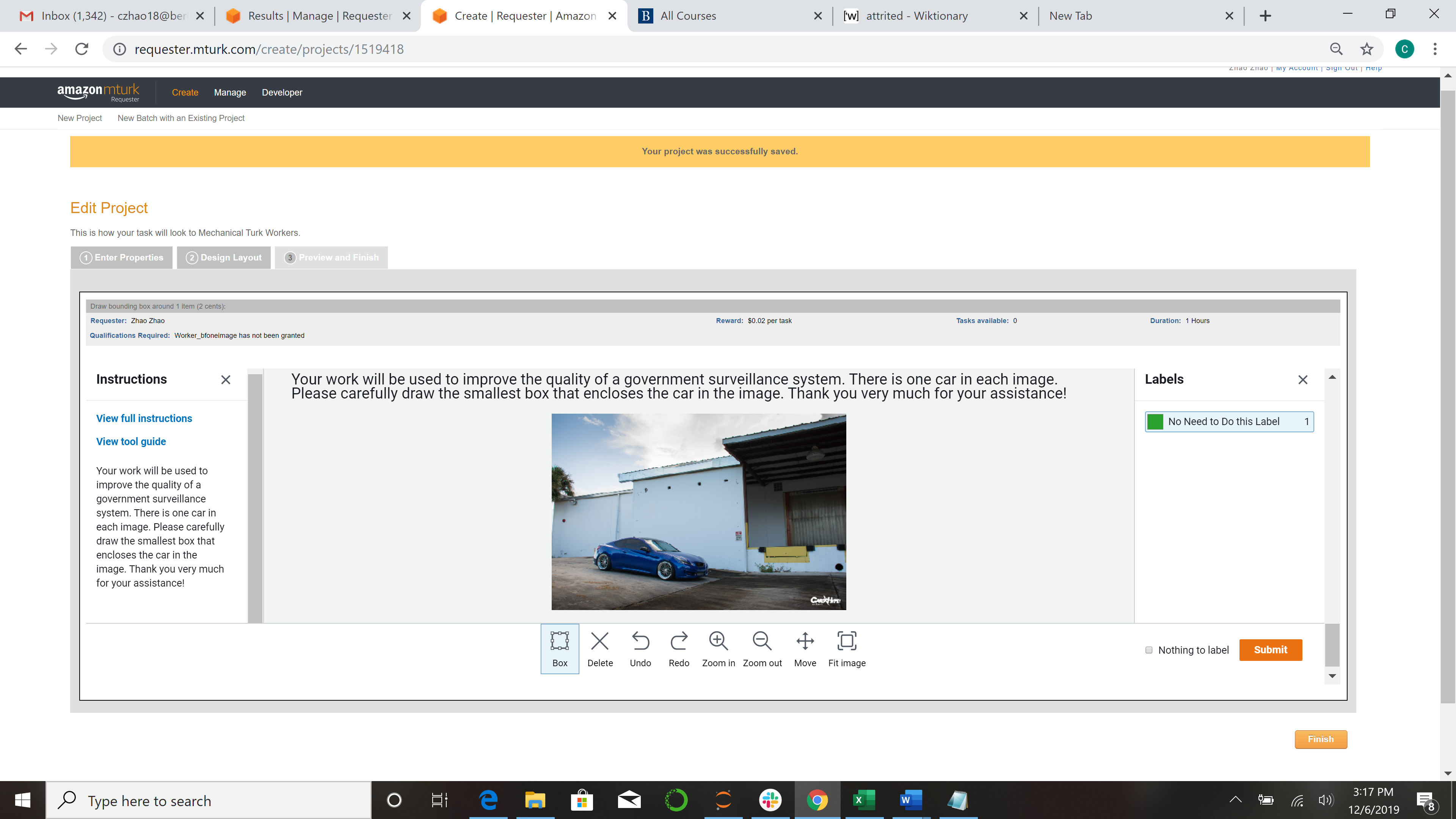
The Survey Screen Shot



**The Experiment Tools**

To label one image





To label multiple images

