Heterogeneity of Experimental Findings: Evidence from Real-Effort Tasks

Pre-Analysis Material – Addendum on Forecasts

July 2018

Stefano DellaVigna, UC Berkeley & NBER

Devin Pope, University of Chicago & NBER

In this project, we consider the heterogeneity of experimental findings. For example, how sensitive are experimental results to the demographics of the sample? How do they depend on the output measure used? How critical is the experimental protocol?

We consider these questions in the context of a real effort task that builds on DellaVigna and Pope (2018) and that we run in May and June 2018. In this 2018 experiment, we run the same group of 16 experimental treatments repeatedly, in different version. Across the different version, we consider different demographic groups, and we vary the task used, the measure of effort, and whether subjects know that the task is an experiment. We examine how much the experimental results change in response to these changes, compared to a pure replication. We also elicit which of these changes experts believe will have the most impact.

In this short document, which builds on the main pre-analysis plan, we outline how we obtain expert forecasts for measures of heterogeneity of experimental results in our experiment. The document is written after the completion of the data collection for the real effort experiment on MTurk, but before the collection of the expert forecast data.

Experiment

We refer to the main pre-analysis plan for a more detailed outline of the experiment itself. In short, we run four versions of a real-effort task.

Version 1 is a direct replication of the task used by DellaVigna and Pope (REStud 2018): participants press two alternating keyboard buttons ('a' and 'b') within a given time period (10 minutes) (Figure 3). There are 16 different randomly-assigned treatments with varying levels and types of incentives, of which 15 are present in DellaVigna and Pope (REStud 2018) and one is an additional treatment which combines a psychological intervention with a piece rate incentive. For example, some participants will be given bonus payments based on points scored, others will raise money for charity, and some will receive no bonus.

Version 2 is similar to Version 1 except that the task consists of coding the occupation from World War II historical cards for 10 minutes (Figure 4). As in version 1, the participants are randomized in one of 16 treatments with varying levels and types of incentives. The treatments are parallel to the ones used in Version 1.

Version 3 involves coding of WWII cards, as in Version 2. However, this time all subjects are required to code 40 cards as their main task (Figure 5). After the coding of the 40 cards is completed, subjects are randomized into 16 different treatments, parallel to the ones used in Version 1 and 2. The treatments consist of different incentives offered to the participants in case they agree to code additional WWII cards, up to 20 extra cards (Figure 6).

Version 4 is the same as version 3, except that, unlike in Versions 1-3, there is no consent form (Figure 2) prior to participating in the tasks, so subjects do not know that the task is an experiment. (Notice that the participants are coding historical data as part of an economic history project, so this is a data coding assignment.)

The key outcome variable is a measure of average effort. The measure of effort differs across the four different versions of the experiment. In versions 1 it is the number of points scored by the subject in 10 minutes, where each point is scored as a result of pressing 'a' then 'b'. In versions 2 it is the number of WWII cards coded by the subject in 10 minutes. In versions 3 and 4 it is the number of extra WWII cards coded by the participants beyond the required 40 cards. Within each version and within each of the 16 treatments that we run in each version, we compute the average of the measure of effort across the subjects in that version-treatment cell.

The details of the different versions and treatments are in Table 1 in the main pre-analysis plan document.

Sample

In the pre-analysis plan, we set out to final sample to exclude subjects that:

(1) do not complete the MTurk task within 30 minutes of starting;

(2) exit and then re-enter the task as a new subject (as these individuals might see multiple treatments);

(3) are not approved for any other reason (e.g. they did not having a valid MTurk ID);

(4) In version 1 (a-b typing) do not complete a single effort unit; there is no need for a parallel requirement for version 2 since the participants have to code a first card to start the task;

(5) in version 1 scored 4000 or more a-b points (since this would indicate cheating);

(6) in version 2 coded 120 or more cards with accuracy below 50% (since this would indicate cheating);

(7) in versions 3 and 4 completed the 40 required cards in less than 3 minutes with accuracy below 50%, or completed the 20 additional cards in less than 1.5 minutes with accuracy below 50% (since this would indicate cheating).

We also planned an ideal number of subjects of 10,000 people completing the tasks. We planned to keep open the task on Amazon Mechanical Turk until either (i) three weeks have passed or (ii) 10,500 subjects have completed the study, whichever comes first.

We followed the pre-registration sample rules. The experiment ran for three weeks. After three weeks, we had 12,983 recorded responses on Qualtrics, from which we first removed 324 observations because they had re-entered the task and therefore may have seen multiple treatments. The largest cut to the sample (2,660 observations) occurred when removing those who had either taken more than 30 minutes to finish or not completed the survey at all. We then 89 individuals who had not been approved for reasons such as an invalid MTurk ID and blatant cheating on the tasks (less than 10% accuracy on the cards). Finally, we removed 40 individuals with no button presses in the a-b typing task and those who coded quickly with less than 50% accuracy.

A final restriction not included in the preregistration were Qualtrics data "glitches." We removed observations with the following data errors: (i) Missing treatment variable; (ii) Negative time stamps; (iii) Descending time stamps; (iv) Time stamps that go beyond 10 minutes in the first task (with a 10 second leeway for early timer starts); (v) More than 10 time stamps than total coded cards. In total, these restrictions removed 59 observations. We do not include these observations because the data is likely not accurate of the surveyor's behavior and probably a result of a "glitch" from Qualtric's end.

In total, we are left with a final valid sample size of 9,811 responses, close to the envisioned sample of 10,000.

Measure of Heterogeneity of Experimental Findings

Our focus is on comparing how different samples and versions of the experiments affect the experimental results. We focus mostly on the 15 treatments which are present both in the 2015 experiment as well as in the 2018 experiment; as additional evidence, we present also the results from the 16th treatment which combines a psychological inducement as well as incentives. In what follows, though, we focus on the 15 treatments, which are also the ones that experts are asked to place forecasts on.

As measure of heterogeneity of experimental findings, we compare the results of the 15 treatments across the different versions. We considered different measures of heterogeneity. Since we compare versions with very different output scales (e.g., coding of WWII conscription cards versus a simple button pushing task), we opted for a measure that is unit-free. We thus considered the Pearson correlation and the rank-order correlation. We opted as main measure for rank-order correlation because a natural measure of stability is that the order of effectiveness of the experimental manipulations should be preserved. The Pearson correlation builds in a stronger assumption of linearity between the treatments in one version versus another.

To be more precise, one can think of the stability of experimental results as follows. Our structural estimates of the effort in the various treatments in DellaVigna and Pope (REStud) depend on two set of parameters: behavioral parameters and incidental parameters. The behavioral parameters are the ones which we can expect to be stable across versions, such as the discounting parameters beta and delta. In contrast, the incidental parameters – curvature of cost of effort, level of cost of effort, and baseline motivation – surely will differ across versions. For example, the level of the cost of effort much be higher for a task that takes longer to execute, such as coding of WWII cards, compared to a simple push of a-b buttons. These tasks likely also may differ in the elasticity of effort to motivation, as well as in the baseline motivation.

We can then define two versions to have stable experimental findings if they share the same behavioral parameters, even if the incidental parameters vary. As simulations show, given how we set up the treatments across version, this translates into the same order of treatments across the different versions, but the average effort for the 15 treatments will vary in a non-linear manner across version; hence, our preference for rank-order correlation, as opposed to Pearson correlation.

An important wrinkle in this is that some of the behavioral parameters are expressed in terms of baseline motivation – such as the increase in baseline motivation due to gift exchange – and thus are likely to vary across versions as the tasks, and thus the baseline motivation, change. We handle this in three ways: (i) by assuming that these behavioral parameters do not change at all across version: (ii) by assuming that they change proportionately to the baseline parameters; (iii) by assuming a Cobb-Douglas transformation combining (i) and (ii).

Expert Forecasts

Given the choice of rank-order correlation to measure the stability of findings, we elicit from the experts the stability in terms of rank-order correlations. More precisely, we ask for 10 rank-order correlations, as we vary sample, task, and measure of output, among other changes. Before we ask any rank-order correlation, we provide examples to illustrate how much changes in the order of treatments are likely to change the rank-order correlation.

Also, for the initial predictions we provide the rank-order correlation which would obtain if the results were completely stable. Consider for example the case of pure replication, comparing the 2015 button pushing results to the 2018 button pushing results. We do Monte Carlo simulations from the 2015 data of new samples of the size of the actual 2018 data and compute the rank-order correlation. This will be less than 1 because the noise in the estimate can flip the order of treatments which are close in average effort. We provide subjects the average rank-order correlation across these simulations, given that each simulation can yield a different correlation number. Similarly, we can also provide the forecasters a benchmark average correlation that would obtain comparing across results for different demographic groups.

We implement only one (randomized) element in the expert survey. When forecasting about the treatments with new outcome variables (Versions 2, 3, and 4), one half of the forecasters will receive information exclusively on the mean of the effort variable and the standard deviation. A second half, instead, will in addition receive information also on the average effort in the three benchmark piece rate treatments: the no-pay treatment, the low-pay treatment, and the high-pay treatment. We indicate the average effort in each of these treatments, as well as the standard error. This provides valuable information on how much the effort in that version responds to motivation, and how much noise there is in the data. We are interested in how much this information influences the expert forecasts.

The attached Appendix reproduces in detail the Qualtrics survey with the 10 rank-order correlations. The two versions, with more and less information, are denoted 1 and V2.

Expert Sample

To draw the main expert sample, we wanted to build on the sample of 200+ experts that provided forecasts for the 2015 experiments, given that these experts were familiar with the original experiment. At the same time, we wanted to scale back the sample given the obvious value of people's time and given that the original forecast sample of 200+ people provided plenty of statistical power: our 2015 forecasting results suggest that a couple dozen respondents are enough to achieve the wisdom-of-the-crowd effect.

Thus, we narrowed the sample as follows: (i) PhD since 2005; (ii) behavioral economics is the main, or second, field of specialization; (iii) the expert provided a forecast in 2015. Out of the resulting sample of 73 expert, we picked 42. In addition, we added 18 behavioral economists with PhD since 2015 (who were not included in the early sample). The latter names were largely drawn from list of attenders and presenters at key conferences in the behavioral area (BEAM and SITE Psychology and Economics).

As additional samples, we also intend to ask a population of PhD students in economics, like we did in 2015, covering at least Berkeley and UChicago. We will also seek the responses of a group of experts working on replication, since the topic studied is related to the issue of conceptual versus exact replication.

Consent Page

THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
Consent for Participation in Research

Principal Investigators: Stefano DellaVigna and Devin Pope

IRB Study Number: 18-0883

DESCRIPTION: This is a study being conducted by researchers at UC Berkeley and the University of Chicago to understand the ability of people to predict experimental results.

RISKS and BENEFITS: This study does not involve any physical or emotional risk to you beyond the risks of daily life. Your involvement in this experiment may benefit the field of economics by helping to advance theories about prediction ability.

CONFIDENTIALITY: The information collected may be published in articles or academic presentations, but your personal identity or your involvement as a research subject will not be published or revealed. Information collected during this study (e.g. time of survey, IP address, responses) will be retained by the researchers and may be used in future research projects, but again, your personal identity or involvement will not be published or revealed.

SUBJECT’S RIGHTS: Participation is on a purely voluntary basis. Your involvement in this study is appreciated, but you may quit participation altogether at any time without receiving any penalty or prejudice.

If you have questions about this project, you may contact us at:

Kristy Kim
University of Chicago Booth School of Business
5807 South Woodlawn Avenue
Chicago, IL 60637
Email: kristykim@chicagobooth.edu

If you have any questions about your rights as a participant in this research, you can contact the following office at the University of Chicago:

Social & Behavioral Sciences Institutional Review Board
University of Chicago,
1155 E 60th Street Room #411
Chicago, IL 60637
Phone: 773-834-7835
Email: sbs-irb@uchicago.edu

Please indicate below that you are at least 18 years old, have read and understand this consent form, and you agree to participate in this online research study.

○ Yes

○ No

Page 1

Thank you for participating in the survey. We will ask you for 10 predictions, which should take 10 - 20 minutes. Our goal is to examine how experimental results change with different designs, and how experts like you anticipate such changes. The current project builds on a large real-effort experiment on Amazon Mechanical Turk (MTurk) that we (Devin and Stefano) ran in 2015, and for which you may have provided forecasts in the past.

The MTurk participants in that study initially agreed to perform a simple task that takes 10 minutes in return for a fixed participation fee of $1.00. They were not given information about the task or about possible bonus money before agreeing to participate. As part of the experiment, they were offered different bonus payments to encourage them to perform well. In bold below is the task exactly as it was described to the MTurk participants:

> **On the next page you will play a simple button-pressing task. The object of this task is to alternately press the 'a' and 'b' buttons on your keyboard as quickly as possible for 10 minutes. Every time you successfully press the 'a' and then the 'b' button, you will receive**

a point. Note that points will only be rewarded when you alternate button pushes: just pressing the 'a' or 'b' button without alternating between the two will not result in points.

Feel free to score as many points as you can.

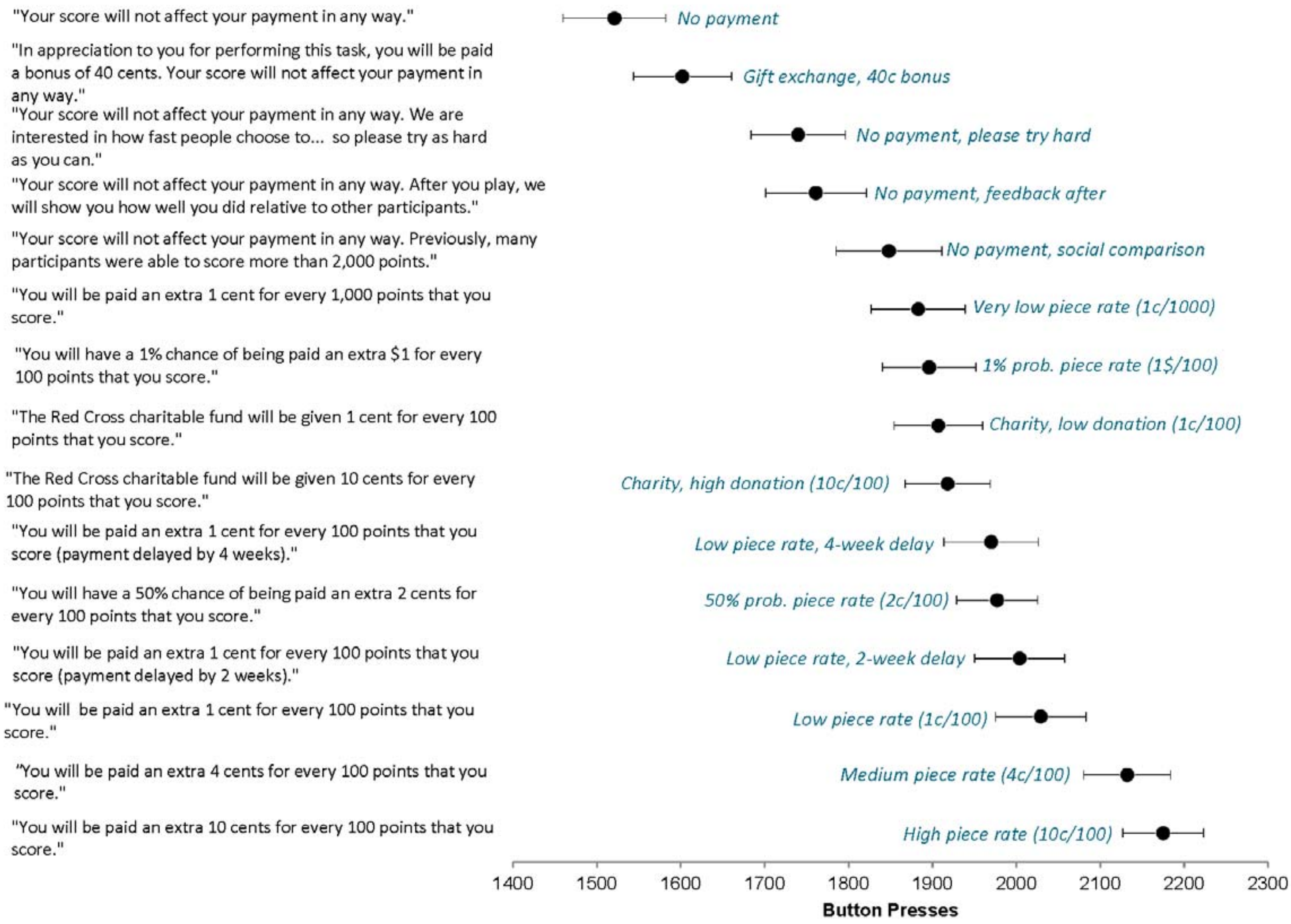
[The participant would then see a different final paragraph depending on the condition to which they were randomly assigned.]

Click here if you would like to sample this button pressing task yourself. Each MTurk participant was randomly assigned to one of 18 conditions. The exact wording for each of the conditions is listed below. The number of participants in each condition was around 550 for a total of 9,861 participants.

After seeing these instructions, MTurk participants saw a screen with a 10-minute countdown timer and were told to start the task. A visible counter allowed participants to keep track of the number of points scored and bonus money received. Below the counter, the participants saw a 1-2 sentence reminder of their incentive condition. Individual scores for this task ranged from 1 to 3,950 points, with an average of 1,936 and a standard deviation of 668.

In the figure below, we show the results from our previous experiment for the 15 treatments that are of interest for today. (We exclude three threshold performance treatments.) Each row displays the key paragraph that distinguished the treatments, as well as the mean effort under the treatment (with confidence intervals).

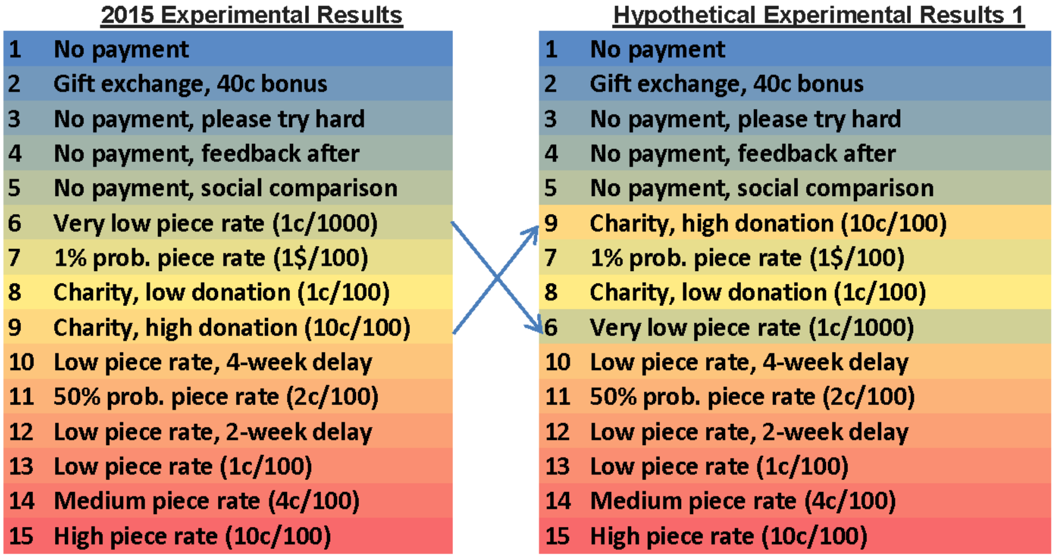
Button Presses by Treatment with 95% Confidence Intervals



Page 2

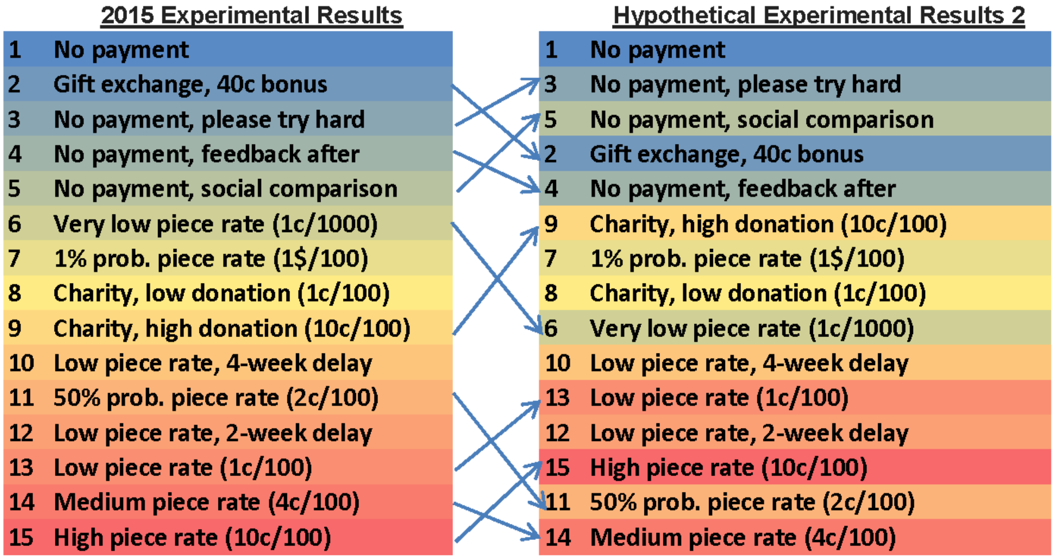
In our new project, in May 2018 we ran a new experiment with four versions. Like the previous experiment, the new experiment was also pre-registered. Each version includes the same 15 treatments highlighted in the previous page. Importantly, the experimental design differs across the four versions. The design changes include using a different effort task, using a different measure of effort, and varying whether the subjects know that they are in an experiment. We are interested in your ability to predict how the effectiveness of the treatments varies from one version of the experiment to the next version.

We will elicit your predictions by asking for the rank-order correlation between the average findings for the treatments in one version, versus in another version. The rank-order correlation is simply the correlation between the ranks of the treatments in the two versions. To familiarize you with this measure, below we show you four hypothetical examples, with their rank-order correlation. In each of the four examples, the left version has the treatments in the order of effectiveness in the 2015 experiment, as you saw from the previous figure. The right version shows hypothetical examples of possible experimental results which would end up changing the order of treatments. (You can see treatments that change rank indicated by arrows.) As you can see, the more the treatments change order between versions, the lower the rank-order correlation.



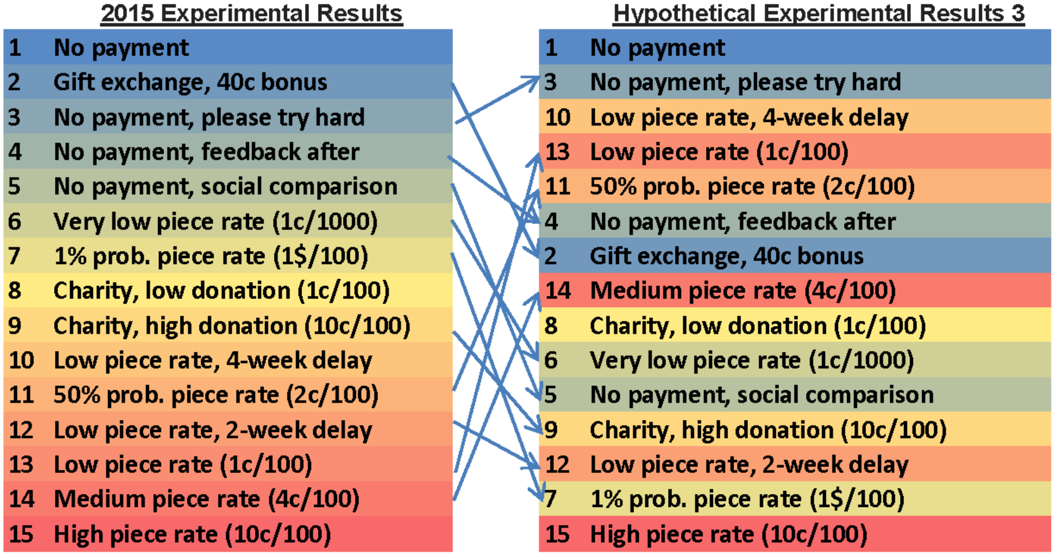
	2015 Experimental Results		Hypothetical Experimental Results 1
1	No payment	1	No payment
2	Gift exchange, 40c bonus	2	Gift exchange, 40c bonus
3	No payment, please try hard	3	No payment, please try hard
4	No payment, feedback after	4	No payment, feedback after
5	No payment, social comparison	5	No payment, social comparison
6	Very low piece rate (1c/1000)	9	Charity, high donation (10c/100)
7	1% prob. piece rate (1$/100)	7	1% prob. piece rate (1$/100)
8	Charity, low donation (1c/100)	8	Charity, low donation (1c/100)
9	Charity, high donation (10c/100)	6	Very low piece rate (1c/1000)
10	Low piece rate, 4-week delay	10	Low piece rate, 4-week delay
11	50% prob. piece rate (2c/100)	11	50% prob. piece rate (2c/100)
12	Low piece rate, 2-week delay	12	Low piece rate, 2-week delay
13	Low piece rate (1c/100)	13	Low piece rate (1c/100)
14	Medium piece rate (4c/100)	14	Medium piece rate (4c/100)
15	High piece rate (10c/100)	15	High piece rate (10c/100)

Rank-order correlation between 2015 results and hypothetical results 1 = 0.968



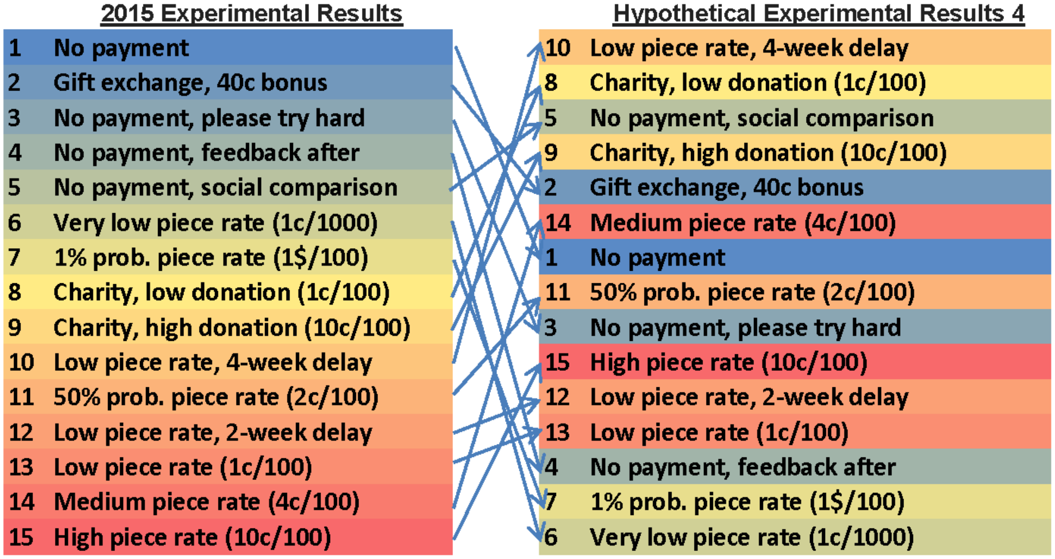
	2015 Experimental Results		Hypothetical Experimental Results 2
1	No payment	1	No payment
2	Gift exchange, 40c bonus	3	No payment, please try hard
3	No payment, please try hard	5	No payment, social comparison
4	No payment, feedback after	2	Gift exchange, 40c bonus
5	No payment, social comparison	4	No payment, feedback after
6	Very low piece rate (1c/1000)	9	Charity, high donation (10c/100)
7	1% prob. piece rate (1$/100)	7	1% prob. piece rate (1$/100)
8	Charity, low donation (1c/100)	8	Charity, low donation (1c/100)
9	Charity, high donation (10c/100)	6	Very low piece rate (1c/1000)
10	Low piece rate, 4-week delay	10	Low piece rate, 4-week delay
11	50% prob. piece rate (2c/100)	13	Low piece rate (1c/100)
12	Low piece rate, 2-week delay	12	Low piece rate, 2-week delay
13	Low piece rate (1c/100)	15	High piece rate (10c/100)
14	Medium piece rate (4c/100)	11	50% prob. piece rate (2c/100)
15	High piece rate (10c/100)	14	Medium piece rate (4c/100)

Rank-order correlation between 2015 results and hypothetical results 2 = 0.918



2015 Experimental Results		Hypothetical Experimental Results 3	
1	No payment	1	No payment
2	Gift exchange, 40c bonus	3	No payment, please try hard
3	No payment, please try hard	10	Low piece rate, 4-week delay
4	No payment, feedback after	13	Low piece rate (1c/100)
5	No payment, social comparison	11	50% prob. piece rate (2c/100)
6	Very low piece rate (1c/1000)	4	No payment, feedback after
7	1% prob. piece rate (1$/100)	2	Gift exchange, 40c bonus
8	Charity, low donation (1c/100)	14	Medium piece rate (4c/100)
9	Charity, high donation (10c/100)	8	Charity, low donation (1c/100)
10	Low piece rate, 4-week delay	6	Very low piece rate (1c/1000)
11	50% prob. piece rate (2c/100)	5	No payment, social comparison
12	Low piece rate, 2-week delay	9	Charity, high donation (10c/100)
13	Low piece rate (1c/100)	12	Low piece rate, 2-week delay
14	Medium piece rate (4c/100)	7	1% prob. piece rate (1$/100)
15	High piece rate (10c/100)	15	High piece rate (10c/100)

Rank-order correlation between 2015 results and hypothetical results 3 = 0.386



2015 Experimental Results		Hypothetical Experimental Results 4	
1	No payment	10	Low piece rate, 4-week delay
2	Gift exchange, 40c bonus	8	Charity, low donation (1c/100)
3	No payment, please try hard	5	No payment, social comparison
4	No payment, feedback after	9	Charity, high donation (10c/100)
5	No payment, social comparison	2	Gift exchange, 40c bonus
6	Very low piece rate (1c/1000)	14	Medium piece rate (4c/100)
7	1% prob. piece rate (1$/100)	1	No payment
8	Charity, low donation (1c/100)	11	50% prob. piece rate (2c/100)
9	Charity, high donation (10c/100)	3	No payment, please try hard
10	Low piece rate, 4-week delay	15	High piece rate (10c/100)
11	50% prob. piece rate (2c/100)	12	Low piece rate, 2-week delay
12	Low piece rate, 2-week delay	13	Low piece rate (1c/100)
13	Low piece rate (1c/100)	4	No payment, feedback after
14	Medium piece rate (4c/100)	7	1% prob. piece rate (1$/100)
15	High piece rate (10c/100)	6	Very low piece rate (1c/1000)

Rank-order correlation between 2015 results and hypothetical results 4 = 0.039

On the next pages, we will describe the various design changes and will ask 10 times for your predictions in terms of rank-order correlation.

Page 3

Pure Replication

Our previous experiment was run in May 2015 with a sample of about 550 MTurk subjects per treatment. In May 2018 we ran a pure replication of the button pressing task using the same instructions and survey material, with 150 subjects per treatment. The 2015 experiment and the 2018 experiment are as close as we could possibly make them, with only very minor changes, such as slight differences in the wording of the consent form. Click here to see the minor differences.

How do the results from our initial experiment compare to the result from the new experiment?

Please note that if these two groups respond to the treatments exactly the same, on average the rank-order correlation would be 0.94 (which is less than 1 due to sampling error).

Prediction 1. What do you think is the rank-order correlation for the 15 treatments between the 2015 experiment and the 2018 experiment?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Demographics

Pooling the data from both the 2015 and the 2018 button pressing task experiment, we are now interested in comparing the effects of the treatments for different demographic groups. For example, in the combined sample, 55% are women and 45% are men. For each of the 15 treatments, we compute the average effort separately for the male workers and for the female workers. We are then interested in the rank-order correlation between the 15 findings for male workers and the 15 findings for female workers. Is it the case that the treatments that lead to higher effort among the male workers also yield higher effort for the female workers, in which case the rank-order correlation would be high? Or are there significant differences in how the two groups of workers respond to the treatments, in which case the rank-order correlation would be lower?

We are going to similarly compare younger vs. older workers, higher-education vs. lower-education workers. For each of these comparisons, the sample is split nearly equally in the two groups. If the two groups – such as men and women, or younger vs. older – react to treatments exactly the same, on average we would expect a rank-order correlation of 0.95 (which is less than 1 due to sampling error).

Prediction 2. What rank-order correlation do you expect between the results of men and of women?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Prediction 3. What rank-order correlation do you expect between the results of MTurkers aged up to 30 versus MTurkers aged 31 and above? (The share aged up to 30 years is 50% in our sample.)

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Prediction 4. What rank-order correlation do you expect between the results of MTurkers without a college degree versus MTurkers with a college, or higher, degree? (The share without a college degree is 44% in our sample.)

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Prediction 5. In our sample, 84% of workers are American and 12% are Indian (based on the IP address). What rank-order correlation do you expect between the results of American MTurkers versus Indian MTurkers? If the two groups react to treatments exactly the same, on average we would expect a rank-order correlation of 0.89 (which is less than 1 due to sampling error, and is less than above due to the different size of the two groups).

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Effort Over Time

Considering once again the data from both the 2015 and the 2018 experiment, we are now interested in examining how the effects of the treatments may change over the 10 minutes of the experiment. Subjects, for example, may grow more tired in the later minutes, which could affect how different treatments perform. For each of the 15 treatments, we compute the average effort in the first 5 minutes and in the next 5 minutes.

Is it the case that the treatments that lead to higher effort in the first 5 minutes also yield higher effort in the next 5 minutes, in which case the rank-order correlation would be high? Or are there significant differences in how the treatments affect motivation in the first minutes versus later, in which case the rank-order correlation would be lower?

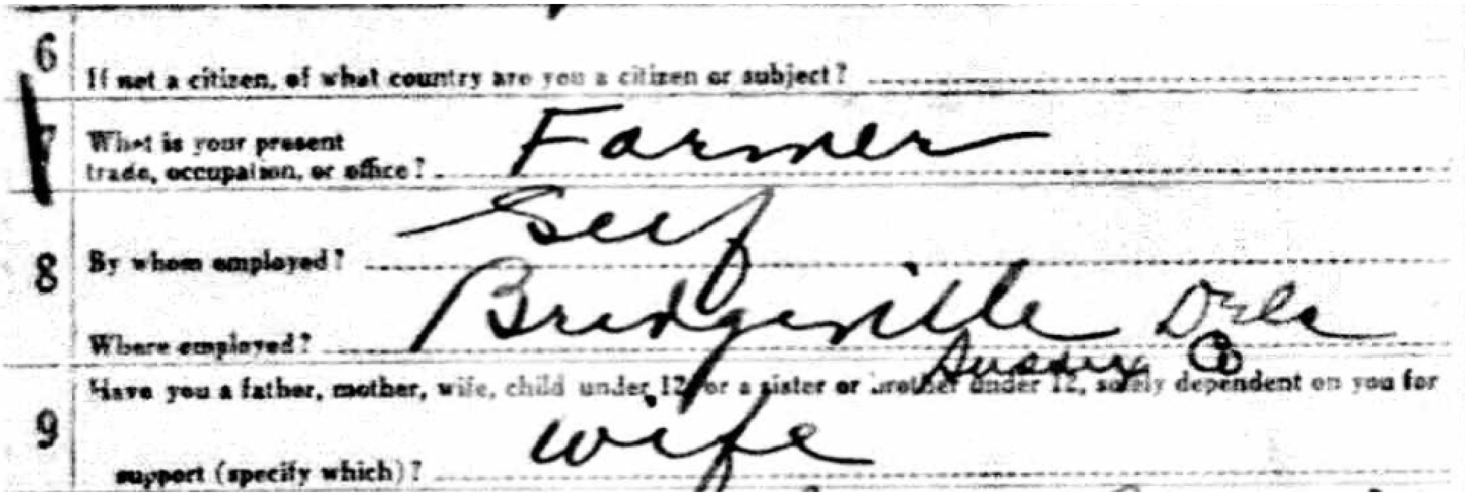
Prediction 6. What rank-order correlation do you expect between the results in the first 5 minutes versus in the next 5 minutes?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Page 4 V1

10-Minute Motivating Task

Still in May 2018, we ran on MTurk a second version (Version 2) of the same treatments, but with a different task that the participants were asked to perform. Rather than pressing the A and B buttons, we asked participants to code conscription records about soldiers in World War II. Their job was to code the occupation for each soldier that could be found in line 7 of pictures like the one below. (In this case, the occupation is "Farmer".)



6 If not a citizen, of what country are you a citizen or subject?

7 What is your present trade, occupation, or office? Farmer

8 By whom employed? Self

Where employed? Bridgeville Dela Sussex Co

9 Have you a father, mother, wife, child under 12, or a sister or brother under 12, solely dependent on you for support (specify which)? wife

Just like the button pressing task, participants were given 10 minutes to code as many cards as they wanted to. You can click here to try this task yourself.

We recruited an average of 170 subjects per treatment, running parallel treatments to the ones for the button pressing task. All payments were scaled to be approximately equivalent to the amount of money that could be made in the button pressing task. For example, the 1-cent-per-100-points treatment for the button pressing task is equivalent to 1 cent per 2 cards in this task. The recruitment of the subjects and the instructions were otherwise as parallel to the button pressing task as possible. Click here for the detailed instructions. We record the average number of cards coded within 10 minutes in a treatment as measure of effort in that treatment.

On average, subjects coded 57 cards in 10 minutes, with a standard deviation of 24.

We are interested in the rank-order correlation between the 15 findings for the button pressing task versus the 15 findings for the card coding. Is it the case that the treatments that lead to higher effort for the button pressing task also yield higher effort in the WWII card coding? Or is the ranking between treatments likely to be quite different?

Prediction 7. What do you think is the rank-order correlation between the treatments in the button pressing task and in the 2018 card coding experiment?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Coding Extra Cards

Still on May 2018, we ran a third version (Version 3) of the treatments that also asked MTurkers to code conscription records about soldiers in World War II (same task as above). In this version, we measured how many cards they were willing to code. We recruited an average of 150 subjects per treatment.

Specifically, participants were *required* to code 40 cards working on the task as part of their base payment. This took on average 9 minutes with an interquartile range of 5 minutes. There was no motivator or incentive for effort for the coding of these 40 cards, aside for the $1 show up fee.

After they finished coding the 40 cards, the subjects were invited to continue working for up to 20 additional cards: "*If you are willing, there are 20 additional cards to be coded. Doing this additional work is not required for your HIT to be approved or for you to receive the $1 promised payment.*" At this stage, subjects are assigned to one of 15 treatments providing monetary or non-monetary incentives to code extra cards, mirroring the ones used in the other versions of the experiment. For example, the 1-cent-per-100-points treatment for the button pressing task is equivalent to 1 cent per 2 additional cards in this task: "*As a bonus, you will be paid an extra 1 cent for every 2 additional cards you complete.*" Click here for the detailed instructions.

We record the effort in this case as the number of *additional* cards coded beyond the required 40, ranging from 0 to 20. (For example, the effort for workers that do not stay longer is coded as 0.)

On average, subjects coded 11 additional cards, with a standard deviation of 9.

Prediction 8. We are interested now in comparing the results from this version of the experiment to the previous versions. We compare it first to Version 2, the card-coding experiment where subjects had 10 minutes to code cards. Both designs involve coding of WWII cards, but in Version 2 we measure the number of cards coded within 10 minutes (the intensive margin), while in Version 3 we measure how many extra cards the workers are willing to do (the extensive margin). What do you think is the rank-order correlation between the two card-coding experiments?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Prediction 9. We are interested now in comparing the results from this version of the experiment (Version 3) to the button pressing task (Version 1). In this case, the experiments differ both because the tasks differ – button pressing versus WWII card coding – and because we measure output differently – button pressing within 10 minutes versus additional cards coded. What do you think is the rank-order correlation between the button pressing task and the coding of additional WWII cards?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Coding Extra Cards with No Consent Form

In all the experimental treatments considered so far, subjects saw a consent form before they participated in the typing task. Click here to see the consent form. In our final version of the experiment, instead, subjects did not see a consent form and moved directly to the task of coding 40 required WWII cards, and coding of additional cards up to 20. (Notice that there is no deception since the subjects are coding historical material for an economic history project.)

This version of the experiment was identical to the version you just considered, other than for the lack of a consent form. Each of the two versions includes on average 150 subjects in each of the treatments.

In each of the versions, we measure output within a treatment as the number of additional cards coded.

Prediction 10. What do you think is the rank-order correlation between the two versions of the same card-coding experiments?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Comment and Feedback

This is the last page of the survey.

You made 10 predictions about rank-order correlations. What is your expectation of how many of your 10 predictions will fall within 0.1 points of the actual correlation? (That is, if the correlation is 0.7 and you guessed 0.75, it will count as correct, but if you guessed 0.85 it will not.)

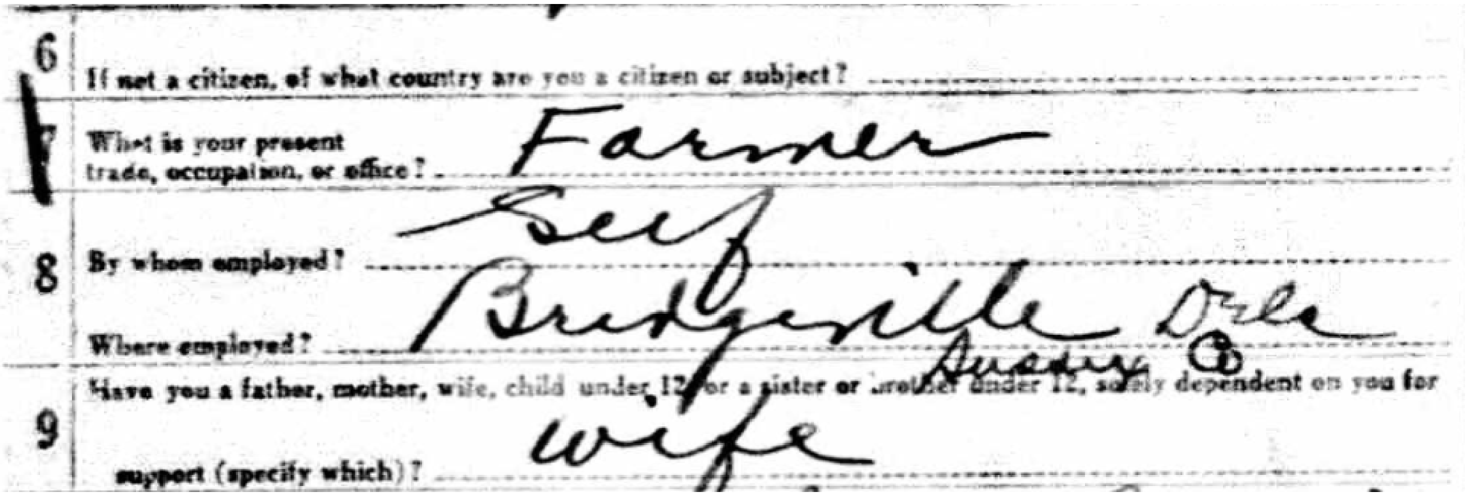
0 1 2 3 4 5 6 7 8 9 10

In case you encountered any issues with the survey or would like to leave a comment, feel free to do so in the text box below.

Page 4 V2

10-Minute Motivating Task

Still in May 2018, we ran on MTurk a second version (Version 2) of the same treatments, but with a different task that the participants were asked to perform. Rather than pressing the A and B buttons, we asked participants to code conscription records about soldiers in World War II. Their job was to code the occupation for each soldier that could be found in line 7 of pictures like the one below. (In this case, the occupation is "Farmer".)



6 If not a citizen, of what country are you a citizen or subject?

7 What is your present trade, occupation, or office? Farmer

8 By whom employed? Self

Where employed? Bridgeville Dela Sussex Co

9 Have you a father, mother, wife, child under 12, or a sister or brother under 12, solely dependent on you for support (specify which)? wife

Just like the button pressing task, participants were given 10 minutes to code as many cards as they wanted to. You can click here to try this task yourself.

We recruited an average of 170 subjects per treatment, running parallel treatments to the ones for the button pressing task. All payments were scaled to be approximately equivalent to the amount of money that could be made in the button pressing task. For example, the 1-cent-per-100-points treatment for the button pressing task is equivalent to 1 cent per 2 cards in this task. The recruitment of the subjects and the instructions were otherwise as parallel to the button pressing task as possible. Click here for the detailed instructions. We record the average number of cards coded within 10 minutes in a treatment as measure of effort in that treatment.

On average, subjects coded 57 cards in 10 minutes, with a standard deviation of 24. To give you an idea of the effect of incentives, on average subjects coded 53.8 cards (standard error 1.8) in 10 minutes in the control treatment (no monetary incentives), they coded 59.4 cards (standard error 1.8) in the low-piece rate treatment (1 cent per 2 cards) and they coded 56.3 cards (standard error 2) in the high-piece rate treatment (5 cents per card).

We are interested in the rank-order correlation between the 15 findings for the button pressing task versus the 15 findings for the card coding. Is it the case that the treatments that lead to higher effort for the button pressing task also yield higher effort in the WWII card coding? Or is the ranking between treatments likely to be quite different?

Prediction 7. What do you think is the rank-order correlation between the treatments in the button pressing task and in the 2018 card coding experiment?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Coding Extra Cards

Still on May 2018, we ran a third version (Version 3) of the treatments that also asked MTurkers to code conscription records about soldiers in World War II (same task as above). In this version, we measured how many cards they were willing to code. We recruited an average of 150 subjects per treatment.

Specifically, participants were *required* to code 40 cards working on the task as part of their base payment. This took on average 9 minutes with an interquartile range of 5 minutes. There was no motivator or incentive for effort for the coding of these 40 cards, aside for the $1 show up fee.

After they finished coding the 40 cards, the subjects were invited to continue working for up to 20 additional cards: "*If you are willing, there are 20 additional cards to be coded. Doing this additional work is not required for your HIT to be approved or for you to receive the $1 promised payment.*" At this stage, subjects are assigned to one of 15 treatments providing monetary or non-monetary incentives to code extra cards, mirroring the ones used in the other versions of the experiment. For example, the 1-cent-per-100-points treatment for the button pressing task is equivalent to 1 cent per 2 additional cards in this task: "*As a bonus, you will be paid an extra 1 cent for every 2 additional cards you complete.*" Click here for the detailed instructions.

We record the effort in this case as the number of *additional* cards coded beyond the required 40, ranging from 0 to 20. (For example, the effort for workers that do not stay longer is coded as 0.)

On average, subjects coded 11 additional cards, with a standard deviation of 9. To give you an idea of the effect of incentives, on average subjects coded 8.6 additional cards (standard error 0.7) in the control treatment (no monetary incentives), they coded 12.6 additional cards (standard error 0.8) in the low-piece rate treatment (1 cent per 2 additional cards) and they coded 17.4 additional cards (standard error 0.5) in the high-piece rate treatment (5 cents per additional card).

Prediction 8. We are interested now in comparing the results from this version of the experiment to the previous versions. We compare it first to Version 2, the card-coding experiment where subjects had 10 minutes to code cards. Both designs involve coding of WWII cards, but in Version 2 we measure the number of cards coded within 10 minutes (the intensive margin), while in Version 3 we measure how many extra cards the workers are willing to do (the extensive margin). What do you think is the rank-order correlation between the two card-coding experiments?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Prediction 9. We are interested now in comparing the results from this version of the experiment (Version 3) to the button pressing task (Version 1). In this case, the experiments differ both because the tasks differ – button pressing versus WWII card coding – and because we measure output differently – button pressing within 10 minutes versus additional cards coded. What do you think is the rank-order correlation between the button pressing task and the coding of additional WWII cards?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Coding Extra Cards with No Consent Form

In all the experimental treatments considered so far, subjects saw a consent form before they participated in the typing task. Click here to see the consent form. In our final version of the experiment, instead, subjects did not see a consent form and moved directly to the task of coding 40 required WWII cards, and coding of additional cards up to 20. (Notice that there is no deception since the subjects are coding historical material for an economic history project.)

This version of the experiment was identical to the version you just considered, other than for the lack of a consent form. Each of the two versions includes on average 150 subjects in each of the treatments. In each of the versions, we measure output within a treatment as the number of additional cards coded.

Prediction 10. What do you think is the rank-order correlation between the two versions of the same card-coding experiments?

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1