NBER WORKING PAPER SERIES

STABILITY OF EXPERIMENTAL RESULTS: FORECASTS AND EVIDENCE

Stefano DellaVigna
Devin Pope

Working Paper 25858
http://www.nber.org/papers/w25858

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2019

We thank Ned Augenblick, Jon de Quidt, Anna Dreber, Magnus Johannesson, Don Moore, Alex Rees-Jones, Joshua Schwartzstein, Dmitry Taubinsky, Kenneth Wolpin, as well as audiences at Harvard University (HBS), Rice University, Stockholm University, the University of Bonn, the University of Toronto, UC Berkeley, Yale University (SOM), 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Stability of Experimental Results: Forecasts and Evidence
Stefano DellaVigna and Devin Pope
NBER Working Paper No. 25858
May 2019
JEL No. C9,C91,C93

ABSTRACT

How robust are experimental results to changes in design? And can researchers anticipate which changes matter most? 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 full stability (which allows for noise), and to expert forecasts. 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 changes to the task and output will make a further impact. We find near perfect replication of the experimental results, and full stability of the results across demographics, significantly higher than the experts expected. The results are quite different across task and output change, mostly because the task change adds noise to the findings. The results are also stable to the lack of consent. Overall, the full stability benchmark is an excellent predictor of the observed stability, while expert forecasts are not that informative. This suggests that researchers' predictions about external validity may not be as informative as they expect. We discuss the implications of both the methods and the results for conceptual replication.

Stefano DellaVigna
University of California, Berkeley
Department of Economics
549 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
sdellavi@econ.berkeley.edu

Devin Pope
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
devin.pope@chicagobooth.edu

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/2987

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: How important is the specific task? Should I worry about a change in consent form that the IRB required? After running the experiment, the researcher is confident that the results would replicate with the same protocol, but less 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. She is concerned about demand effects, given that the subjects knew they were part of an experiment, and also about the specificity of the setting in rural Brazil. These concerns lead her to recommend rejection for the paper. The editor is unsure how informative the referee assessment of external validity is.

A third researcher reads about the replication of psychology and economic experiments (Open Science Collaboration, 2015; Camerer et al., 2016, 2018) and wonders: If we move beyond pure replication to conceptual replication, will the experimental results replicate? How do we even measure replication, if for example the units of measure in the replication differ from the original units?

These three researchers are concerned about how experimental results vary as the design changes. This concern, depending on the specific literature, is labeled as being about *stability of results*, *conceptual replication*, or *external validity*. A number of papers examine the stability of experimental results with respect to specific design choices, such as for example the use of the strategy method or direct choice (Brandts and Charness, 2011) and the impact of demand effects (de Quidt, Haushofer, and Roth, 2018). In a field setting, for example Allcott (2015) studies the heterogeneous effects by demographics of the OPower electricity reports and Vivalt (2017) the heterogeneous effects of interventions in development economics.

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 as a roadmap for conceptual replication in experiments with multiple treatments arms (e.g., Gerber and Green, 2000, Bertrand et al., 2010 and Bhargava and Manoli, 2015). 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, as in the referee and researcher examples above, we collect forecasts 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 others may be important: (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 change with a different measure; (vi) *(consent form)* it may matter that subjects know that it is an experiment.

The initial task is a typing task employed in DellaVigna and Pope (2018a,b): subjects on MTurk have 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. 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 effort motivators.

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

While we make the above comparisons 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 measure not the number of cards coded in a fixed amount of time, but the number of extra cards coded beyond a required amount.[1] Finally, 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 in 10 minutes is 1,800 points, but in the WWII coding task, the average output in 10 minutes is 58 cards. One could make the two designs comparable by rescaling the effect sizes by 1,800/58. But this rescaling does not account for differences in the elasticity of effort to motivation: a 30 percent increase in effort in the a-b task, which we observe

[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] Since subjects are coding historical data, this is a natural framing and does not require any deception on our part (and was approved by the IRB board).

in response to piece rate variation, may not be achievable in the WWII card coding task.

With these considerations in mind, we use the rank-order correlation of the average effort in the 15 treatments as our benchmark measure of stability. 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 the other treatments keep similar ranks; in this case, the rank-order correlation will be high. If instead those treatments move to positions 7, 4, and 10, and the other treatments also move rank, the rank-order correlation will be low. While this measure is not without drawbacks, it performs well also when the underlying model predicts a non-linear transformation, as in the output change. Importantly, we compare the observed rank-order correlation to the average rank-order correlation under a *full-stability benchmark*, in which the only variation in rank is due to idiosyncratic noise in the realized effort. For some design changes, we can generate this benchmark with bootstraps from the data; in other cases, we need to use structural estimates of the behavioral parameters to make predictions that account for the task-specific degree of noise and effort elasticity.

Having identified the design changes and the measure of stability, following DellaVigna and Pope (2018b) 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 full-stability benchmark. 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 be fairly close to full replication (0.82 correlation, compared to 0.94 under full stability); (ii) the results will differ sizably for different demographics (age/gender/education) (0.73 correlation, compared to 0.95 under full stability), (iii) the results will differ for the India and US sample (0.63 correlation, compared to 0.89 under full stability); (iv) the task and output changes will have a sizable impact (0.50 to 0.70 correlation); (v) the disclosure of 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 or by 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 full stability. We find (ii) strikingly high stability across demographics—correlations of 0.96 for gender, 0.97 for education, and 0.98 for age—significantly higher than the experts expected (0.73 on average). Interestingly, the demographic groups *do* differ in the average effort and even in the sensitivity to financial incentives. Once we control for that, though, as rank-order correlation does, the various groups respond very similarly to the behavioral treatments, also in comparison to the response 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 this lower correlation is partly due to noise (given that Indian workers are just 12 percent of the data). We find near-perfect correlation (0.96) in the results for workers from "blue states" as opposed to "red states".

Comparing across tasks (iv), the rank-order correlation between the 10-minute a-b typing task

versus WWII card coding is 0.59, close to the expert forecast of 0.66. We then compare (v), two designs with the same task—coding WWII cards—but different output measures: the number of cards coded in 10 minutes, versus the number of extra cards coded after completion of the required cards. The rank-order correlation is just 0.27, compared to the expert prediction of 0.61. Changes in the task and measure of output are the factors that lead to the most instability of the results.

This instability has two possible explanations. First, changes in task and output may have truly changed the impact of behavioral motivators. Second, effort in the 10-minute WWII task, unlike in the a-b task, may just be a very noisy measure of motivation, and the noise may be swamping the motivational effects. Consistent with this second interpretation, the 10-minute WWII task is especially noisy on two grounds: output is barely responsive to incentives, and the between-subject standard deviation of effort (the noise term) is large. Indeed, the full-stability benchmark for the task change built from the structural estimates is 0.50, similar to the observed correlation of 0.59.

We confirm this interpretation with a combined output/task comparison of the a-b 10-minute task to the WWII extra-cards coding. This latter task is highly responsive to incentives. If the lack of stability is mostly tied to noise, we would expect a sizable correlation, as both tasks are responsive. Indeed, the correlation for the joint task/output change, 0.65, is higher than for just the output change, 0.27.

Interestingly, the experts appear to miss the role for noise, since they instead predict a lower correlation for the joint task/output change, 0.54, than for just the output change, 0.63. Of course, the degree of noise in the different tasks was not obvious to the forecasters. To address this issue, we 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 neglect for the role of noise.

Lastly, we analyze dimension (vi) comparing the same extra-card WWII coding task with, and without, a consent form. The rank-order correlation is 0.84, which is close to the expert prediction (0.78) and to the full-stability measure (0.88). Thus, in our context it does not matter much whether we disclose that the task is an experiment.

Altogether, we draw five main lessons. First, we find an encouraging degree of stability of experimental results across design changes. Nine out of ten planned comparisons have a correlation above 0.60, and six comparisons have a correlation above 0.80. This conclusion is not affected by the metric used to compute the stability, and is not contaminated by selective reporting, as all the comparisons are pre-specified. Indeed, our full-stability benchmark, which assumes full replication but allows for the role of noise, is an excellent predictor of the observed correlations.

Second, the experts have at best a mixed record in their ability to predict how much design changes affect the results. This contrasts with recent evidence that experts are able to predict quite accurately replication in pure-replication studies (Dreber et al., 2015; Camerer et al., 2016) as well as the effect of behavioral motivators (DellaVigna and Pope, 2018a,b). This confirms the anecdotal impression that design choices are a somewhat unpredictable part of the experimenter

toolbox, and suggests that external validity judgments may be more tentative than we realize.

Turning to two specific results, our third take-away is the remarkable stability of the results with respect to the demographic composition of the sample, in contrast to the view of the experts, who expected a larger role for demographics. Selective publication may explain this discrepancy: while null results on demographic differences may not get published (Franco, Malhotra, and Simonovits, 2014), differences that are statistically significant draw attention and may thus be more salient.

Fourth, the degree of noise in the experimental results is a first-order determinant of stability of the results, in a way that the experts do not appear to readily anticipate, even when provided with diagnostic information. This finding is reminiscent of Tversky and Kahneman (1971)'s findings from a survey of psychologists and may also be related to 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 is a methodological contribution to conceptual replication. We demonstrate how rank-order correlation can serve as a useful metric for experiments with multiple treatment arms. We also introduce a benchmark of how much the experimental results would change purely due to noise. As we show, taking noise into account is critical to the evaluation of stability.

Related to our paper is the work on 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. Perhaps the closest large-scale effort is the "Many Labs" projects (Klein et al., 2014, 2018) which consist of pure replications for dozens of psychological findings, but in different labs around the world. Relatedly, Landy et al. (2018) crowdsource the test of five psychological hypotheses. Consistent with our findings on stability by demographics and geography, Klein et al. (2014, 2018) find limited evidence of heterogeneity in replication success across labs; Landy et al. (2018) find larger heterogeneity in results when the experimental design is left to the different researchers, leading presumably to larger design differences.

An example of conceptual replication is the comparison of experimental results run across different platforms, such as comparing laboratory experiments versus on MTurk (e.g., Horton, Rand, and Zeckhauser, 2011 and Snowberg and Yariv, 2018) or laboratory experiments versus field experiments (e.g, Falk and Heckman (2009)). Our design does not include a platform comparison in part because this was an aspect that already attracted significant attention in the literature.

2 Design and Measure of Stability

2.1 Experimental Design

2.1.1 2015 Experiment and Model

The starting point for the design is the real-effort task in DellaVigna and Pope (2018a,b) which we ran in May 2015 on Amazon Mechanical Turk (MTurk). MTurk is an online platform that allows

researchers and businesses to post small tasks (referred to as HITs). 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 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).

We recruited subjects on MTurk for a $1 pay for an "*academic study regarding performance in a simple task.*" Subjects interested in participating signed a consent form, entered their MTurk ID, answered three demographic questions, and then saw the instructions, reproduced in Online Appendix Figure 1b: "*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 final paragraph (bold and underlined) depended on the treatment condition. 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.*" 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."*

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 summarized the condition for earning a bonus (if any) in that particular treatment. At the end of the 10 minutes, the subjects were presented with the total points and the payout, were thanked for their participation and given a validation code to redeem the earnings. After applying the sample restrictions detailed in DellaVigna and Pope (2018a), the final sample included 9,861 subjects, about 550 per treatment.

The 18 treatments aim 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 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) $e^* = c'^{-1}(s+p)$. A useful special case, discussed further in DellaVigna et al. (2015), is the exponential cost of effort function $C(e) = exp(k)exp(\gamma e)/\gamma$, which has elasticity of effort $1/(\gamma e)$ with respect to the value of effort. Under this assumption, we obtain

$$e^* = \frac{1}{\gamma} ln\,(s+p) - \frac{1}{\gamma}k. \tag{2}$$

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.

Behavioral Treatments. 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 intrinsic motivation. In the next two social preferences treatments, 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). In the third social-preference treatment, on gift exchange (as in Gneezy and List, 2006), subjects receive an unconditional 40 cent bonus.

We model these treatments as follows. For the paying-too-little treatment and for the gift-exchange treatment 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 charity treatments, we allow for both a pure-altruism parameter α 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, in the Beckerian pure altruism world, the return to the charity is important, while in the "warm glow" model it is not, as the individual is motivated by the "warm glow" of working for the charity, not by the exact return.

In two treatments motivated by the research on present bias (Laibson, 1997; O'Donoghue and Rabin, 1999), the piece rate is 1 cent per 100 points, but the bonus will be deposited "*two weeks from today*" or, in a second case, "*four weeks from today*". We model the motivation 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.

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*" while in the second treatment it is "*a 50% chance of being paid an extra 2 cents for every 100 points*". The expected value of the piece rate in these two treatments is the same as in the low-piece-rate 1-cent treatment, but the piece rate is stochastic. We model the motivation as $(s+\pi(P)p)e$, with $P = 0.01$ or $P = 0.5$. Under risk neutrality and no probability weighting, we should estimate $\pi(P) = P$. Under the typical prospect theory parametrizations, small probabilities are overweighted as in, e.g., Prelec (1998) ($\pi(0.01) > 0.01$), and thus, provided subjects are not too risk averse, we expect higher effort in the 1-percent-of-$1 treatment than in the 1-cent treatment. The treatment with a 50 percent

probability of a 2-cent piece rate provides evidence on the concavity of the value function, i.e., risk aversion, which we capture in reduced-form as $\pi(0.5) < 0.5$.

The final three treatments do not involve incentives and are more directly borrowed from psychology, with wording aimed to boost effort with social comparisons ("*many participants were able to score more than 2,000 points*"), rankings ("*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 model these psychological treatments as increasing the baseline motivation by a term Δs.

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 the 15 treatments.

2.1.2 2018 Experiment

In May of 2018 we ran a new round of experiments on MTurk following a pre-analysis plan, with design variants aimed at testing the stability of the earlier experimental results. Other than the design changes, we kept the experimental material as close as possible to the earlier experiment.

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). Workers that clicked on the ad were randomized to one of four versions of the experiment, with versions 2, 3, and 4 oversampled by 15 percent. We designed the oversampling in light of higher attrition (15 percent higher in pilot data) for the task used in version 2-4. This higher attrition is likely due to difficulty for some workers in reading cursive writing (employed in these cards).

Within each version, the workers were randomized into 1 of 16 treatments with equal weights.[4] In addition to the 15 treatments from the earlier experiment, an additional 16th treatment combined 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, but we return to it in an out-of-sample prediction.

We now describe in detail the four versions of the 2018 experiment.

Pure Replication. The first version 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]

10-Minute WWII Coding. The second version is also a 10-minute task, but subjects are assigned to code the occupation in World War II enrollment cards[6]: "*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*

[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 differences: (i) the advertising screen in 2015 mentioned a 15-minute "*academic study regarding performance in a simple task*"; in 2018 we mentioned an 11-12 minute "*typing task*", to be consistent across the four versions; (ii) in 2018 the IRB required a longer consent form; (iii) the demographic questions are at the beginning of the survey in 2015 and at the end in 2018; (iv) the final pay-out page has slightly different wording.

[6]Bruno Caprettini and Joachim Voth provided us with cards to be coded as part of a historical project.

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 state "*Please be as careful as possible (we will check the accuracy of your work).*" For each card, the subjects type the occupation 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.

The 16 treatments in this second version, with the wording displayed in Column 2 of Table 1 in brackets, are as close to those in the first version as possible, expect for the piece rates. In pilot data, on average subjects coded 50-60 cards in 10 minutes, compared to 1,500-2,000 a-b presses in 10 minutes. Based on this ratio of productivity, and in order to set incentives at round numbers, we multiply the piece rates by a factor of 50. Thus, 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*". The implied average pay 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, the subjects still code World War II cards, but with a different design. In versions 1 and 2, we measure productivity—the number of units produced within a given time—, as in most real-effort experiments, including Gneezy, Niederle, and Rustichini (2003) and Gneezy and List (2006). Yet, an alternative margin of effort is a form of labor supply—how much *extra work* one is willing to do. Abeler et al. (2011) pioneered this design: after 4 minutes of required work, subjects were asked if they would do more work, for up to 60 minutes. The outcome of interest is how long subjects work.

In our third version, the subjects first code the occupation field for 40 WWII cards (Online Appendix Figure 1d), with no experimental manipulations, that is, with no extra incentive. 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 incentives as in the second version, though this implies that the average total payment will tend to be lower in this version, compared to the 10-minute WWII card coding version. 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 version is identical to the third version, except that it lacks a consent form. While in all other versions the workers see a consent form after

[7] In this version, we removed the demographic questions, since we did not want demographic questions in the next version, and wanted to keep the two versions parallel.

clicking on the MTurk HIT, in this version they are taken directly to the description of the task. Given that the task involves the coding of historical documents—a common job on platforms like MTurk, the absence of a consent form should not be a surprise. This condition provides evidence on whether it matters if subjects know they are participating in an experiment. This aspect is often debated, for example in the Harrison and List (2004) classification of a natural field experiment. Yet surprisingly, there is little evidence on whether this matters for the results of experiments.

Sample. In the pre-analysis plan, we set out to exclude subjects that: (1) do not complete the 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 have 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. Next, we eliminate likely cases of cheating: (5) in version 1 scored 4000 or more a-b points; (6) in version 2 coded 120 or more cards with accuracy below 50%; (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%. We set a target of 10,000 subjects completing the tasks and kept open the task on MTurk until either (i) three weeks had passed or (ii) 10,500 subjects had completed the study, whichever came first.

We followed the pre-registration sample rules. The experiment ran for three weeks, at which point we had 12,983 subjects who started the task on Qualtrics. We then removed 324 workers who had re-entered the task, 2,660 workers who had either taken more than 30 minutes to finish or not completed the survey at all (restrictions 1 and 2), 68 individuals who had not been approved (restriction 3), and 40 individuals who violated restriction 4-7. Two final restrictions not included in the preregistration were excluding 21 MTurkers who blatantly cheated on the card-coding task in ways not covered above and 59 observations due to Qualtrics data "glitches".[8]

The final sample is 9,811 responses, close to the envisioned sample of 10,000, with similar sample sizes of 2,330-2,390 subjects in Versions 1, 3, and 4. The oversampling (by 15 percent) of Versions 3 and 4 thus succeeded in approximately equating the sample size. Version 2 has a larger sample size, with 2,708 subjects, due to the oversampling and no offsetting increase in attrition.

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 dimensions, listed in Table 2.

Dimension 1. Pure Replication. We compare the results of the a-b task experiments run in 2015 and in 2018. The two experiments have nearly identical design, with slight changes in 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%

[8]The 21 MTurkers eliminated for cheating had less than 10% accuracy and gave, for example, multiple one-letter responses and multiple responses of "I don't know." The 59 observations with glitches had: (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) 10 time stamps more than the total coded cards.

versus 54.8%). Also, the 2018 experiment has a smaller sample size—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 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). Pooling the 2015 and 2018 data, 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 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, we compare workers in the US (N=8,803) versus workers in India (N=1,225).[9] For an additional comparison, we compare workers in "red states" versus "blue states" according to the state-level 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. The two 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: Version 2 in which output is the number of cards coded in 10 minutes, and Version 3 in which output is 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 compare 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, with consent form (Version 3) and without (Version 4).

2.3 Measure of Stability

In each of these dimensions, we want to compare the average effort for the 15 treatments in the two different designs to measure the stability of the results. This seemingly simple comparison raises three issues. First, how do we compute the stability given that there are multiple treatments to compare? Second, how do we account for the role of noise? Third, how do we measure stability when effort is not comparable across versions, e.g., across tasks?

The first issue arises because our experiment has multiple treatment arms, covering incentives and multiple behavioral motivators. Having multiple treatment comparisons is obviously neither better nor worse than the more typical binary treatment-control comparison, as in most of the replication literature so far (e.g., Camerer et al., 2016). As large data sets become increasingly available to firms, policy-makers, and researchers, we conjecture that horse-races of multiple interventions will become increasingly common. Examples in the literature include Bertrand et al. (2010) with firm data (loan take-up) and Bhargava and Manoli (2015) from a policy experiment (EITC take-up). In such cases, one is typically interested not only in how each treatment compares

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

to a baseline group, but also in comparisons of the effectiveness across treatment arms. To capture these multiple comparisons with a simple metric, we use the rank-order correlation between the treatment effectiveness in one version, versus the treatment effectiveness in another version.

As an example, consider the hypothetical set of results in Online Appendix Figure 3a for the pure replication case: in the first panel, only two treatments switch order, and the rank-order correlation is very high (0.97). In the next examples, the treatments change position more, and the rank-order correlation is lower. While we considered different measures of stability, such as the Pearson correlation, we opted for the rank-order correlation because it is stable to non-linear transformations. It is also relatively easy to explain, including to the forecasters.

Second, the rank-order correlation will not be perfect (that is, 1) even if the treatment effects are perfectly stable, because noise in the experimental results will lead to switches in the treatment ranks. To partial out the impact of noise from actual instability in the treatments, we define a *full-stability benchmark*: the average rank-order correlation which we would expect if the treatment effects were fully stable, but allowing for noise in the data. For the design changes that take place within one task, we use a simple bootstrap procedure. For example, in the pure replication case, (i) we bootstrap from the 2015 sample (with replacement) 150 observations from each of the 15 treatments, mirroring the sample size for the 2018 experiment, (ii) we compute the average effort in the 15 simulated cells, (iii) we compute the rank-order correlation of the 15 bootstrapped means with the actual 2015 results for the 15 treatments, and (iv) we repeat this 1,000 times. The average rank-order correlation across the 1,000 iterations, 0.94, is the full-stability benchmark.[10]

Similarly, we compute a bootstrap for the demographic comparisons in the pooled 2015-2018 a-b task: (i) in each of the 15 treatments, we randomly assign a subject to demographics A or B, with the share assigned to group A matching the empirical one; (ii) we compute the average effort in each of the 15*2 cells, and (iii) the rank-order correlation; (iv) we repeat 1,000 times.

This bootstrap procedure however is not feasible when comparing two versions with effort measured in different units, such as going from the a-b task to the WWII card coding. This is the third issue raised above. We thus add some additional modeling structure, as in the *structural behavioral economics* approach (DellaVigna, 2018), to compute the needed counterfactual.

Specifically, the effort in the various treatments depends on behavioral and incidental parameters. The *behavioral parameters*, such as the social preference or the probability weighting ones, are the ones which we can expect to be stable across versions. In contrast, the *incidental parameters*—the curvature and level of cost of effort, the baseline motivation, and the standard deviation of noise—surely will differ across versions, for example between the a-b task and WWII cards coding. We define two versions to have stable experimental findings if they share the same behavioral parameters, even if the incidental parameters vary. Thus, we compute the full-stability benchmark as follows. Taking as example the task comparison (a-b task versus WWII coding), (i) we structurally estimate the (behavioral and incidental) parameters in both the a-b task and in WWII

[10]In this bootstrapping procedure, we do hold the 2015 results as given and do not bootstrap it at each iteration; the results are very similar if we do that as well.

coding; (ii) for the a-b task, we draw a sample of 700 observations per treatment given the a-b task structural estimates; (iii) for the WWII card coding, we draw a sample of 150 observations per treatment assuming the *incidental* parameters for the WWII coding task, but taking the *structural* parameters for the a-b task; (iv) we compute the rank-order correlation of the sample means; and (v) we repeat 1,000 times. This procedure allows for different incidental parameter, but in step (ii) embeds full stability of the structural parameters. We revisit this in Section 4.3.

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? Following DellaVigna and Pope (2018a,b), we contact a group of researchers to collect their forecasts about the importance of design changes.

Sample. 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 our 2015 forecasting results suggest that a couple dozen responses would provide sufficient statistical power.

Thus, we narrowed the sample to the 73 experts with (i) PhD year between 2005 and 2015, and (ii) behavioral economics as a main field of specialization; we contacted 42 out of the 73 experts. We then added 18 behavioral economists with PhD in 2015-2018 (who were not included in the earlier sample), drawing names from lists of attenders and presenters at key behavioral conferences (BEAM and SITE Psychology and Economics). In addition, we identified 10 experts working on 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.

We also contacted, like we did in 2015, a group of PhD students in economics at UC Berkeley and the University of Chicago, yielding 33 responses. Finally, we posted the survey (for a $1 payment) on MTurk and collected 109 valid responses.[11]

Survey. The survey, which was expected to take 15-20 minutes, walked the forecasters through four steps. First, we briefly summarized the design in the 2015 experiment and displayed the results on average effort by treatment using Online Appendix Figure 2. Second, we introduced the concept of rank-order correlation using four graphical examples, displayed in Online Appendix Figure 3a.

Third, we asked for ten forecasts, listed in Table 2, of rank-order correlation (Online Appendix Figure 3b displays the slider): (i) 1 forecast about pure replication; (ii) 3 forecasts about demographics (gender/education/age); (iii) 1 forecast about geography/culture comparing MTurkers in US versus in India; (iv) 1 forecast about task change; (v) 3 forecasts about output change, comparing first the 10-minute WWII coding to the extra-work WWII coding; then comparing the a-b

[11] We recruited 150 MTurkers. In order to prevent bots and inattentive survey-takers, we introduced a captcha verification and an attention question. We dropped 18 MTurkers who failed the attention check, 21 MTurkers who took the survey in under 5 minutes, and 2 MTurkers with duplicate IP addresses.

task to the extra-work WWII coding; and finally, comparing, within the a-b task, effort in the first 5 minutes versus in the last 5 minutes; and (vi) 1 forecast about the impact of the consent form, comparing Version 3 to 4.

In some of these comparisons, we provide the full-stability benchmark, that is, what rank-order correlation we would expect to observe on average if the results did not change (Column 1 in Table 2), as discussed in Section 2.3. Specifically, we report it for the pure replication (0.94), the demographic comparisons (0.95 for the gender/age/education splits) and the US-India comparison (0.89). We did not report a full-stability benchmark comparing across different tasks or output, given that this requires a full set of structural estimates. Finally, the table reports the full-stability benchmark for the comparison of output in the first 5 minutes and next 5 minutes (0.99) and for the consent form (0.88), but we did not report such benchmarks to the subjects.[12]

In the fourth step of the forecasting survey, respondents indicated their 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 3-5 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 7). The cdf plot in Figure 1a shows that 75 percent of experts expect a correlation above 0.80, with only 18 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 0.73 (gender), 0.71 (education), and 0.74 (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 that is lower than 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 0.63.

Turning to the task and output correlations, the experts on average expect a correlation of 0.66 for the change in task (a-b typing versus WWII card coding) and a similar correlation of 0.61 comparing across output margins (effort within-10-minutes versus extra work) within a task. In the forecast about the joint task/output change (comparing the 10-minute a-b typing to the extra-work WWII coding), the experts are most pessimistic, with an average forecast of 0.53. In another output comparison—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. Finally, regarding the impact of the consent form, or absence thereof, the experts on average

[12] We did not compute these benchmarks as we wanted to remain as blinded to the full 2018 experimental data as possible. Notice that the full-stability benchmark for the pure replication uses only the 2015 data. The full-stability benchmark for the demographic comparisons does require the 2018 a-b task data.

expect a correlation of 0.78, compared to the full-stability benchmark of 0.88.

How confident are the experts? They predicted that on average they would guess 3.99 correlations (out of 10) within 0.1 of the realized value. We revisit this prediction in Section 5.

Turning to the predictions of two other groups, PhD students and MTurk workers, we find that 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 with the pure replication in 2018 of the 2015 a-b typing task. Before we turn to the results by treatment, we document the overall distribution of effort and the response to piece rates. As Online Appendix Figure 4a-b shows, the distribution of effort across the 15 treatments is very similar in 2015 and 2018, if somewhat noisier in 2018, given the smaller sample size. Figures 2a-b shows that effort also responds similarly to the piece rate incentives.

What about then the behavioral treatments? As Figure 3 shows, the results by treatment for 2018 line up very nicely with the 2015 results, only slightly below the 45-degree line (dotted line). Just one treatment deviates by more than 100 points from the interpolating line (continuous line), the probability weighting treatment, which yields higher effort in 2018 than predicted based on the 2015 results. 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 9). Thus, our pure replication produces very similar results 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.

Figure 4a displays the treatment results separately for male and female subjects.[13] The data suggests two striking patterns. First, men and women *do* differ: male subjects are more responsive to incentives, varying their effort from 1,450 points to nearly 2,300 points, while female subjects increase effort from 1,500 points to 2,050 points.[14] And yet, 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. Thus there is no gender difference in the response to the different behavioral motivators, and in the response to the behavioral motivators compared to the

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

[14]In a meta-analysis of 17 studies, Bandiera et al. (2018) show that, on average, women response to incentives similarly to men. Our focus differs as we focus on how women respond to behavioral motivators, *conditional* on their overall response to piece rate incentives (which we find to be flatter).

financial motivators. This leads to a very high rank-order correlation of 0.96, a correlation that is statistically significantly different from the average expert forecast of 0.73.

Is this result unique to the gender comparison? In Figure 4b we compute the treatment effects separately for subjects with a completed college degree and for subjects without. The two groups of subjects differ in the level of effort: higher-education subjects exert less effort in any given treatment. But once we control for this difference, the treatment effects line up very nicely. Indeed, the rank-order correlation in effectiveness is 0.97, much larger than the average forecast of 0.71, a difference that again is statistically highly significant. Similarly, splitting the results by age in Figure 4c, we see that 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. 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 with 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. Adjusting for this difference, the behavioral and incentive treatments show a sizable correlation of 0.65. This correlation is statistically lower than the full-stability benchmark (p=0.049) and nearly identical to the average forecast (0.63).

Task. In the fourth comparison, we compare the results in the 10-minute a-b typing task, still pooling the 2015 and 2018 experiments, to the 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 effort measure in this new task, the number of cards coded in 10 minutes, is approximately normally distributed, with a median around 60 cards. Figure 2c shows that the new task is very unresponsive to financial incentives: we cannot reject that the high-piece rate treatment and the baseline no-piece-rate treatment yield the same effort.[15]

In light of this, it is not surprising that the correlation between the two tasks is not particularly high. Figure 6 shows that the rank-order correlation is 0.59, in line with the average expert forecast of 0.66. In fact, given the noise, 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, for a given task, affect the experimental findings. First, we compare 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 decide how many extra cards to code (from 0 to 20), after completing a required batch of 40 cards. As Online Appendix Figure 4d shows, the distribution of extra cards coded is highly bimodal: the large majority of subjects code 0 extra cards, or all 20 extra cards. Importantly, as Figure 2d shows, the output measure in this task is highly responsive to incentives: 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). Each of the increases is statistically significant. Thus, this

[15]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. Online Appendix Table 3 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.

design is well-suited to capture variation in motivation.

Figure 7a shows that the treatment effects with this output measure have a low correlation of 0.27 with the treatment effects in the 10-minute WWII coding task; this correlation is much lower than in the expert forecasts (0.61). This (relative) instability has two possible explanations. First, changes in output may have truly changed the impact of behavioral motivators. Second, productivity in the 10-minute WWII task, unlike in the a-b task, may just be a very noisy measure of motivation, and the noise in the realized effort may be swamping the motivational effects.

In order to provide some evidence on the first explanation, we do a combined output/task comparison of the a-b 10-minute task to the WWII extra-cards coding. Since both of these tasks are responsive to incentives, the comparison should not be too affected by noise. As Figure 7b shows, the correlation for the joint task/output change, 0.65, is higher than for just the output change, 0.27. Interestingly, the expert forecasters instead expect the correlation to be higher for just the output change, 0.61, than for the task/output change, 0.53.

As an additional output comparison, we return to the a-b typing task (pooling 2015 and 2018) and compare the effort by treatment for the first 5 minutes versus the next 5 minutes. As Figure 7c shows, the rank-order correlation is very high at 0.97, close to the full-stability benchmark of 0.99 and statistically significantly higher than the average forecast of 0.72.

Consent. Finally, we compare two versions of the extra-cards WWII task, which only differ in that the first one (discussed above) has a consent form, while the second one does not. As Figure 8 shows, the two versions yield very similar results, with a rank-order correlation of 0.84, close to the full-stability benchmark of 0.88 and to the average forecast of 0.78.

4.2 Robustness

Alternative Measure of Stability. A legitimate question is whether these results on the impact of design changes depend on the specific measure of stability, the rank-order correlation. In Online Appendix Table 4, we replicate the results using alternative measures of stability, though of course for these alternative measures we do not have expert forecasts.

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 effort measures across designs. The results are very similar using this alternative measure.

In Columns 3 and 4, for each treatment (other than the baseline one), we compare the difference in effort in log points, relative to 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. The pure replication and the different demographic versions are associated with fairly small log point changes, close to the full-stability benchmark, with larger log point changes for the task and output changes. Columns 5 and 6 show that the calculations are fairly similar if we use the same log point measure, but using the high-pay treatment as comparison point. In Columns 7-10 we obtain parallel results measuring the changes from the baseline in standard deviation units (z scores), instead of in log point units.

Alternative Comparisons of Designs. A separate robustness issue is that so far we focused on 10 (pre-specified) design changes. In Table 3 we consider 12 additional design comparisons.

The first three comparisons present the familiar demographic comparisons, but for the 10-minute WWII card coding task.[16] The rank-order correlations across the demographics are clearly lower than for our benchmark comparisons, given the smaller sample and noisiness of the WWII card estimates, but the correlations are sizable and close to the full-stability benchmarks.

We also revisit the geographic/culture comparisons. First, comparing between the India and US sample, but for the extra-card WWII card coding task[17], we find a correlation of 0.72, close to the full-stability benchmark of 0.76. Second, we consider a different measure of geographic and cultural differentiation, comparing between Mturkers with an IP address in "red states" versus "blue states", attributing a state depending on the winner of the vote share in the 2016 presidential election. We estimate a very high correlation of 0.96, close to the full-stability benchmark of 0.94. These results further reinforce the message that the results are stable to demographic and geographic variation.

Next, we consider a different measure of output: instead of using the mean effort by treatment, we take the 25th and 75th percentile of effort. Within the a-b task, in treatments where the 25th percentile worker exerts high effort, so does the 75th percentile worker. In Online Appendix Figures 6 and 7 and Online Appendix Table 5 we present the pure replication, the gender comparison, and the task comparison using the 25th or 75th percentile of effort, instead of the average effort. This comparison replicates the high correlation for the first two comparisons and estimates a low rank-order correlation around 0.3 to 0.4 for the task comparison.

Finally, we consider two further forms of sample selection which do not fit neatly into either of the other dimensions, but which have been identified 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 worker motivation; and (ii) whether the subjects perform the test during the day or during the night. Comparing the results along these two dimensions for our three tasks, we find rank-order correlations that are close to the full-stability benchmarks, providing another example of stability of results.

4.3 Structural Estimates

We now present estimates of the model in Section 2.1.1 with four purposes: (i) to quantify the elasticity of effort in the various designs; (ii) to present an alternative measure of stability, stability of the underlying structural parameters; (iii) to form out-of-sample predictions for the 16th treatment, which we have not discussed so far; and (iv) to create a full-stability benchmark for design changes in which such benchmark cannot be created with bootstraps from the data.

Estimation. We take the model in Section 2.1.1 with an exponential cost of effort function, which conveniently implies a specification that expresses effort as function of the motivation

[16] We cannot make this comparison for the extra-card WWII coding 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.

[17] We pool across Versions 3 and 4. We do not do such comparison for the 10-minute WWII task given the noisiness of the estimates, since the Indian workers constitute only 12% of the sample.

parameters. To bring the model to the data, we need to specify the source of heterogeneity. Building on DellaVigna et al. (2015), we assume that the heterogeneity across subjects j takes form $c_j(e_j) = exp(k - \gamma\varepsilon_j)exp(\gamma e_j)\gamma^{-1}$, with ε_j normally distributed $\varepsilon_j \sim N(0, {\sigma_\varepsilon}^2)$. This assumption ensures positive realizations for the marginal cost of effort. This implies the first-order condition $s + p - exp(k - \gamma\varepsilon_j)exp(\gamma e_j) = 0$ and, taking logs and transforming,

$$e_j = \frac{1}{\gamma}[log(s + p) - k] + \varepsilon_j. \tag{3}$$

Equation (3) can be estimated with non-linear least squares (NLS). The three parameters $\hat{s}$, $\hat{k}$, and $\hat{\gamma}$ are over-identified given the four piece rates. 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 task as the number of cards coded, including the required 40 cards. We incorporate the behavioral treatments as discussed in Section 2.1.1.

In Table 4 we present estimates of the parameters using all the 15 treatments. Since our estimation allows for one parameter for each behavioral treatment, 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 4 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 NLS model on, respectively, the button-pressing data for 2015 and for 2018. The estimates for the 2015 experiment replicate the ones in DellaVigna and Pope (2018a) and are close to the estimates for the 2018 experiment: in both data sets, the elasticity of effort is precisely estimated to be about 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 estimate for the social comparison treatment $\Delta\hat{s}_{SC} = 0.06$ indicates an impact equivalent to an incentive of 0.06 cents per 100 presses. Indeed, this treatment, which is the most effective of the psychological treatments, is clearly less effective than the low-piece rate treatment, which we code as an incentive $p = 1$. There is no evidence that the paying-too-little treatment crowds out motivation, and thus $\Delta\hat{s}_{CO} \approx 0$.

We estimate a precisely-estimated 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. 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 (i) there is no response in worker effort to the return to the charity, but (ii) subjects work harder when there is a charitable giving, compared to the baseline condition.

The probability weighting parameter in 2015 is estimated to be $\pi(0.01) < 0.01$, that is, *underweighting* of small probabilities, while in 2018 we estimate $\pi(0.01) = 0.01$. In neither case do we find overweighting of small probabilities.

Estimates, Demographics. We pool the 2015 and 2018 a-b task data and present estimates

split by gender (Columns 3 and 4), education (Columns 5 and 6) and age (Columns 7 and 8).[18] There are some differences across the groups in the incidental parameters, though 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. 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, are very consistent across demographics.

Estimates, 10-min. WWII Task. For the 10-minute WWII task (Columns 9), we estimate an elasticity smaller than 0.01, really tiny, consistent with the very limited response to incentives.[19] 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.

Estimates, Extra Work. For the extra-work WWII coding (Columns 10 and 11), we estimate the model by maximum likelihood, accounting for censoring at 0 cards coded and at 20 cards coded. Just as we expected based on the results in Abeler et al. (2011) and Gneezy et al. (2017), the elasticity of effort to incentives, 0.45 in Column 10, is among the highest in the literature (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 relatively high elasticity implies that this design yields good statistical power for the behavioral estimates. Figures 2d-e show that the structural estimates capture well the curvature observed effort under the different piece rates. The estimates for the behavioral parameters are in line with the estimates for the other designs, except for a much larger gift exchange parameter.

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. The bottom row in the table shows that the model does quite well in the prediction.

Full-Stability Benchmark. We also use these estimates to compute a structural full-stability benchmark, as opposed to a bootstrap-based benchmark. For this structural benchmark, we assume that the behavioral parameters remain constant across design changes, but that the incidental parameters change. Consider for example the task change. As we summarize in Section 2.3, we compare a simulated sample from the estimated a-b task parameters with a simulated sample that combines the structural parameters from the a-b task and the incidental parameters from the WWII task. This second combination is the full-stability counterfactual for the WWII task: the effort we would observe if the task had the same behavioral parameters as in the a-b task—e.g., full stability—but its own cost of effort function and noise term. We draw 1,000 such samples, each with number of observations by treatment matching the actual ones. As Column 1 in Table 2 shows, the mean structural full-stability rank-order correlation for the task change is 0.50 (s.e. 0.19), in fact slightly *lower* than the observed correlation. Thus, the relatively low stability for the task change is entirely explained by the noise in the WWII task.

We similarly do a structural full-stability benchmark for the output comparisons. For the com-

[18] The Online Appendix Table 6 reports the estimates on the pooled 2015 and 2018 sample.

[19] We impose an upper bound for $\gamma = 2$ and the estimate converges to this upper bound. Without a bound, the estimator achieves a slightly better fit for even higher values of γ (that is, lower elasticity), but convergence is poor.

parison between the two WWII cards treatments, the estimated structural full-stability benchmark is 0.58 (s.e. 0.17), indicating that the observed low correlation of 0.27 is largely (but not fully) explained by the noise. For the comparison between the a-b task and the extra-work WWII task, the full-stability benchmark is instead much higher at 0.85 (s.e. 0.07), not surprisingly since both tasks are characterized by relatively low noise. To provide further evidence on this case, in Online Appendix Figure 8b we plot the actual treatment effects for the extra-work WWII task versus the ones implied under full stability from the structural estimates in the a-b task. (In this case, we take a large sample of subjects per treatment to illustrate the role of mean deviations, as opposed to the role of noise.) The results for 13 of the 15 treatments are remarkably lined up, with two treatments (probability weighting and gift exchange) clearly off the line.

A legitimate question is whether the structural full-stability benchmarks are similar to the bootstrap-based ones for the cases in which it is possible to obtain both. Table 2 shows that the two benchmarks are nearly identical for the pure replication, the demographics comparison and the geographic comparison, validating the structural measure.

4.4 Summary: Predictors of Stability

To summarize the results, we compare the predictive power of forecasts versus of the full-stability benchmark. In Figure 9a we plot for each of the 10 design changes the average expert forecast of correlation versus the actual correlation. In Figure 9b, instead, we plot the full-stability benchmark versus the actual correlation including not only the 10 main comparisons in Table 2, but also (with a different dot size) the additional comparisons in Table 3. As the figures make clear, the expert forecasts display only a weak correlation with the measured stability, while the full-stability benchmark is a very strong predictor. In our setting at least, the behavioral findings appear to be really stable (provided one adjusts for noise), more than experts expect.

5 Revisiting the Forecasts

We return to the expert forecasts to further probe some of the findings and interpretations.

Impact of Noise on Stability. The high degree of noise in the 10-minute WWII task largely explains the lower stability across tasks and output measures. The forecasters do not anticipate this pattern but, in fairness, it was not obvious that the 10-minute WWII card task would be much noisier than the other designs. Would the experts respond to information on noise if they had it?

To address this issue, we randomized the provision of additional information. For one half of the forecasters, we provided the mean effort (and s.e.) under the three key piece rate treatments, indicating a flat and non-monotonic response to incentives in the 10-minute WWII task, and in contrast a clear and monotonic response in the extra-work WWII task.

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). The forecasters respond very little to the additional information. 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, effort. As we document in DellaVigna and Pope (2018b), 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). We thus split the forecasters depending on whether they clicked on at least one link with additional information on the experimental design (Columns 4 and 5) and by the time taken to do the survey (Columns 6 and 7). Under either dimension, we find little evidence that effort is correlated with higher accuracy.

Confidence. A relevant question too is whether confidence in the forecasts predicts accuracy, as it does, to some extent, in DellaVigna and Pope (2018b). The forecasters indicated how many of their forecasts they expected to be within 0.1 of the truth. Forecasters with higher confidence (Column 9 versus 8) are on average closer to the truth for the pure replication, the impact of demographics, and of consent. They, however, have a mixed record on the output forecasts.

Figure 10 presents more detailed evidence on confidence as predictor of accuracy, showing 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. While accuracy does increase with the confidence, the slope is too flat. In particular, individuals with higher confidence overstate their accuracy.[20] 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.

Vertical Expertise and Wisdom-of-Crowds. In DellaVigna and Pope (2018b) we showed that there was no obvious impact on forecast accuracy of "vertical expertise"—faculty did not do better than PhDs—and that there was a large "wisdom-of-the-crowds" effect—the average forecast outperformed 97 percent of individual forecasts. We strongly replicate the first result: as the bottom of Table 2 shows, PhD forecasters do slightly better than faculty forecasters in accuracy. As for the second result, while the wisdom-of-crowd accuracy is higher than for the average of individual forecasts, the difference is smaller, and 42 percent of the 55 expert forecasters outperformed the wisdom-of-crowd forecast. The smaller wisdom-of-crowd advantage is likely due to the smaller dispersion of forecasts in this setting (e.g., the model in DellaVigna and Pope, 2018b).

Superforecasters. In DellaVigna and Pope (2018b) forecasters who do a better job forecasting a group of treatment also have higher accuracy in forecasting other groups of treatments, as in the superforecasters literature (Tetlock and Gardner, 2015; Mellers et al., 2015). But does this forecasting ability translate across experiments? For the 35 individuals who made forecasts both in 2015 and in 2018, in Figure 11a we compare their average absolute error (in terms of point) in 2015 and (in terms of rank-order correlation) in 2018. We find no evidence of positive correlation between accuracy across the two experiments. We conjecture that reliably detecting superforecasters will require tracking forecasts over a larger sample of forecasters and experiments.

Explaining the 2015 Forecasts Errors. Finally, we reinterpret forecast errors in 2015 in light of the newer data. As we document in DellaVigna and Pope (2018b), while the average (i.e.,

[20] Online Appendix Figure 9 displays the same evidence with respect to the absolute forecast error.

wisdom-of-crowd) forecast for a treatment generally does a good job of predicting the average effort in that treatment, the average forecast is sizably off for some treatments: it under-predicts effort in the very-low-pay treatment and over-predicts 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 examine if the treatments where experts had the larger forecast error in 2015 are more aligned in the new 2018 runs with the original 2015 forecasts. The x axis in Figure 11b indicates the average forecast error in 2015 for a treatment, while on the y axis we plot, for each of the four 2018 versions of the experiment, how much a treatment shifted in rank from the 2015 experiment to the 2018 experiment. The probability weighting treatment, which experts had overpredicted in 2015, indeed moves up by 3, 4, 5, and 6 ranks in the four 2018 runs compared to the 2015 results. However, the very-low-pay treatment does not move down in ranks, as one would predict based on the 2015 forecast error. All in all, we find 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 experiment—a real effort task with a dozen 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 experimental 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 design changes. One needs a measure of stability that accounts for the role of noise, as well as the fact that design changes may alter the units of measure of the results. We proposed rank-order correlation, in comparison to a full-stability benchmark, as a measure of stability with desirable properties. 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 motivation 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, 2018) 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*, Vol. 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*, Vol. 2(1), 1–12.

Bandiera, Oriana, Greg Fischer, Andrea Prat and Erina Ytsma. 2018. "Do women respond less to performance pay? Building evidence from multiple experiments" Working paper.

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman. 2010. "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment," *Quarterly Journal of Economics*, Vol. 125(1), 263–306.

Bhargava, Saurabh and Day Manoli. 2015. "Psychological frictions and the incomplete take-up of social benefits: Evidence from an IRS field experiment," *American Economic Review,* Vol. 105(11), 3489-3529.

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

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

Camerer, Colin 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*, 1-15.

Christensen, Garret S., and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature,* Vol. 56, 920-980.

DellaVigna, Stefano. 2018. "Structural Behavioral Economics" in *Handbook of Behavioral Economics*, Volume 1 (eds. Doug Bernheim, Stefano DellaVigna, and David Laibson), Elsevier.

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. 2018a. "What Motivates Effort? Evidence and Expert Forecasts", *Review of Economic Studies*, Vol. 85, 1029–1069.

DellaVigna, Stefano, and Devin Pope. 2018b. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy,* Vol. 126, 1410-2456.

de Quidt, Jonathan, Johannes Haushofer, Christopher Roth. 2018. "Measuring and Bounding Experimenter Demand," *American Economic Review,* Vol. 108, 3266-3302.

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(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), 535-538.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science*, Vol. 345(6203), 1502-1505.

Gerber, Alan S. and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment," *American Political Science Review*, Vol. 94(3), 653-663.

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), 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), 1049–1074.

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

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

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

Klein et al. 2014. "Investigating Variation in Replicability: A "Many Labs" Replication Project." *Social Psychology*, Vol. 45(3), 142-152.

Klein et al. 2018. "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings." *Advances in Methods and Practices in Psychological Science*, Vol. 1(4).

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(2), 443-477.

Landy, Justin et al. 2018. "Crowdsourcing Hypothesis Test: Making transparent how design choices shape research results" Working paper.

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,* Vol. 10(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(5), 411-419.

Prelec, Drazen. 1998. "The Probability Weighting Function." *Econometrica,* Vol. 66(3), 497-527.

Snowberg, Erik and Leeat Yariv. 2018. "Testing the Waters: Behavior across Participant Pools," Working paper.

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

Tversky, Amos and Daniel Kahneman. 1971. "Belief in the Law of Small Numbers." *Psychological Bulletin.* 76(2), 105-110.

Vivalt, Eva. 2017. "How Much Can We Generalize from Impact Evaluations?" Working paper.

Figure 1. Expert Forecasts, CDFs

Figure 1a. Forecasts of Replication and Demographics

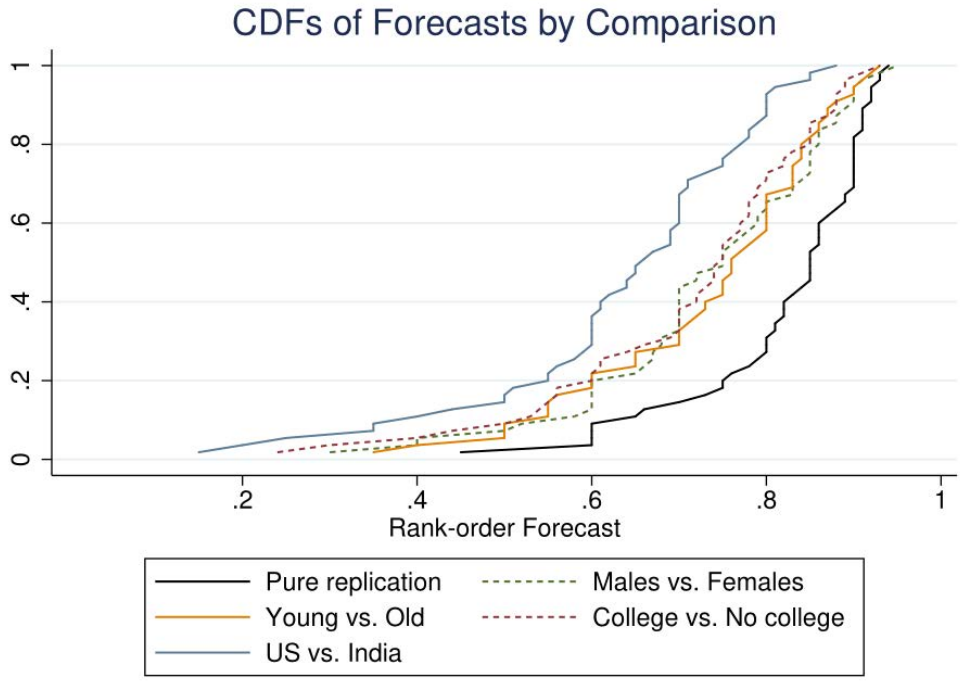
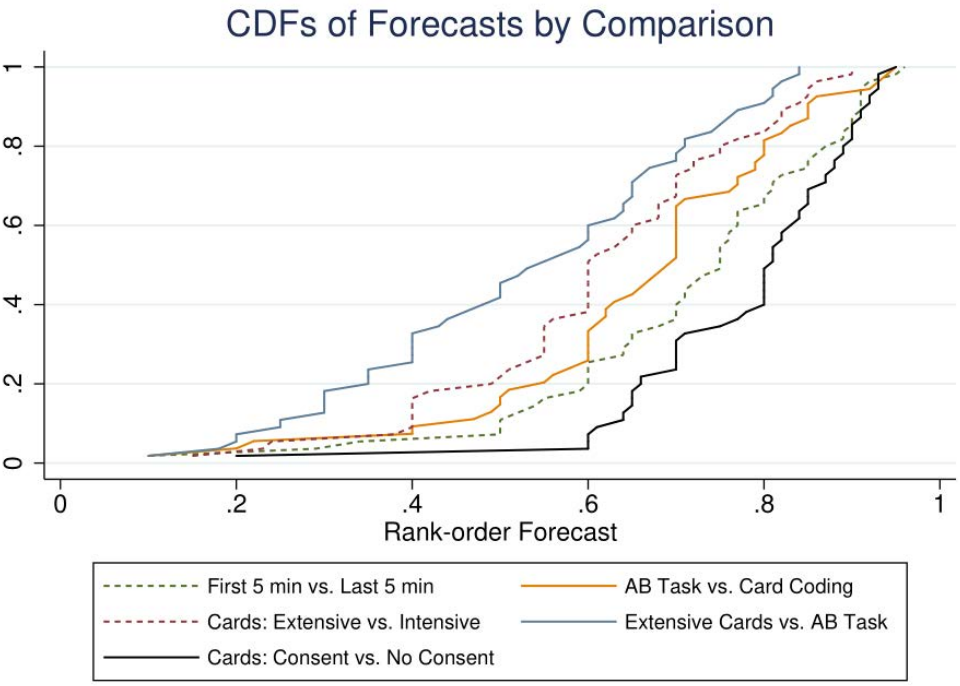


Figure 1b. Forecasts of Output, Task, and Context



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

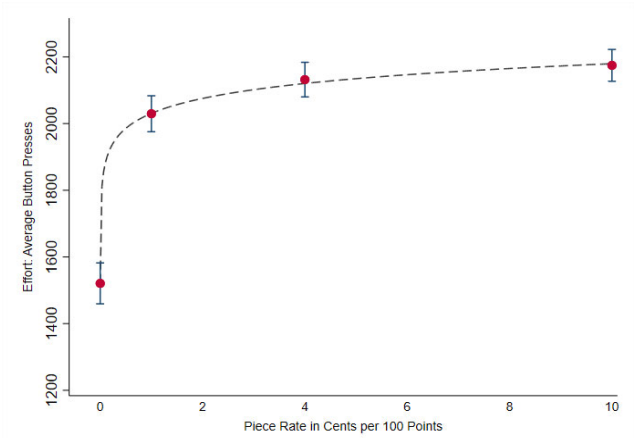


Figure 2b. 2018 Button Pushing Task

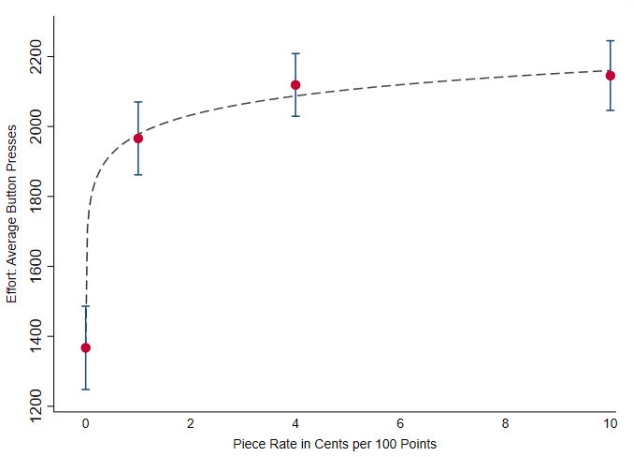


Figure 2c. 2018 10-Minute WWII Card Coding Task

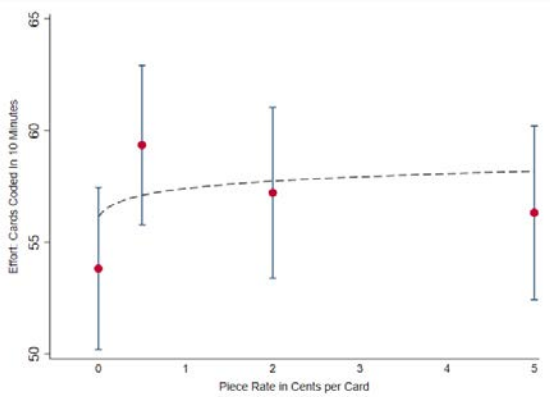


Figure 2d. 2018 Extra Card Coding Task

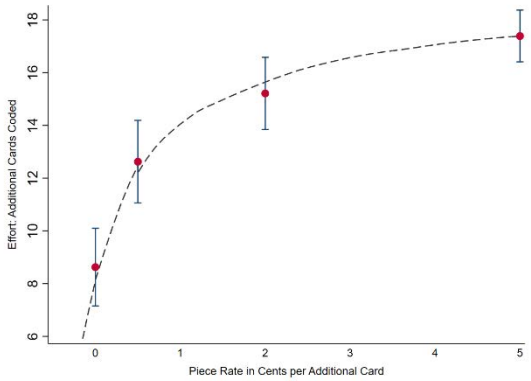
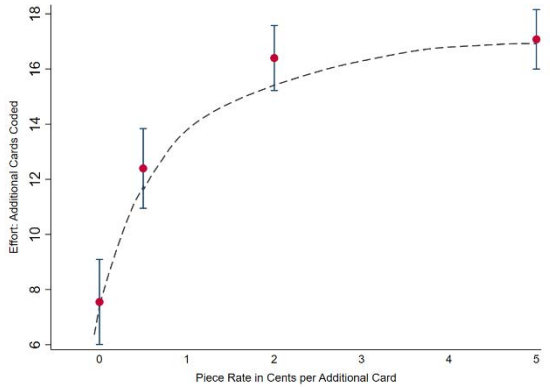
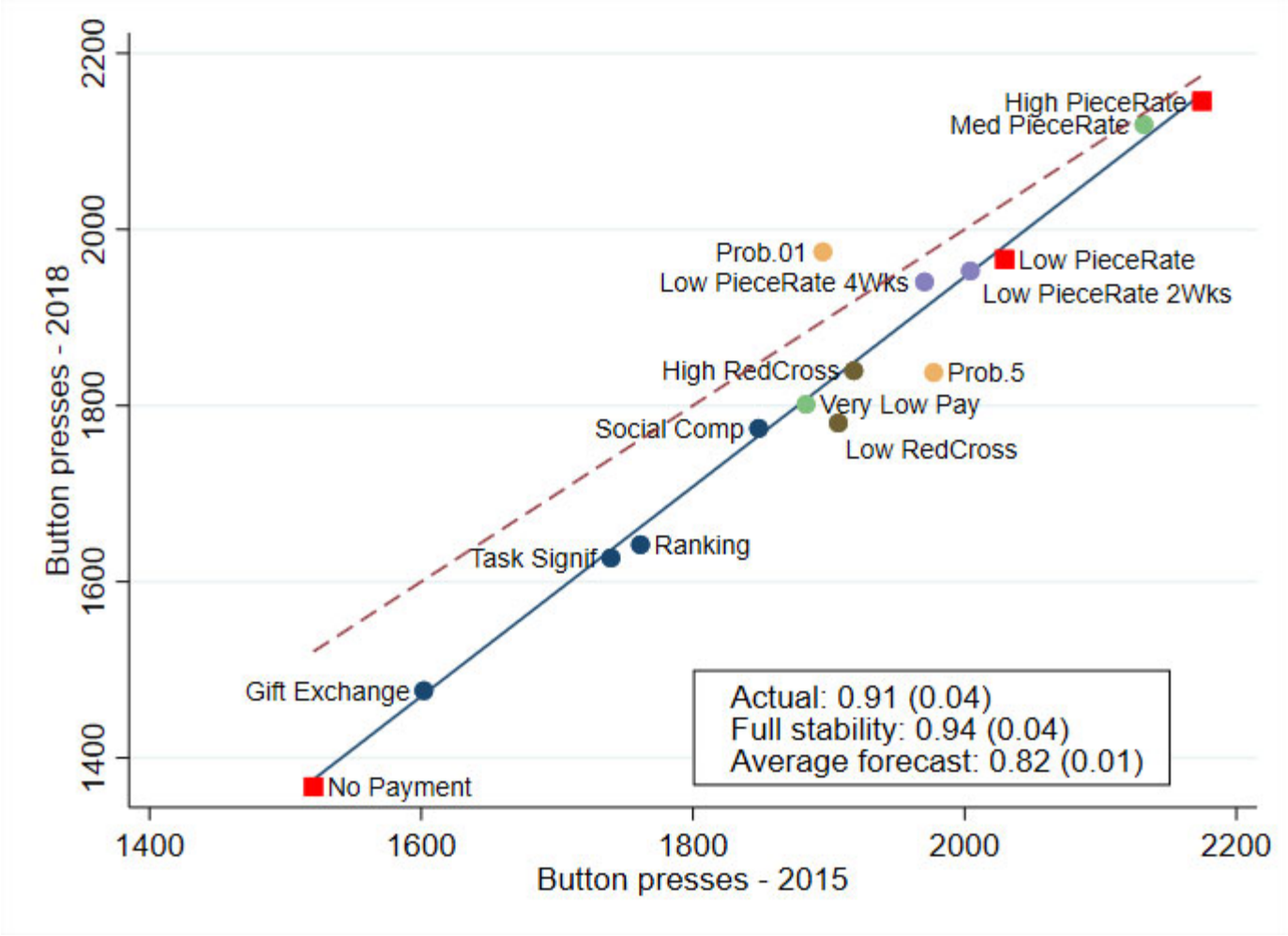


Figure 2e. 2018 Extra Card Coding Task, No Consent



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

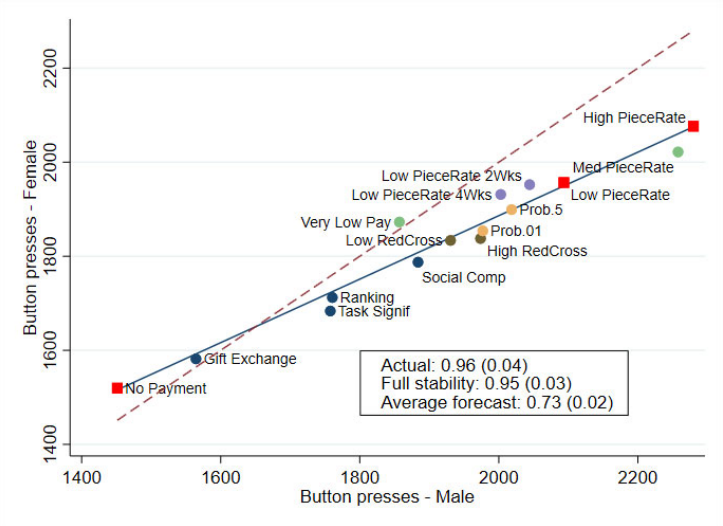


Figure 4b. Education

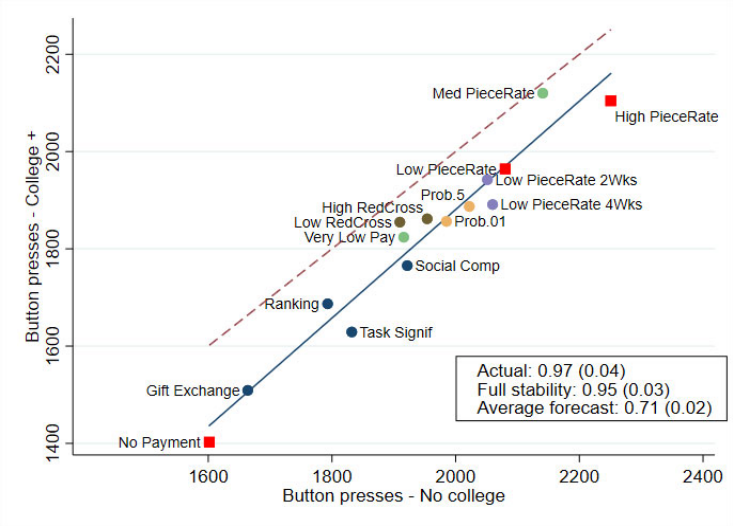
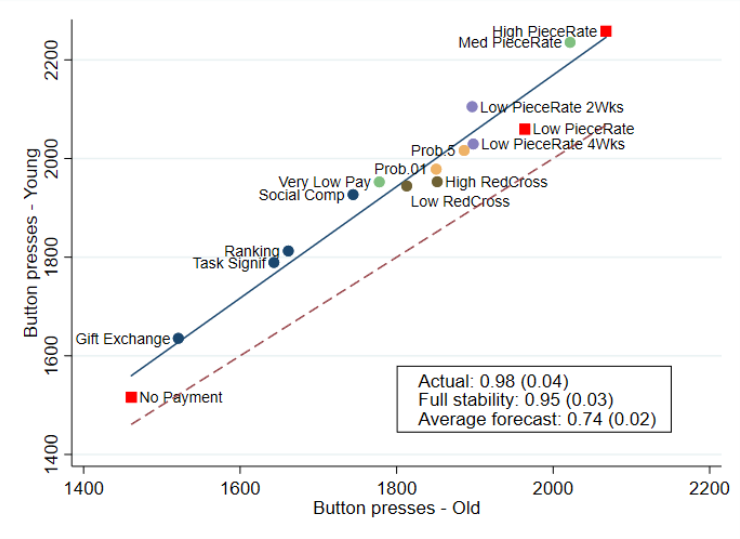
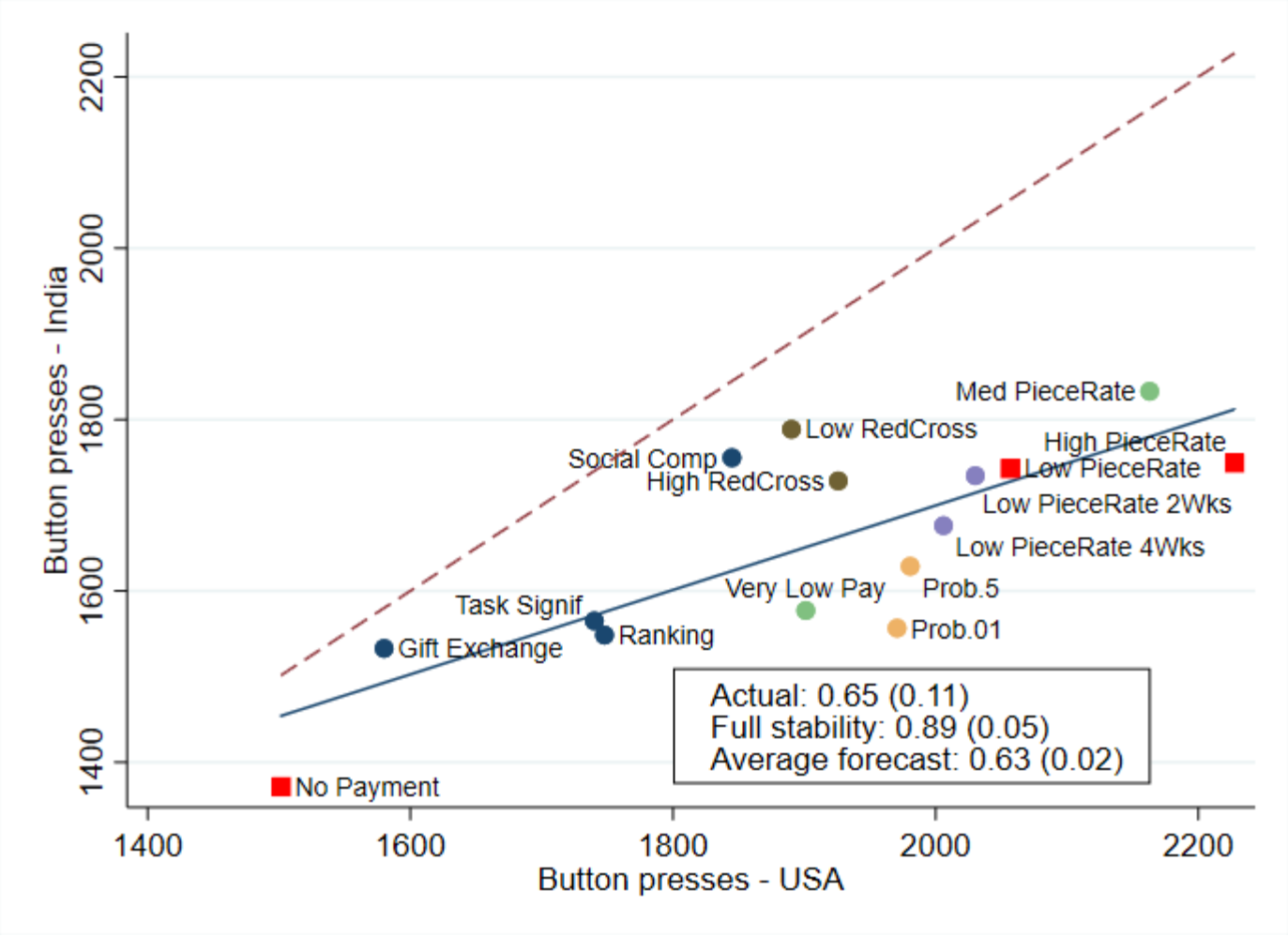


Figure 4c. Age



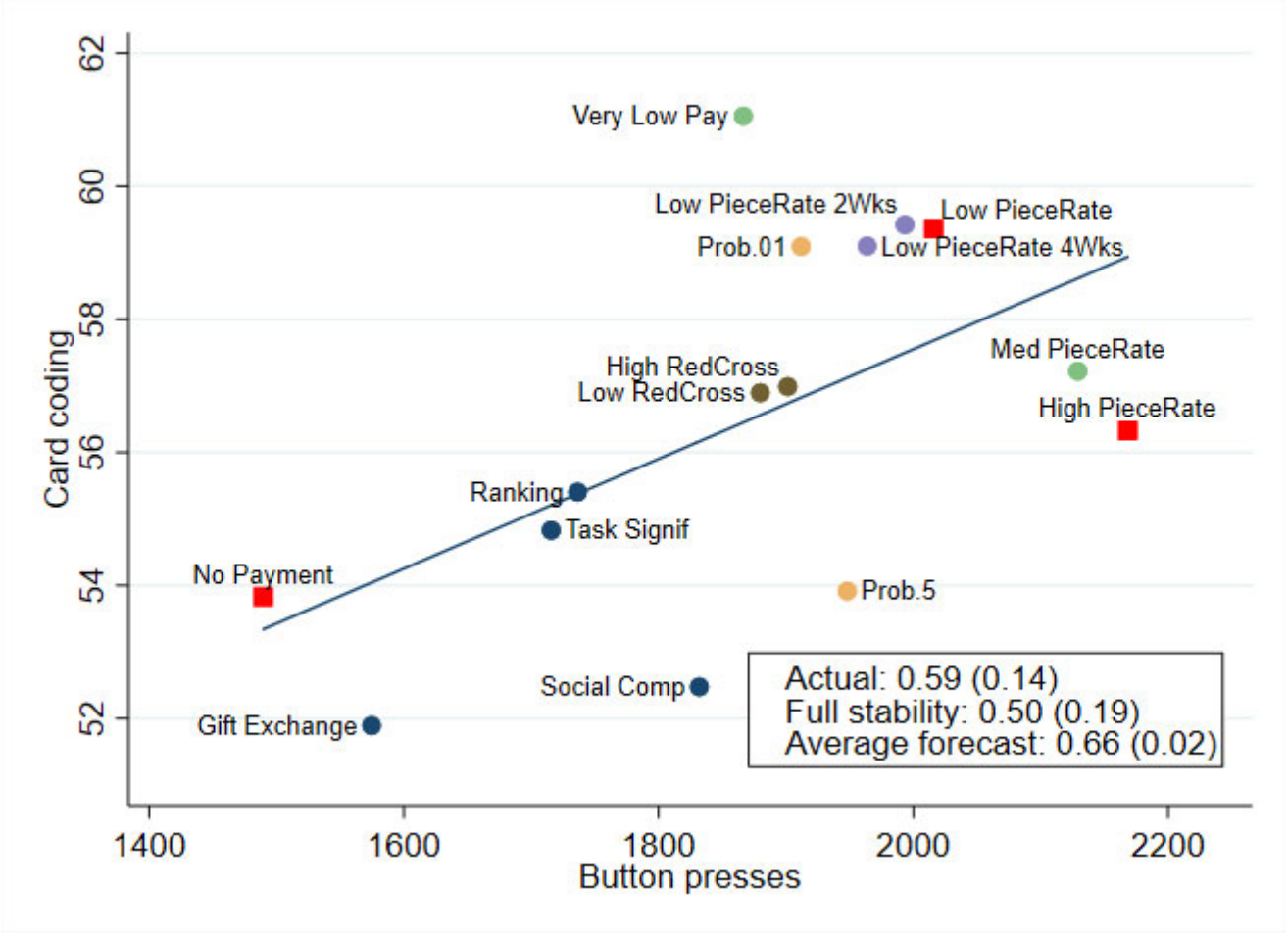
Notes: Figures4a-c display, for each one of 15 treatments, the average effort for the button pushing task (pooling the 2015 and 2018 experiments) across different demographics of the subjects, splitting by gender (Figure 4a), by education (Figure 4b), and by age (Figure 4c). See notes to Figure 3 for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

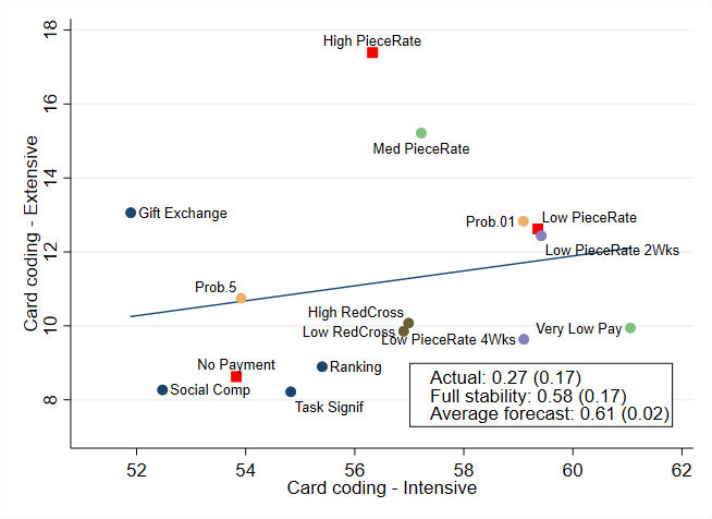


Figure 7b. WWII Coding, Ext. Margin vs. Button Pushing Task

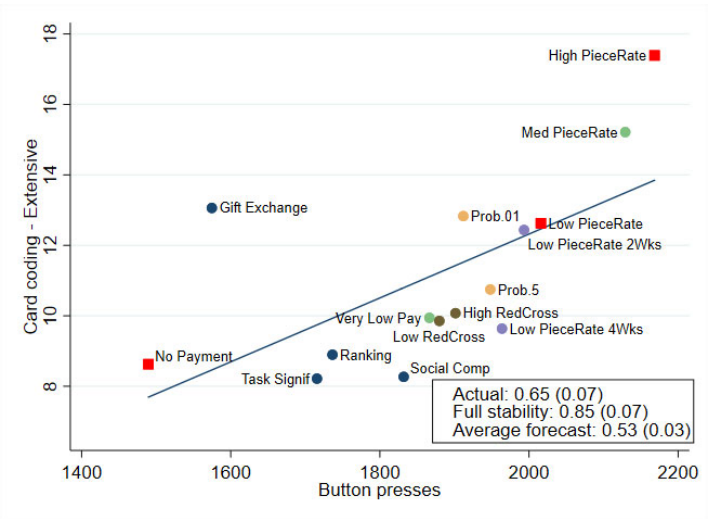
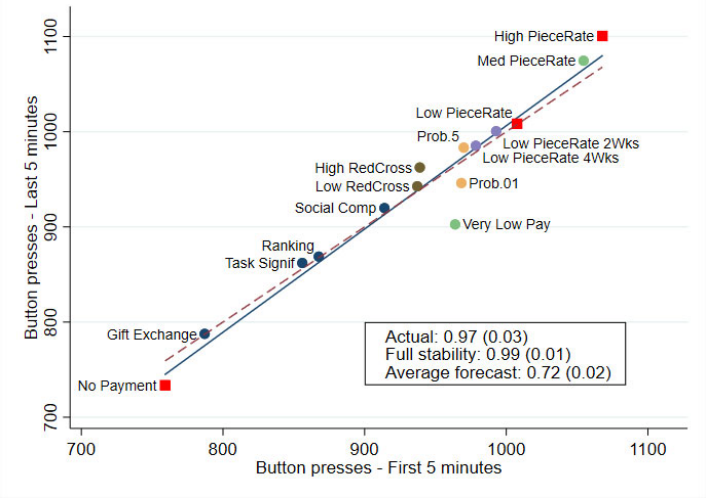
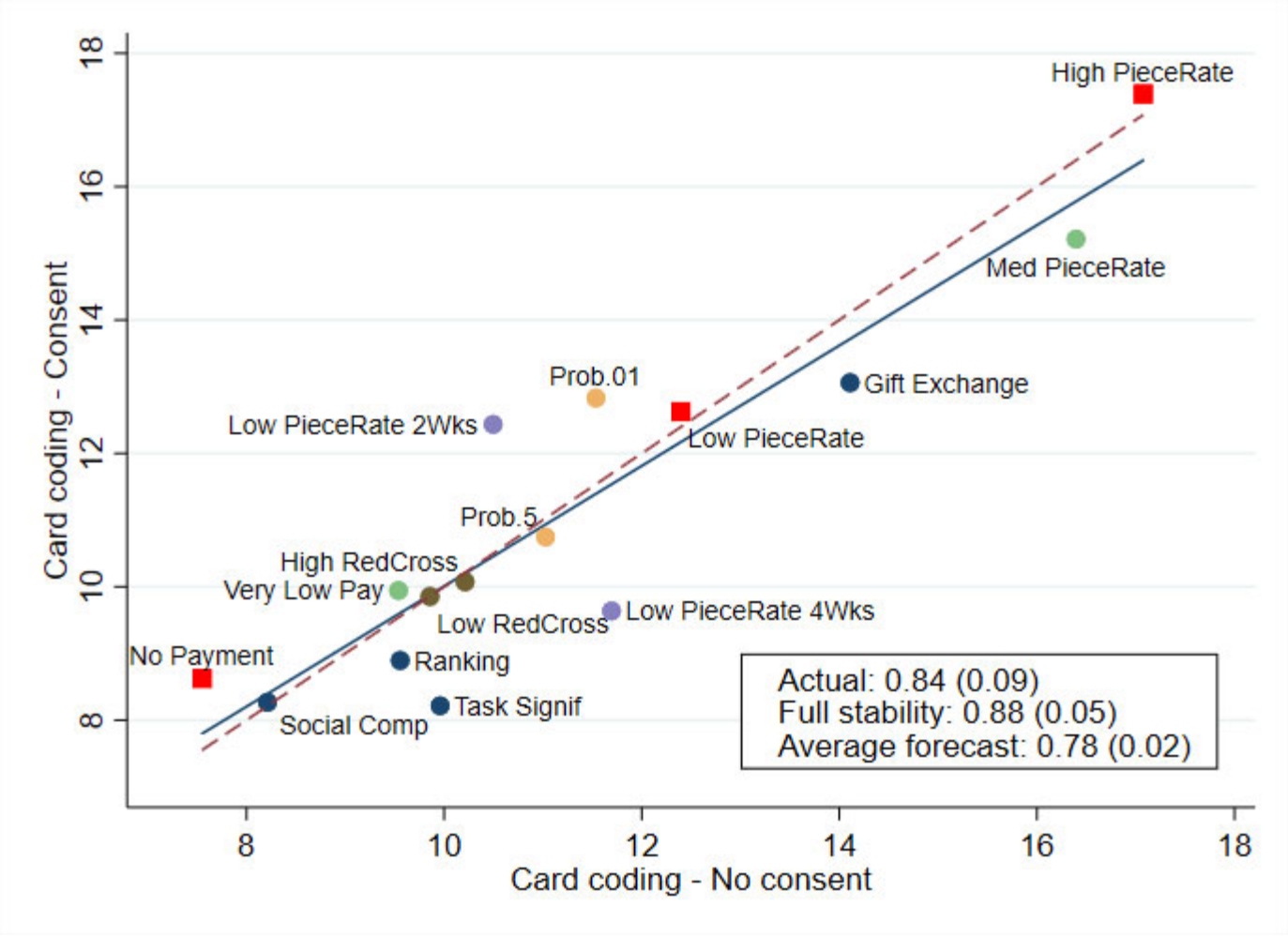


Figure 7c. Output in First 5 Minutes vs. Later 5 Minutes, Button Pushing Task



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. Actual Rank-Order Correlation, Average Forecast, and Full Stability Benchmark

Figure 9a. Actual Rank-Order Correlation and Average Expert Forecast

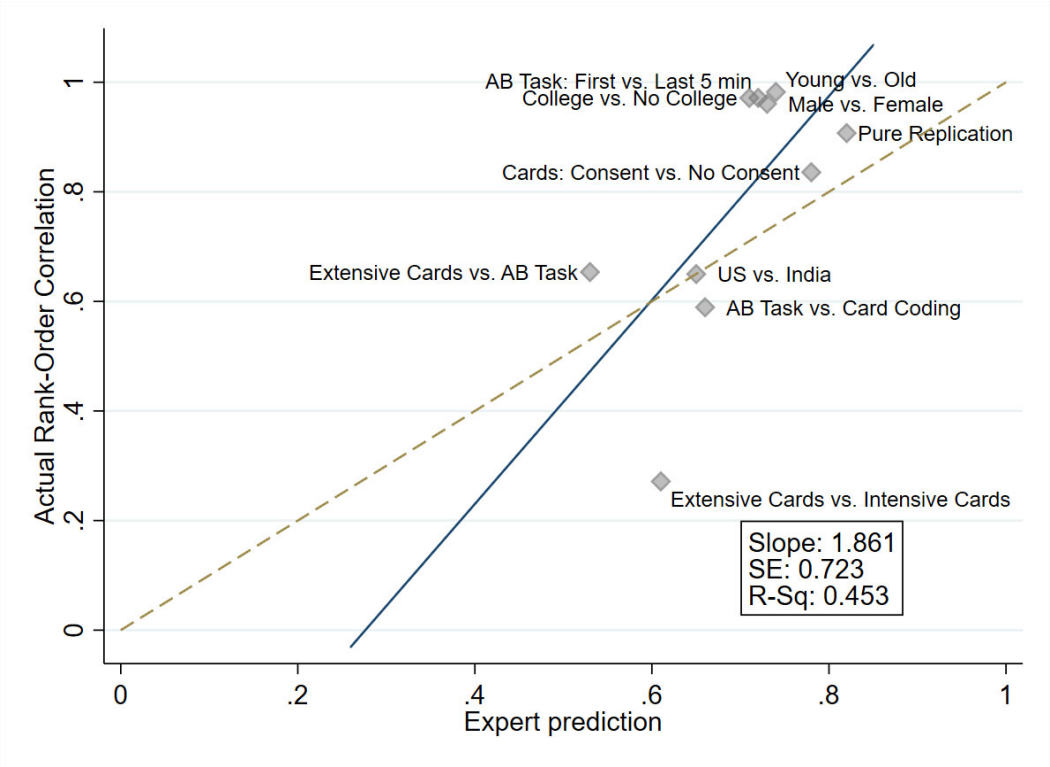
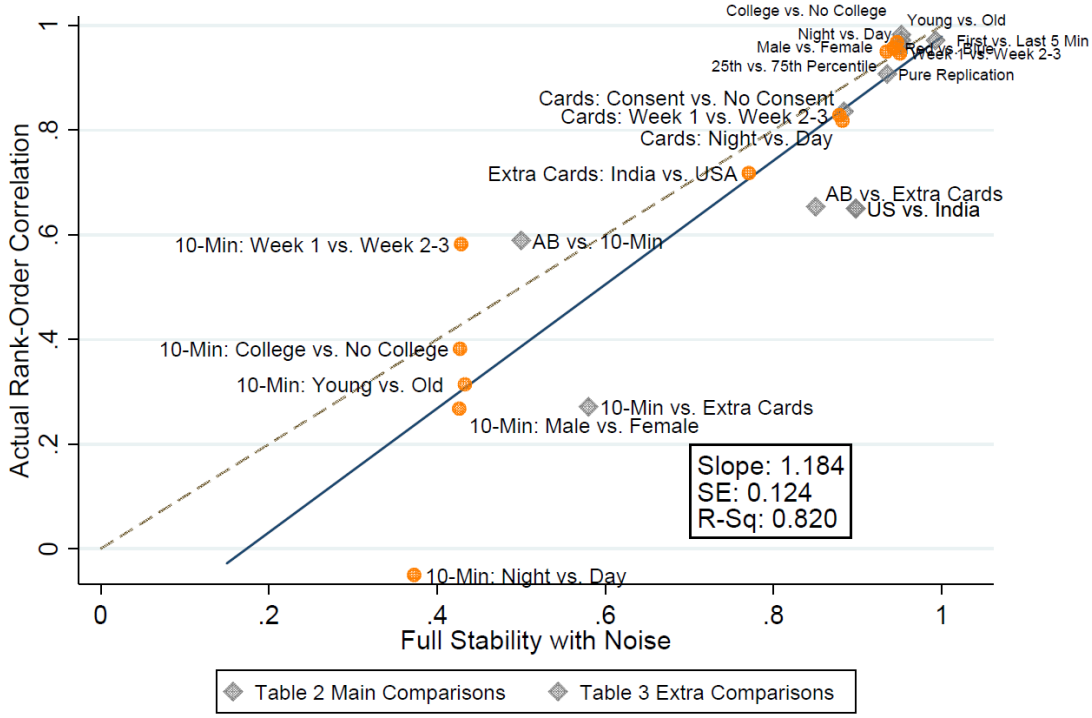
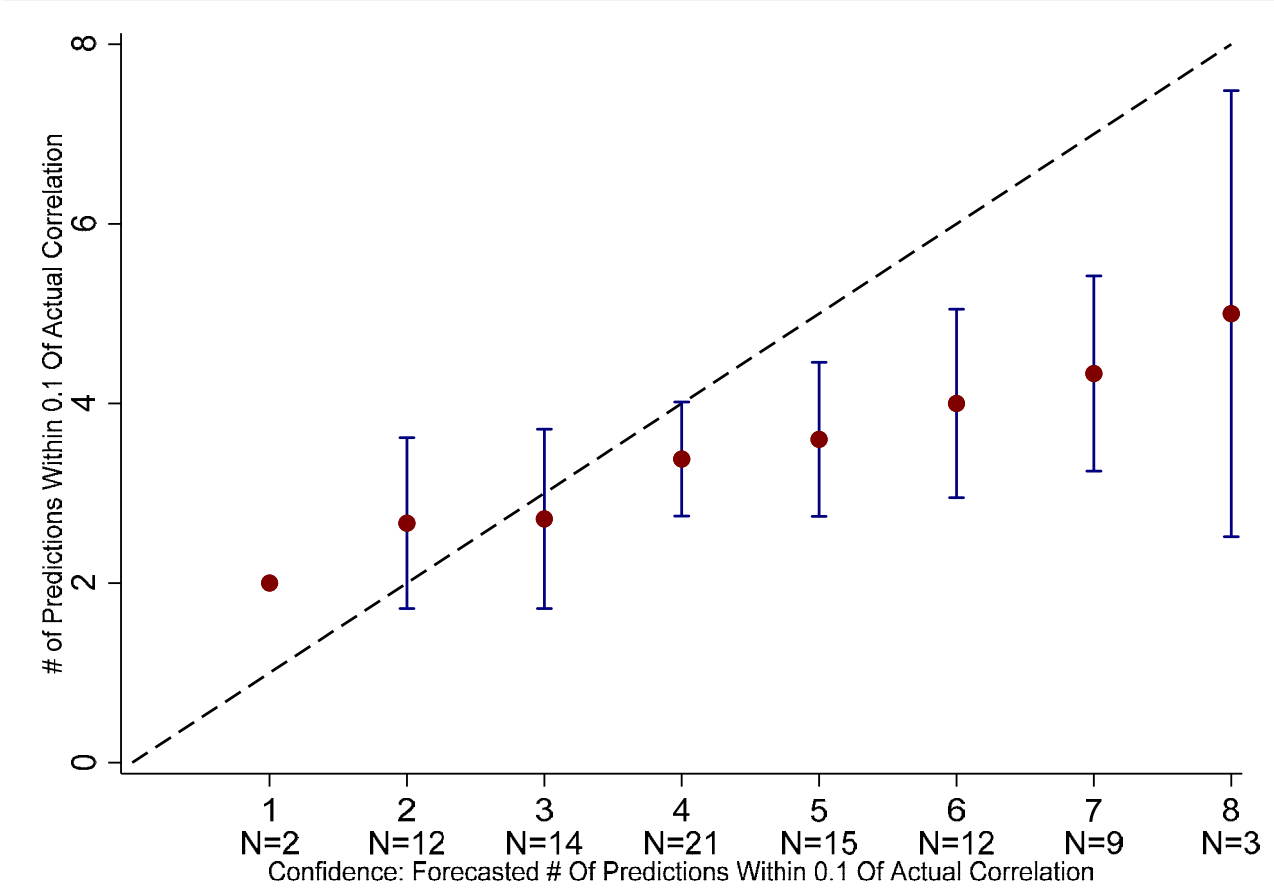


Figure 9b. Actual Rank-Order Correlation and Full Stability Benchmark



Notes: Figure 9a 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 9b presents the actual rank-order correlation versus the full-stability benchmark. In Figure 9b we plot both the 10 benchmark version changes (Table 2) as well as the additional version changes (Table 3).

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

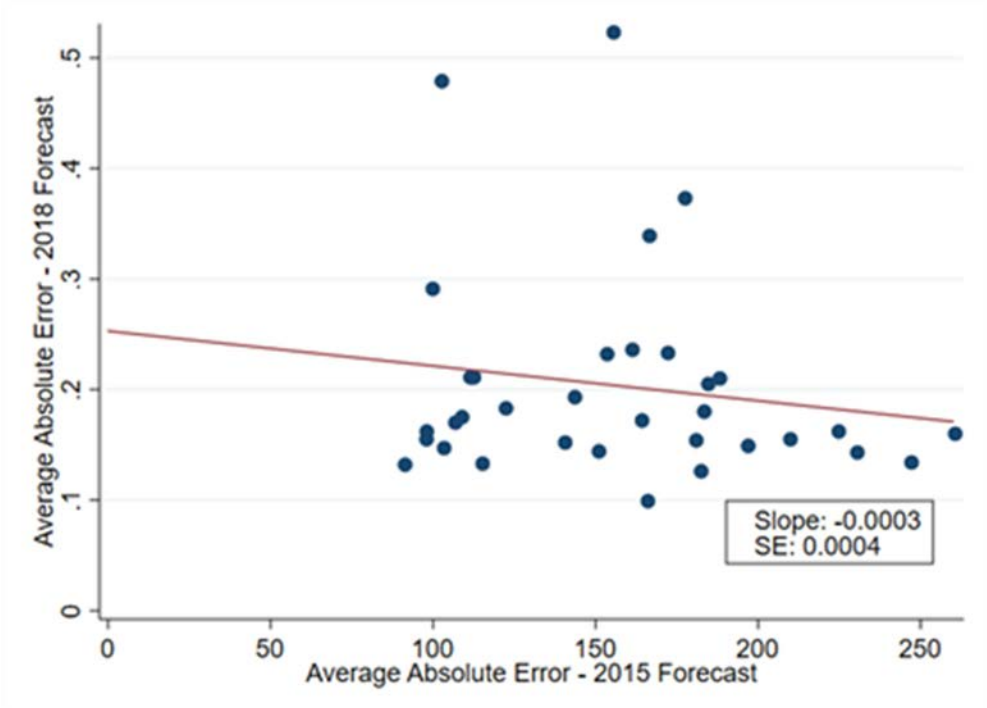
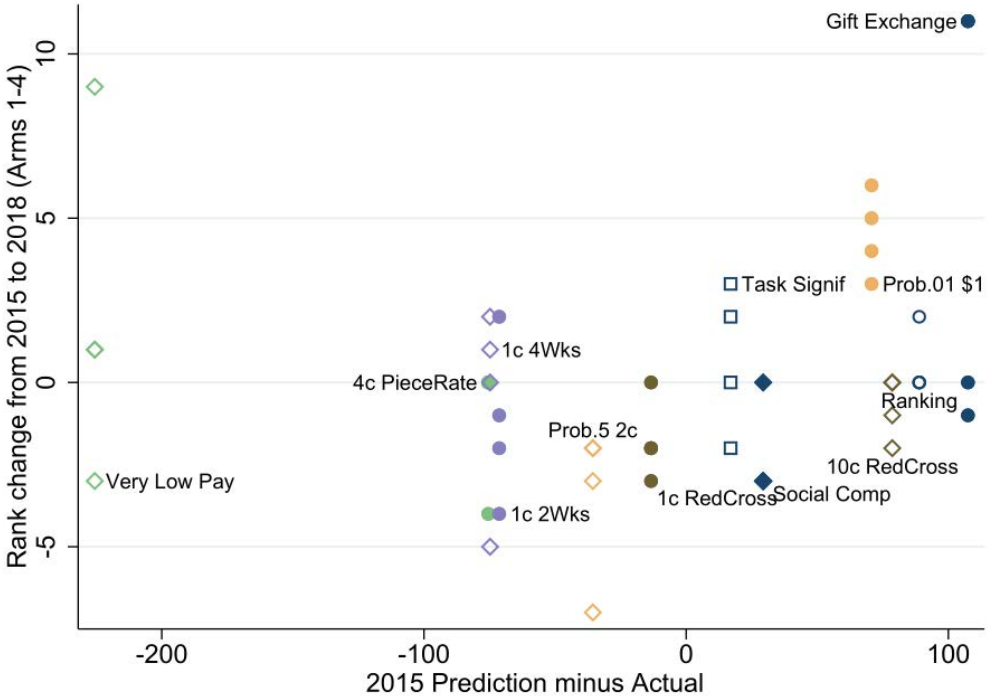


Figure 11b. Errors in 2015 Forecasts and Changes of Treatment Rank in 2018 Experiments



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 vs. Full-Stability

Category	Design Comparison	Rank-Ord. Correl. Full-Stability		Average Forecast of Rank-Order Correlation			Rank-Ord. Correl.	p-value for Difference		
		Boostrap from Data	Structural	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)	(9)
Pure Repl.	2015 AB Task vs. 2018 AB Task (n=8,252; n=2,219)	0.94 (0.04)	0.94 (0.03)	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)	0.94 (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)	0.94 (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)	0.89 (0.05)	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)	0.83 (0.08)	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=10,471; n=2,537)	-	0.50 (0.19)	0.66 (0.02)	0.63 (0.03)	0.64 (0.02)	0.59 (0.14)	0.392	0.569	0.619
Output	10-min Cards vs. Extra Cards (n=2,537; n=2,188)	-	0.58 (0.17)	0.61 (0.02)	0.61 (0.03)	0.62 (0.02)	0.27 (0.17)	0.831	0.202	0.052
	AB Task vs. Extra Cards (n=10,471; n=2,188)	-	0.85 (0.07)	0.53 (0.03)	0.56 (0.04)	0.58 (0.02)	0.65 (0.07)	0.000	0.166	0.110
	AB Task: First 5 min vs. Last 5 min (n=10,471)	0.99 (0.01)	0.96 (0.02)	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)	0.88 (0.06)	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 constitute the focus of the paper. For example, in row 1 we compare the estimate of effort in the 15 treatments in the a-b button pushing task, comparing the results in 2015 versus in 2018, using rank-order correlation of the average effort in the 15 treatments across versions as measure. In Columns 1-2 we report the average correlation under a benchmark of full-stability, that is, if the results do not change with the change in design, but allowing for noise in the eralized effort. This benchmark is derived from a data-based bootstrap in Column 1 while it uses strucural estimates of the parameters (see Table 4) in Column 2. Columns 3-5 report the average forecast of rank-order correlation for the population of academic experts (Column 3), PhD students (Column 4), and MTurkers (Column 5). Column 6 reports the actual rank-order correlation. Columns 7-9 report the p-value for the difference between the relevant columns. For the full-stability benchmark, we use the value the data-based bootstrap (Column 1) when available.

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)	0.72 (0.10)	0.782
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
Output, AB Task	25th Percentile vs. 75th Percentile (n=10,471)	0.93 (0.03)	0.95 (0.04)	0.730
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 Column 1 in Table 2) and in Column 2 we present the actual rank-order correlation.

Table 4. Structural Estimates

Category	Parameters	Button Pushing		Demographics, Typing Task, Pooled 2015-2018						2018 WWII Cards Coding Task		
		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 ɣ	0.016	0.012	0.012	0.019	0.015	0.014	0.011	0.022	2.000*	0.045	0.055
		(0.004)	(0.005)	(0.003)	(0.007)	(0.005)	(0.004)	(0.003)	(0.010)		(0.014)	(0.014)
	Implied Elasticity	0.034	0.044	0.043	0.028	0.035	0.037	0.046	0.025	0.009	0.430	0.354
		(0.008)	(0.017)	(0.012)	(0.009)	(0.011)	(0.011)	(0.010)	(0.011)		(0.134)	(0.087)
	Level of Cost of Effort *k*	-36.427	-29.183	-29.828	-42.500	-35.260	-33.447	-27.798	-48.420	-114.743	-3.470	-4.312
		(8.283)	(10.160)	(7.375)	(13.252)	(9.628)	(8.703)	(5.595)	(19.659)	(2.205)	(1.428)	(1.255)
	Baseline Motivation *s*	3.3e-04	5.1e-04	3.9e-04	2.2e-04	1.3e-04	0.001	0.002	1.4e-05	0.084	0.203	0.097
		(7.9e-04)	(0.002)	(0.001)	(7.3e-04)	(4.0e-04)	(0.003)	(0.004)	(7.6e-05)	(0.362)	(0.192)	(0.083)
Pay Enough or	Δs_{CO}	-0.005	0.011	-0.051	0.104	-0.015	0.017	0.156	-0.084	1.6e+05	0.068	0.052
		(0.099)	(0.177)	(0.066)	(0.244)	(0.110)	(0.134)	(0.197)	(0.041)	(6.9e+05)	(0.102)	(0.076)
Social Pref. Parameters	Pure Altruism *alpha*	0.003	0.010	0.009	8.7e-04	0.002	0.010	0.003	0.005	0.017	0.011	0.007
		(0.010)	(0.017)	(0.013)	(0.009)	(0.011)	(0.014)	(0.015)	(0.011)	(0.518)	(0.028)	(0.020)
	Warm Glow *a*	0.135	0.075	0.111	0.094	0.136	0.097	0.231	0.030	0.754	0.205	0.267
		(0.133)	(0.132)	(0.128)	(0.134)	(0.158)	(0.122)	(0.184)	(0.071)	(3.535)	(0.222)	(0.182)
Social Pref.: Gift Exch.	Δs_{GE}	8.4e-04	0.001	0.001	5.1e-04	5.3e-04	0.002	0.005	3.9e-05	-0.083	0.831	0.908
		(0.002)	(0.004)	(0.002)	(0.001)	(0.001)	(0.004)	(0.008)	(1.9e-04)	(0.360)	(0.348)	(0.356)
Discounting	*Beta*	1.075	0.855	0.769	1.422	1.171	0.692	3.320	0.222	228.893	5.395	0.215
		(1.122)	(1.306)	(0.931)	(1.861)	(1.435)	(0.852)	(3.112)	(0.450)	(2.1e+03)	(6.030)	(0.227)
	Delta (Weekly)	0.767	0.926	0.780	0.817	0.673	1.061	0.654	1.016	0.726	0.435	1.317
		(0.243)	(0.416)	(0.280)	(0.327)	(0.260)	(0.372)	(0.202)	(0.514)	(1.971)	(0.201)	(0.380)
Social Comparisons	Δs_{SC}	0.055	0.079	0.068	0.039	0.034	0.127	0.192	0.008	-0.079	-0.018	0.024
		(0.065)	(0.132)	(0.086)	(0.066)	(0.053)	(0.146)	(0.161)	(0.022)	(0.357)	(0.071)	(0.044)
Ranking	Δs_{R}	0.014	0.015	0.015	0.009	0.010	0.020	0.052	0.001	1.864	0.001	0.103
		(0.020)	(0.033)	(0.025)	(0.019)	(0.019)	(0.031)	(0.056)	(0.004)	(7.787)	(0.071)	(0.075)
Task Significance	Δs_{TS}	0.010	0.012	0.015	0.005	0.004	0.035	0.040	7.9e-04	0.533	-0.007	0.143
		(0.015)	(0.028)	(0.024)	(0.011)	(0.009)	(0.051)	(0.044)	(0.003)	(2.569)	(0.070)	(0.092)
Probability Weighting Parameters	Pi (0.01)	0.001	0.010	0.002	0.001	0.001	0.003	0.003	8.2e-04	0.626	0.013	0.005
		(0.001)	(0.008)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(2.518)	(0.006)	(0.003)
	Pi (0.50)	0.207	0.087	0.171	0.169	0.113	0.259	0.263	0.091	1.5e-04	0.177	0.249
		(0.147)	(0.121)	(0.145)	(0.171)	(0.113)	(0.201)	(0.164)	(0.138)	(0.005)	(0.127)	(0.135)
No. of Obs.		8252	2219	4686	5785	5842	4629	5212	5259	2537	2188	2246
Avg effort		1893	1836	1931	1839	1824	1951	1955	1806	56.51	11.19	11.31
Root MSE		659.20	643.40	724.14	591.92	666.19	637.69	686.06	618.03	24.21	56.23	51.91
Extra Treat.: Incentive + Please try	Out-of-Sample Pred.		1979							57.03	12.86	13.75
			(47)							(1.089)		
	Actual		2056							56.18	10.81	13.30
			(46)							(1.756)	(0.785)	(0.738)

Notes: The Table shows structural estimates of the incidental parameters (ɣ, 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, while Cols (10)-(11) are estimated with maximum likelihood due to censoring. Col (1) refers to the 2015 typing task, Col (2) to the 2018 typing task. Cols (2)-(8) pool the 2015 and 2018 typing tasks but restrict the sample to a demographic subset. Cols (9)-(11) show estimates on the 2018 card coding treatments. Standard errors in parantheses.

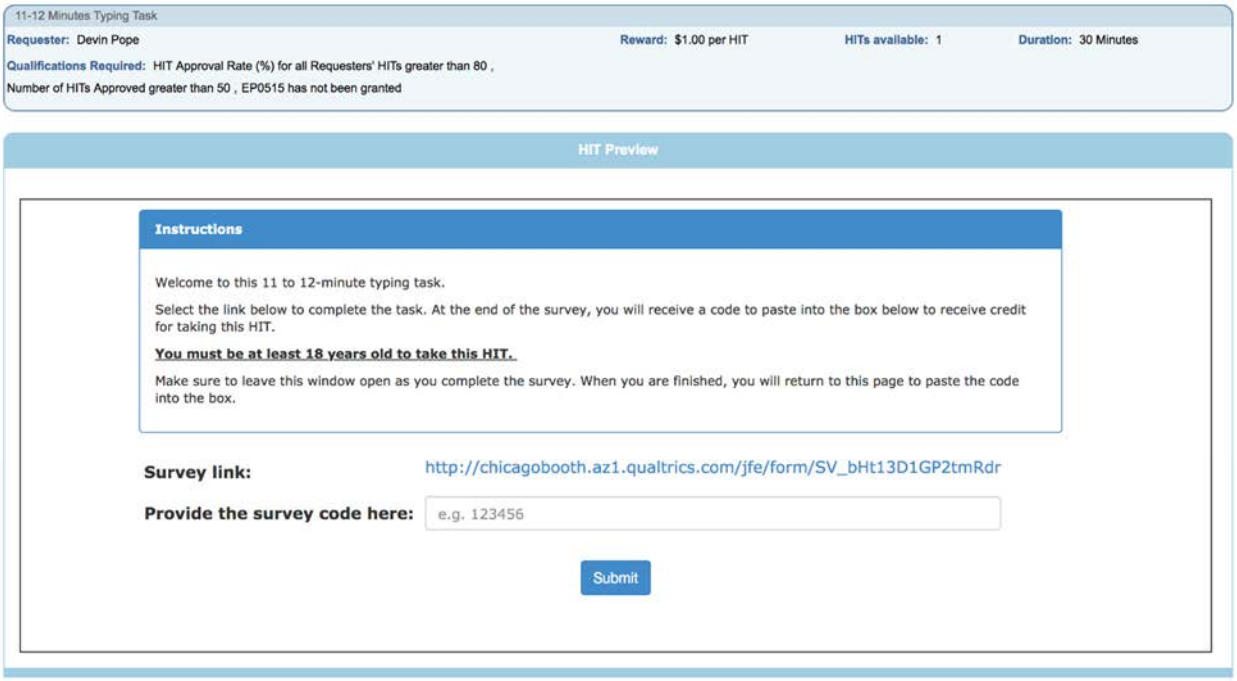
Table 5. Forecasts of Rank-Order Correlations by Different Forecasters

		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)	0.85 (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)	0.73 (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)**	**0.70 (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)	0.62 (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)**	**0.56 (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)**	**0.45 (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)	**0.74 (0.02)**	**0.66 (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)

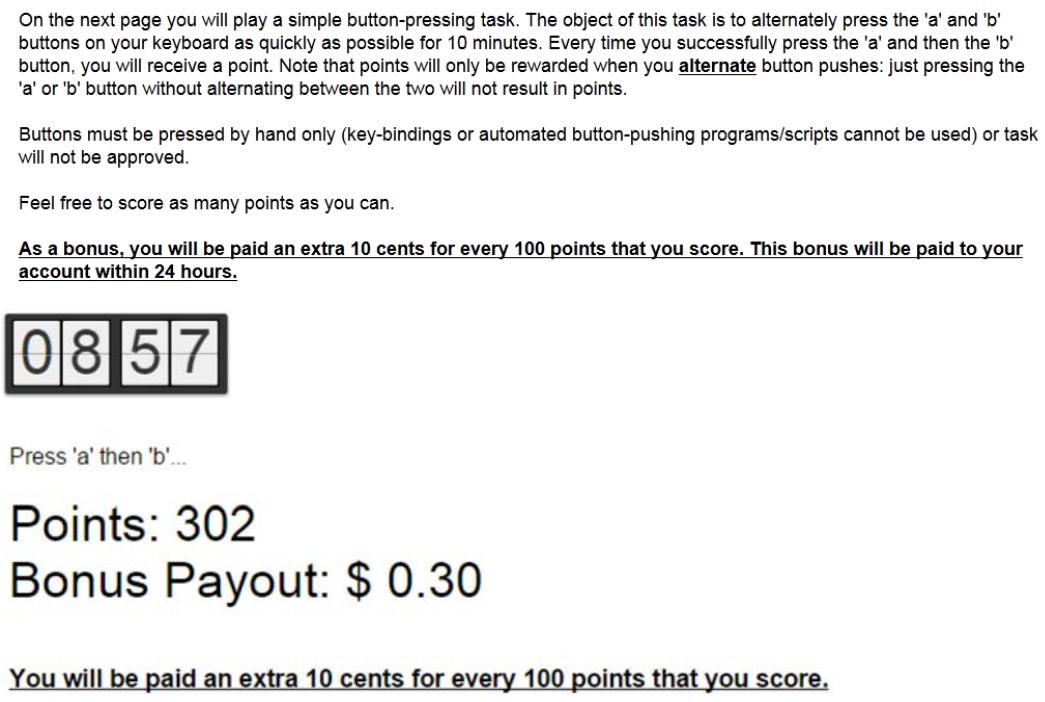
Notes: The Table considers the forecasts of sub-groups. Column 1 presents the results for the overall group of academic experts ad PhDs. In Columns 2 and 3 we split this group depending on whether the respondents were randomized to be provided information on the aerage effort by piece rate or not. In 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. Bolded values are significantly different from one another at the 95% confidence level.

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



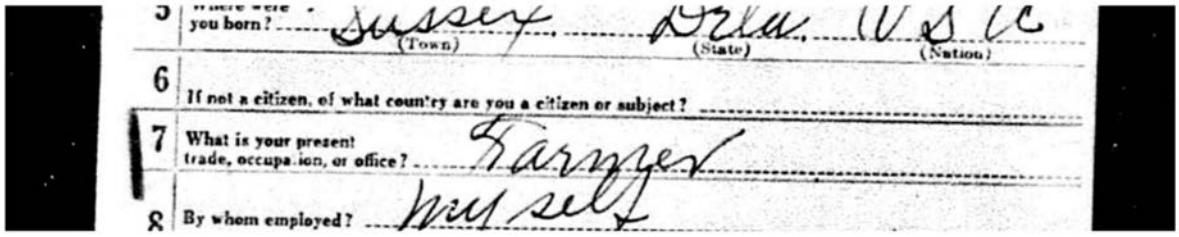
Online Appendix Figure 1c. Screenshot for WWII 10-minute Card Coding Task, Example

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.

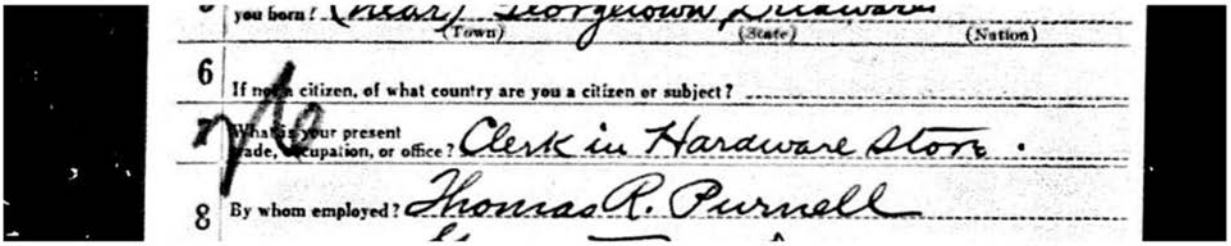
Type occupation here:

→

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

You have completed 3 of 40 required cards.

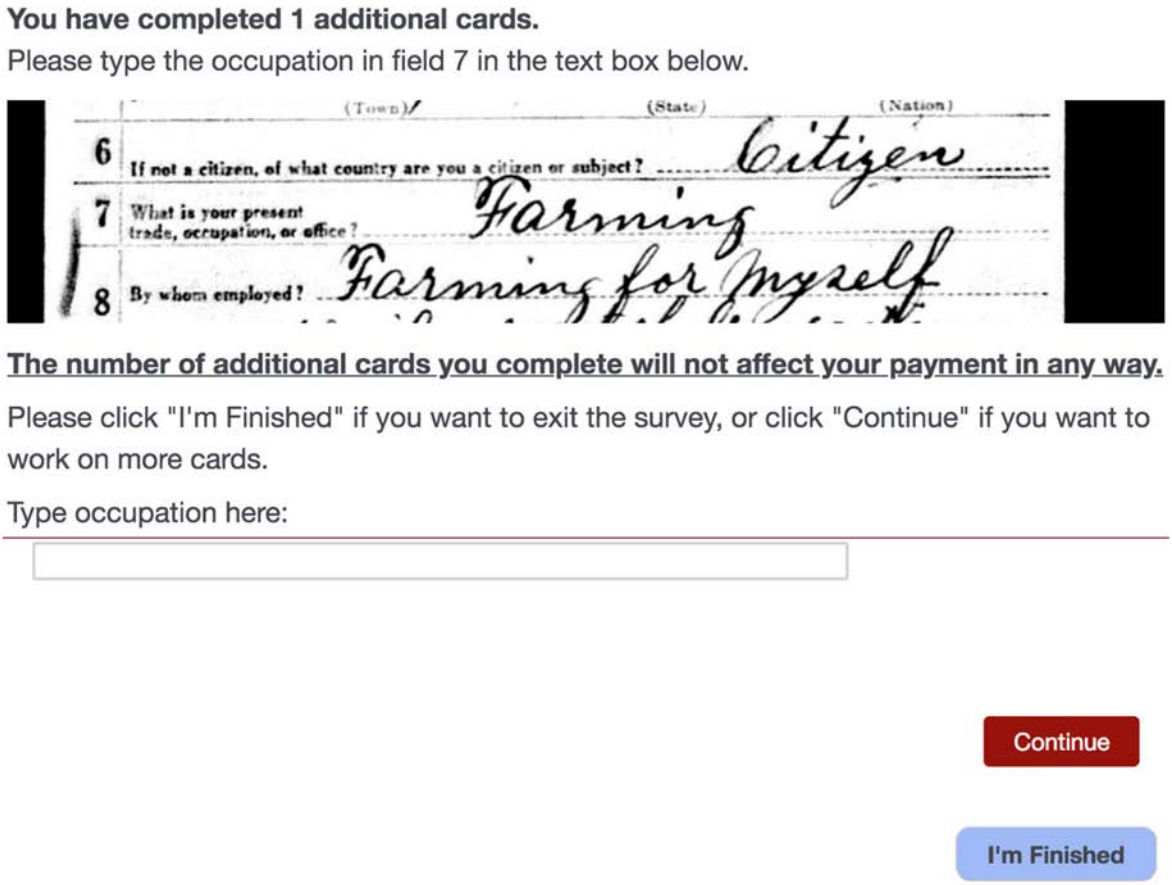
Please type the occupation in field 7 in the text box below.



Type occupation here:

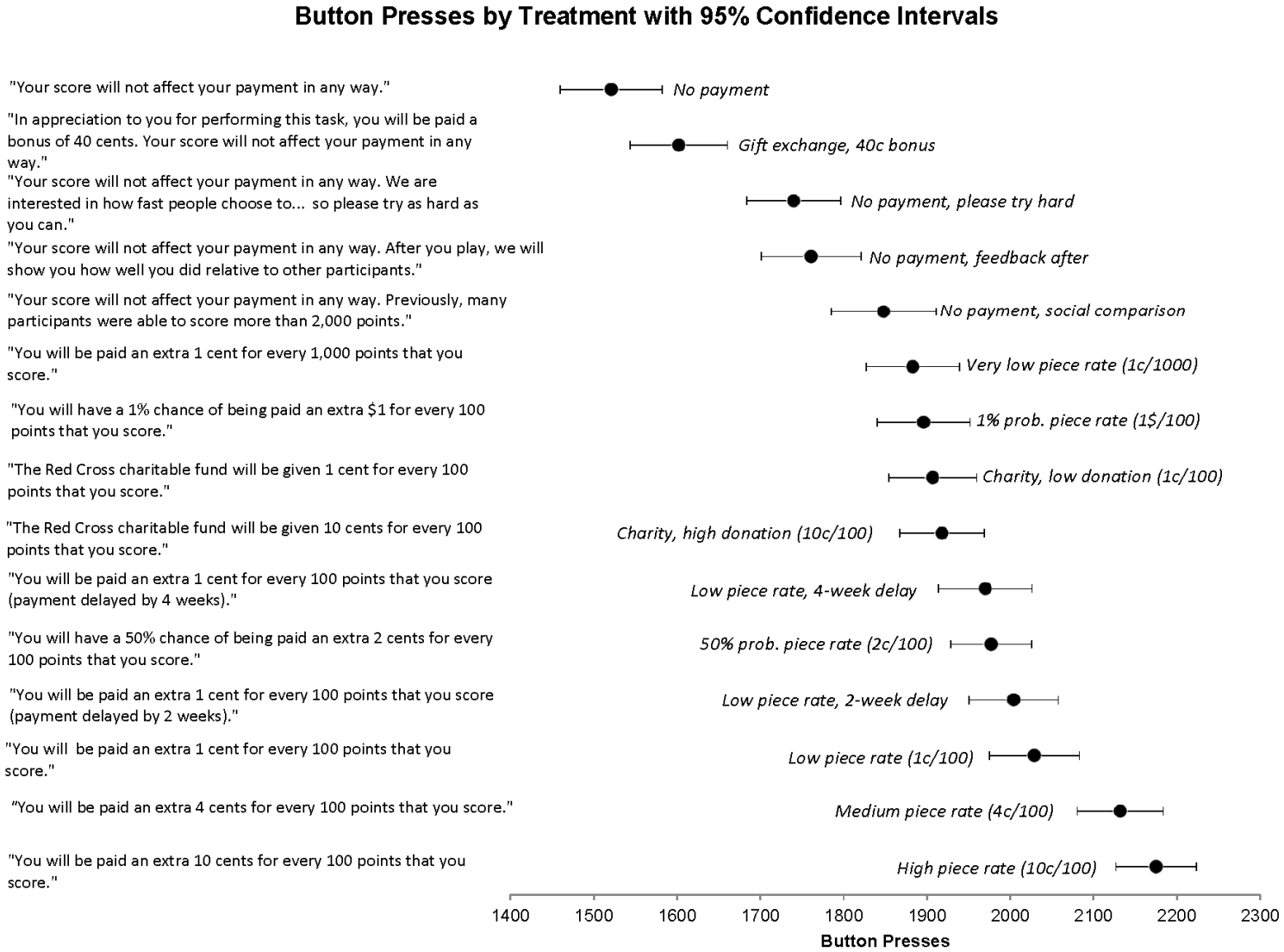
→

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.

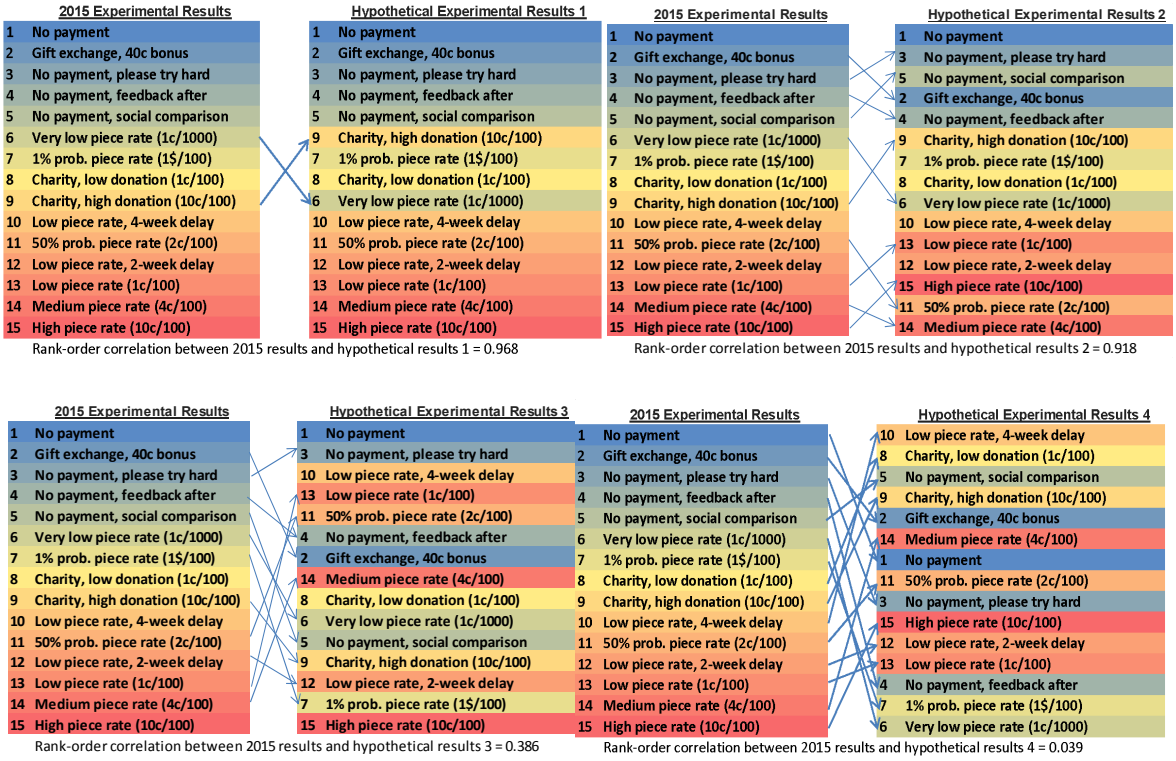
Online Appendix Figure 2. Summary of Treatments and Results from DellaVigna and Pope (2018)



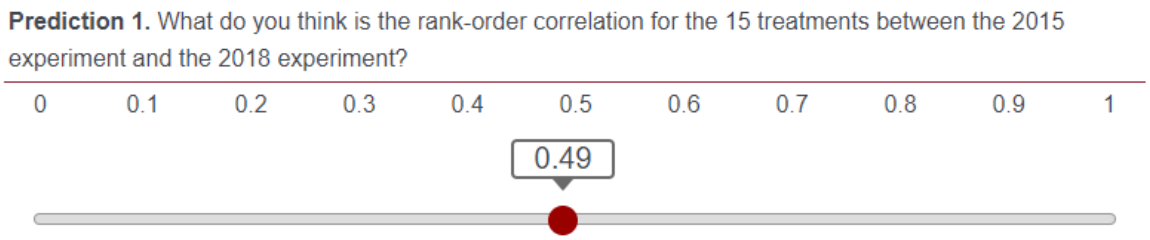
Notes: The figure summarizes the key wording as well as the average effort and standard error for the mean effort in the 2015 experimental results of DellaVigna and Pope (2018) for the 15 treatments which we replicate. This image is as presented to the forecasters.

Online Appendix Figure 3. Expert Survey, Screenshots

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



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 the four 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 4a. 2015 MTurk Button Pushing Task

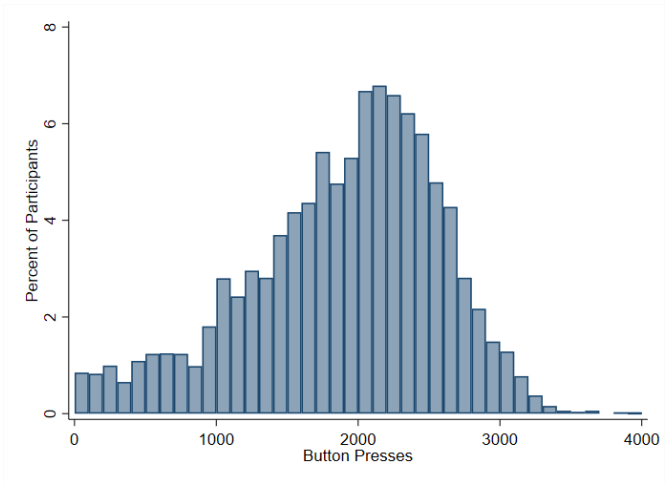


Figure 4b. 2018 MTurk Button Pushing Task

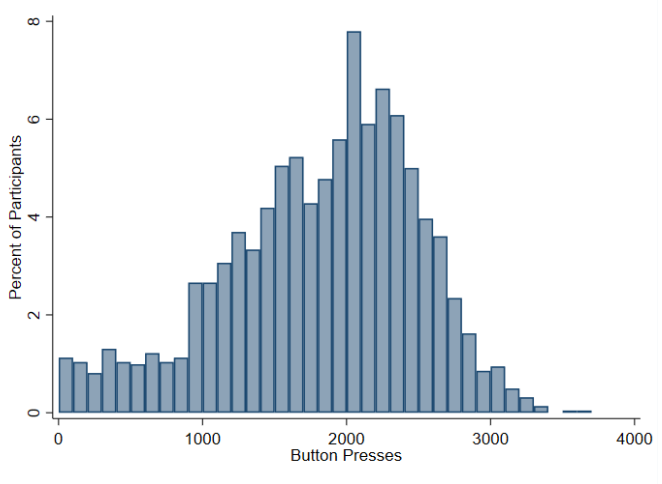


Figure 4c. 2018 10-Minute Card Coding Task

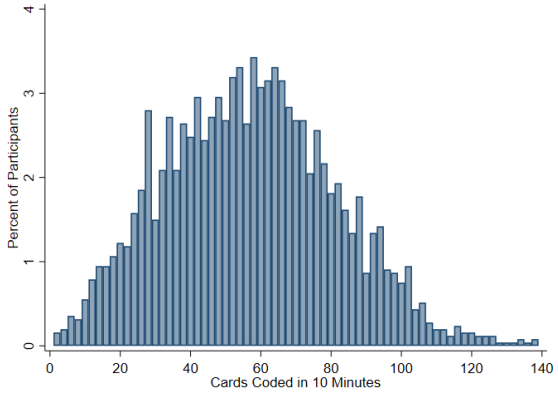


Figure 4d. 2018 Extra Card Coding Task

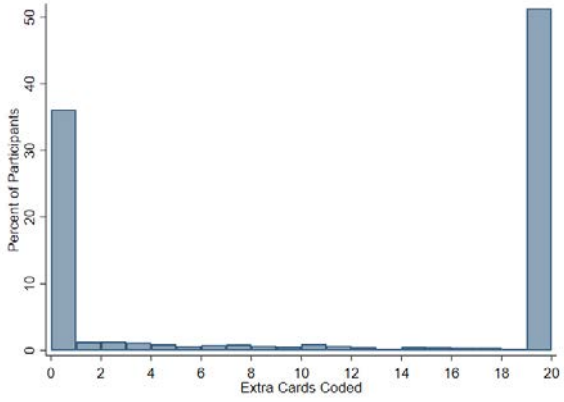
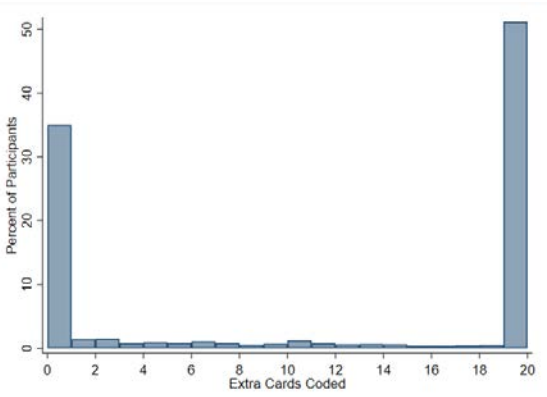
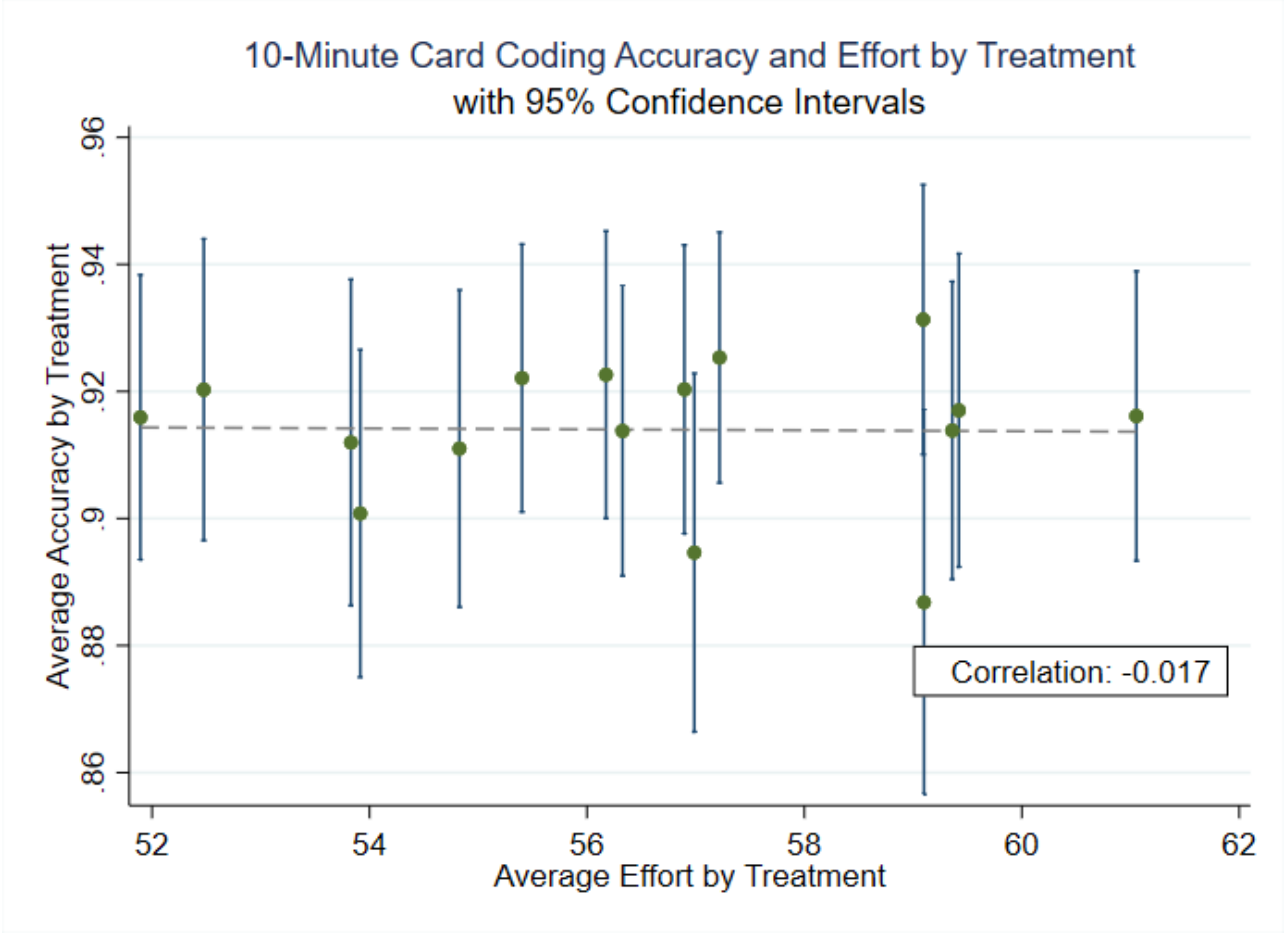


Figure 4e. 2018 Extra-Card Coding Task,No Consent



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.

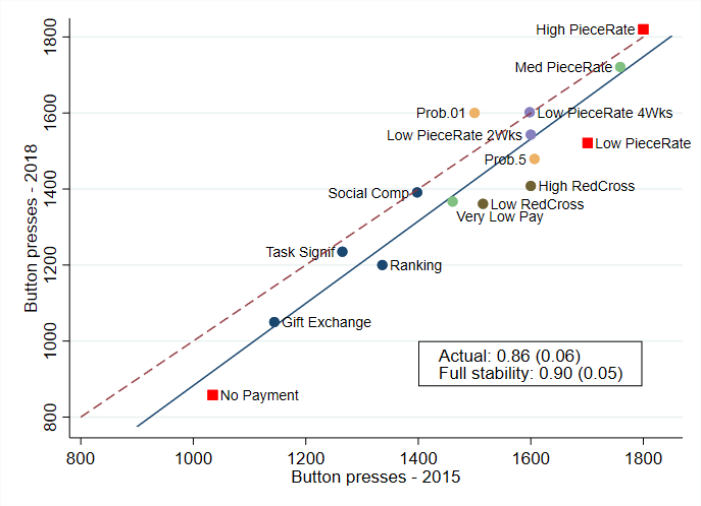
Online Appendix Figure 5. Average Accuracy and Effort by Treatment in the 10-Minute Card Coding Experiment



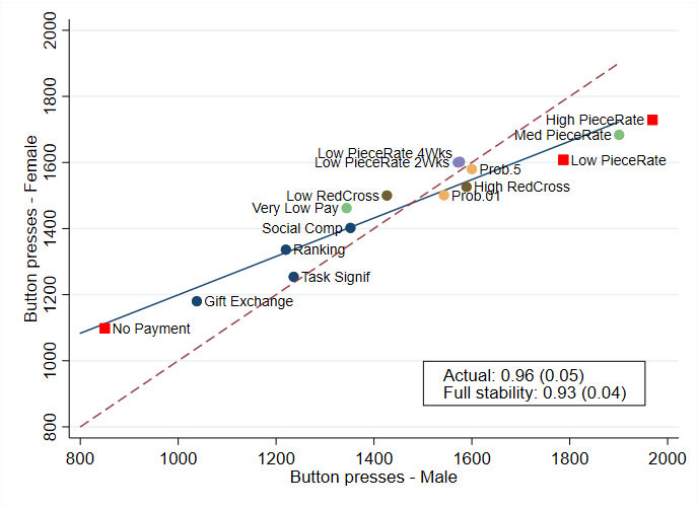
Notes: The figure displays evidence on accuracy for the 10-minute WWII coding task. The graph plots the average effort by treatment (on the x axis) against the average accuracy of coding (on the y axis). The measure of accuracy is the share of cards coded correctly, where we only considered cards for 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 Figure 6. Comparison Across Versions, 25th Percentile of Effort

Onl. App. Figure 6a. Pure Replication, Button Pushing Task



Onl. App. Figure 6b. Impact of Demographics (Gender), Button Pushing

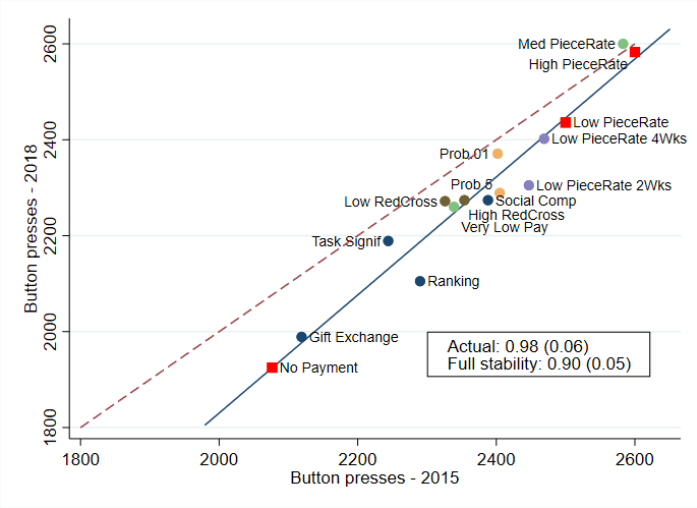


Onl. App. Figure 6c. Impact of Task

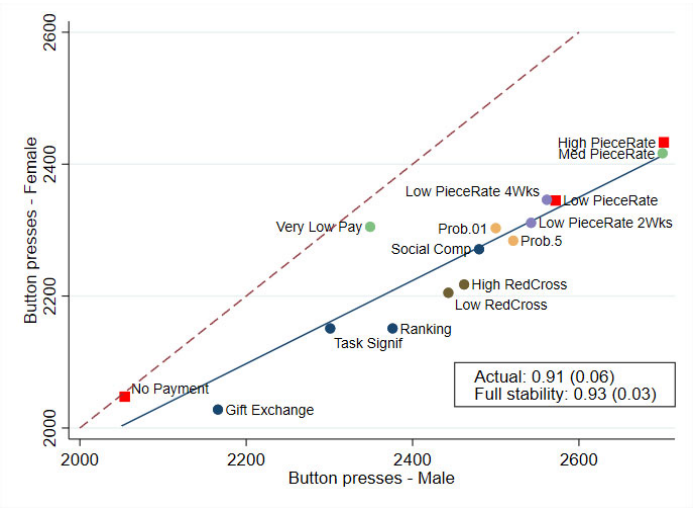
Notes: Online Appendix Figure 6a-c presents the equivalent material as in Figures 3, 4a, and 6, except that for each treatment we plot the 25th percentile of effort, instead of the mean effort as in the original figures. We do not plot these figures for comparisons involving the extra-work card task, since in this task the 25th percentile is almost always a corner solution (0 or 20), making the plot less informative.

Online Appendix Figure 7. Comparison Across Versions, 75th Percentile of Effort

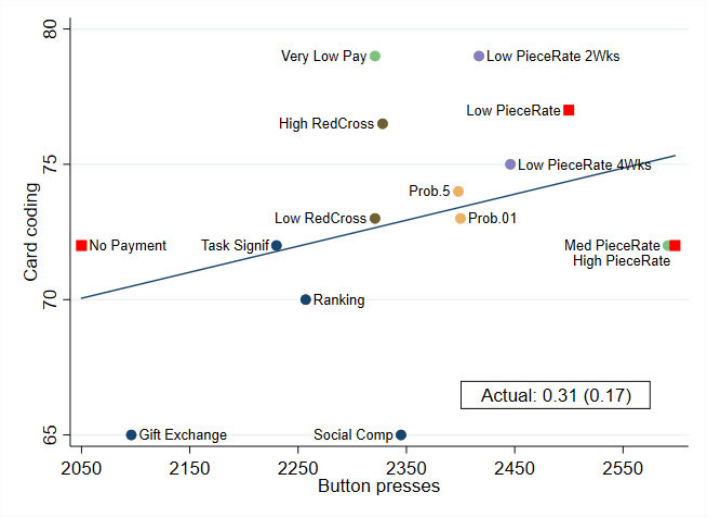
Onl. App. Figure 7a. Pure Replication, Button Pushing Task



Onl. App. Figure 7a. Impact of Demographics (Gender), Button Pushing



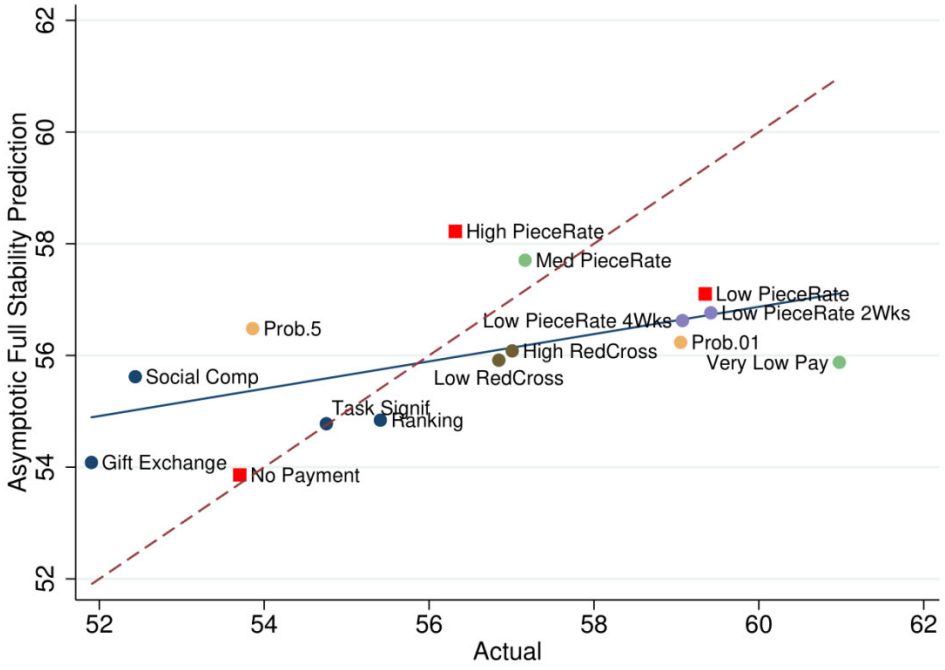
Onl. App. Figure 7c. Impact of Task



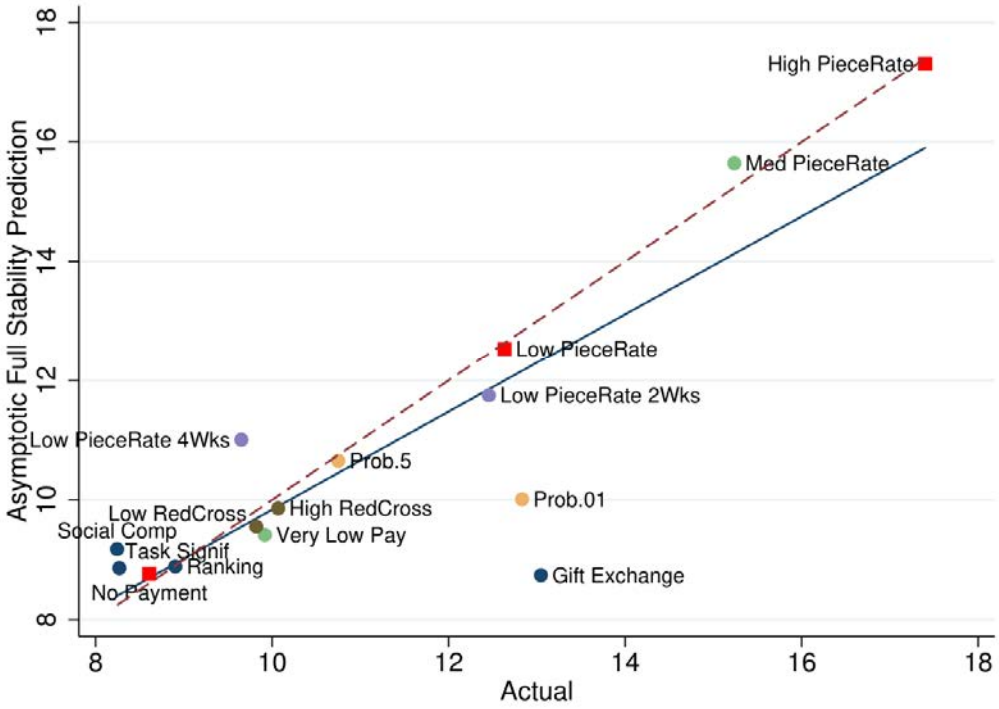
Notes: Online Appendix Figure 7a-c presents the equivalent material as in Figures 3, 4a, and 6, except that for each treatment we plot the 75th percentile of effort, instead of the mean effort as in the original figures. We do not plot these figures for comparisons involving the extra-work card task, since in this task the 75th percentile is almost always a corner solution (0 or 20), making the plot less informative.

Online Appendix Figure 8. Illustration of Full-Stability Benchmark Prediction

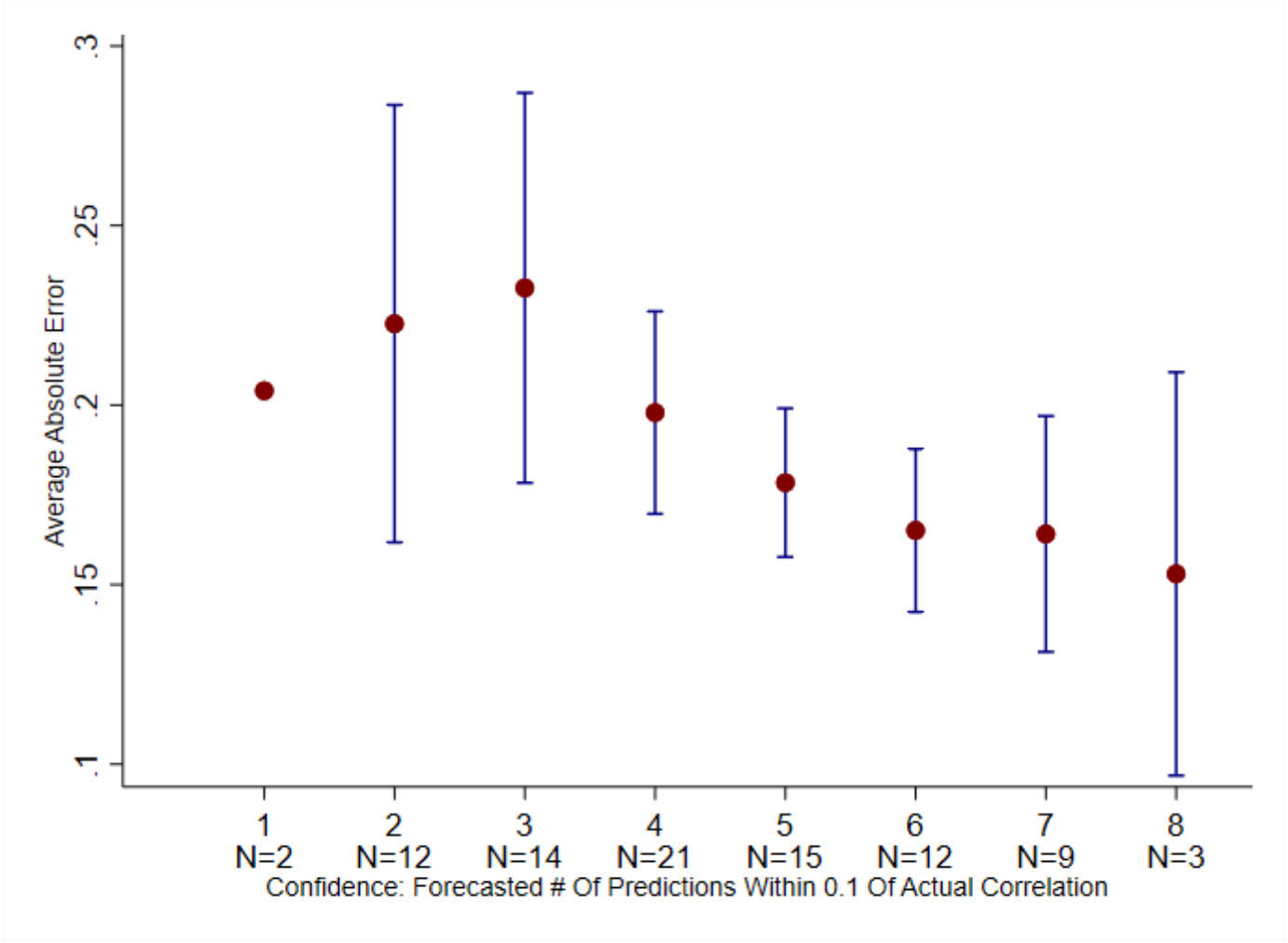
Onl. App. Figure 8a. WWII 10-min: Actual versus Full-Stability Prediction (Asymptotic) from a-b Task



Onl. App. Figure 8b. WII Extra Cards: Actual versus Full-Stability Prediction (Asymptotic) from a-b Task



Online Appendix Figure 9. Confidence (in the Forecast of Rank-Order Correlation) and Average Absolute Error



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 10 we plot the average absolute error in the forecast, 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.

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. 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.

Online Appendix Table 4. Comparison Across Designs, Alternative Measures

		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	
Category	**Version Comparison**	**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	0.97	0.04	0.07	0.03	0.04	0.11	0.13	0.10	0.10
		(0.02)	(0.02)	(0.02)	(0.04)	(0.01)	(0.01)	(0.05)	(0.07)	(0.04)	(0.04)
Demographics	Male vs. Female	0.97	0.98	0.04	0.10	0.03	0.06	0.15	0.12	0.12	0.09
		(0.01)	(0.02)	(0.02)	(0.03)	(0.01)	(0.02)	(0.06)	(0.06)	(0.04)	(0.04)
	College vs. No College	0.97	0.97	0.04	0.07	0.03	0.02	0.09	0.09	0.08	0.07
		(0.01)	(0.02)	(0.02)	(0.03)	(0.01)	(0.01)	(0.04)	(0.05)	(0.03)	(0.03)
	Young vs. Old	0.97	0.98	0.04	0.04	0.03	0.02	0.11	0.08	0.09	0.05
		(0.01)	(0.02)	(0.02)	(0.03)	(0.01)	(0.01)	(0.05)	(0.05)	(0.04)	(0.03)
Geography/ Culture	US vs. India	0.92	0.78	0.06	0.07	0.04	0.11	0.15	0.20	0.13	0.34
		(0.03)	(0.09)	(0.03)	(0.03)	(0.01)	(0.02)	(0.07)	(0.06)	(0.05)	(0.10)
Task	AB Task vs. Card Coding	-	0.55	-	0.19	-	0.16	-	0.49	-	0.47
			(0.14)		(0.04)		(0.04)		(0.09)		(0.09)
Output	Extensive Cards vs. Intensive Cards	-	0.21	-	0.21	-	0.50	-	0.23	-	0.71
			(0.17)		(0.06)		(0.03)		(0.05)		(0.10)
	Extensive Cards vs. AB Task	-	0.63	-	0.16	-	0.34	-	0.37	-	0.31
			(0.07)		(0.03)		(0.03)		(0.07)		(0.05)
	AB Task: First 5 min vs. Last 5 min	0.99	0.98	0.03	0.04	0.02	0.03	0.09	0.04	0.08	0.04
		(0.00)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.04)	(0.02)	(0.03)	(0.02)
Ecological validity	Cards: Consent vs. No Consent	0.92	0.92	0.13	0.16	0.09	0.08	0.13	0.15	0.11	0.10
		(0.03)	(0.04)	(0.06)	(0.08)	(0.02)	(0.02)	(0.05)	(0.07)	(0.03)	(0.03)

Notes: The Table presents alternative measures of stability of experimental results for the version comparisons of Table 2. There are no values that are significantly different from the full stability measure (p value <0.05).

Online Appendix Table 5. Stability Across Designs: Rank-Order Correlations

Category	Design Comparison	25th Percentile		75th Percentile	
		Full Stability w/ Noise	Actual	Full Stability w/ Noise	Actual
		(1)	(2)	(3)	(4)
Pure Repl.	2015 AB Task vs. 2018 AB Task (n=8,252; n=2,219)	0.90 (0.05)	0.86 (0.06)	0.90 (0.05)	0.98 (0.06)
Demogr., Typing Task	Male vs. Female (n=4,686; n=5,785)	0.93 (0.04)	0.96 (0.05)	0.93 (0.03)	0.91 (0.06)
	College vs. No College (n=5,842; n=4,629)	0.93 (0.04)	0.94 (0.07)	0.93 (0.03)	0.92 (0.05)
	Young (=<30) vs. Old (30+) (n=5,259; n=5,212)	0.93 (0.04)	0.93 (0.05)	0.93 (0.03)	0.90 (0.06)
Geogr./ Culture	US vs. India (n=8,803; n=1,225)	0.85 (0.08)	0.71 (0.16)	0.85 (0.07)	0.59 (0.15)
Task	AB Task vs. 10-min Card Coding (n=10,471; n=2,537)	-	0.37 (0.18)	-	0.31 (0.17)
Output	10-min Cards vs. Extra Cards (n=2,537; n=2,188)	-	-	-	-
	Extra Cards vs. AB Task (n=2,188; n=2,219)	-	-	-	-
	AB Task: First 5 min vs. Last 5 min (n=10,471)	0.95 (0.03)	0.93 (0.05)	0.96 (0.02)	0.97 (0.04)
Consent	Cards: Consent vs. No Consent (n=2,188; n=2,246)	-	-	-	-

Notes: The Table lists the 10 design changes to the experiment which constitute 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 in 2015 versus in 2018. We report the actual rank-order correlation, as well as the results under a full-stability benchmark (see Table 2). These results differ from the benchmark ones in Table 2 because we compute the effort estimate using the 25th and 75th percentile of effort instead of the mean effort. We do not report these measures of comparsons involving the extra-work task in which the 25th or 75th percentile effort is typically a corner solution (0 or 20).

Onl. App. Table 6. Structural Estimates, Additional Specifications

		Button Pushing Task, 10 Min	Demographics, Typing, Pooled '15-'18		2018 WWII Cards Coding
Category	**Parameters**	**2015 + 2018 Pooled Exp.**	**USA**	**India**	**Extra Work, Pooled**
		(1)	(2)	(3)	(4)
Incidental Parameters	Curvature of Cost of Effort ɣ	0.015	0.013	0.064	0.051
		(0.003)	(0.003)	(0.169)	(0.010)
	Implied Elasticity	0.036	0.039	0.010	0.384
		(0.008)	(0.008)	(0.025)	(0.074)
	Level of Cost of Effort *k*	-34.591	-31.993	-101.300	-3.954
		(6.619)	(5.688)	(259.826)	(0.934)
	Baseline Motivation *s*	3.7e-04	6.7e-04	4.9e-20	0.137
		(7.5e-04)	(0.001)	(3.4e-16)	(0.085)
Pay Enough or	Δs_{CO}	-1.2e-04	0.021	-0.100	0.059
		(0.087)	(0.095)	(2.3e-05)	(0.061)
Social Pref. Parameters	Pure Altruism *alpha*	0.005	0.007	-8.1e-06	0.008
		(0.008)	(0.008)	(2.3e-04)	(0.017)
	Warm Glow *a*	0.117	0.107	8.8e-05	0.241
		(0.100)	(0.089)	(0.003)	(0.139)
Social Pref.: Gift Exch.	Δs_{GE}	0.001	0.001	6.9e-11	0.857
		(0.002)	(0.002)	(5.0e-09)	(0.245)
Discounting	*Beta*	1.021	1.041	4.477	0.995
		(0.888)	(0.861)	(43.931)	(0.674)
	Delta (Weekly)	0.803	0.828	0.156	0.789
		(0.210)	(0.206)	(0.898)	(0.169)
Social Comparisons	Δs_{SC}	0.060	0.059	2.1e-05	0.007
		(0.058)	(0.055)	(6.6e-04)	(0.037)
Ranking	Δs_{R}	0.014	0.019	2.4e-10	0.056
		(0.018)	(0.021)	(1.5e-08)	(0.047)
Task Significance	Δs_{TS}	0.010	0.014	7.8e-10	0.069
		(0.013)	(0.016)	(4.6e-08)	(0.050)
Probability Weighting Parameters	Pi (0.01)	0.002	0.003	4.3e-12	0.008
		(0.001)	(0.002)	(2.6e-10)	(0.003)
	Pi (0.50)	0.168	0.223	2.1e-08	0.212
		(0.107)	(0.124)	(1.0e-06)	(0.091)
No. of Obs.		10471	9246	1225	4434
Avg effort		1880	1910	1653	11.25
Root MSE		656.59	654.47	715.74	54.08
Extra Treat.: Incentive + Please try	Out-of-Sample Pred.				13.13
	Actual				12.069
					(0.543)

Notes: The Table shows structural estimates of the incidental parameters (ɣ, k, and s) and psychological parameters estimated using all 15 treatments across 11 different samples. All models assume an exponential cost function. Cols (1)-(3) are estimated using nonlinear least squares for the a-b yping task, while Col 4 is estimated on the extra-work task using maximum likelihood due to censoring. Standard errors in parantheses.