Stability of Experimental Results: Forecasts and Evidence*

Stefano DellaVigna
UC Berkeley and NBER

Devin Pope
University of Chicago and NBER

September 2018

Abstract

How robust are experimental results to changes in design? Researchers designing an experiment expect some design choices to matter more than other ones. The stability of results is also critical to conceptual, as opposed to exact, replication. We consider a specific context, a real-effort task with multiple behavioral treatments, and examine the stability along six dimensions: (i) pure replication; (ii) demographics; (iii) geography and culture; (iv) the task; (v) the output measure; (vi) the presence of a consent form. We use rank-order correlation across the treatments as measure of stability, and compare the observed correlation to the one under a benchmark of perfect stability. We also collect expert forecasts of the correlation along each dimension. The academic experts expect that the pure replication will be close to perfect, that the results will differ sizably across demographic groups (age/gender/education), and that the task and output will make a further impact. We find near perfect replication of the experimental results, and significantly higher stability across demographics than the experts expect. Specifically, the demographic groups differ in average effort and sensitivity to incentives, but have a stable response to the behavioral treatments. The results are quite different across tasks, mostly because the task change introduces added noise in the findings. The experts are insensitive to this source of instability, even when it is made clear. We discuss the implications for measures of conceptual replication.

*Preliminary and incomplete, please do not cite without permission. We thank Ned Augenblick, Jon de Quidt, Alex Rees-Jones, Dmitry Taubinsky, as well as audiences at Rice University, at the University of Bonn, at UC Berkeley, and at the 2018 SITE Conference for Psychology and Economics for comments and suggestions. We thank Kristy Kim, Maxim Massenkoff, Jihong Song, and Ao Wang for outstanding research assistance. Our survey was approved by University of Chicago IRB, protocol IRB18-0144 and pre-registered as trial AEARCTR-0002987.

1 Introduction

A researcher has designed an experiment to test a model of reciprocity. The key elements of the design are set, and yet the researcher wonders: Will it matter if I run the experiment in a laboratory, or on MTurk? How important is the choice of the specific task? Should I worry about a change in consent form that the IRB required? After running the experiment, the researcher is not confident that the results would be similar if the experiment was run with different design choices.

Another researcher is evaluating a field experiment as a journal referee. While the results in the paper are statistically significant and internally valid, the researcher worries about external validity. He is concerned about demand effects, given that the subjects knew they were part of an experiment, and also about the specificity of the geographic setting in rural Brazil. These concerns about external validity lead him to recommend rejection for the paper under review.

A third researcher reads about the exact replications of economic experiments (Camerer et al., 2016) and wonders: If we move beyond exact replication to conceptual replication, how do we even measure replication, if for example the units of measure in the replication differ from the units in the original experiments?

These three researchers are concerned about how experimental results vary as the design changes. This is a key concern in the literature: a number of papers examine the stability of experimental results with respect to specific design choices. Classical examples include the difference between the strategy method versus direct choice (Brandts and Charness, 2011), the debate on within-subject versus between-subject designs (Greenwald, 1976), and the impact of anonymity (Hoffman, McCabe, and Smith, 1996). Among the more recent examples, de Quidt, Haushofer, and Roth (forthcoming) consider the impact of demand effects and Allcott (2015) studies the heterogeneity by demographic group of the effect of the mailing of the OPower electricity report letter.

Most of these papers consider in depth the impact of *one* particular design aspect, such as the degree of anonymity, demand effects, or the demographic groups. Surprisingly, there has been little work instead comparing the robustness of one experimental result to a *battery* of design changes. And yet, this is a question that often preoccupies researchers at the design or review stage: within a set of plausible design changes, which ones would affect the results substantially, and which ones not? This assessment requires a comparison across different designs, holding constant one setting.

In this paper, we consider a specific setting, a real-effort task with multiple behavioral treatments, and we examine the stability of the results across several design variants. We use this specific case to illustrate a roadmap for how to think about conceptual replication more broadly. Since some of the design changes produce results with different units of measurement, we propose rank-order correlation as a suitable way to compare treatment effects. Further, since we are interested not only in how the results change, but also in how researchers *expect* the results to change, we collect forecasts from academic experts about the stability of the experimental results for each design change.

Which design changes are of interest? We single out six of them, although clearly other ones also play a role: (i) (*pure replication*) the results may change even if we re-run the experiment as similarly as possible to the original; (ii) (*demographics*) the results may change with a sample with a different share of women or, say, college-educated respondents; (iii) (*geography and culture*) the

results may be specific to a geographic or cultural setting; (iv) *(task chosen)* the result may be specific to a task; (v) *(output measure)* the results may be different with a different measure; (vi) *(consent form)* it may matter that subjects know that it is an experiment.

The experimental task that we take as starting place is a typing task employed in DellaVigna and Pope (2018, forthcoming): subjects on MTurk were given 10 minutes to alternatively press the 'a' and 'b' buttons on their keyboards as quickly as possible. While the task is not meaningful per se, it lends itself to study motivation since the typing exercise becomes tiresome. In DellaVigna and Pope (2018, forthcoming), we recruited nearly 10,000 MTurk subjects and compared effort under 18 treatments which included, among others, 4 piece rate incentives, 3 social preference treatments, 2 time preferences treatments, 2 probability weighting treatments, 3 purely psychological manipulations, and a paying-too-little treatment. The experiment was designed to be a microcosm of behavioral economics, comparing the effectiveness of different motivators in inducing effort.

We build on this experiment by considering several design variants, covering the six dimensions singled out above. In each design variant we include 15 of the original treatments, following a pre-registered design. Overall, we collect data on nearly 10,000 new MTurk subjects. First, we re-run, 3 years later, the same experiment, to examine the extent of *pure replication*. Second, taking advantage of the substantial *demographic* heterogeneity in the MTurk sample, we compare the experimental results along three key demographic cleavages: gender, education, and age. Third, we consider the *geographic and cultural* component comparing the results for US subjects versus subjects from India, as well as results in "red states" versus "blue states".

While all the above comparisons take place for the same typing task, for our fourth comparison we use a more motivating *task*—coding World-War II conscription cards— and measure the number of cards coded within 10 minutes. Fifth, we consider alternative measures of *output*. Inspired by Abeler et al. (2011), we repeat the WWII card coding, but we now measure not how many cards subjects code in a fixed amount of time, but how many extra cards they code beyond a minimum required amount.[1] Finally, for our sixth dimension, we run a version of the WWII card coding in which, unlike in all previous versions, subjects are not given a consent form and are thus plausibly unaware that they are part of the experiment.[2]

Moving from one design to the next, we are interested in the stability of the findings on effort for the 15 treatments. But what is the right metric of stability? For example, consider the task change: in the a-b typing task, the average output within 10 minutes is 1,800 points, but in the WWII coding task, the average output within 10 minutes is about 58 cards. One could make the two designs comparable by rescaling the effect sizes by 1,800/58. But this rescaling does not take into account differences in the elasticity of effort to motivation: a 30 percent increase in effort in the a-b task, which we observe in response to piece rate variation, may not be achievable in the WWII card coding case. Importantly, like in most real-effort experiments, we do want to control for the responsiveness to motivation, since the focus is on comparing across different motivating treatments.

With these considerations in mind, we use the rank-order correlation of the average effort in the

[1] As another change in the output measure, returning to the a-b typing task, we compare the performance in the first 5 minutes of the task versus the later 5 minutes.

[2] Notice that, since subjects are coding historical data, this should be a natural framing. As elsewhere in our experiment, there is no deception.

15 treatments as our benchmark measure of stability of experimental results. To illustrate, consider a case in which treatments ranked by effort, respectively, 3, 8, and 14 out of 15 in context A are ranked 4, 8, and 15 in context B, and similarly the other treatments keep the same rank; in this case, the rank-order correlation will be high. If instead those treatments move to positions 7, 4, and 10, the rank-order correlation will be low. While this measure is not without draw-backs, it performs well also in cases in which the underlying model predicts a non-linear transformation, as in the output change. Importantly, wherever possible, we compare the observed rank-order correlation to the average rank-order correlation under a benchmark of perfect stability, in which the only variation in rank is due to idiosyncratic noise in the realization of effort in the treatments.

Having identified the design changes and the measure of stability, following DellaVigna and Pope (forthcoming) we collect forecasts. We contact 70 behavioral experts or experts on replication, yielding 55 responses. Each expert sees a description of the task, of the design changes, and an illustration of how rank-order correlation works; whenever possible, we also provide information on the rank-order correlation under full stability. The experts then forecast the rank-order correlation for 10 design changes. We also collect forecasts from PhD students and MTurk respondents.

The experts expect that: (i) the pure replication will not be perfect, but will be fairly close to exact replication (0.82 correlation, compared to 0.94 under full stability); (ii) the results will differ sizably for different demographic groups (age/gender/education) (0.75 correlation, compared to 0.95 under full stability), (iii) the results will also be different for the India and US sample (0.65 correlation, compared to 0.89 under full stability); (iv) the change in task will have a similar, sizable impact (0.65 correlation); (v) similarly for the change in output (0.5 to 0.6 correlation); (vi) the disclosure of experimental consent will have a modest impact (0.78 correlation, compared to 0.88 under full stability). There is very little heterogeneity in the forecasts, whether comparing experts, PhDs, and MTurks, or splitting by confidence in the forecasts, or by measures of effort (e.g., time spent) in making forecasts.

We then compare the forecasts to the experimental results. We find (i) near perfect replication of the a-b task (correlation of 0.91), within the confidence interval of the full-stability benchmark. We find (ii) strikingly high stability across demographics—correlation of 0.96 for gender, 0.97 for education, and 0.98 for age—, significantly higher than the experts expected (0.75 on average). Interestingly, the demographic groups *do* differ in the average effort and even in the sensitivity to incentives. Once we control for that, though, as the rank-order correlation measure does, there is no difference across the demographic groups in the response to the behavioral treatments, and in how the behavioral treatments compare to the incentive treatments. We find a lower correlation for our geographic comparison (iii) between US subjects and Indian subjects (0.65), just as the experts predicted, though we cannot reject that this lower correlation could be due to noise (given that Indian workers are just 12 percent of the data). In another geographic comparison, we find near-perfect correlation (0.94) in the results for workers from "blue states" as opposed to "red states".

We then compare across tasks (iv): the rank-order correlation between the 10-minute productivity in a-b typing versus in WWII card coding is 0.64, close to the expert forecast of 0.66. We also compare across output, (v), by comparing two designs with the same task—coding WWII cards—but different output measures: the number of cards coded within 10 minutes, versus the number of extra cards

that the workers are willing to code after completing the required cards. The rank-order correlation in this output dimension is just 0.27, compared to the expert prediction of 0.61. Changing the task and output measures, thus, has a quite large impact on the results, more than the experts expected.

This (relative) instability has two possible explanations. First, changes in task and output may have affected the impact of behavioral and financial motivators. Second and more simply, the 10-minute WWII task, unlike the a-b task, may be quite insensitive to motivation; if this is the case, we would expect a lower correlation, as the noise in the realized effort by treatment would swamp the motivational effects. Indeed, in the 10-minute WWII task, output is barely responsive to incentives, with an elasticity of effort of less than 0.01, compared to 0.04 for the a-b task, likely because this task is highly motivating to start with, limiting the impact of additional motivators.

We confirm this interpretation with a combined output/task comparison of the a-b 10-minute task to the WWII coding with extra cards. This latter task is responsive to incentives with an effort elasticity of 0.4, among the highest for any real-effort task in the literature. If changes in task and output matter, we would expect a low correlation between the two tasks, as both task and output measure change. If the lack of stability is mostly tied to noise, we would expect a relatively high correlation, as effort is clearly responsive to incentives in both tasks.

The correlation for the joint task/output change, 0.65, is higher than for just the output change, 0.27, consistent with the role of noise. Interestingly, the experts appear to miss the role for noise, since they predict a higher correlation for just the output change, 0.63, than for the joint task/output change, 0.54. Of course, it is plausible that the degree of noise in the different tasks was not obvious to the forecasters. To address this issue, we randomly provided half of forecasters with information on the mean effort (and s.e.) under three piece rate treatments, indicating a flat and non-monotonic response to incentives in the 10-minute WWII task, and in contrast a precisely-estimated responsiveness in the extra-work WWII task. This additional information has little impact on the expert forecasts, indicating a deeper neglect for the role of noise.

Lastly, we analyze dimension (vi) using the same extra-card WWII coding task, but without a consent form at the beginning of the experiment. Thus, participants are arguably unaware that they are taking part in an experiment, but rather think they are doing a coding job, which is not uncommon on MTurk. The rank-order correlation across treatments for this dimension is 0.84, which is not significantly different from the expert prediction (0.78) or the full-stability measure (0.88).

Taking this altogether, we draw five main lessons. First, we find a remarkable degree of stability of experimental results across design changes. Nine out of ten planned comparisons have a rank-order correlation above 0.60, and six comparisons have a rank-order correlation above 0.80. This conclusion is not affected by the particular choice of metric to compute the stability, and is not contaminated by selective reporting, as we report all the comparisons as pre-specified.

Second, we find mixed evidence on the ability of experts to predict which design changes will affect the results the most. The experts are qualitatively accurate that the results would be stable to pure replication and to the omission of a consent form, and would be less stable in response to a task change. However, they are incorrect in expecting an important role for demographic composition and in failing to anticipate the role of noise. These results confirm the anecdotal impression that design choices are a difficult and somewhat unpredictable part of the experimenter toolbox.

Third, we find remarkable stability of the results with respect to the demographic composition of the sample, or even geographic and cultural differences. In contrast, nearly all the experts expected a larger role for the demographic composition. Selective publication may explain some of this discrepancy: while null results on demographic differences typically do not get published, differences that are statistically significant draw attention and may thus be salient in the mind of experts.

Fourth, the degree of noise in the experimental results is a first-order determinant of stability of the results: the only two instances of low replication are due to a task with very inelastic output. The experts do not appear to anticipate this important role for noise, even when provided with diagnostic information. The neglect for noise may again have to do with publication bias, as experimental designs with noisy results are typically not published. And yet, predicting which designs will yield noisy results is an important component of design choice.

A final lesson relates to conceptual replication more broadly. We demonstrate how rank-order correlation can serve as a useful metric when analyzing the stability of results across design changes. Further, we illustrate the importance of thinking about the elasticity of the outcome with respect to the treatment (in our case, effort with respect to motivation). Design changes may lead to "non-replicable results" not because the treatments are ineffective at changing intentions, but because intentions no longer translate to outcomes in such a clean manner. Underlying structural models can provide insight when thinking about conceptual replication with major design changes.

Related to our paper is the work on direct/exact/pure replication, including the recent open-science work on large-scale replication of experiments (Open Science Collaboration, 2015; Camerer et al., 2016, 2018). To our knowledge, there has not been a similar, systematic effort to test for the conceptual replication of a large group of studies. As we discussed above, many papers have studied the stability of results to specific changes in experimental design.

2 Design and Measure of Stability

2.1 Experimental Design

2.1.1 2015 Experiment and Model

The starting point for the design in this paper is the real-effort task in DellaVigna and Pope (2018, forthcoming) which we ran in May 2015 on the Amazon Mechanical Turk (MTurk) platform. MTurk is an online platform that allows researchers and businesses to post small tasks (referred to as HITs) that require a human to perform. Potential workers browse the postings and choose whether to complete a task for the amount offered. MTurk has become a popular platform to run experiments in marketing and psychology (Paolacci, 2010) and is also used increasingly in economics (e.g., Kuziemko, Norton, Saez, and Stantcheva, 2015). The evidence suggests that the findings of studies run on MTurk are similar to the results in more standard laboratory or field settings (Horton, Rand, and Zeckhauser (2011)).

The task involves alternating presses of 'a' and 'b' on a computer keyboard for 10 minutes, achieving a point for each a-b alternation (see Online Appendix Figure 1b). While the task is not meaningful per se, it does have features that parallel clerical jobs: it involves repetition and it gets

tiring, thus testing the motivation of the workers. It is also simple to explain to both subjects and experts.

In May 2015, we recruited subjects on MTurk for a $1 pay for an "*academic study regarding performance in a simple task.*" Subjects interested in participating sign a consent form, enter their MTurk ID, answer three demographic questions, and then saw the instructions, reproduced in Online Appendix Figure 1b, indicating that "*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. [...] Feel free to score as many points as you can.*" The participants then saw a different final paragraph (bold and underlined) depending on which one of 18 treatment conditions they were assigned to. For example, in the high-piece rate treatment, the sentence read *"As a bonus, you will be paid an extra 10 cents for every 100 points that you score. This bonus will be paid to your account within 24 hours.*" To give another example, in the high-return charity condition, the return is the same, but it accrues to the Red Cross: *"As a bonus, the Red Cross charitable fund will be given 10 cents for every 100 points that you score."*

The subjects could try the task before moving on to the real task. As subjects pressed digits, the page showed a clock with a 10-minute countdown, the current points, and any earnings accumulated. The final sentence on the page summarizes the condition for earning a bonus (if any) in that particular treatment. At the end of the 10 minutes, the subjects are presented with the total points and the payout, are thanked for their participation and given a validation code to redeem the earnings.

The experiment ran for three weeks in May 2015. After applying the sample restrictions detailed in DellaVigna and Pope (2018), the final sample included 9,861 subjects, about 550 per treatment.

The 18 treatments were selected to compare the impact of traditional piece-rate incentives and of behavioral and psychological motivators. Table 1 lists 15 of the 18 treatments run in this initial sample, plus a 16th additional treatment. The treatments differ in only three ways: the main paragraph in the instructions explaining the condition, summarized in Column 2 of Table 1, the one-line reminder on the task screen, and the rate at which earnings (if any) accumulate on the task screen.

The first four treatments in Table 1 are piece-rate treatments, with the piece rate varying from no-piece-rate to low-piece-rate (1 cent per 100 points) to mid-piece-rate (4 cents per 100 points) to high-piece-rate (10 cents per 100 points). These treatments capture the response to financial motivations and thus allow us to back out the baseline motivation and the cost of effort curvature.

Model. Assume that participants in the experiment maximize the return from effort e net of the cost of effort, where e denotes the number of points (that is, alternating a-b presses). For each point e, the individual receives a piece-rate p as well as a non-monetary reward, $s > 0$. The parameter s captures, in reduced form, intrinsic motivation, personal competitiveness, or sense of duty to put in effort for an employer. This motivation is important because otherwise, for $s = 0$, effort would equal zero in the no-piece rate treatment, counterfactually. Assume also a convex cost of effort function $c(e)$: $c'(e) > 0$ and $c''(e) > 0$ for all $e > 0$. Assuming risk-neutrality, an individual solves

$$max_{e \geq 0}(s + p)e - c(e), \tag{1}$$

leading to the solution (when interior) $e^* = c'^{-1}(s+p)$. Optimal effort e^* is increasing in the piece rate p and in the motivation s. A useful special case for the cost function, discussed further in DellaVigna et al. (2015) is the power cost function $c(e) = ke^{1+\gamma}/(1+\gamma)$, characterized by a constant elasticity of effort $1/\gamma$ with respect to the value of effort. Under this assumption, we obtain

$$e^* = \left(\frac{s+p}{k}\right)^{1/\gamma}. \tag{2}$$

A plausible alternative is that the elasticity decreases as effort increases. A function with this feature is the exponential cost function, $C(e) = kexp(\gamma e)/\gamma$, which has elasticity $1/(\gamma e)$. This cost function leads to solution

$$e^* = \frac{1}{\gamma} ln\left(\frac{s+p}{k}\right). \tag{3}$$

Under either function, the solution for effort has three unknowns, s, k, and γ which we can back out from the observed effort at different piece rates. Three piece rates are in principle enough, but we incorporate four piece rates to build in over-identification. We present the estimation details in Section 4.3.

Returning to the list of treatments in Table 1, the next treatments are motivated by behavioral research. In the paying-too-little treatment, we set a very low piece rate, 1 cent for very 1,000 points, to test whether this crowds out internal motivation. The next three treatments focus on social preferences. In the first two, subjects earn a return for a charity by working (as in Imas, 2014), with either a low return to the charity (1 cent per 100 points) or a high return (10 cents per 100 points). The third social-preference treatment, on gift exchange (as in Gneezy and List (2006)), provides an unconditional payment of 40 cents to the subjects to test whether the subjects responds with higher effort, as in the gift exchange hypothesis.

Next, we consider two time-discounting treatments motivated by the research on present bias (Laibson, 1997; O'Donoghue and Rabin, 1999). In both cases the piece rate is 1 cent per 100 points, but in one case the bonus will be deposited "*two weeks from today*", while in a second case the bonus will be deposited "*four weeks from today*". These two treatments allow, under some assumptions, to back out β and δ.

The next two treatments consider probability weighting and risk aversion. In the first treatment, subjects have "*a 1% chance of being paid an extra $1 for every 100 points*". The expected value of this piece rate is the same as in the low-piece-rate treatment, but the piece rate is now stochastic. Under the typical parametrizations of the probability weighting function in prospect theory, we would expect a higher effort under this treatment, provided subjects are not too risk averse, given that the probability weighting function magnifies small probabilities. The next treatment aims to capture in a simple way risk aversion and offers "*a 50% chance of being paid an extra 2 cents for every 100 points*". Once again, the expected value is the same as the 1 cent piece rate, but in this case we do not expect the probability weighting to play a role.

The next three treatments do not involve any incentive and are more directly borrowed from psychology, with wording aimed to boost effort, either by introducing social comparisons ("*many*

participants were able to score more than 2,000 points"), ranking of the subjects ("*we will show you how well you did relative to other participants*"), or a task significance manipulation ("*your work is very valuable for us*").

We can generalize the model in (1) to incorporate the impact of behavioral motivators, as we discuss in more detail in DellaVigna and Pope (2018). For example, we can model the psychological treatments, and the gift exchange treatment, as increasing the baseline motivation by a term Δs, such that the individual maximizes $(s + \Delta s + p)e - c(e)$.

The 2015 experiment also included three treatments focused on gain and loss framing, which we decided not to replicate in 2018, leaving 15 treatments.[3] Column 3 of Table 1 and Online Appendix Figure 2 summarize the average effort in each of these 15 treatments that resulted from the experiment that we ran in 2015.

2.1.2 2018 Experiment

In May of 2018 we ran a new round of experiments in MTurk following a pre-analysis plan, with design variants aimed at testing the stability of the earlier experimental results. Other than the emphasized design changes, we aimed to keep the experimental material as close as possible to the earlier experiment, to be able to attribute any differences just to the design changes.

We ran the experiment for 3 weeks, advertising the task as an "*11 to 12-minute typing task*" paying $1, the same pay as in the 2015 experiment (see the screenshot in Online Appendix Figure 1a). Subjects that clicked on the ad on MTurk were randomized to one of four different versions of the new experiment and, within each version they were randomized into 1 of 16 treatments. In addition to the 15 treatments from the earlier experiment, the new experiment also include an additional 16th treatment, listed at the bottom of the table, combining a piece rate and a psychological manipulation. We do not use this treatment for the main comparisons given that we did not run it in 2015; we return to this treatment later in an out-of-sample comparison of the model.

The assignment of subjects into versions and treatments is as follows. The subjects are assigned into one of the four versions randomly, with versions 2, 3, and 4 oversampled by 15 percent. This is in anticipation of the fact that the historical task used in version 2-4 will likely have a higher share of subjects not complete the task due to, for example, difficulty in reading cursive writing (employed in these cards). In pilot data, we observed higher attrition in these versions by about 15 percent. The overweighting is designed to equate as much as possible the post-attrition sample size across the four versions. Within a version, we randomize participants into one of the 16 treatments with equal weights.[4] We now describe in detail the four versions of the 2018 experiment.

Exact Replication. The first version, summarized in Column 4 of Table 1, is an exact replication of the 2015 experiment, with the same 10-minute a-b typing task and the same wording for the 15 treatments as detailed above.[5]

[3] These three treatments turned out to be under-powered to identify the reference dependence parameters, making a replication less meaningful. In addition, these were the only treatments based on a threshold payoff (e.g., 40c for reaching 2,000 points), and a model-based prediction of the effort for these treatments requires information on the full distribution of effort, unlike for the other treatments. This made it particularly tricky to compare across contexts.

[4] Online Appendix Table 1 reports the number of observations in each cell.

[5] There are four small difference: (i) the advertising screen in 2015 mentioned a 15-minute "*academic study regarding performance in a simple task*"; we changed this in 2018 and mention an 11-12 minutes "*typing task*", in order for

10-Minute WWII Coding. The second version, summarized in Column 5 of Table 1, is also a 10-minute task, but subjects, instead of doing a-b button presses, are instead assigned to coding the occupation in World War II enrollment cards[6] We introduce the task as follows: "*In this task you will be coding up conscription records about soldiers in World War II. You will have 10 minutes to complete as many cards as you can. Your job is to identify the occupation in field 7 of each record and to type it into the text box below each card. If you are unable to determine what the occupation is, or if field 7 is missing from the card, please type "unclear".*" We then show the subjects an example of a card and then state "*Please be as careful as possible (we will check the accuracy of your work).*". For each card, the subjects thus have to type the occupation as they read it, and click to load the next card (see Online Appendix Figure 1c). We randomly draw cards out of a sample of over 3,353 cards.

This second version of the experiment has the same 16 treatments as the first version with piece rate variation and behavioral and psychological treatments. We aim to make the fewest changes in the wording possible, so as to keep the design parallel, other than the change in task. Column 2 of Table 1 displays in bracket the wording used for this second version. The most important change is the change in units for the treatments with incentives. Based on pilot data, we determined that on average subjects coded 50-60 cards in 10 minutes, compared to 1,500-2,000 a-b presses. Based on this ratio of productivity, and in order to set incentives at round numbers, we multiply the a-b payoffs by a factor of 50. So for example, the low-piece-rate treatment yields a bonus of "*an extra 1 cent for every 2 cards that you complete*" and the high-piece-rate treatment yields a bonus of "*an extra 5 cents for every card that you complete*". This yielded an average pay that is somewhat higher than, but comparable to, the pay in the a-b task. We apply a similar conversion to the other payoffs, keeping the unconditional gift exchange payment to 40 cents.

Extra-Work WWII Coding. For the third version, summarized in Column 6 of Table 1, the subjects still code World War II cards, but with a different design and output measure. In versions 1 and 2, the task of the subjects is to produce as many units as possible within a given time limit. This is the typical structure of real-effort experiments, including Gneezy, Niederle, and Rustichini (2003), Gneezy and List (2006), and Gill and Prowse (2012). Yet, an alternative margin of effort is not how hard one works in a given unit of time, but how much work one is willing to do. Abeler et al. (2011) pioneered this design for the study of forward-looking reference points. In their design, subjects have 4 minutes to count as many tables as possible. Afterwards, subjects are asked if they would do more of the work and stay longer, for up to 60 minutes, and how long subjects stay is the key outcome of interest.

In our third version, we similarly adopt the margin of how long workers decide to work, after a minimum required. Namely, after the initial sign-up screen, the subjects randomized into the third

this to be consistent across the different 2018 experiments; (ii) the consent form was longer, as required by the IRB; (iii) in 2015 we asked the demographic questions at the beginning of the survey, while in 2018 we asked them at the end of the survey; (iv) the formatting of the final pay-out page changed from "*Points: XXX, Bonus Payout: $XXX, Total Payout: $XXX, Any bonus payment must be approved before it is given*" in 2015 to "*Thank you for your participation. You will be paid $1.00 for this HIT.*" and, if the participant received a bonus, also "*You will also be paid a bonus of XXX for every XXX points that you scored. Since you scored XXX points, your total bonus will be XXX.*"

[6]We coded this cards as part of an ongoing historical project by Bruno Caprettini and Joachim Voth, who provided us with the cards to be coded.

version are asked to code the occupation field for 40 WWII cards (Online Appendix Figure 1d). The task is as in the previous version, and there is no manipulation at this stage: all workers are asked to do this for no extra payoff.

After they are done with the 40 cards, all subjects see "*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. Please feel free to complete any number of additional cards, up to 20.*" At this point, the randomization into the 16 treatments kicks in. Subjects in the control group read "*The number of additional cards you complete will not affect your payment in any way*," while subjects in the low piece rate, for example, are informed "*as a bonus, you will be paid an extra 1 cent for every 2 additional cards you complete. This bonus will be paid to your account within 24 hours.*" Column 2 in Table 1 shows the key wording for the treatments in double brackets. We keep the same exact incentives as in the second version; this keeps the marginal incentives the same, though it implies that the average total payment will tend to be lower in this version, compared to the 10-minute WWII card coding version, given that subjects can at most code 20 extra cards. To partially compensate for this, we set the required number of cards to code in this version, 40 cards, such that most subjects would finish earlier than in 10 minutes.[7]

No-Consent WWII Coding. The fourth and final version, featured in Column 7 of Table 1, is identical to the third version, except that the subjects do not see a consent form. In all other versions, the workers see a consent form right after clicking on the MTurk HIT. In this version, instead, they are taken directly to the description of the task. Given that the task is quite similar to the coding of historical documents that are common on platforms like MTurk, the absence of a consent form should not come as a surprise. We consider this condition given the debate on whether it matters if subjects know that it is an experiment. For example, in the Harrison and List (2004) classification, a natural field experiment has as requirement that subjects are unaware that it is an experiment. Surprisingly, there is little evidence on whether this matters for the results of experiments other than in List (2006), where knowing that one is part of an experiment makes a difference.

Sample. In the pre-analysis plan, we set out 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

[7] In this version, we removed the demographic questions, since it was awkward to ask them after already asking for extra work; in addition, we did not want demographic questions in the next version, and wanted to keep the two versions parallel.

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 dropped 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.

In total, we are left with a final valid sample size of 9,811 responses, close to the envisioned sample of 10,000. The sample size is similar in Versions 1, 3, and 4, with a range of 2,330-2,390 subjects in the three versions. The oversampling (by 15 percent) of Versions 3 and 4, as mentioned above, thus succeeded in approximately equating the sample size. Version 2 has a larger sample size, with 2,708 subjects, due to the oversampling.

2.2 Design Changes

Using the data from both the 2015 and the 2018 real-effort experiments, we measure the change in experimental results with respect to six main dimensions, listed in Table 2.

Dimension 1. Pure Replication. We compare the results from the 2015 a-b task experiment and the 2018 a-b task experiment. The two experiments have nearly identical design; the key difference is the year in which subjects are recruited, which had a small effect on the make-up of the MTurk sample. The 2018 sample has more female workers (59.2% versus 54.4%), more older workers (55.4% above the age of 30, compared to 48.5%) and more college-educated workers (58.8% versus 54.8%). Also, the 2018 experiment has a smaller sample size of about 150 subjects per treatment, compared to 550 subjects in 2015, given that the subjects in 2018 are split across four versions.

Dimension 2. Demographics. We take advantage of the demographic heterogeneity in the Mturk population and compare across three different demographic break-downs, splitting subjects into two groups of approximate size (to maximize the statistical power of the comparison). We do these splits pooling the 2015 and the 2018 data in order to maximize the sample size. We compare: (i) male workers (N=4,686) versus female workers (N=5,785); (ii) workers with a completed college degree (N=5,842) to other workers (N=4,629); (iii) workers who are up to 30 years old (N=5,259) versus workers who older than 30 (N=5,212).

Dimension 3. Geography/Culture. Using the latitude and longitude inferred from the IP address, we can geo-code the likely location of the workers (barring say the use of a VPN). Still

pooling the 2015 and 2018 a-b task data, for our main geographic/cultural comparison, we compare workers in the US (N=8,803) versus workers in India (N=1,225).[8] For an additional comparison, we compare workers in "red states" versus "blue states" according to the vote share in the 2016 presidential election.

Dimension 4. Task. We compare the pooled 2015-18 results for the a-b task to the results for the 10-minute WWII card coding task in 2018. As we discussed above, the two experimental designs are as close as possible, including keeping marginal incentives for effort close, except for a different, more motivating task.

Dimension 5. Output. We compare two versions of the WWII coding experiment, comparing Version 2 in which output is coded as the number of units of output coded within 10 minutes to Version 3 in which output is coded as the number of extra cards coded (between 0 and 20). As a second output comparison, returning to the 2015-18 a-b coding task, we compute the output in the first 5 minutes versus in the last 5 minutes.

Dimension 6. Consent. As our final comparison, we estimate the impact of awareness of participation in an experiment by comparing two versions of the extra-work WWII card coding experiment: Version 3 in which subjects see a consent form and Version 4 in which subjects do not see any consent form. There is no other difference between the two versions.

2.3 Measure of Stability

Across all the dimensions listed above, we compare the average measure of effort for the 15 treatments, across the two different experimental designs. 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

[8] We exclude the workers with geo-location in neither of these countries.

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.

3 Expert Forecasts of Stability

3.1 Design

Can academic experts predict how stable the experimental results will be to each of the six dimensions listed above? The ability of researchers to predict the importance of various design changes is an important factor for how they choose to implement experiments and how they evaluate the projects of other researchers. Following DellaVigna and Pope (2018, forthcoming), we contact a group of researchers to collect their forecasts and test their ability to make predictions about the importance of design changes.

Sample. To determine the sample of forecasters, we build on the sample of 208 experts that provided forecasts for the 2015 experiments, given that these experts are familiar with the original experiment. At the same time, we wanted to scale back the sample given the value of people's time and given that the original forecast sample of 208 respondents 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 year between 2005 and 2015; (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 experts, we picked 42. In addition, we added 18 behavioral economists with PhD in 2015-2018 (who were not included in the earlier 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). In addition, we identified 10 experts working on replication, since the topic studied is related to the issue of conceptual replication. Out of the 70 experts contacted, we received 55 responses, 50 from the behavioral experts and 5 from the replication experts, for an overall response rate of 79 percent.

As additional samples, we also contacted a group of PhD students in economics, like we did in 2015, at UC Berkeley and the University of Chicago. We obtained a total of 33 responses. Finally, we posted the same survey (for a $1 payment) on MTurk for a maximum sample of 150; we collected a total of 109 valid responses.[9]

Survey. The survey, which was expected to take 15-20 minutes, walked the forecasters through four steps. In the first step, the survey briefly summarized the design in the 2015 experiment and the key results using Online Appendix Figure 2 which lists each treatment, together with the average effort in 2015. In the second step, the survey introduced the concept of rank-order correlation, using four graphical examples, two of which are displayed in Online Appendix Figure 3.

[9] We recruited 150 MTurkers to take a forecasting survey on Qualtrics. In order to prevent bots and inattentive survey-takers, two features were implemented in the survey: a captcha verification and an attention question. Those who failed the attention check (18 MTurkers) were dropped. Furthermore, 21 Mturkers who had taken the survey in under 5 minutes were dropped as they were also likely to be inattentive survey-takers. Lastly, we removed MTurkers with the same IP address as it may indicate duplicate users (2 MTurkers). In total, we were left with a sample size of 109 MTurk forecasters.

In the third step, the forecasters are asked to make ten forecasts, as listed in Table 2. For each forecast, they simply predict the rank-order correlation between 0 and 1 using a slider (see Online Appendix Figure 3b): (i) 1 forecast about exact replication; (ii) 3 forecasts about demographics, along the gender, education, and age lines; (iii) 1 forecast about geography/culture regarding differences between the workers in US versus in India; (iv) 1 forecast about task change; (v) 3 forecasts about output change, comparing first effort in the 10-minute WWII coding to effort in the extra-work WWII coding; then comparing effort in the a-b task to effort in the extra-work WWII coding; and finally, comparing effort in the first 5 minutes of the a-b task to effort in the last 5 minutes of the a-b task; and (vi) 1 forecast about the importance of the consent form, comparing Version 3 to Version 4.

In some of these comparisons above, to ease the forecast, we provide as a benchmark the rank-order correlation under full stability, that is, what rank-order correlation we would expect to observe if the results did not change (Column 1 in Table 2). This number will be less than 1, due to sampling noise. For example, in the case of dimension (i) (pure replication), we bootstrap from the 2015 experimental sample, drawing (with replacement) from each of the 15 treatments 150 observations, to mirror the smaller sample size in the 2018 experiment, compute the average effort in each of 15 cells, and compute the rank order correlation with the 2015 results. We repeat this 300 times, and report to the forecasters the average rank-order correlation of 0.94. Column 1 in Table 2 reports also the standard deviation of such bootstrap (0.04), which we did not report to the forecasters.

We also report a boot-strap for the demographic comparisons. In this case, we pool the 2015 and 2018 sample. In each of the 15 treatments, we randomly assign a subject to either demographic group A or demographic group B. We then compute the average effort in each of the 15*2 cells, and thus the rank-order correlation. We report the average rank-order correlation across 300 bootstraps to the forecasters, which is 0.95. We also report a bootstrap for the US-India comparison, which is built the same way as the demographic bootstrap, except that we explicitly model that one of the two groups is just 12 percent of the data and the other 88 percent; this leads to a lower average rank-order correlation of 0.89.

We did not report a full-stability benchmark comparing across different tasks or output, given that comparing across experiments with different units makes the stability benchmark much less obvious. We could compute the stability benchmark for the comparison of output in the first 5 minutes and next 5 minutes (0.99), but we did not report this to the forecasters. Finally, we can compute the stability benchmark for the last comparison of the consent form (0.88), but again we did not report it.

In the fourth and final step of the forecasting survey, respondents indicated their overall confidence in their response accuracy by predicting, once again with a slider scale, how many of the 10 responses would fall within 0.1 of the correct rank-order correlation. This last question ended the survey.

3.2 Forecasts of Correlation

Figures 1a-b and Columns 2-4 in Table 2 report the results from the forecasts. On average, the experts expected that the rank-order correlation for the pure replication would be quite high (0.83), though lower than the full stability one (0.94), a difference that is statistically significant (p=0.004, Column 6). The cdf plot in Figure 1a shows that 75 percent of experts expect a correlation above 0.80, with only 10 percent of experts expecting a correlation above 0.9.

The forecasts of correlation are sizably lower for the three demographic variables, with average forecasted rank-order correlation of 0.75 (gender), 0.73 (education), and 0.75 (age). As Figure 1a shows, the cdfs for the three demographic forecasts are quite similar. Only 20 percent of experts expect a correlation of 0.85 or higher, and only 5 percent of experts expect a correlation higher than 0.9. That is, nearly all experts expect a rank-order correlation below the average rank-order correlation under full stability. The forecast of rank-order correlation for the geographic/cultural difference is further shifted down, to a correlation of 0.65. For both the demographic difference and the geographic differences, the average expert expects a rank-order correlation that is statistically significantly lower than the benchmark of full stability.

Turning the task and output correlations, the experts on average expect a correlation of 0.67 for the change in task (a-b typing versus WWII card coding) and a similar correlation of 0.63 when comparing within a task (WWII card coding) two different output margins, the effort within 10 minutes as opposed to the number of extra cards coded. Across the two forecasts, only 5 percent of experts expect a correlation lower than 0.4, with a nearly uniform distribution of forecasts between a correlation of 0.4 and a correlation of 0.8. We also elicit forecasts about the joint task/output change, comparing the 10-minute a-b typing to the extra-work WWI coding. The experts are most pessimistic about the rank-order correlation in this scenario, with an average forecast of 0.54 and 25 percent of expert expecting a correlation lower than 0.4. We also elicit a different output comparison, comparing the typing in the a-b task in the first 5 minutes, versus in the last 5 minutes; the experts on average expect a correlation of 0.72, quite a bit lower than the full-stability benchmark of 0.99. The final comparison is for the presence, or absence, of the consent form. The experts on average expect a correlation of 0.78, compared to the full-stability benchmark of 0.88.

How confident are the 55 experts about their forecasts? We asked each forecaster for a prediction of the expected number of their forecasts, out of 10, which would end up being within 0.1 of the realized value. As we display at the bottom of Table 2, the experts are only mildly confident about their accuracy, expecting 3.99 "correct" correlation forecasts. We return to the accuracy of this forecast below.

We can compare the predictions of the experts to the predictions of two other groups, PhD students at UC Berkeley and at the University of Chicago, and MTurk workers. The predictions of the PhD students track closely the predictions of the experts; we cannot reject that the two predictions are the same in each of the 10 comparisons. The PhD students express higher confidence, expecting 4.95 correct predictions out of 10. The forecasts of the MTurk subjects are on average somewhat lower, but exhibit similar patterns. Thus, the expectations do not vary much with the population at hand; we present the evidence on further splits in Section 5.

4 Stability of Experimental Results

4.1 Main Results on Stability

We now compare the results along each of the key six design comparisons.

Pure Replication. We start by comparing the results for the a-b typing task in the 2015 experiment and in the 2018 experiment. This is a pure replication, as there is essentially no change in design, other than the year it was run in. Online Appendix Figure 4a and 4b compare the distribution of effort, pooling across the 15 treatments. The distribution in the two years is very similar, if somewhat noisier in 2018, given the smaller sample size. In Figure 2a-b we plot the average effort in the 3 piece-rate treatments, comparing to the baseline no-piece rate treatment. The estimates are very similar, with some difference just in the baseline effort, which is somewhat lower in the 2018 run. But overall, the elasticity of effort to incentives is very similar, and similarly precisely estimated.

What about the other behavioral treatments? Figure 3 shows that the results for the behavioral treatments replicate very nicely as well. The treatments stack up on a line (continuous line) that is only slightly lower than the 45-degree line (dotted line). Just one treatment deviates by more than 100 points from the interpolating line, the probability weighting treatment, which yields higher effort in 2018 than one would have predicted based on the 2015 results. Overall, the rank-order correlation is very high, at 0.91, and close to the full-stability benchmark of 0.94, and higher than the average forecast at 0.82 (p=0.068 for the difference, Column 8). Thus, our first result is that doing a pure replication produces results that are remarkably similar to the original experiment.

Demographics. Next, we consider the impact of demographic differences in the subject pool, along gender/age/education lines. To maximize statistical power (and given the evidence of nearly perfect replication), we consider such differences in the pooled 2015/2018 data. In Figure 4a we display the evidence splitting male and female respondents.[10] The data suggests two striking patterns. First, men and women *do* differ: male subjects are more responsive to incentive, varying their effort from 1,450 to nearly 2,300 from the baseline treatment to the high-piece-rate treatment. Female subjects, in contrast, increase effort from 1,500 to 2,050. And yet, the second finding is that, conditional on this difference in elasticity of effort to motivation, the experimental results in the two demographic groups are remarkably lined up, as the continuous line shows. This indicates that there is no difference between men and women in how they respond to the different behavioral motivators, and in how they respond to the behavioral motivators compared to the financial motivators. This leads to a very high rank-order correlation of 0.96, which is much higher than the average expert forecast of 0.73, a difference that is highly statistically significant.

Is this result unique to the gender comparison? In Figure 4b we do a median split on the education variable, comparing subjects with a completed college degree with subjects without. The two groups of subjects do differ, but this time only in the level of effort, as opposed to the elasticity: higher-education subjects display less effort, for any given treatment. Once again, once we control for this difference, the behavioral treatment effects lined up nicely. Indeed, the rank-order correlation

[10] Online Appendix Table 2 presents the average effort for each treatment-demographic combination.

in effectiveness is 0.97, much larger than the average forecast of 0.71, a difference that again is statistically highly significant.

In Figure 4c we present the last demographic split, by age. Subjects younger than 30 years of age display higher effort than subjects that are older, but once again the rank–order of the behavioral treatments is very high (0.98).

Geography/Culture. We now turn to our third comparison: geographical and cultural lines. In particular, while the previous demographic features are self-reported, we now take advantage of the geo-location due to the IP address. We compare the average effort by treatment among the 12% of subjects that have an IP in India versus the subjects with an IP in the US. As Figure 5 shows, there is a sizable difference in the average effort, and in the elasticity, with the subjects in India displaying lower average effort and lower elasticity. Still, adjusting for this difference, the behavioral and incentive treatments are quite nicely lined up, for a rank-order correlation of 0.65. This correlation is statistically lower than the full-stability benchmark (p=0.049).

Task. In the fourth comparison, we compare the effort results in the 10-minute a-b typing task, still pooling the 2015 and 2018 experiments, to the effort results in a 10-minute task of coding the occupation in WWII enrollment cards, which we envisioned would be more motivating. Online Appendix Figure 4c shows that the distribution of the effort measure in this new task, the number of cards coded, is approximately normally distributed, with a mode and median around 60 cards. How responsive is this task to financial incentives? Figure 2c shows that the task is very unresponsive to these incentives, and in fact the effort with respect to the piece rate is not monotonic; indeed, we cannot reject that the high-piece rate treatment and the baseline no-piece-rate treatment yield the same effort.[11]

In light of this, it is not surprising that the correlation between the results across the two tasks is not particularly high. Figure 6 shows that the rank-order correlation is 0.64, in line with the average expert forecast of 0.66. In fact, given the noise in this task, we cannot reject a rank-order correlation as low as 0.34.

Output Measure. In our fifth comparison, we consider how changes in measures of output, even for a given task, may change the experimental findings. We start by comparing two versions of the WWII card-coding task: the one described above, with a 10-minute time limit, and a second one, in which subjects can code as many extra cards as they decide from 0 to 20, after completing a first required batch of 40 cards. Online Appendix Figure 4d shows that the distribution of extra cards coded in this task is highly bimodal: pooling across the 15 treatments, the large majority of subjects code 0 extra cards, or all 20 extra cards, with only a small number of subjects coding a number of cards between 1 and 19. Most importantly, the output measure in this task is highly responsive to incentives. As Figure 2d shows, the average number of extra cards coded rises from 8.6 (no piece rate) to 12.6 (low piece rate) to 15.2 (mid piece rate) to 17.4 (high piece rate). This increase is highly significant; importantly, the increase in effort is statistically significant even moving from the mid-piece rate to the high-piece rate, while this increase is not significant in the a-b task. Thus, this

[11] While it is not the focus of the experiment, a legitimate question is whether the incentive conditions induce differences in accuracy in the coding of cards, in addition to differences in quality. Online Appendix Table 4 and Online Appendix Figure 5 show that there is no systematic relationship between the number of units coded in the different treatments and the accuracy of the coding.

extra-work task appears well-suited to capture variation in motivation.

Figure 7a shows that the effort recorded with this output measure only has a low correlation of 0.27 with the effort in the 10-minute WWII coding task. In particular, the correlation is much lower than in the expert forecasts (0.61).

This (relative) instability has two possible explanations. First, changes in task and output may have affected the impact of behavioral and financial motivators. Second and more simply, the 10-minute WWII coding task, unlike the a-b task, is quite insensitive to motivation, as we saw. Thus, purely due to noise we would expect a lower correlation, as the noise in the realized effort by treatment would swamp the treatment effects.

In order to provide some evidence on whether the task change *per se* changed the behavioral results, setting aside the noise, we compare output in the a-b 10-minute task to output in the WWII coding with extra cards. As we showed, both of these tasks are responsive to incentives and thus the comparison should not be too affected by noise. If, instead, changes in task and output measure change the behavioral effects, even aside from noise, we would expect the correlation between the a-b 10-minute task to output in the WWII coding with extra cards to be relatively low, as both task and output measure change.

Figure 7b shows that the correlation for the joint task/output change, 0.70, is quite high, and much higher than for just the output change, 0.27, consistent with the important role of noise in the experimental results. Interestingly, the expert forecasters instead expect the correlation to be higher for just the output change, 0.63, than for the task/output change, 0.54.

As a final output comparison, in Figure 7c we return to the a-b typing task (pooling 2015 and 2018) and compute the results using as a first measure of output the number of cards coded in the first 5 minutes and as a second measure the number of cards coded in the next 5 minutes. As Figure 7c shows, the two measures are very highly correlated, with a rank-order correlation of 0.97, close to the full-stability benchmark of 0.99 and clearly higher than the average forecast of 0.72.

Consent. As our final comparison, we estimate the impact of awareness of participation in an experiment by comparing two versions of the extra-work WWII card coding experiment, Version 3 in which subjects see a consent form and Version 4 in which subjects do not see any consent form. There is no other difference between the two versions. As Figure 8 shows, the two versions yield very similar results, with all treatments close to the 45 degree line. The rank-order correlation between these two tasks is 0.84, close to the full-stability benchmark of 0.88 and higher, but not too dissimilar to the average forecast of 0.78.

Overall Assessment. To summarize the results across the different versions, also summarized in Table 2, in Figure 9 we plot for each of the 10 rank-order correlation predictions the average expert forecast of correlation versus the actual correlation. As the figure makes clear, the two measures display only a weak correlation.

4.2 Robustness

Alternative Measure of Stability. A legitimate worry is that the results displayed so far on the impact of design changes depend on the specific measure used, the rank-order correlation. In Online

Appendix Table 3, we replicate the results on the stability of the empirical results across different versions using alternative measures. For these measures we do not have, of course, expert forecasts, but we can still compute the actual measure of stability and, when possible, the measure under full stability.

In Columns 1 and 2 we present the results using the Pearson correlation which is related to the rank-order correlation but imposes a linearity assumption between the two version being compared. The table shows that the results are very similar using this alternative measure.

In the next columns, we move away from correlation and compute the effect of the treatment in one of the quantitative scales, relative to a baseline. In Columns 3 and 4, for each treatment (other than the baseline one), we compare the difference in effort in log points, compared to the effort in the baseline group. We then compute, for each treatment, the absolute difference in this log-point effect across the two versions–say, between male subjects and female subjects—and then average across the 14 treatments. This measure shows that the pure replication and the different demographic versions are associated with fairly small log point changes, and in any case within the confidence interval of the full-stability benchmark. The log point change is larger for the task and output changes, not surprisingly since this measure essentially assume the same elasticity—same log point response in effort—across different versions. Columns 5 and 6 show that the calculations are fairly similar if we use a different treatment as comparison point, in this case the high-pay treatment.

In Columns 7-10 we repeat the same exercise, but we measure the changes from the baseline in standard deviation units (z scores), instead of in log point units. As for the other columns, the pure replication, demographics, and geography changes all yield results within the full-stability benchmark, with larger differences for the task and output changes.

Alternative Comparisons of Designs. A separate robustness issue is that we focus on ten rank-order correlation comparisons to estimate the degree of stability with respect to the various dimensions. What if we use other comparisons within a particular dimension?

In Table 3 we consider 11 additional comparisons which expand along our motivating dimensions. The first three comparisons present the familiar demographic comparisons, but instead of making the comparison for the a-b typing task, we compare along demographic dimensions for the 10-minute WWII card coding task.[12] The rank-order correlation across the two versions is clearly lower than for our benchmark comparisons, given the smaller sample and noisiness of the results in this sample, but the correlation is sizable and close to the one computed under full stability.

Next, we revisit the geographic/culture comparison between the India and US sample for the extra-card WWII card coding task.[13] We obtain a rank-order correlation between the results for the two samples of 0.68, which is close to the full-stability benchmark of 0.77.

In the next step, we provide a different measure of geographic and cultural differentiation, comparing between Mturkers with an IP address in "Red states" versus "Blue States", which we determine using the geo-coding of the IP address and attributing state to either group depending

[12]We cannot make this comparison for the extra-card WWII coding task, since we did not collect demographics for that task, since we did not want to collect demographics for a task that, in Version 4, we run as an actual data coding job, with no consent form.

[13]We do not do such comparison for the 10-minute WWII task given the noisiness of the estimates, given the Indian workers constitute only 12% of the sample.

on the winner of the vote share in the 2016 presidential election. In this case, we obtain a very strong estimate of stability of the results, with a rank-order correlation of 0.94, very close to the full-stability benchmark of 0.97. Overall, these results further reinforce the message that demographic and geographic variation in the result are small and the results are close to full stability along these dimensions.

In the final six comparisons, we consider two further forms of sample selection which do not fit neatly into either of the other dimensions, but which have been identified by previous papers as potentially important for the productivity of MTurk workers (Case et al., 2017): (i) whether subjects sign up early on in a experimental study, or later on, as this could be a proxy for how motivated the worker are; and (ii) whether the subjects perform the test during the day or during the night. We compare along these two dimensions for the a-b typing task, for the 10-minute WWII card-coding task, and for the extra-cards WWII coding task. For five of the six comparisons, the actual rank-order correlation is close to the one under full stability, providing another example of stability of the results.

4.3 Structural Estimates

We return the model which we briefly described in Section 2.1.1 and present estimates of the model parameters. We use such estimates to quantify the elasticity of effort in the various design versions, and to present an alternative measure of the stability of the results across different design versions: the stability of the underlying structural parameters. In addition, we discuss the results of the 16th treatment, which we have omitted so far, as an out-of-sample model validation.

Estimation. To bring the model in Section 2.1.1 to the data, we need to specify the source of heterogeneity in the data. As benchmark model, we take the specification with exponential cost of effort function, since it implies a specification that conveniently expresses effort as function of the motivation parameters; we show below that the results are similar assuming a power cost of effort function. Building on DellaVigna et al. (2015), we assume that the cost of effort parameter k has a log-normal distribution across subjects j, implying a cost of effort $c_j(e_j) = kexp(\gamma e_j)\gamma^{-1}exp(-\gamma\varepsilon_j)$, with ε_j normally distributed $\varepsilon_j \sim N(0, {\sigma_\varepsilon}^2)$. This assumption ensures positive realizations for the marginal cost of effort. Given that the agent maximizes $(s+p)e - c(e)$, this implies the first-order condition $s + p - kexp(\gamma e_j)exp(-\gamma\varepsilon_j) = 0$ and, taking logs and transforming, yields

$$e_j = \frac{1}{\gamma}[log(s+p) - log(k)] + \varepsilon_j. \tag{4}$$

Equation (4) can be estimated with non-linear least squares (NLS). We estimate the three parameters $\hat{s}$, $\hat{k}$, and $\hat{\gamma}$, taking advantage of the four piece rate treatments. In the a-b button pushing task, we specify effort e_j as the number of button presses, in the 10-minute WWII coding as the number of cards coded, and in the extra-work experiment as the number of cards coded, including in the count the required 40 cards.

In order to accommodate the behavioral treatments, we generalize the model of motivation, allowing for an additional parameter for each behavioral treatment. Specifically, for the paying-

too-little treatment, for the gift-exchange treatment, and for the three psychological treatments, we allow for additive motivation shifters Δs such that motivation becomes $s + \Delta s$. For example, the null hypothesis of no crowd out due to paying too little entails $\Delta s_{CO} = 0$.

For the two social preference treatments, we allow for both a pure-altruism parameter α a la Becker (1974) and for a "warm glow" parameter a: the altruism parameter α multiplies the actual return to the charity while the warm glow term a multiplies the return to the charity for the low-return treatment (1 cent per 100 presses for the a-b task). Intuitively, the difference between the two models is that in the Beckerian pure altruism world, the return to the charity is important, and we expect the effort to be significantly higher when the return to the charity is high (10 cents per 100 points) versus when it is low (1 cent per 100 points). In the "warm glow" model, instead, the return to the charity does not matter, as the individual exerts extra effort for the charity in response to a "warm glow", not in response to the exact return.

For the two delayed-payment treatments, we model the motivation part as in the present-bias model (Laibson, 1997; O'Donoghue and Rabin, 1999) as $(s + \beta\delta^t p)e$, with t denoting the weeks of delay, β the present bias parameter, and δ the (weekly) discount factor.

Finally, for the probability weighting treatments, we model the motivation $(s + \pi(P)p)e$, where P is the probability of receiving the piece rate, that is, $P = 0.01$or $P = 0.5$. Under risk neutrality, we should estimate $\pi(P) = P$. The evidence on probability weighting (e.g., Prelec, 1998) suggests that small probabilities are overweighted by a factor of 3 to 6, with a probability of 50 percent is slightly downweighted. The treatment with a 1 percent probability of a \$1 piece rate allows us to test for such overweighting of small probability and estimate $\pi(0.01)$, while the treatment with 50 percent probability of a 2-cent piece rate to provide evidence on the concavity of the value function, i.e., the risk aversion, which in this case we capture as reduced-form in $\pi(0.5)$.

In Table 2 we present estimates of the parameters of the model, using all the 15 treatments used for the design comparisons. Since our estimation allows for one parameter for each behavioral treatment, effectively the identification of the incidental parameters is given by the piece-rate treatments, while the identification of the behavioral parameters is given by the behavioral treatments. That is, the incidental parameters in Table 2 are essentially identical if we estimate them including only the piece rate treatments.

Estimates, Button Pushing. Columns 1 and 2 report the estimates of the model using NLS on, respectively, the 2015 button-pressing data and the 2018 button-pressing data. The estimates for the 2015 experiment replicate the ones in DellaVigna and Pope (2018) and the estimates for the 2018 experiment are very close: in both data sets, the elasticity of effort is quite precisely estimated to be $1/\hat{\gamma} = 0.04$. Figures 2a-b display the predicted effort given the parameter estimates and show that the model fit is near perfect. This is not obvious given that the model fits 4 piece rates with 3 parameters.

The next rows show the estimates of the behavioral parameters. The estimates for the motivation terms—that is, s and Δs—are displayed in units for cents per 100 presses. For example, the estimate for the social comparison treatment $\hat{\Delta s_{SC}} = 0.06$ indicates an impact equivalent to an incentive of 0.06 cents per 100 presses. Indeed, this treatment, which is the most effective of all the psychological treatments, has an effect which is smaller than even the paying-too-little treatment, which we code

as having an incentive $p = 0.1$. As we noted in DellaVigna and Pope (2018), there is no evidence that the paying-too-little treatment crowds out motivation, with estimates for $\Delta\hat{s_{CO}}$ very close to zero.

The estimates for the social preference parameters are the most informative ones. They indicate a precisely-estimate zero effect for the altruism parameter, with point estimates $\hat{\alpha} = 0.003$ (s.e. 0.010) for 2015 and $\hat{\alpha} = 0.010$ (s.e. 0.017) for 2018. This means that in both years we can reject a pure altruism coefficient as low as $\alpha = 0.05$; for comparison, full altruism (equal weight on the recipient) is $\alpha = 1$. The estimates indicate instead a warm-glow weight $\hat{a}$ around 0.1. This is consistent with the fact that we do not find any response in worker effort to the return to the charity, but we do find that subjects work harder when there is a charitable giving, compared to the baseline condition. Turning to the time-preference point estimates, unfortunately they are too imprecisely estimated in this design.

Estimates, Demographics. With these estimates at hand, we revisit the key finding above that the results are very stable with respect to demographic shifts. We present the result of estimates split by gender (Columns 3 and 4), but education (Columns 5 and 6) and by age (Columns 7 and 8). In all cases, we pool the 2015 and 2018 experiments, consistent with the results in Column 1 and 2 suggesting no differences, and so as to maximize statistical power. The point estimates indicate that there are some differences across the groups in the incidental parameters, especially the curvature, even as the differences are not quite statistically significant. For example the estimated cost-of-effort curvature $\hat{\gamma}$ equals 0.012 (se 0.003) for males but 0.019 (se 0.007) for females, and it equals 0.011 (se 0.003) for younger workers but 0.022 (se 0.009) for older workers. Among the behavioral parameters, however, there is no evidence of any difference. In particular, the social preference parameters, which are among the most precisely estimated, indicate very consistent evidence supporting the warm glow model, as opposed to the pure altruism model.

Estimates, WWII Task. We then turn in Columns 9 to the estimate of the 10-minute WWII task. The estimates for the curvature parameter $\hat{\gamma} = 1.909$ imply an elasticity smaller than 0.01, really tiny, consistent with the very limited response to incentives. Figure 2c shows that we capture to some extent the response to incentives in the data, but the fit is imperfect, as expected given the observed non-monotonicity in the response of effort to piece rate. Given the very small elasticity, the estimates of the key parameters are necessarily noisy; nonetheless, qualitatively the key parameters are aligned.

In Columns 10 and 11 we turn to the extra-work treatments for the WWII coding, with experimental consent (Column 10) and without (Column 11). In this case, we have to explicitly model the censoring at both 0 cards coded and at 20 cards coded, and we thus turn to maximum-likelihood estimation. Otherwise, the model estimation is the same as for the other specifications.

The estimates indicate a much higher elasticity of effort to incentives, just as we expected based on the results in Abeler et al. (2011) and Gneezy et al. (2017). Indeed, the estimated $\hat{\gamma} = 0.047$ (se 0.0015) in Column 10 implies an elasticity of $1/(0.047 * (40 + 11.1) = 0.42$, which is among the highest real-effort elasticities recorded, higher than in all cases in the literature we are aware of where effort is measured as units produced in a fixed amount of time (e.g., 0.1 for stuffing envelopes in DellaVigna et al., 2015 and 0.025 for the slider task in Araujo et al., 2016). This higher elasticity

implies that this design yields good statistical power for the behavioral estimates. Figures 2d-e show that the structural estimates for the cost of effort function capture well the curvature observed effort under the different piece rate conditions.

Importantly, the structural estimates for the various treatments are in line with what we find for the other designs, subject to recognizing differences in the baseline level of intrinsic motivation s (now measure in terms of cents per extra card). The one difference is the higher intrinsic motivation due to gift exchange.

Out-of-Sample Prediction. These estimates also allow us to make predictions about out 16th treatment, which combines the low-piece rate incentive with a "please try" psychological inducement. Our structural estimates make a prediction on what the observed effort should be in this treatment – how close do we come in the prediction? The bottom row in the table shows that the model does quite well in the prediction.

5 Revisiting the Forecasts

In this section, we return to the expert forecasts to further probe some of the findings and interpretations from the analysis.

Impact of Noise on Stability. A key theme for the results on task and output is the impact of noise in the experimental results. As Figures 2a-b show, the 10-minute WWII task is very insensitive to incentive, unlike the button pushing task and the extra-cards WWII task, both of which clearly respond to motivation. As a result of this, the experimental findings in the 10-minute WWII task are much more imprecisely estimated than in the other design versions, since the average output in essentially all treatments varies just between 52 and 61 cards coded within 10 minutes, with a substantial between-subject dispersion (Online Appendix Figure 4c). As we discussed above, this additional degree of noise largely explains the lower degree of stability of the results across tasks and across output, when one of the comparisons involves the 10-minute WWII task.

As we discussed, the forecasters do not appear to anticipate this pattern. Of course, it was not obvious that the 10-minute WWII card task would have substantially more noise than any of the other designs. Precisely to address this issue, we randomized the provision of additional information. All the forecasters were informed about the overall mean and standard deviation of effort in the two WWII card experiments. For one half of the forecasters, in addition, we provided the mean effort (and s.e.) under the three key piece rate treatments, indicating a flat and non-monotonic pattern with respect to incentives in the 10-minute WWII task, and in contrast a precisely-estimated responsiveness to incentives in the extra-work WWII task. Does this additional information have an impact on the forecasts for the task and output comparisons?

In Table 5, we compare the forecasts by the two groups in Columns 2 and 3, using the pooled sample of academic experts and PhDs (Column 1), given that the two groups make very similar forecasts. Columns 2 and 3 show that the experts respond very little to the additional information on the noisiness of the 10-minute WWII task. Thus, the forecasters do not appear to take much into account an important determinant of the stability of experimental results, the noisiness of an experimental set-up.

Forecaster Effort. In Table 5, we also consider another determinant of possible differences in forecaster accuracy. As we document in DellaVigna and Pope (forthcoming), forecasters who appear to put more effort by taking longer time and by clicking on links do a bit better in the their forecasts (at least in some conditions). In Columns 4 and 5 we split the forecasters depending on whether the forecasters clicked on at least one link to gather additional information on the experimental design. In Columns 6 and 7 similarly we split by the time taken to do the survey. Under either dimension, we find little evidence that the forecasts differ in the direction of higher accuracy.

Confidence. A relevant question too is whether confidence in the answers is a relevant predictor of the forecasts, as it is, to some extent, in DellaVigna and Pope (forthcoming). In the last question of the survey, we asked the forecasters how many of their rank-order correlation forecasts they expect to be within 0.1 of the truth. In Columns 8 and 9 we present the average forecast for individuals with higher, versus lower, confidence. The individuals with higher confidence are closer to the truth on average for the exact replication and for the impact of demographics, the impact of the task, and the impact of consent. They are, however, not closer to the truth for the two first output forecasts.

Overall, confidence is a predictor of accuracy, as we show in an alternative form in Figure 10. The figure shows the actual number of forecasts within 0.1 of the truth for the group of forecasters making that forecast. Unbiased forecasts should lie on the 45-degree line. The plot shows that the accuracy does increase with the confidence, but the slope is too flat. In particular, while the individuals with lower confidence are unbiased, the individuals with higher confidence overstate their precision. For example, the 12 forecasters who on average expect to get 6 answers correctly get just 4 answers correctly, a mean forecast that is statistically different from 6. This suggests that experimenters with higher confidence in the design have real information about the stability of the results, but probably not as much as they think they have.

Expert Accuracy. A related question is whether there are sizable differences in the ability to predict experimental results across forecasters. In DellaVigna and Pope (forthcoming) we showed that some of the variables that one would have expected to be high predictors of accuracy – like the professorial rank or citations – do not predict accuracy. This is indeed also the case in our setting, given that the accuracy of PhD students is at least as high as the accuracy of the faculty forecasters (Table 2). But in DellaVigna and Pope (forthcoming) we find some evidence suggestive of individual differences in forecasting ability, as in the superforecasters literature (Tetlock and Gardner (2015)). In DellaVigna and Pope (forthcoming) forecasters who do a better job forecasting a group of treatment also have higher accuracy in forecasting other groups of treatments within the experiment.

But does this forecasting ability translate across experiments? For the 35 individuals who made forecasts both in 2015 and in 2018, we can compare the accuracy of their two forecasts. Figure 11a displays the accuracy of each of these 35 forecasters, with their average absolute error (in terms of point) in the 2015 forecasts on the x axis and their average absolute error (in terms of rank-order correlation) in the 2018 forecasts. As Figure 11a shows, there is no correlation, or in fact the hint of a negative correlation, between accuracy across the two experiments. This finding suggests that the correlation in accuracy in forecasts may be small.

Explaining the 2015 Forecasts Errors. Finally, we return to the 2015 forecasts to reinterpret

some patterns of the forecasts in light of the newer data. As we document in DellaVigna and Pope (forthcoming), the average (i.e., wisdom-of-crowd) forecast for each treatment does a good job of predicting the average effort in that treatment. Yet, there are some treatments where the experts are sizably off in their forecast: the experts on average under-predict effort in the very-low-pay treatment and over-predict effort for the probability weighting treatment and for the ranking treatment. One interpretation of these results is that the experts were not wrong: their forecasts are on average accurate, but the specific experimental design that we ran in 2015 provides a result that may not be representative of the result over a range of different designs.

Thus, we can revisit the findings in light of the 2018 experiment and ask if the treatments where experts had the larger forecast error in 2015 are such that the treatments do better in the 2018 new runs than they did in 2015. Figure 11b presents this evidence. The x axis indicates for each treatment the average forecast error, while on the y axis we plot, for each of the four 2018 new versions of the experiment, how much a treatment shifted in rank from the 2015 experiment to the 2018 experiment. For example, Figure 11b shows that the probability weighting treatment indeed moves up by 3, 4, 5, and 6 ranks in the four 2018 runs compared to the 2015 results. On the other hand, there is no evidence that the very-low-pay treatment moves down in ranks, as one would predict based on the 2015 forecast error. All in all, Figure 11b provides just suggestive evidence that the 2015 forecast errors could be explained by alternative versions of the design.

6 Conclusion

In this paper, we have considered a particular experimental setting—a real effort task with a dozen of treatments corresponding to behavioral and financial motivators—and we have examined the stability of the findings to several design changes. We considered pure replication, changes in the demographic groups and in the geographic/cultural mix of subjects, changes in the task and in the output measure, and changes in whether subjects are aware that they are part of an experiment. We compared the results on stability to both the forecasts of experts and to a benchmark of full stability, which accounts for noise in the results. While we stress that any lessons are to some extent specific to the experimental set-up we consider, we highlight two main implications.

The first implication is methodological. We highlight, and attempt to address, the issues that arise when examining the stability of an experimental finding to substantial changes in design. When one compares across different tasks and output measures, one needs a measure of stability that accounts for the fact that the units of measure may not be compatible. We proposed rank-order correlation as a measure of stability with desirable properties when one compares several treatments. The measure is simple enough that it is possible to also elicit forecasts of stability.

The second implication is in the substance. We find a remarkable degree of stability of experimental results with respect to changes in the demographic composition of the sample, or even geographic and cultural differences, in contrast to the beliefs of nearly all the experts, who expected larger differences in results due to the demographic composition. We also find that the degree of noise in the experimental results is, in our setting, the main determinant of stability: the only two instances of low replication are due to a task with very inelastic output, limiting the role of moti-

vation compared to the role for noise. The experts do not appear to fully appreciate the important role for noise, even when provided with diagnostic information.

What can explain the divergence between the replication results and the expectations of experts? We conjecture that selective publication (Christensen and Miguel, forthcoming) may provide at least a partial explanation: while null results on demographic differences typically do not get published, or even remarked upon in a paper, differences that are statistically significant draw attention. Similarly, experimental designs with (ex post) noisy results are typically not published.

References

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. 2011. "Reference Points and Effort Provision" *American Economic Review,* Vol. 101(2), 470-492.

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation", *Quarterly Journal of Economics*, Volume 130(3), 1117–1165.

Araujo, Felipe A., Erin Carbone, Lynn Conell-Price, Marli W. Dunietz, Ania Jaroszewicz, Rachel Landsman, Diego Lamé, Lise Vesterlund, Stephanie W. Wang, Alistair J. Wilson. 2016. "The slider task: an example of restricted inference on incentive effects". *Journal of the Economic Science Association*, Volume 2(1), pp 1–12.

Becker, Gary S. 1974. "A Theory of Social Interactions" *Journal of Political Economy*, 82(6), 1063-1093.

Brandts, Jordi and Gary Charness. 2011. "The strategy versus the direct-response method: a first survey of experimental comparisons," *Experimental Economics*, Volume 14(3), 375–398.

Camerer, Colin et al.. 2016. "Evaluating Replicability of Laboratory Experiments in Economics" *Science*, 10.1126.

Camerer et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015". *Nature Human Behavior*.

Case, Logan; Chandler, Jesse; Levine, Adam; Proctor, Andrew; and Strolovitch, Dara; 2017; "Intertemporal Differences Among MTurk Workers: Time-Based Sample Variations and Implications for Online Data Collection", *Sage Open*, pp. 1-15.

Christensen, Garret S., and Edward Miguel. Forthcoming. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature*.

DellaVigna, Stefano, John List, Ulrike Malmendier, and Gautam Rao. 2015. "Estimating Social Preferences and Gift Exchange at Work" Working paper.

DellaVigna, Stefano, and Devin Pope. 2018. "What Motivates Effort? Evidence and Expert Forecasts", *Review of Economic Studies*, April 2018, Vol. 85, 1029–1069.

DellaVigna, Stefano, and Devin Pope. Forthcoming. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy.*

de Quidt, Jonathan, Johannes Haushofer, Christopher Roth. Forthcoming. "Measuring and Bounding Experimenter Demand," *American Economic Review.*

Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. 2015. "Using prediction markets to estimate the reproducibility of scientific research", *PNAS*, Vol. 112 no. 50, 15343-15347.

Falk, Armin, James J. Heckman. 2009. "Lab Experiments Are a Major Source of Knowledge in the Social Sciences" *Science*, Vol. 326(5952), pp. 535-538.

Gill, David and Victoria Prowse. 2012. "A structural analysis of disappointment aversion in a real effort competition" *American Economic Review*, 102(1), 469-503.

Gneezy, Uri, Lorenz Goette, Charles Sprenger, and Florian Zimmermann. 2017. "The Limits of Expectations-Based Reference Dependence" *Journal of the European Economic Association*, Vol. 15(4), pp. 861–876.

Gneezy, Uri and John A List. 2006. "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments," *Econometrica*, Vol. 74(5), 1365-1384.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. "Performance in Competitive Environments: Gender Differences," *Quarterly Journal of Economics*, Vol. 118(3), pp. 1049–1074.

Harrison, Glenn W. and John A. List. 2004. "Field Experiments," Journal of Economic Literature, Vol. 42(4), pp. 1009-1055

Hoffman, Elizabeth, Kevin McCabe and Vernon L. Smith. 1996. "Social Distance and Other-Regarding Behavior in Dictator Games." *American Economic Review,* 86(3), 653-660.

Horton, John J. and Chilton, Lydia B. 2010. "The Labor Economics of Paid Crowdsourcing" *Proceedings of the 11th ACM Conference on Electronic Commerce.*

Horton, John J., David Rand, and Richard Zeckhauser. 2011. "The online laboratory: conducting experiments in a real labor market" *Experimental Economics*, Vol. 14(3), pp 399-425.

Imas, Alex. 2014. "Working for the "warm glow": On the benefits and limits of prosocial incentives" *Journal of Public Economics*, Vol. 114, pp. 14--18.

Ipeirotis, Panagiotis G. "Analyzing the Amazon Mechanical Turk Marketplace. 2010. " *XRDS: Crossroads, The ACM Magazine for Students* Vol. 17, No. 2: 16-21.

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva. 2015. "How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments." *American Economic Review* 105(4): 1478-1508.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* Vol. 112, No. 2: 443-477.

List, John A. 2006. "The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions," *Journal of political Economy,* 114(1), 1-37.

Mellers, Barbara, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar, Philip Tetlock. 2015. "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions," *Perspectives on Psychological Science* May 2015 vol. 10 no. 3 267-281.

O'Donoghue, Edward and Matthew Rabin. 1999. "Doing It Now or Later". *American Economic Review*, Vol. 89(1), 103-124.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349, aac4716.

Paolacci, Gabriele. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgement and Decision Making* Vol. 5, No. 5: 411-419.

Prelec, Drazen. 1998. "The Probability Weighting Function." *Econometrica* Vol. 66, No. 3: 497-527.

Tetlock, Philip E., Dan Gardner. 2015 *Superforecasting: The Art and Science of Prediction*, Crown Publisher.

Figure 1. Expert Forecasts, CDFs

Figure 1a. Forecasts of Replication and Demographics

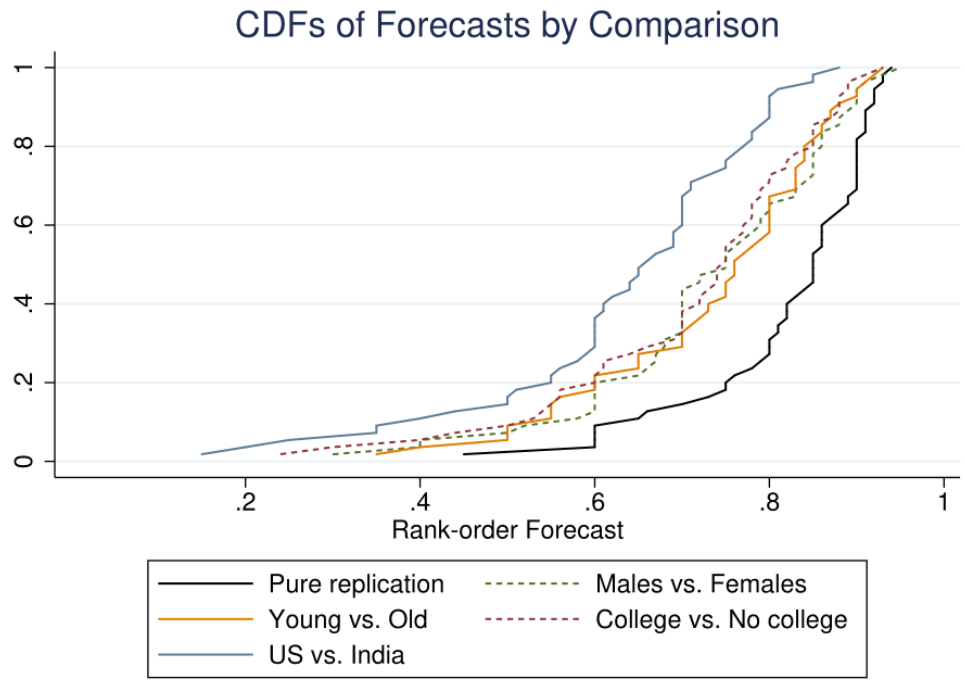
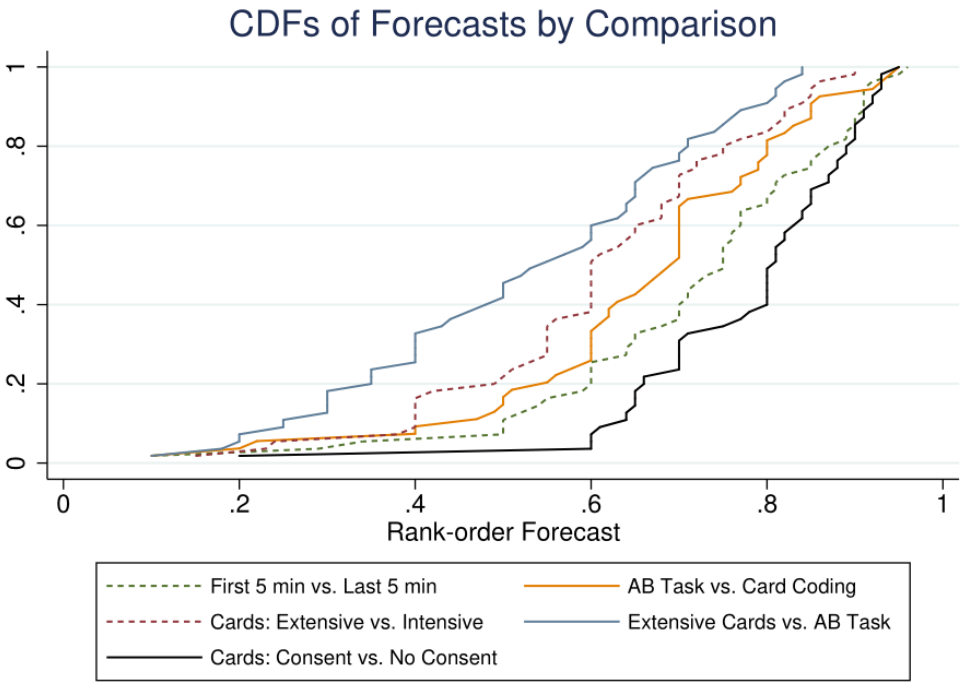


Figure 1b. Forecasts of Output, Task, and Context



Notes: Figures 1a-b present the c.d.f. of the forecasts by the 55 academic experts. Each expert made forecasts about rank-order correlation with respect to 10 design changes. We split the 10 forecasts into Figure 1a and Figure 1b.

Figure 2. Average Effort in Piece-Rate Treatments

Figure2a. 2015 Button Pushing Task

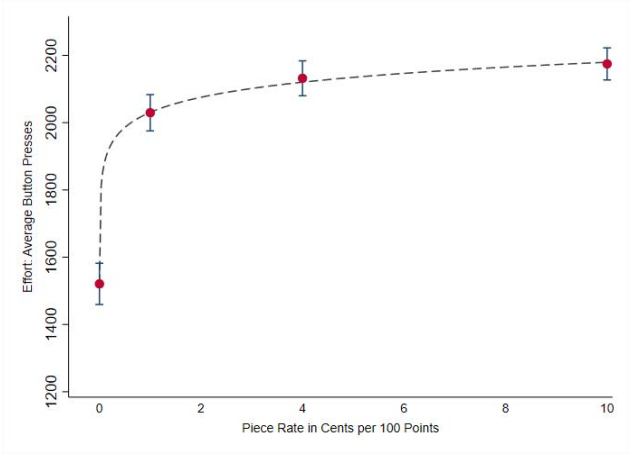


Figure 2b. 2018 Button Pushing Task

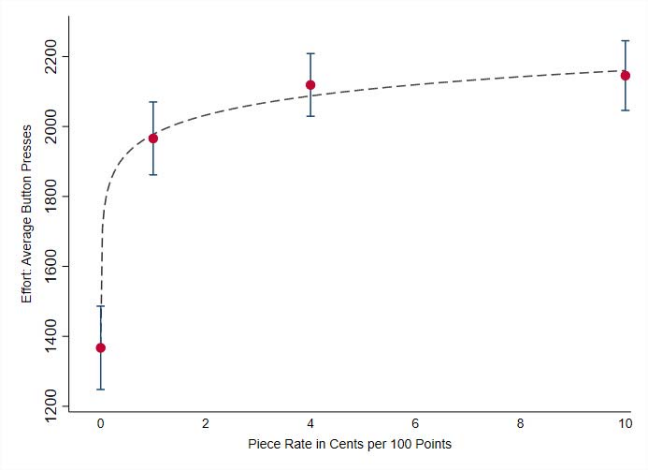


Figure 2c. 2018 10-Minute WWII Card Coding Task

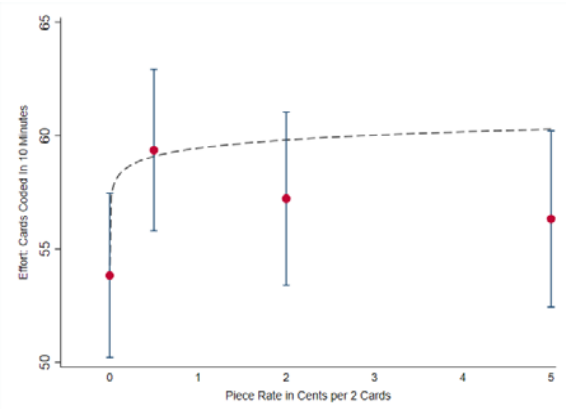


Figure 2d. 2018 Extra Card Coding Task

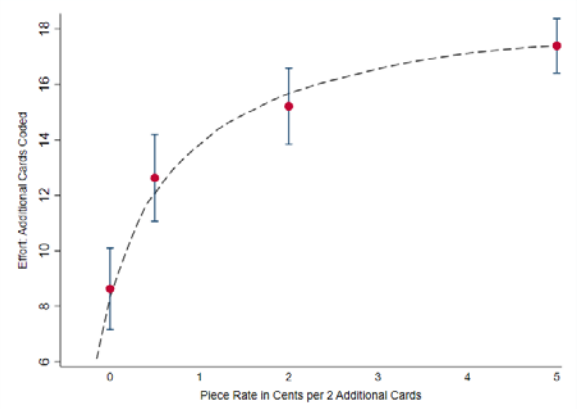
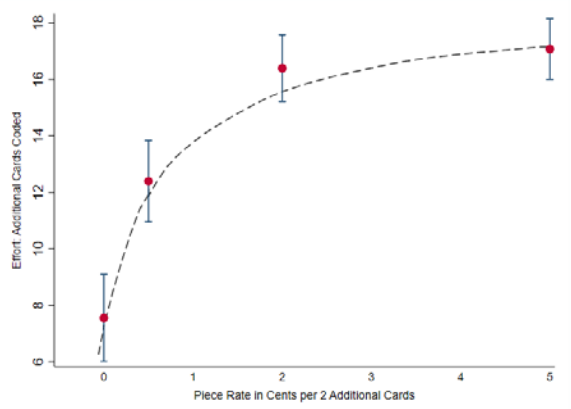
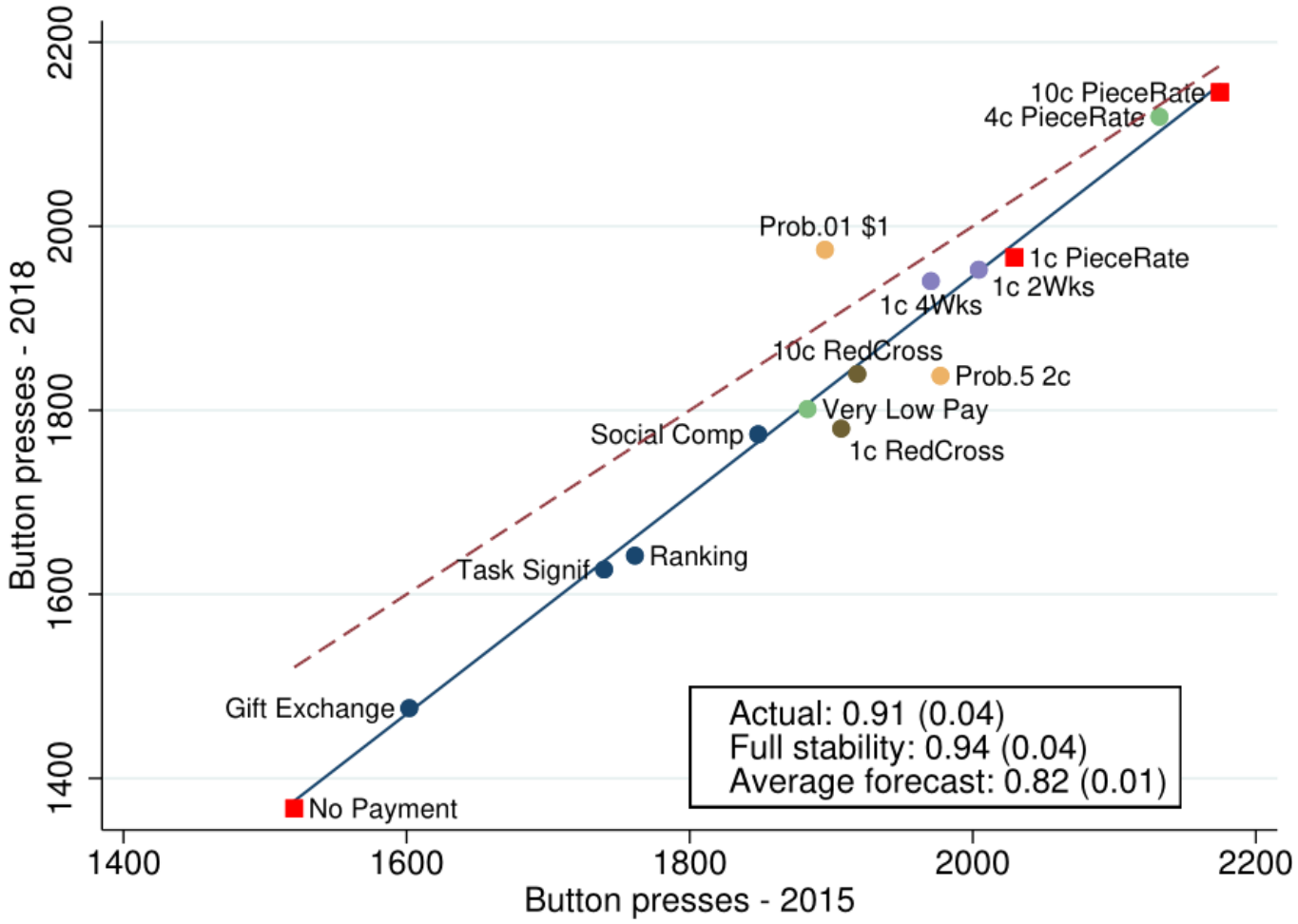


Figure 2e. 2018 Extra Card Coding Task, No Consent



Notes: Figures 2a-e displays the average effort in four piece rate conditions (including the no-piece-rate baseline), separately each of five experiments: the 2015 button press (Figure 2a), the 2018 button press (Figure 2b), the 2018 10-minute card coding (Figure 2c), the 2018 extra card coding (Figure 2d), and the 2018 extra card coding with no consent form (Figure 2e).The figures display a 95% confidence interval around the mean effort. The figure also displays with a dotted line the predicted effort from the structural estimates in Table 4, Columns 1, 2, 9, 10, and 11.

Figure 3. Pure Replication, Button Pushing Task



Notes: Figure 3 displays, for each one of 15 treatments, the average effort across two experimental versions: on the x axis the average effort in the 2015 button pushing task, on the y axis the average effort in the 2018 button pushing task. The 15 treatments are denoted with dots of different shape and color to indicate different groups of treatments: e.g., the square red dots denote the baseline and piece-rate treatments. The dotted line indicates the 45-degree line, while the continuous blue line is the best-fit line. The figure also indicates the rank-order correlation across the two versions, the rank-order correlation under a benchmark of stable results (see text for details), and the average forecast of rank-order correlation by the experts.

Figure 4. Impact of Demographics, Button Pushing Task

Figure 4a. Gender

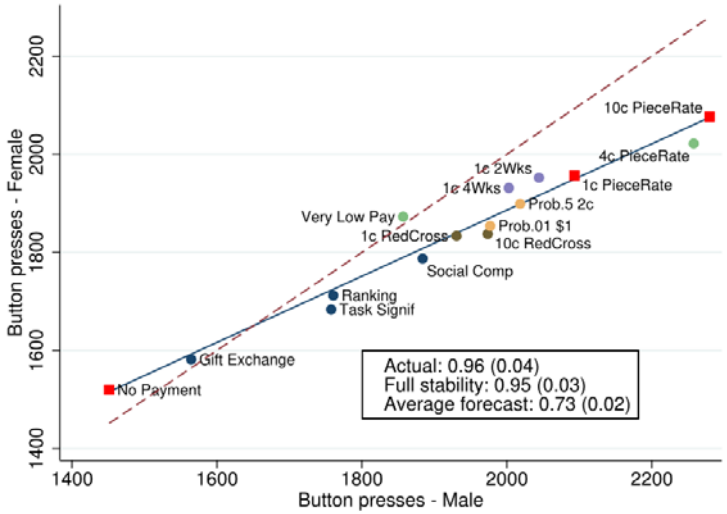


Figure 4b. Education

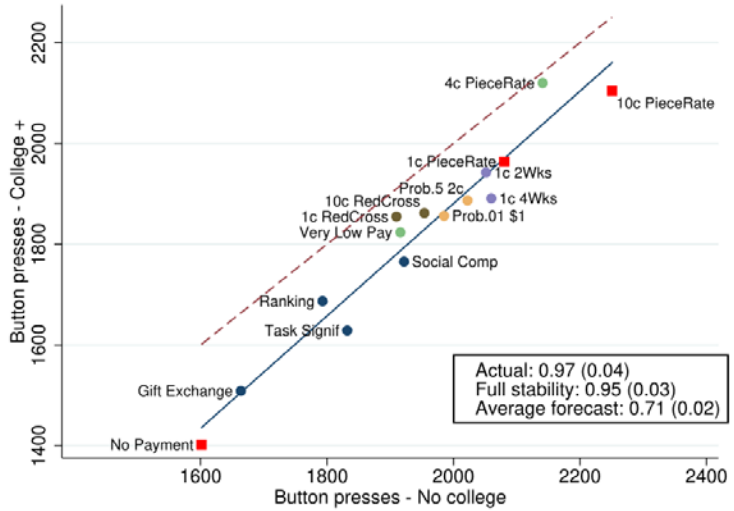
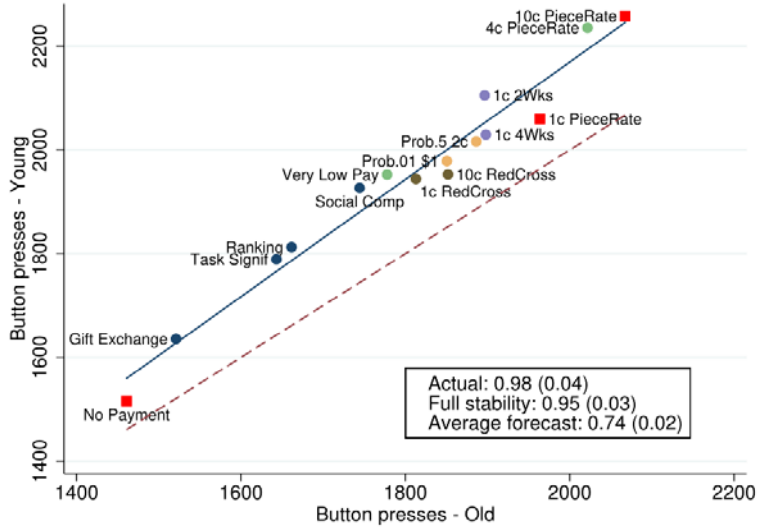
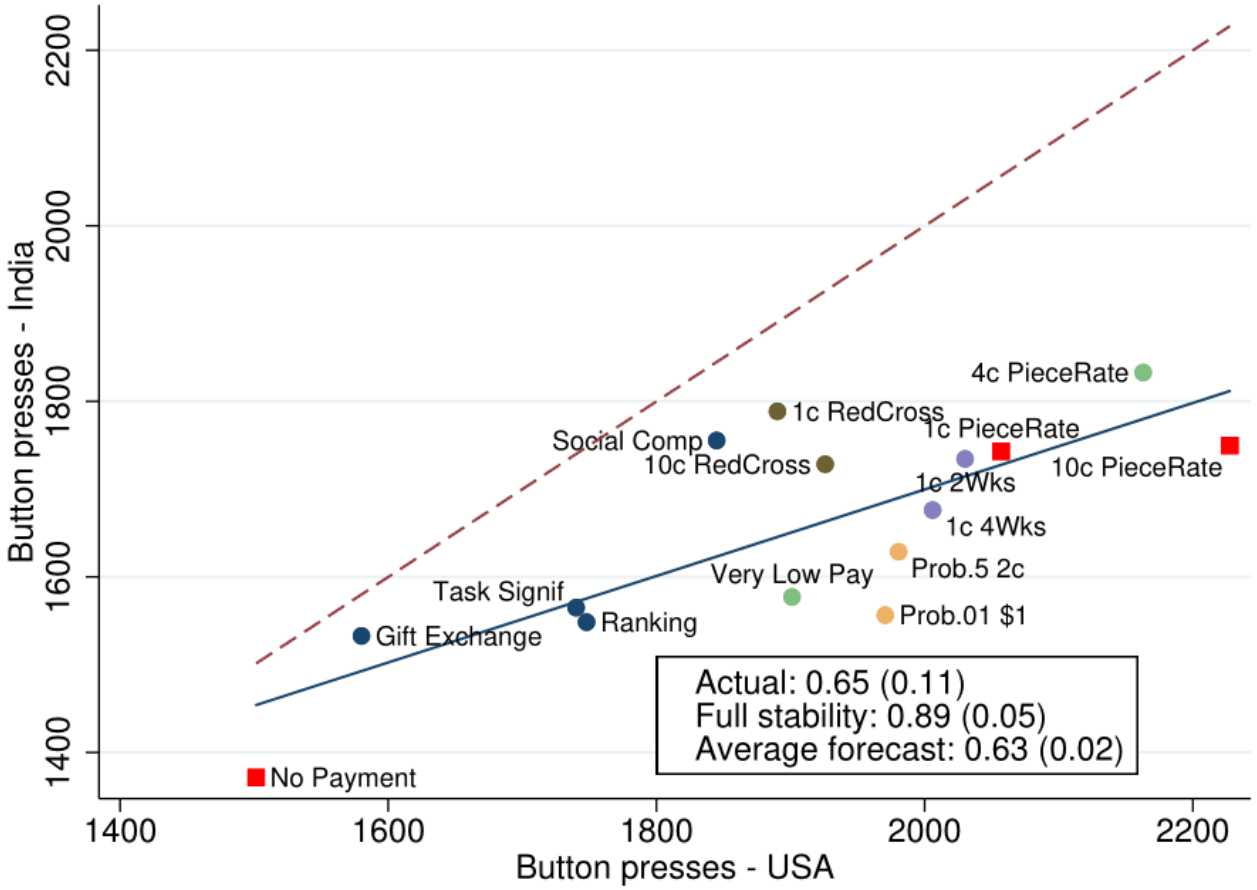


Figure 4c. Age



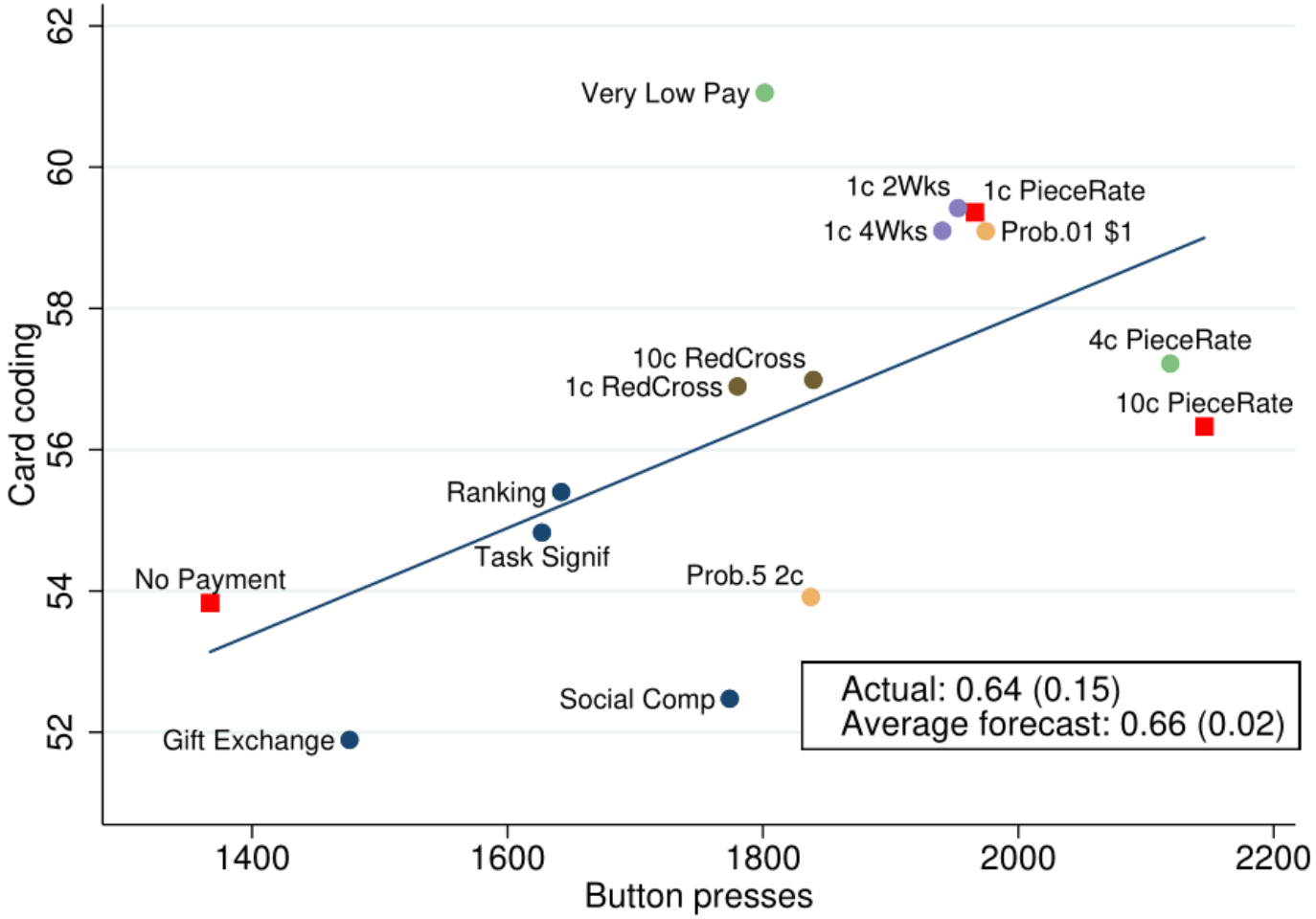
Notes: Figures4a-c display, for each one of 15 treatments, the average effort for the button pushing task (pooling the 2015 and 2018 experiments) across different demographics of the subjects, splitting by gender (Figure 4a), by education (Figure 4b), and by age (Figure 4c). See notes to Figure 3 for more detail.

Figure 5. Impact of Geography/Culture, Button Pushing Task



Notes: Figure 5 displays, for each one of 15 treatments, the average effort for the button pushing task (pooling the 2015 and 2018 experiments), splitting subjects by whether the respondents have an IP address associated with a S location (x axis) or with a location in India (y axis). See notes to Figure 3 for more detail.

Figure 6. Impact of Task, Button Pushing Task vs. WWII Card Coding Task



Notes: Figure 6 displays, for each one of 15 treatments, the average effort across two different tasks. On the x axis is the effort for the a-b typing task (pooling the 2015 and 2018 experiments), while on the y axis is the effort for the 2018 WWII 10-minute card coding task. See notes to Figure 3 for more detail.

Figure 7. Impact of Output

Figure 7a. WWII Coding, Ext. Margin vs. WWII, Int. Margin

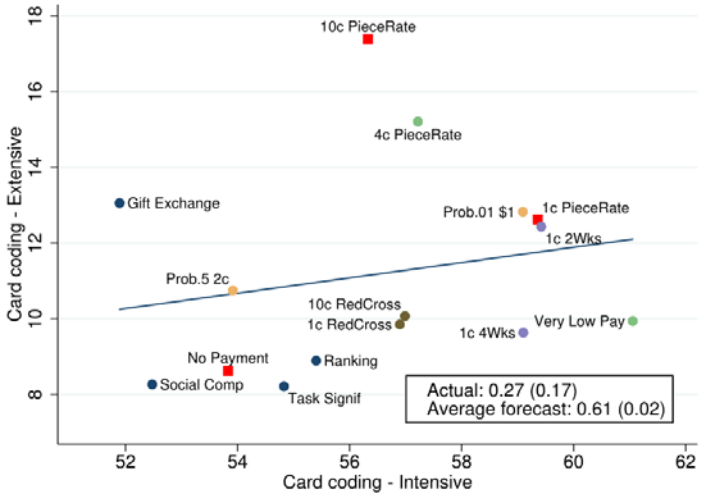


Figure 7b. WWII Coding, Ext. Margin vs. Button Pushing Task

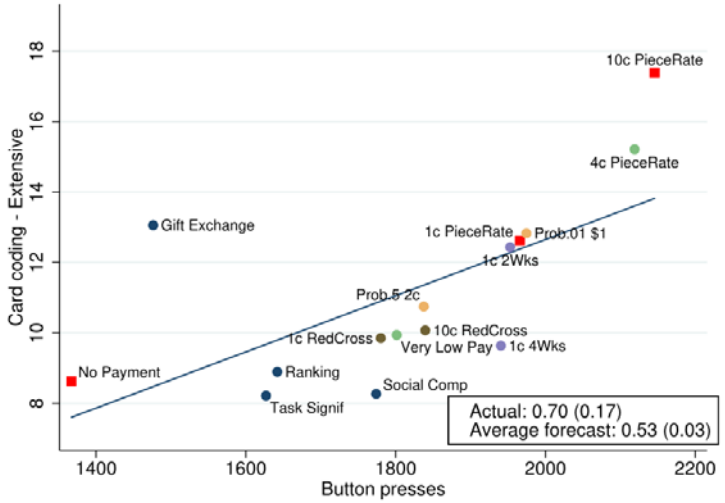
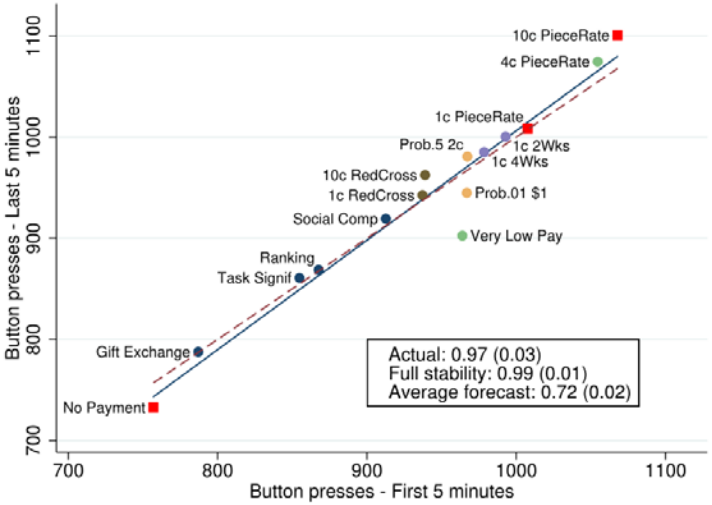
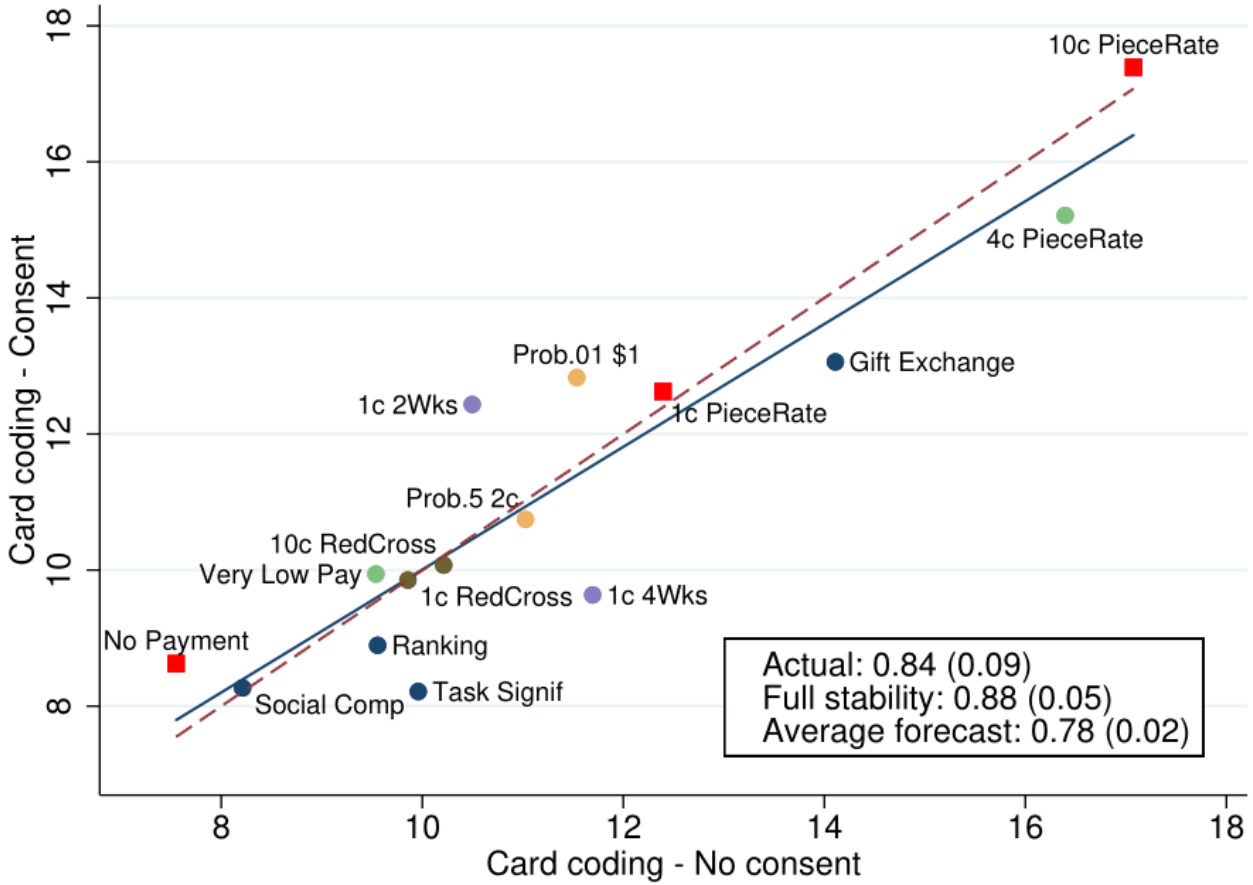


Figure 7c. Output in First 5 Minutes vs. Later 5 Minutes, Button Pushing Task



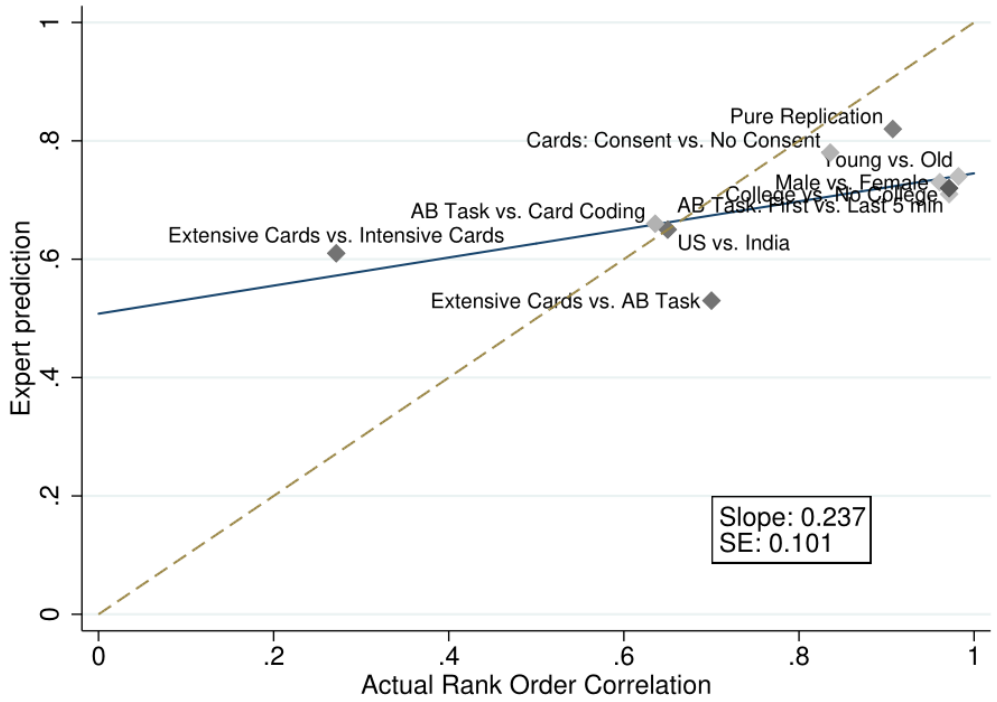
Notes: Figures 7a-c display, for each one of 15 treatments, the average effort across two different output measures. In Figure 7a we compare the cards coded in the 10-minute WWII card coding task to the extra cards coded in the extra-work WWII card task. In Figure 7b we compare the a-b points in the 10-minute button pushing taskto the extra cards coded in the extra-work WWII card task. In Figure 7c we compare, within the button pushing task (pooling 2015 and 2018), productivity in the first 5 minutes versus in the next 5 minutes. See notes to Figure 3 for more detail.

Figure 8. Impact of Consent, WWII Coding Task



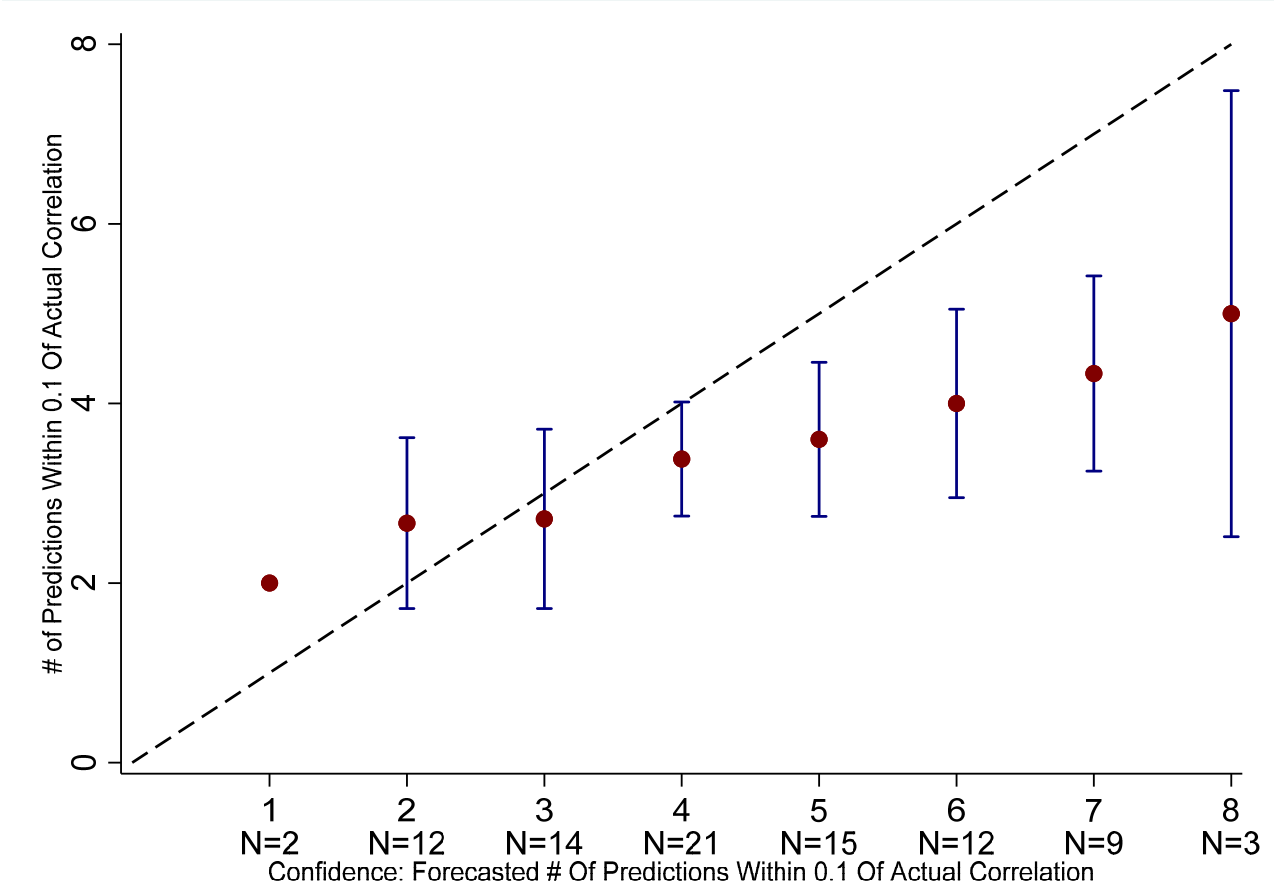
Notes: Figure 8 displays, for each one of 15 treatments, the average effort for two versions of the same extra-work WWII card coding experiment. In the version on the x axis, subjects are not displayed a consent form (and thus are presumably unaware of being part of an experiment) while in the version on the y axis, subjects are shown a consent form. See notes to Figure 3 for more detail.

Figure 9. Comparing the Expert Forecasts of Rank Correlation to the Actual Correlation, Across 10 Design Changes



Notes: Figure 9 displays, for each of 10 version changes, the actual rank-order correlation and the average expert prediction for that same rank-order correlation. For example, the Pure Replication dot indicates that the actual rank-order correlation on Pure Replication (Figure 3) is 0.91, while the average expert prediction is 0.82.

Figure 10. Confidence (in the Forecast of Rank-Order Correlation) and Accuracy



Notes: In the survey of forecasters, as last question we asked the expected number of forecasts of rank-order correlation which the forecasters expected to get within 0.1 of the correct answer. In Figure 10we plot the actual share of answers about rank-order correlation that were within 0.1 of the correct answer, splitting by the measure of confidence, that is, the forecast (rounded to the closest round number) of the number of "correct" predictions. The sample includes academic experts, as well as PhDs. The dotted line is the 45-degree line indicating an unbiased estimate.

Figure 11. How much Information is in Expert Forecasts? Revisiting the 2015 Expert Forecasts

Figure 11a. Accuracy of 2015 Forecasts vs. 2018 Forecasts

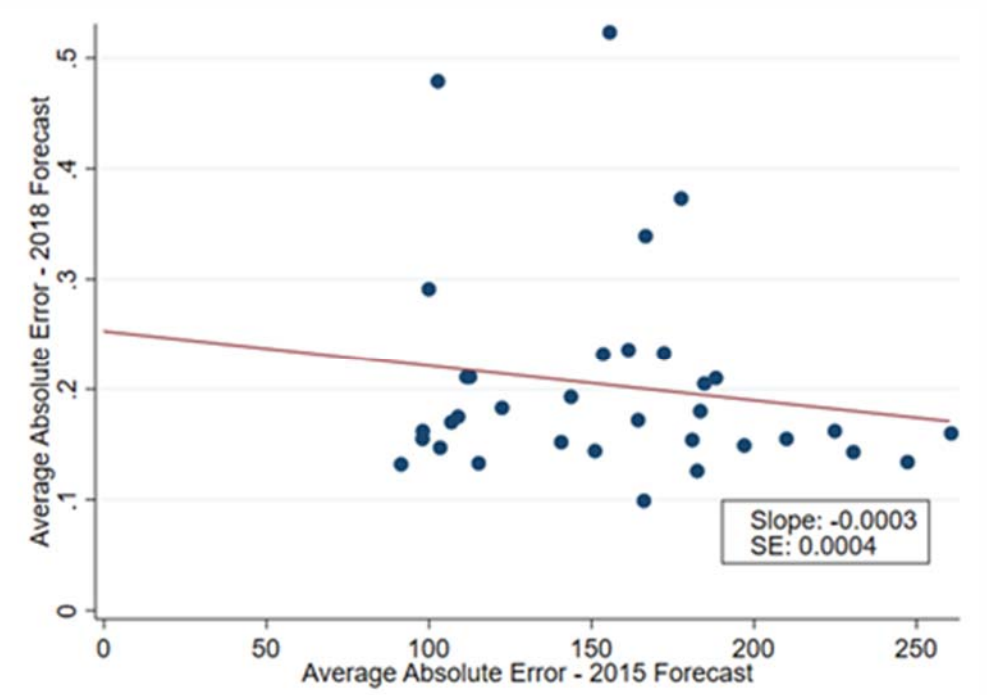
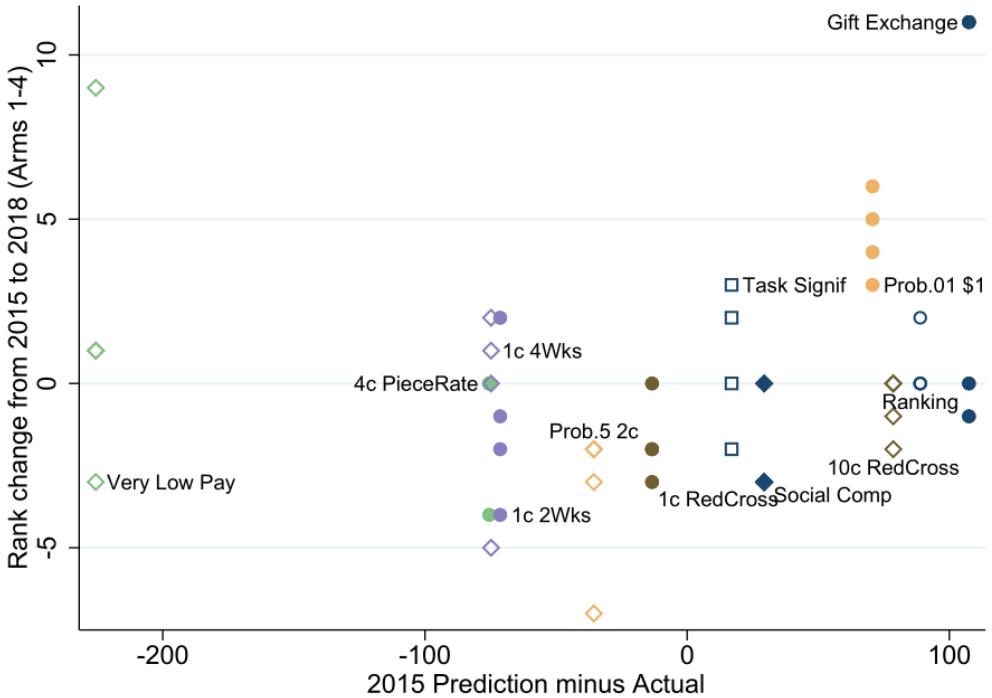


Figure 11b. Errors in 2015 Forecasts and Changes of Treatment Rank in 2018 Experiments



Notes: For the 35 individuals who made forecasts both in 2015 and in 2018, in Figure 11a we compare the accuracy of their two forecasts, displaying the average absolute error (in terms of point) in the 2015 forecasts on the x axis and the average absolute error (in terms of rank-order correlation) in the 2018 forecasts. In Figure 11b, the x axis indicates for each treatment the average forecast error in 2015, while on the y axis we plot, for each of the four 2018 new versions of the experiment, how much a treatment shifted in rank from the 2015 experiment to the 2018 experiment.

Table 1. Findings by Treatment: Effort in Different Versions

		Mean Effort (s.e.)				
	Task:	**Button Pushing, 10 Min**		**2018 WWII Cards Coding Task**		
Category	**Treatment Wording**	**2015 Exp.**	**2018 Exp.**	**10-Min**	**Extra Work**	**Extra Work, No Consent**
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Piece Rate	"Your score [The number of [additional] cards you complete] will not affect your payment in any way."	1521 (31)	1367 (60)	53.83 (1.84)	8.63 (0.75)	7.55 (0.78)
	"As a bonus, you will be paid an extra **1 cent** for every **100 points** that you score [**2** [**additional**] **cards** that you complete]"	2029 (27)	1966 (53)	59.36 (1.81)	12.63 (0.79)	12.39 (0.73)
	"As a bonus, you will be paid an extra **4 cents** for every **100 points** that you score [**2 cents** for every [**additional**] **card** that you complete]."	2132 (26)	2119 (45)	57.22 (1.93)	15.21 (0.69)	16.40 (0.60)
	"As a bonus, you will be paid an extra **10 cents** for every **100 points** that you score [**5 cents** for every [**additional**] **card** that you complete]."	2175 (24)	2146 (50)	56.33 (1.97)	17.39 (0.50)	17.08 (0.55)
Pay Enough or Don't Pay	"As a bonus, you will be paid an extra 1 cent for every **1,000 points** that you score [**20** [**additional**] **cards** you complete]."	1883 (29)	1801 (60)	61.05 (1.87)	9.94 (0.78)	9.54 (0.81)
Social Preferences: Charity	"As a bonus, the Red Cross charitable fund will be given **1 cent** for every 100 points that you score [2 [additional] cards you complete]."	1907 (27)	1780 (50)	56.90 (1.80)	9.85 (0.84)	9.86 (0.71)
	"As a bonus, the Red Cross charitable fund will be given **10 cents** for every 100 points that you score [5 cents for every [additional] card you complete]."	1918 (26)	1839 (51)	56.99 (2.00)	10.07 (0.81)	10.21 (0.73)
Social Preferences: Gift Exchange	"In appreciation to you for performing this task, you will be paid a **bonus** of **40 cents**. Your score will not affect your payment in any way [The number of cards you complete will not affect your payment in any way / You will receive this bonus even if you choose not to complete any additional cards]."	1602 (30)	1476 (54)	51.89 (1.76)	13.06 (0.73)	14.11 (0.70)
Discounting	"As a bonus, you will be paid an extra 1 cent for every 100 points that you score [every 2 [additional] cards you complete]. This bonus will be paid to your account **two weeks** from today."	2004 (27)	1953 (48)	59.42 (2.01)	12.44 (0.77)	10.50 (0.80)
	"As a bonus, you will be paid an extra 1 cent for every 100 points that you score [every 2 [additional] cards you complete]. This bonus will be paid to your account **four weeks** from today."	1970 (29)	1940 (53)	59.10 (1.83)	9.64 (0.76)	11.70 (0.82)
Risk Aversion and Probability Weighting	"As a bonus, you will have a **1% chance** of being paid an **extra $1** for every 100 points that you score [**extra 50 cents** for every [additional] card you complete]."	1896 (28)	1975 (47)	59.09 (1.68)	12.83 (0.76)	11.54 (0.79)
	"As a bonus, you will have a **50% chance** of being paid an **extra 2 cents** for every 100 points that you score [**extra 1 cents** for every [additional] card you complete]."	1977 (25)	1837 (51)	53.92 (1.95)	10.75 (0.80)	11.03 (0.78)
Social Comparisons	"Your score [The number of [additional] cards you complete] will not affect your payment in any way. In a previous version of this task, **many participants** [**workers**] were able to score more than **2,000 points** [completed more than **70 cards** [the additional cards]]."	1848 (32)	1774 (54)	52.48 (1.90)	8.27 (0.79)	8.21 (0.75)
Ranking	"Your score [The number of [additional] cards you complete] will not affect your payment in any way. After you play [finish], we will show you **how well you did** [**how many** [**additional**] **cards you completed**] **relative** to other participants [workers] who have previously done this task."	1761 (31)	1642 (56)	55.40 (1.70)	8.90 (0.78)	9.56 (0.77)
Task Significance	"Your score [The number of [additional] cards you complete] will not affect your payment in any way [, but your work is very valuable for us, and we would really appreciate your help]. We are interested in how fast people choose to press digits and we would like you to do your very best. So **please try as hard** [**do as many**] **as you can**."	1740 (29)	1627 (58)	54.83 (1.83)	8.22 (0.77)	9.96 (0.77)
Piece Rate + Task Significance	"We are interested in how fast people choose to press digits and we would like you to do your very best [Your work is very valuable for us, and we would really appreciate your help]. So **please try as hard** [do as many [additional] cards] as you can. As a bonus, you will be paid an **extra 1 cent** for every 100 points that you score [2 [additional] cards you complete]."	-	2056 (46)	56.18 (1.76)	10.81 (0.79)	13.3 (0.74)
	Number of Observations	8,252	2,380	2,708	2,331	2,392

Notes: The Table lists the 16 treatments in the Mturk experiment; the main analysis focuses on the first 15 treatments which are run in all experiments. Column 1 reports the conceptual grouping of the treaments and Column 2 reports the exact wording that distinguishes the treatments. The treatments differ just in one paragraph explaining the task and in the vizualization of the points earned. Column (2) reports the key part of the wording of the paragraph. For brevity, we omit from the description the sentence "This bonus will be paid to your account within 24 hours" which applies to all treatments with incentives other than in the Time Preference ones where the payment is delayed. Notice that the bolding is for the benefit of the reader of the Table. In the actual description to the MTurk workers, the whole paragraph was bolded and underlined. The main wording applies to the Button Pushing task (Columns 3 and 4), which we run in 2015 (Column 3) and replicate in 2018 (Column 4). The wording in brackets applies to the experiments on WWII card coding, in Columns 5-7. Columns 3-7 report the mean output and the standard error of the output in each treatment.

Table 2. Stability Across Designs: Rank-Order Correlations, Forecasts vs. Actual

Category	Design Comparison	Rank-Ord. Correl.	Average Forecast of Rank-Order Correlation			Rank-Ord. Correl.	p-value for Difference		
		Full Stability w/ Noise	Faculty Experts	PhD Students	Mturkers	Actual	Experts vs. Full Stability	Actual vs. Full Stability	Actual vs. Experts
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Pure Repl.	2015 AB Task vs. 2018 AB Task (n=8,252; n=2,219)	0.94 (0.04)	0.82 (0.01)	0.87 (0.01)	0.75 (0.02)	0.91 (0.04)	0.004	0.630	0.068
Demogr., Typing Task	Male vs. Female (n=4,686; n=5,785)	0.95 (0.03)	0.73 (0.02)	0.77 (0.02)	0.73 (0.02)	0.96 (0.04)	0.000	0.856	0.000
	College vs. No College (n=5,842; n=4,629)	0.95 (0.03)	0.71 (0.02)	0.74 (0.02)	0.67 (0.02)	0.97 (0.04)	0.000	0.691	0.000
	Young (=<30) vs. Old (30+) (n=5,259; n=5,212)	0.95 (0.03)	0.74 (0.02)	0.76 (0.02)	0.66 (0.02)	0.98 (0.04)	0.000	0.527	0.000
Geogr./ Culture	US vs. India (n=8,803; n=1,225)	0.89 (0.05)	0.63 (0.02)	0.67 (0.03)	0.68 (0.02)	0.65 (0.11)	0.000	0.049	0.897
Task	AB Task vs. 10-min Card Coding (n=2,219; n=2,537)	-	0.66 (0.02)	0.63 (0.03)	0.64 (0.02)	0.64 (0.15)	-	-	0.866
Output	10-min Cards vs. Extra Cards (n=2,537; n=2,188)	-	0.61 (0.02)	0.61 (0.03)	0.62 (0.02)	0.27 (0.17)	-	-	0.052
	Extra Cards vs. AB Task (n=2,188; n=2,219)	-	0.53 (0.03)	0.56 (0.04)	0.58 (0.02)	0.70 (0.17)	-	-	0.046
	AB Task: First 5 min vs. Last 5 min (n=10,471)	0.99 (0.01)	0.72 (0.02)	0.70 (0.03)	0.64 (0.02)	0.97 (0.03)	0.000	0.543	0.000
Consent	Cards: Consent vs. No Consent (n=2,188; n=2,246)	0.88 (0.05)	0.78 (0.02)	0.81 (0.02)	0.70 (0.02)	0.84 (0.09)	0.067	0.645	0.552
N			N=55	N=33	N=109				
Average Individual Abs. Error			0.20 (0.01)	0.19 (0.01)	0.24 (0.01)				
Wisdom of Crowd Error			0.17 (0.03)	0.15 (0.04)	0.20 (0.04)				
Average Forecast of No. Rank-o. Corr w/in 0.1 of Truth			3.99 (0.24)	4.95 (0.25)	4.66 (0.22)				
Average Actual No. Rank-o. Corr w/in 0.1 of Truth			3.35 (0.24)	3.55 (0.23)	3.02 (0.16)				

Notes: The Table lists the 10 design changes to the experiment which sonstitute the focus of the paper. For example, in row 1 we compare the estimate of effort in the 15 treatments in the button pushing task, comparing the results for male subjects versus for female subjects. To compare the stability of results across versions, we compute the rank-order correlation of the average effort in the 15 treatments across versions. In Column 1 we report the average correlation under a benchmark of full-stability, that is, if the results do not change with the change in design. This correlation, which is the average over a series of bootstraps, is lower than 1 due to measurement error. Columns 2-4 report the average forecast of rank-order correlation for the population of academic experts (Column 2), PhD students (Column 3), and MTurkers (Column 4). Column 5 reports the actual rank-order correlation. Columns 6 -8 report the p-value for the difference between the relevant columns.

Table 3. Stability Across Designs, Additional Comparisons

Rank-Order Correlations Across Designs

Category	Version Comparison	Full Stability w/ Noise	Actual	p-value for Difference
		(1)	(2)	(3)
Demographics, 10-minute WWII Coding Task	Male vs. Female (n=1,014; n=1,523)	0.43 (0.18)	0.27 (0.22)	0.573
	College vs. No College (n=1,478; n=1,059)	0.44 (0.18)	0.38 (0.21)	0.845
	Young vs. Old (n=1,128; n=1,409)	0.43 (0.17)	0.31 (0.21)	0.680
Geography/ Culture, Extra-Cards WWII Coding	US vs. India (n=3,668; n=492)	0.76 (0.11)	0.65 (0.10)	0.479
Geography/ Culture, AB Typing Task	Red States vs. Blue States (n=5,062; n=3,464)	0.94 (0.03)	0.96 (0.04)	0.748
Other Selection, AB Task	Enrollment in Week 1 vs. Weeks 2-3 (n=6,359; n=4,112)	0.95 (0.03)	0.95 (0.04)	0.947
	Night vs. Day (n=4,556; n=5,195)	0.94 (0.03)	0.97 (0.04)	0.624
Other Selection, 10-minute WWII Coding Task	Enrollment in Week 1 vs. Weeks 2-3 (n=1,569; n=968)	0.43 (0.18)	0.58 (0.21)	0.570
	Night vs. Day (n=949; n=1,338)	0.37 (0.18)	-0.05 (0.22)	0.138
Other Selection, WWII Coding Extra Cards	Enrollment in Week 1 vs. Weeks 2-3 (n=2,641; n=1,793)	0.88 (0.06)	0.83 (0.10)	0.634
	Night vs. Day (n=1,600; n=2,428)	0.88 (0.06)	0.82 (0.09)	0.564

Notes: The Table lists additional design changes which we did not present to the forecasters. In Column (1) we report the results under a full-stability benchmark (see notes to Table 2) and in Column 2 we present the actual rank-order correlation.

Table 4. Structural Estimates

Exponential cost of Effort Function

		Button Pushing Task, 10 Min		Demographics, Typing Task, Pooled 2015-2018						2018 WWII Cards Coding Task		
Category	**Parameters**	**2015 Exp.**	**2018 Exp.**	**Male**	**Female**	**College**	**No College**	**Young (=<30)**	**Old (30+)**	**10-Min**	**Extra Work**	**Extra Work, No Consent**
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Incidental Parameters	Curvature of Cost of Effort γ	0.015 (0.004)	0.013 (0.005)	0.012 (0.003)	0.019 (0.007)	0.015 (0.005)	0.014 (0.004)	0.011 (0.003)	0.022 (0.009)	1.909 (4.180)	0.047 (0.015)	0.057 (0.014)
	Implied Elasticity	0.034	0.04	0.043	0.028	0.036	0.037	0.046	0.025	0.01	0.42	0.34
	Level of Cost of Effort *k*	2.39E-16 (1.93E-15)	1.46E-13 (1.53E-12)	1.34E-13 (9.74E-13)	4.43E-19 (5.82E-18)	5.10E-16 (4.90E-15)	4.37E-15 (3.73E-14)	8.63E-13 (4.83E-12)	2.34E-21 (4.43E-20)	5.11E-50 (9.37E-31)	0.028 (0.041)	0.014 (0.015)
	Baseline Motivation *s*	3.87E-4 (9.03E-4)	4.77E-4 (1.75E-3)	4.25E-4 (0.001)	2.47E-04 (8.21E-04)	1.363E-4 (4.275E-4)	0.002 (0.004)	0.002 (0.004)	1.90E-5 (1.00E-04)	1.79E-5 (4.52E-4)	0.203 (0.191)	0.097 (0.083)
Pay Enough or Don't Pay	Δs_{CO}	0.003 (0.104)	0.009 (0.177)	-0.049 (0.067)	0.121 (0.259)	-0.013 (0.112)	0.028 (0.142)	0.158 (0.199)	-0.080 (0.047)	-0.009 (0.370)	0.068 (0.102	0.052 (0.076)
Social Pref. Parameters	Pure Altruism *alpha* (1 is full altruism)	0.003 (0.010)	0.010 (0.017)	0.009 (0.013)	0.001 (0.009)	0.002 (0.110)	0.010 (0.014)	0.004 (0.015)	0.006 (0.011)	0.001 (0.014)	0.011 (0.028)	0.007 (0.020)
	Warm Glow *a*	0.138 (0.134)	0.070 (0.128)	0.112 (0.128)	0.094 (0.133)	0.131 (0.154)	0.102 (0.126)	0.224 (0.180)	0.034 (0.077)	0.014 (0.621)	0.205 (0.220)	0.267 (0.180)
Social Pref.: Gift Exchange	Δs_{GE}	2.39E-5 (4.74E-5)	3.56E-5 (0.0001)	2.99E-5 (6.45E-5)	1.43E-5 (4.06E-5)	1.38E-5 (3.56E-5)	5.87E-5 (1.12E-4)	0.0001 (0.0002)	1.34E-06 (6.14E-06)	-1.24E-5 (3.45E-4)	2.073 (0.865)	2.271 (0.884)
Discounting	*Beta*	1.163 (1.194)	0.828 (1.292)	0.805 (0.967)	1.516 (1.970)	1.242 (1.519)	0.729 (0.884)	3.412 (3.203)	0.254 (0.498)	1.283 (11.431)	5.407 (6.018)	0.215 (0.225)
	Delta (Weekly)	0.757 (0.237)	0.930 (0.425)	0.771 (0.276)	0.810 (0.323)	0.665 (0.258)	1.048 (0.363)	0.651 (0.201)	0.992 (0.493)	0.690 (1.875)	0.435 (0.200)	1.318 (0.378)
Social Comparisons	Δs_{SC}	0.060 (0.070)	0.076 (0.130)	0.069 (0.086)	.044 (0.072)	0.036 (0.055)	0.137 (0.153)	0.194 (0.162)	0.009 (0.025)	-3.73e-06 (1.184E-4)	-0.018 (0.07)	0.024 (0.044)
Ranking	Δs_{R}	0.015 (0.021)	0.015 (0.033)	0.015 (0.025)	0.010 (0.020)	0.010 (0.019)	0.021 (0.032)	0.053 (0.056)	0.001 (0.005)	5.51E-6 (1.59E-4)	0.001 (0.070)	0.103 (0.074)
Task Significance	Δs_{TS}	0.011 (0.016)	0.011 (0.026)	0.015 (0.024)	0.005 (0.012)	0.004 (0.009)	0.037 (0.052)	0.040 (0.045)	0.001 (0.003)	3.33E-6 (1.23E-4)	-0.007 (0.070)	0.143 (0.091)
No. of Obs.		7,129	1,925	4,061	4,993	5,051	4,003	4,535	4,519	2,200	1,895	1,950
Avg effort		1,886	1,807	1,918	1,829	1,813	1,940	1,944	1,794	56.53	11.10	11.31
Root MSE		664.29	651.85	730.53	596.67	670.53	645.21	692.40	622.22	24.36	53.62	49.60
Incentive + Please try	Out-of-Sample Pred.	-	1978	2164	1952	1982	2069	2076	1959	59.09	12.62	13.00
	Actual	-	2056	2065	2049	2011	2106	2178	1910	56.18	10.81	13.30

Notes: The Table shows structural estimates of the incidental parameters (γ, k, and s) and psychological parameters estimated using all 15 treatments across 11 different samples. All models assume an exponential cost function. Cols (1)-(9) are estimated using nonlinear least squares using the individual effort of MTurkers (rounded to the nearest 100). Cols (10)-(11) are estimated with maximum likelihood due to censoring. Col (1) shows estimates using all 2015 typing task conditions, Col (2) shows the estimates using all 2018 typing task conditions. Cols (2)-(8) pool the typing task conditions from both years and estimate parameters restricting to a demographic subset. Cols (9)-(11) show estimates on the 2018 card coding treatments. Standard errors in parantheses.

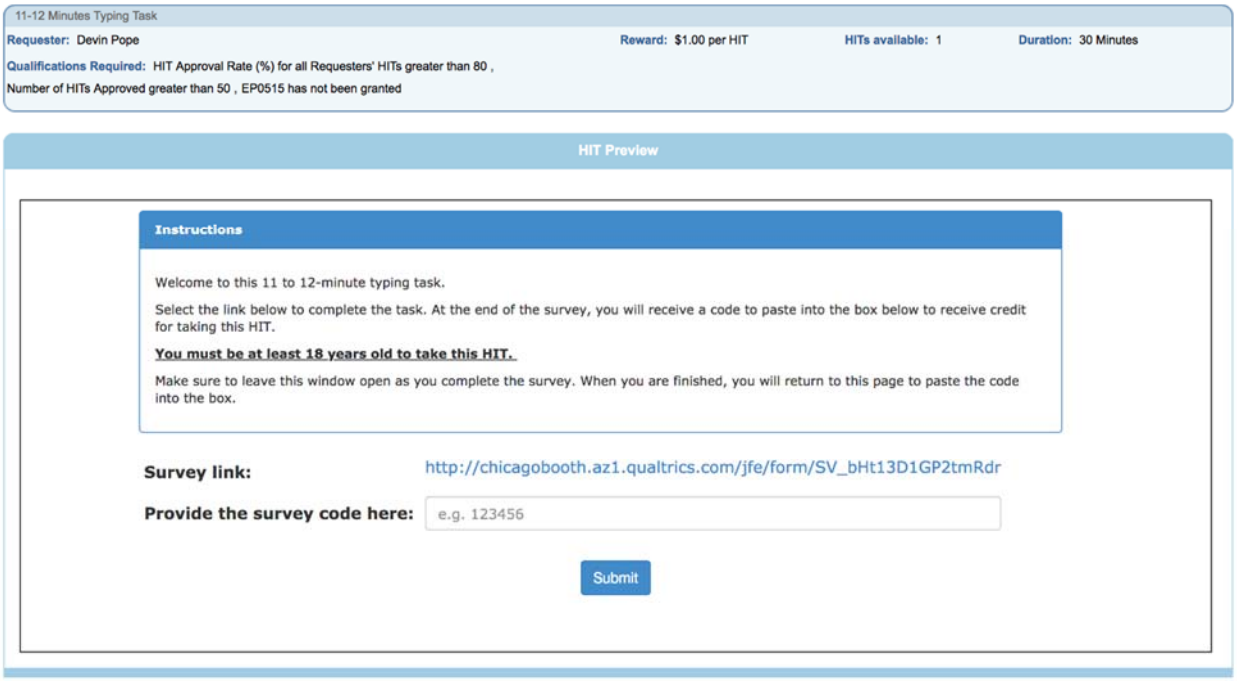
Table 5. Forecasts of Rank-Order Correlations by Different Forecasters

		Average Forecast of Rank-Order Correlation for the 15 Treatments Across Designs								
			Version		Clicked a Link		Time Spent on Survey		Confidence	
Category	**Version Comparison**	**Pooled Experts and PhDs**	**Info on Piece Rate**	**No Info on Piece Rate**	**Yes**	**No**	**Long (18 mins+)**	**Short (<18 mins)**	**High (4+ corr. w/in 0.1)**	**Low (<4 corr. w/in 0.1)**
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Pure Repl.	2015 AB Task vs. 2018 AB Task	0.84 (0.01)	0.84 (0.01)	0.83 (0.02)	0.86 (0.01)	0.82 (0.02)	0.84 (0.01)	0.83 (0.02)	0.86 (0.01)	0.81 (0.02)
Demogr., Typing Task	Male vs. Female	0.75 (0.01)	0.76 (0.02)	0.74 (0.02)	0.76 (0.02)	0.72 (0.02)	0.76 (0.02)	0.73 (0.02)	0.77 (0.02)	0.71 (0.03)
	College vs. No College	0.72 (0.01)	0.72 (0.02)	0.73 (0.02)	0.75 (0.01)	0.69 (0.02)	0.75 (0.02)	0.70 (0.02)	0.76 (0.02)	0.67 (0.03)
	Young (=<30) vs. Old (30+)	0.74 (0.01)	0.76 (0.02)	0.73 (0.02)	0.77 (0.02)	0.72 (0.02)	0.77 (0.02)	0.72 (0.02)	0.77 (0.02)	0.71 (0.03)
Geogr. / Culture	US vs. India	0.65 (0.02)	0.65 (0.02)	0.65 (0.03)	0.65 (0.02)	0.65 (0.03)	0.67 (0.02)	0.62 (0.03)	0.69 (0.02)	0.60 (0.03)
Task	AB Task vs. Card Coding	0.65 (0.02)	0.62 (0.03)	0.69 (0.03)	0.64 (0.03)	0.66 (0.03)	0.66 (0.03)	0.64 (0.03)	0.67 (0.02)	0.49 (0.04)
Output	10-min Cards vs. Extra Cards	0.61 (0.02)	0.60 (0.03)	0.63 (0.03)	0.60 (0.03)	0.64 (0.03)	0.62 (0.03)	0.61 (0.02)	0.65 (0.02)	0.66 (0.04)
	Extra Cards vs. AB Task	0.54 (0.02)	0.54 (0.03)	0.54 (0.03)	0.53 (0.03)	0.55 (0.03)	0.57 (0.03)	0.51 (0.03)	0.60 (0.02)	0.65 (0.03)
	AB Task: First 5 min vs. Last 5 min	0.71 (0.02)	0.72 (0.02)	0.70 (0.03)	0.70 (0.02)	0.73 (0.03)	0.71 (0.03)	0.71 (0.03)	0.69 (0.02)	0.58 (0.03)
Consent	Cards: Consent vs. No Consent	0.79 (0.01)	0.81 (0.02)	0.78 (0.02)	0.78 (0.02)	0.80 (0.02)	0.80 (0.02)	0.79 (0.02)	0.81 (0.01)	0.76 (0.03)
N		N=88	N=48	N=40	N=45	N=43	N=44	N=44	N=54	N=34
Average Ind. Abs. Error		0.19 (0.01)	0.19 (0.01)	0.20 (0.01)	0.19 (0.01)	0.20 (0.01)	0.18 (0.01)	0.20 (0.01)	0.18 (0.01)	0.22 (0.01)
Wisdom-of-Crowd Error		0.16 (0.04)	0.15 (0.03)	0.17 (0.04)	0.15 (0.03)	0.17 (0.04)	0.15 (0.03)	0.17 (0.04)	0.15 (0.03)	0.20 (0.03)

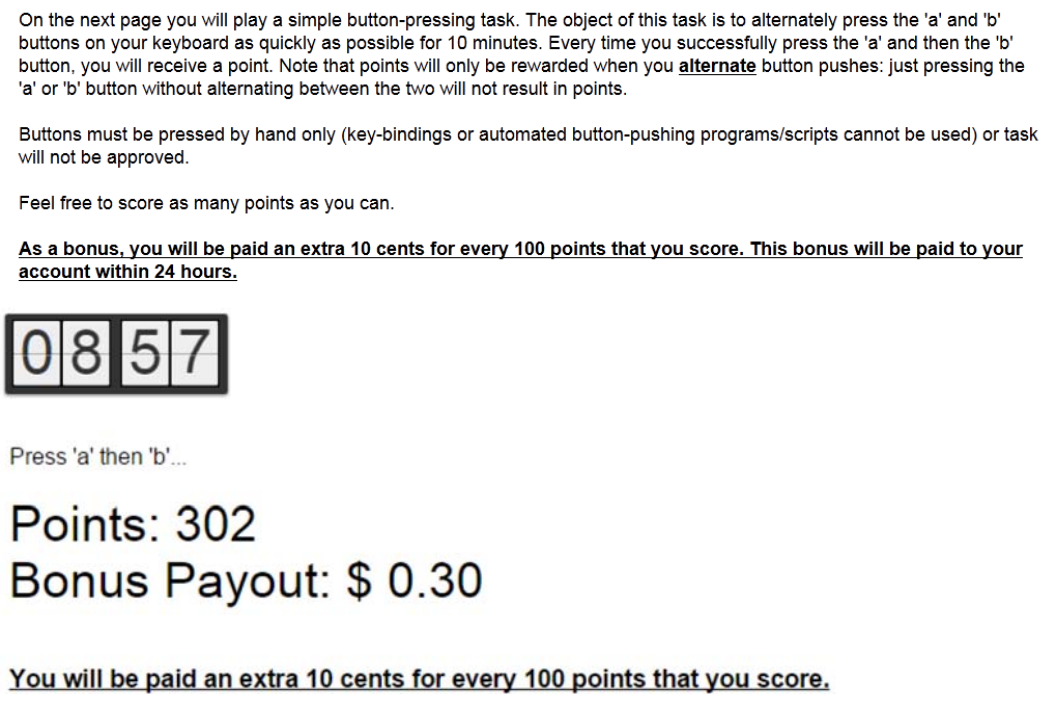
Notes: The Table considers the forecasts of sub-groups. Column 1 presents the results for the overall group of academic experts ad PhDs. In Columns 2 and 3 we split this group depending on whether the respondents were randomized to be provided information on the aerage effort by piece rate or not. In Columns 4 and 5 we split by whether the subjects clicked on at least one link for additional information. In Columns 6 and 7 we split by the time taken to complete the survey. In Columns 8 and 9 we split by the expressed degree of confidence in the forecast.

Online Appendix Figures 1a-e. MTurk Task, Examples of Screenshots

Online Appendix Figure 1a. Recruitment Ad on MTurk



Online Appendix Figure 1b. Screenshot for Button Pushing Task, Example



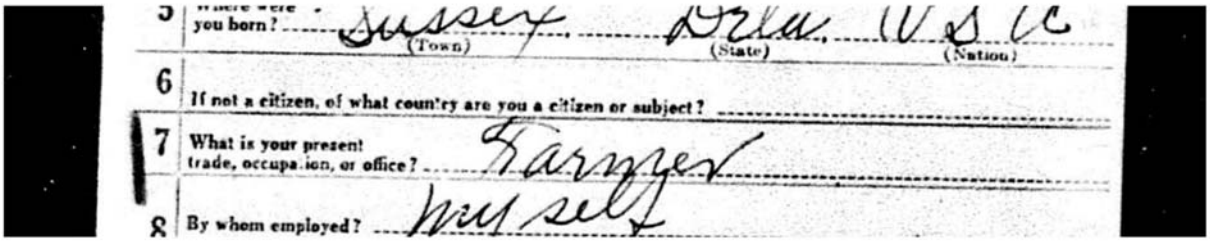
Online Appendix Figure 1c. Screenshot for WWII 10-minute Card Coding Task, Example

Time remaining: 9 Minutes, 55 Seconds

You have completed 4 cards.

Your current bonus is $0.02.

Please type the occupation in field 7 in the text box below.



You will be paid an extra 1 cent for every 2 cards you complete. This bonus will be paid to your account two weeks from today.

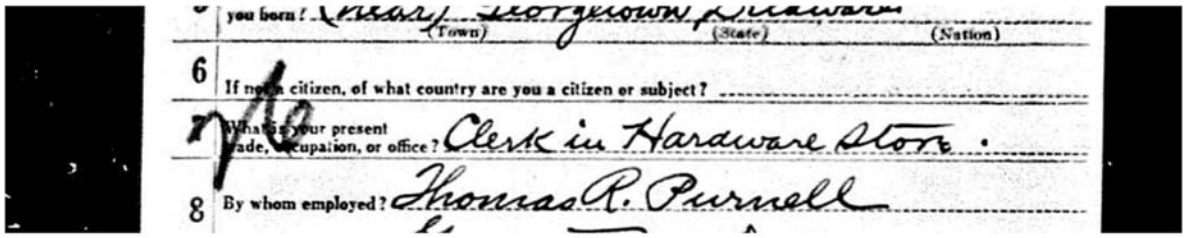
Type occupation here:

→

Online Appendix Figure 1d. Screenshot for Extra-Cards WWII Coding Task, Example I

You have completed 3 of 40 required cards.

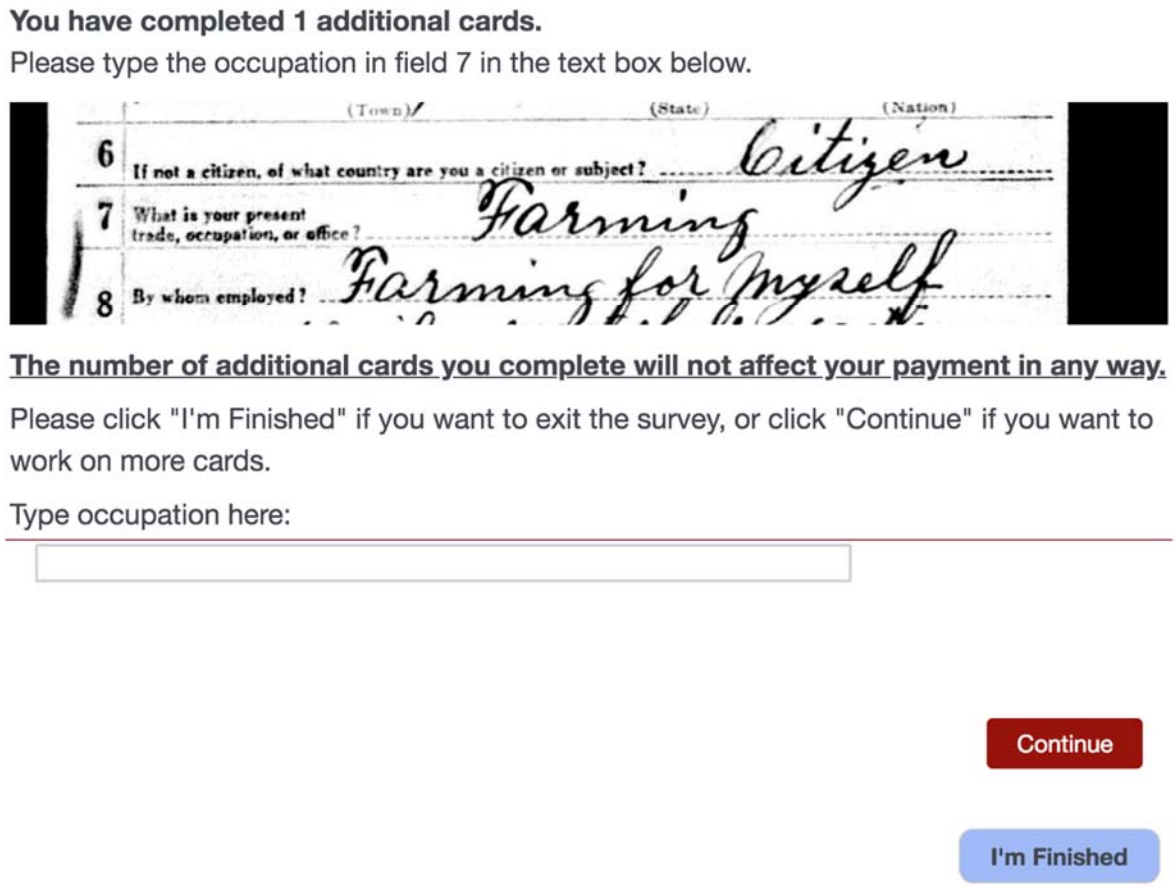
Please type the occupation in field 7 in the text box below.



Type occupation here:

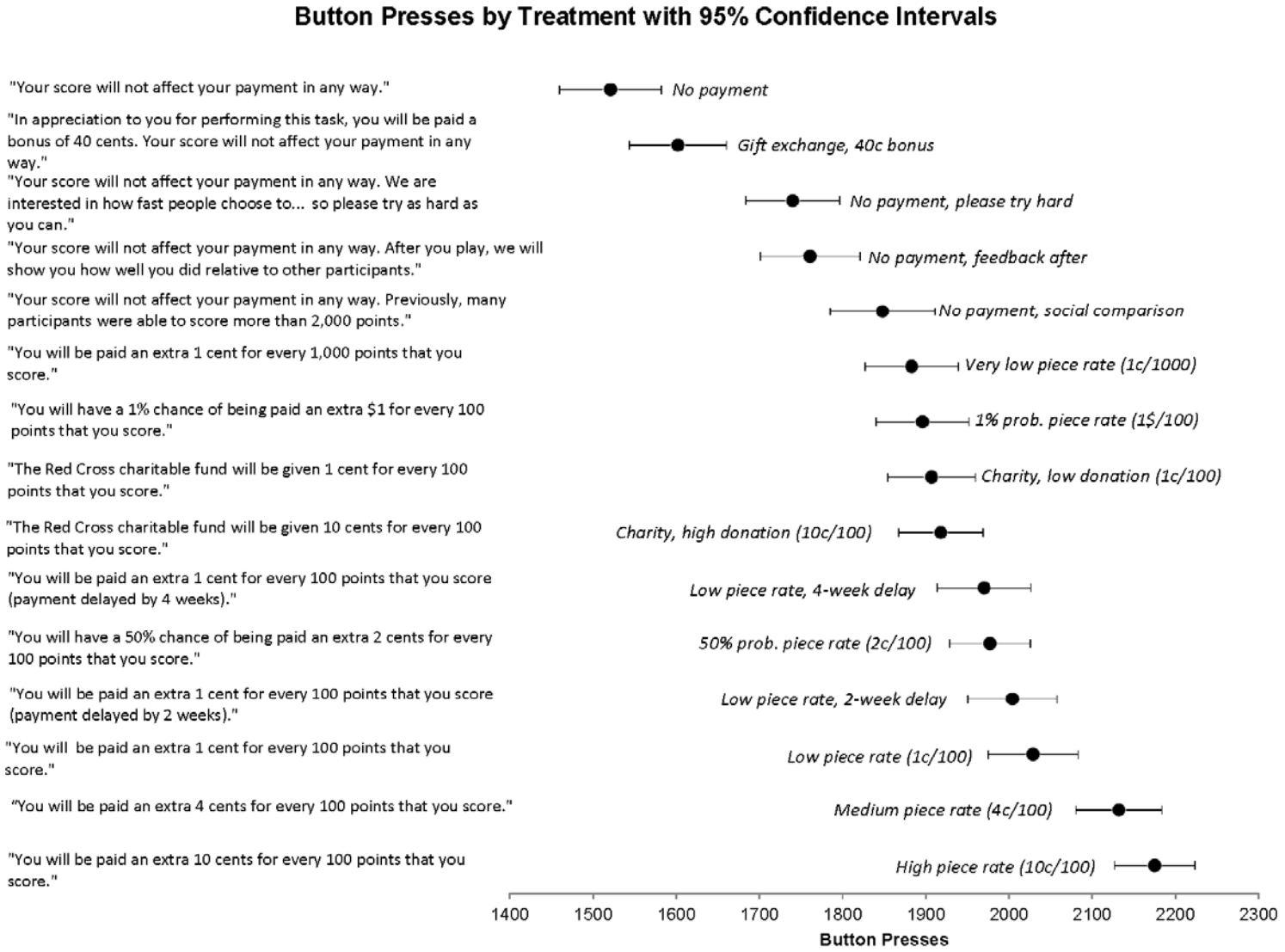
→

Online Appendix Figure 1e. Screenshot for Extra-Cards WWII Coding Task, Example II



Notes: Online Appendix Figures 1a-e plot excerpts of the MTurk real-effort task. Figure 1a displays the advertising for the task on MTurk, whereas the next figures display the key screen for the different experimental designs run in the 2018 experiment.

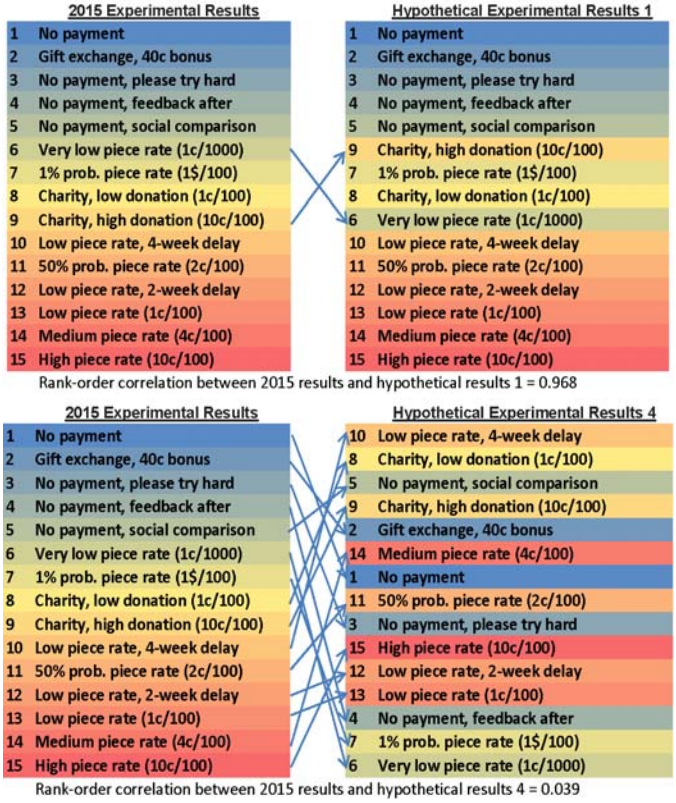
Online Appendix Figure 2. Summary of Treatments and Results from DellaVigna and Pope (2018)



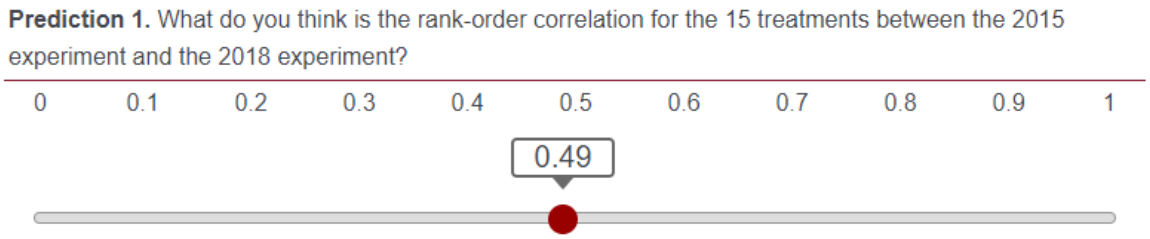
Notes: The figure summarizes the key wording as well as the average effort and standard error for the mean effort in the 2015 experimental results of DellaVigna and Pope (2018) for the 15 treatments which we replicate. This image is as presented to the forecasters.

Online Appendix Figure 3. Expert Survey, Screenshots

Online Appendix Figure 3a. Examples of Rank-order Correlation



Online Appendix Figure 3b. Example of Slider for Expert Forecast



Notes: The figure shows two screenshots reproducing portions of the Qualtrics survey eliciting forecasts. The first screenshot reproduces two of the examples of rank-order correlation as treatments change effectiveness across two versions. The second screenshot shows one of the 10 sliders that the forecasters used to make forecasts.

Online Appendix Figure 4. Distribution of Effort Across All Treatments

Online Appendix Figure 4. 2015 MTurk Button Pushing Task

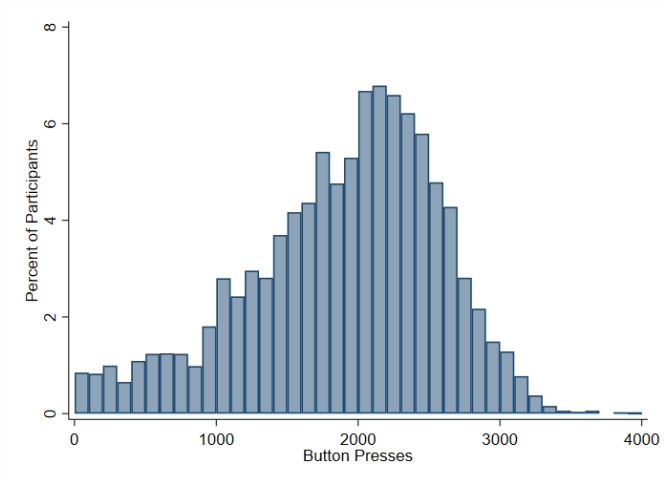


Figure 4b. 2018 MTurk Button Pushing Task

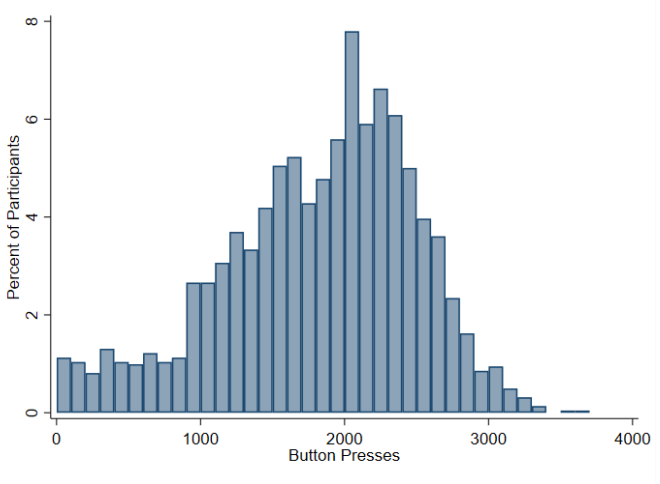


Figure 4c. 2018 10-Minute Card Coding Task

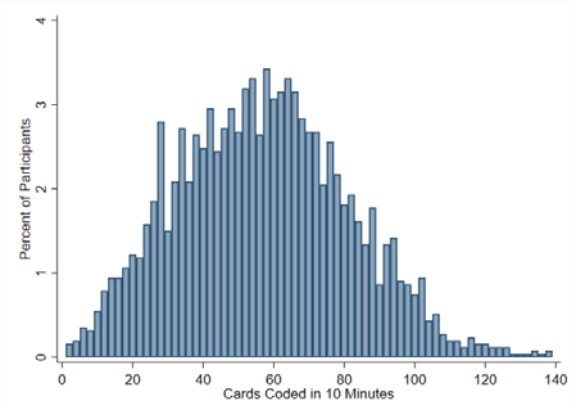


Figure 4d. 2018 Extra Card Coding Task

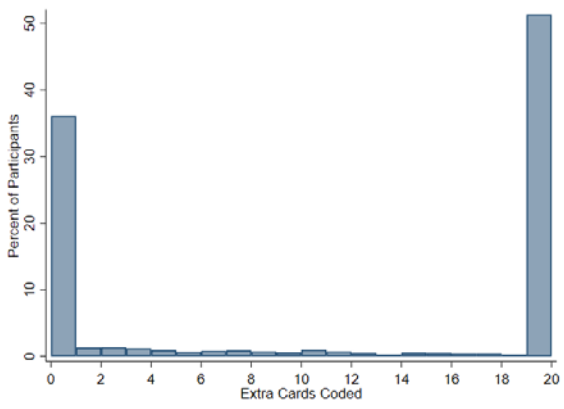
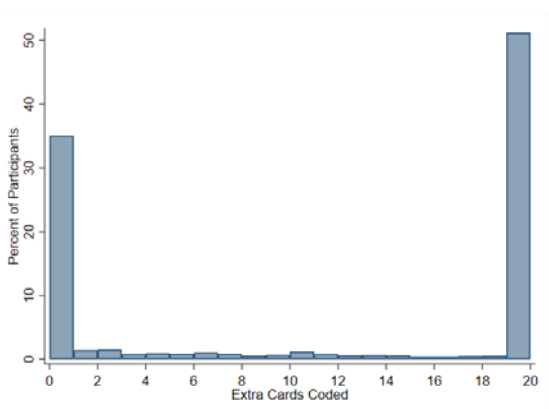
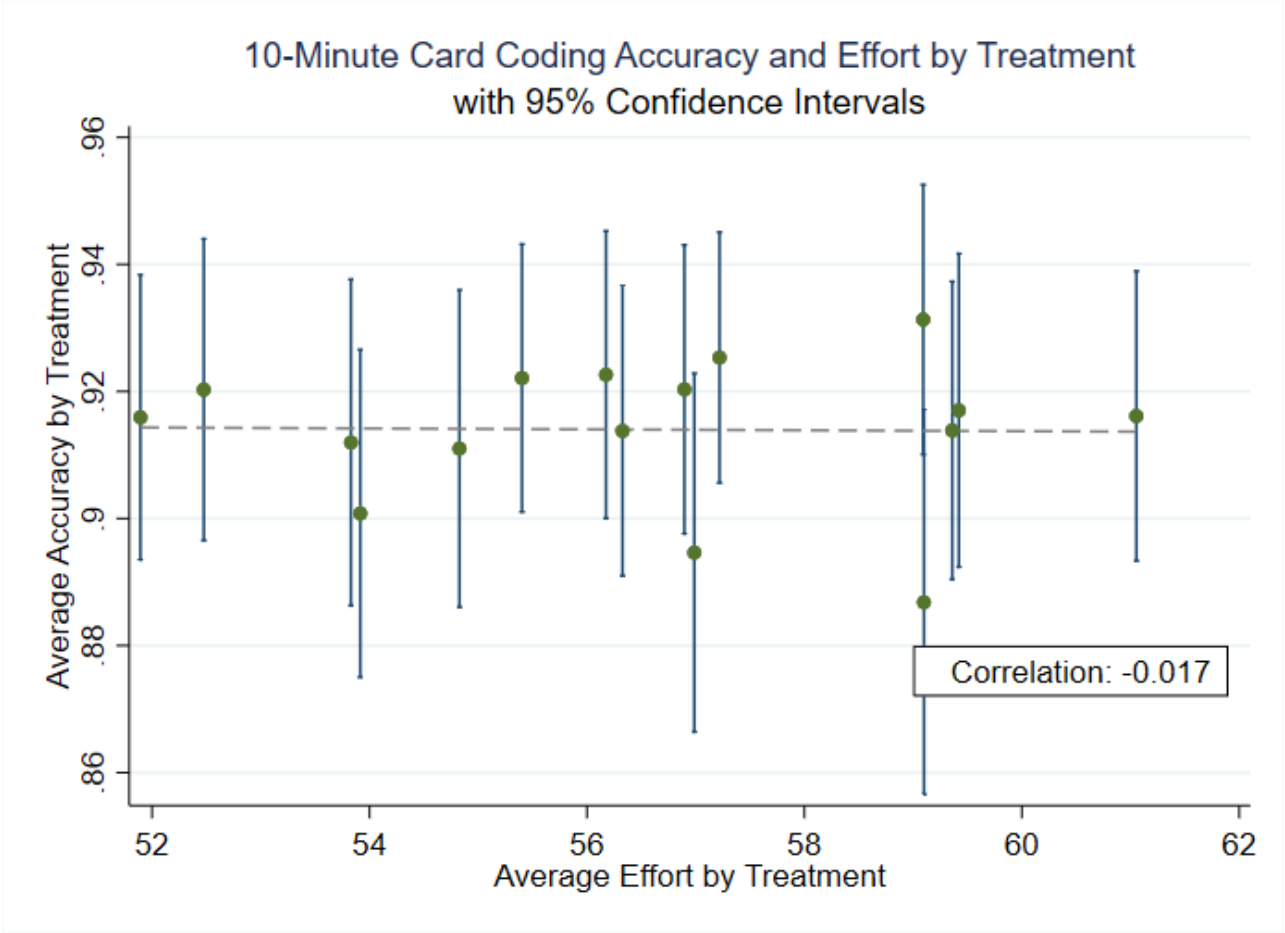


Figure 4e. 2018 Extra-Card Coding Task,No Consent



Notes: Online Appendix Figures 4a-e plot the distribution of the effort measure across the 2015 experimental results (Figure 4a) and for the four versions of the 2018 experimental results (Figures 4b-e). The distributions include all 15 treatments of focus in the paper.

Online Appendix Figure 5. Average Accuracy and Effort by Treatment in the 10-Minute Card Coding Experiment



Notes: Online Appendix Figures 5 displays evidence on accuracy for the 10-minute WWII coding task. The graph plots the average effort by treatment (on the x axis) against the average accuracy of coding (on the y axis). The measure of accuracy is the share of cards coded correctly, where we only considered cards for which 80% or higher of respondents provide the same answer (considering only the alphabetical letters of the responses) and cards that were formatted correctly (some cards did not have the right fields for respondents to code).

Online Appendix Table 1. Observation Counts by Treatment

		Number of Observations				
	Task:	**Typing Task, 10**		**2018 WWII Cards Coding Task**		
Category	**Treatment Description**	**2015 Exp.**	**2018 Exp.**	**10-Min**	**Extra Work**	**Extra Work, No Consent**
		(1)	(2)	(3)	(4)	(5)
Piece Rate	No payment	540	137	170	158	138
	Low piece rate	558	151	175	136	157
	Medium piece rate	562	150	173	136	154
	High piece rate	566	155	174	154	145
Pay Enough or Don't Pay	Very low piece rate	538	138	167	155	143
Social Preferences: Charity	Charity, low donation	554	151	164	130	168
	Charity, high donation	549	151	168	135	160
Social Preferences: Gift Exchange	Gift exchange, 40c bonus	545	151	168	150	146
Discounting	Low piece rate, 2-week delay	544	145	164	154	145
	Low piece rate, 4-week delay	550	155	170	154	141
Risk Aversion and Probability Weighting	1% prob. Piece rate	555	145	172	147	149
	50% prob. Piece rate	568	149	165	146	147
Social Comparisons	No payment, social comparison	526	149	164	142	151
Ranking	No payment, feedback after	543	143	169	143	153
Task Significance	No payment, please try hard	554	149	174	148	149
Piece Rate + Task Significance	Low piece rate, please try hard	-	161	171	143	146
	Number of Observations	8,252	2,380	2,708	2,331	2,392

Notes: The Table lists the number of observations in each treatment cell. Because treatment randomization occurred in the 2018 Extra Coding Consent (version 3) and No Consent (version 4) as one unit, the survey platform evenly presented the different treatments using all participants in these two versions. Therefore, there is a tradeoff between Column (4) and Column (5). For additional information on effort and treatments, see Table 2.

Online Appendix Table 2. Findings by Treatment: Effort in Different Versions of Experiment

		Mean Effort (s.e.)									
Task:		**Buttob-Pushing a-b Typing Task**									
Category	**Treatment Wording**	**Male**	**Female**	**College**	**No College**	**Young (=<30)**	**Old (30+)**	**USA**	**India**	**First 5 Mins**	**Last 5 Mins**
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Piece Rate	No payment	1451	1520	1403	1602	1516	1461	1502	1371	734	759
		(46)	(34)	(37)	(42)	(42)	(36)	(31)	(66)	(16)	(14)
	Low piece rate	2094	1957	1964	2080	2060	1964	2057	1743	1008	1008
		(40)	(30)	(33)	(36)	(35)	(33)	(26)	(68)	(14)	(12)
	Medium piece rate	2258	2022	2120	2141	2235	2022	2163	1833	1075	1055
		(35)	(29)	(30)	(35)	(33)	(31)	(25)	(71)	(13)	(12)
	High piece rate	2280	2076	2104	2251	2258	2067	2228	1750	1101	1068
		(36)	(26)	(30)	(31)	(30)	(31)	(22)	(69)	(12)	(11)
Pay Enough or Don't Pay	Very low piece rate	1857	1873	1824	1916	1953	1778	1901	1577	903	964
		(45)	(31)	(37)	(36)	(37)	(35)	(28)	(74)	(16)	(13)
Social Preferences: Charity	Charity, low donation	1931	1834	1855	1910	1944	1813	1890	1789	943	937
		(39)	(28)	(31)	(37)	(34)	(32)	(26)	(65)	(14)	(12)
	Charity, high donation	1974	1838	1862	1954	1953	1852	1926	1728	962	939
		(37)	(29)	(31)	(34)	(34)	(31)	(25)	(64)	(14)	(12)
Social Preferences: Gift Exchange	Gift exchange, 40c bonus	1564	1582	1509	1664	1635	1521	1580	1533	788	787
		(45)	(31)	(35)	(39)	(42)	(33)	(29)	(71)	(15)	(13)
Discounting	Low piece rate, 2-week delay	2044	1952	1942	2051	2105	1896	2030	1734	1001	993
		(41)	(28)	(33)	(35)	(36)	(31)	(26)	(67)	(14)	(12)
	Low piece rate, 4-week delay	2003	1931	1891	2060	2029	1898	2006	1676	985	979
		(43)	(30)	(35)	(36)	(38)	(33)	(27)	(65)	(14)	(13)
Risk Aversion and Probability Weighting	1% prob. Piece rate	1977	1854	1856	1985	1978	1851	1971	1557	946	968
		(39)	(31)	(34)	(35)	(37)	(33)	(26)	(64)	(15)	(12)
	50% prob. Piece rate	2018	1899	1887	2022	2016	1886	1981	1629	983	970
		(39)	(26)	(31)	(32)	(34)	(29)	(24)	(65)	(13)	(12)
Social Comparisons	No payment, social comparison	1884	1787	1765	1922	1927	1744	1845	1755	920	914
		(45)	(34)	(38)	(40)	(40)	(38)	(31)	(77)	(16)	(14)
Ranking	No payment, feedback after	1761	1712	1687	1793	1813	1662	1748	1548	869	868
		(43)	(32)	(37)	(39)	(40)	(36)	(30)	(73)	(15)	(13)
Task Significance	No payment, please try hard	1758	1684	1629	1832	1789	1643	1740	1565	862	856
		(42)	(32)	(35)	(37)	(39)	(34)	(28)	(72)	(15)	(12)
Piece Rate + Task Significance	Low piece rate, please try hard	2065	2049	2011	2106	2178	1910	2131	1686	1038	1019
		(85)	(50)	(64)	(65)	(62)	(65)	(49)	(125)	(23)	(26)
	Number of Observations	4,754	5,878	5,927	4,705	5,300	5,332	8,926	1,247	10,632	10,632

Notes: The Table presents the average output for each treatment cel, split by the dimensions listed in the column headings. See Table 2 for more information.

Online Appendix Table 3. Comparison Across Designs, Alternative Measures

Category	Version Comparison	Pearson Correlations Across Versions		Average Log Point Difference From Baseline Treatment				Average Absolute z-Score Difference from Baseline Treatment			
				Baseline Treatment: No Payment		Baseline Treatment: 10 Cent		Baseline Treatment: No Payment		Baseline Treatment: 10 Cent	
		Full Stability w/ Noise	Actual	Full Stability w/ Noise	Actual	Full Stability w/ Noise	Actual	Full Stability w/ Noise	Actual	Full Stability w/ Noise	Actual
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Pure Replication	2015 AB Task vs. 2018 AB Task	0.96 (0.02)	0.97 (0.02)	0.04 (0.02)	0.07 (0.04)	0.03 (0.01)	0.04 (0.01)	0.33 (0.14)	0.13 (0.07)	0.32 (0.12)	0.10 (0.04)
Demographics	Male vs. Female	0.97 (0.01)	0.98 (0.02)	0.04 (0.02)	0.10 (0.03)	0.03 (0.01)	0.06 (0.02)	0.28 (0.11)	0.12 (0.06)	0.26 (0.09)	0.09 (0.04)
	College vs. No College	0.97 (0.01)	0.97 (0.02)	0.04 (0.02)	0.07 (0.03)	0.03 (0.01)	0.02 (0.01)	0.28 (0.10)	0.09 (0.05)	0.26 (0.09)	**0.07 (0.03)**
	Young vs. Old	0.97 (0.01)	0.98 (0.02)	0.04 (0.02)	0.04 (0.03)	0.03 (0.01)	0.02 (0.01)	0.28 (0.11)	0.08 (0.05)	0.27 (0.10)	**0.05 (0.03)**
Geography/ Culture	US vs. India	0.92 (0.03)	0.78 (0.09)	0.06 (0.03)	0.07 (0.03)	0.04 (0.01)	0.11 (0.02)	0.42 (0.18)	0.20 (0.06)	0.41 (0.14)	0.34 (0.10)
Task	AB Task vs. Card Coding	-	0.60 (0.14)	-	0.25 (0.06)	-	0.19 (0.04)	-	0.59 (0.12)	-	0.53 (0.11)
Output	Extensive Cards vs. Intensive Cards	-	0.21 (0.17)	-	0.21 (0.06)	-	0.50 (0.05)	-	0.23 (0.05)	-	0.71 (0.10)
	Extensive Cards vs. AB Task	-	0.65 (0.07)	-	0.16 (0.04)	-	0.33 (0.04)	-	0.47 (0.11)	-	0.29 (0.07)
	AB Task: First 5 min vs. Last 5 min	0.99 (0.00)	0.98 (0.01)	0.03 (0.01)	0.04 (0.02)	0.02 (0.01)	0.03 (0.01)	0.21 (0.07)	**0.04 (0.02)**	0.21 (0.07)	**0.04 (0.02)**
Ecological validity	Cards: Consent vs. No Consent	0.92 (0.03)	0.92 (0.04)	0.13 (0.06)	0.16 (0.08)	0.09 (0.02)	0.08 (0.02)	0.43 (0.17)	0.15 (0.07)	0.36 (0.10)	**0.10 (0.03)**

Notes: The Table presents alternative measures of stability of experimental results for the version comparisons of Table 2. Values that are bolded are significantly different from the full stability measure (p value <0.05).

Online Appendix Table 4. Accuracy in the 2018 Card-Coding Task

Category	Treatment Wording	10-Minute Card Coding	Required Cards, Pooled	Extra Cards, Pooled
		(1)	(2)	(3)
Piece Rate	No payment	0.912	0.928	0.920
		(0.013)	(0.009)	(0.018)
	Low piece rate	0.914	0.912	0.922
		(0.012)	(0.010)	(0.014)
	Medium piece rate	0.925	0.921	0.936
		(0.010)	(0.009)	(0.011)
	High piece rate	0.914	0.896	0.884
		(0.012)	(0.011)	(0.016)
Pay Enough or Don't Pay	Very low piece rate	0.916	0.919	0.898
		(0.012)	(0.009)	(0.02)
Social Preferences: Charity	Charity, low donation	0.920	0.932	0.906
		(0.012)	(0.008)	(0.017)
	Charity, high donation	0.895	0.920	0.929
		(0.014)	(0.009)	(0.015)
Social Pref: Gift Exchange	Gift exchange, 40c bonus	0.916	0.928	0.934
		(0.011)	(0.009)	(0.013)
Discounting	Low piece rate, 2-week delay	0.917	0.922	0.920
		(0.013)	(0.01)	(0.015)
	Low piece rate, 4-week delay	0.887	0.906	0.899
		(0.015)	(0.01)	(0.017)
Risk Aversion and Probability Weighting	1% prob. Piece rate	0.931	0.929	0.943
		(0.011)	(0.009)	(0.011)
	50% prob. Piece rate	0.901	0.914	0.920
		(0.013)	(0.01)	(0.015)
Social Comparisons	No payment, social comparison	0.920	0.909	0.896
		(0.012)	(0.010)	(0.019)
Ranking	No payment, feedback after	0.922	0.918	0.921
		(0.011)	(0.009)	(0.016)
Task Significance	No payment, please try hard	0.911	0.927	0.922
		(0.013)	(0.009)	(0.016)
Piece Rate + Task Significance	Low piece rate, please try hard	0.923	0.918	0.904
		(0.012)	(0.009)	(0.016)
	Number of Observations	2,708	4,723	3,026
	Average Accuracy	0.914	0.919	0.916
		(0.003)	(0.002)	(0.004)
	Prob > F	0.750	0.477	0.188

Notes: The Table presents the average accuracy of coding of occupation in WWII cards. The accuracy is defined as follows: We consider only cards for which 80% or higher of respondents provide the same answer (considering only the alphabetical letters of the responses) and cards that were formatted correctly (some cards did not have the right fields for respondents to code). This restricts the sample from 3,353 cards to 2,588 cards. Restricting the analysis to such cards, we compute the share of cards that an individual computed correctly, and then average across the individuals in a treatment. Column 1 refers the 10-minute card-coding experiment, Column 2 refers to the required-cards experiment, and Column 3 refers to the coding of the extra cards.