American Economic Journal: Microeconomics 2022, 14(3): 889–925
https://doi.org/10.1257/mic.20200129

Stability of Experimental Results: Forecasts and Evidence[†]

By Stefano DellaVigna and Devin Pope*

How robust are experimental results to changes in design? And can researchers anticipate which changes matter most? We consider 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, and* (*vi*) *the presence of a consent form. We find near-perfect replication of the experimental results and full stability of the results across demographics, significantly higher than a group of experts expected. The results differ instead across task and output change, mostly because the task change adds noise to the findings.* (*JEL* C90, D82, D91)

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 Institutional Review Board (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 were run with different design choices.

Another researcher is evaluating a field experiment as a journal referee. While the results in the paper are 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. 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

* DellaVigna: Department of Economics, UC Berkeley and NBER (email: sdellavi@econ.berkeley.edu); Pope: Booth School of Business, University of Chicago and NBER (email: devin.pope@chicagobooth.edu). John Asker was coeditor for this article. We thank Ned Augenblick, Jon de Quidt, Anna Dreber, Magnus Johannesson, Don Moore, Alex Rees-Jones, Joshua Schwartzstein, Dmitry Taubinsky, and 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 the 2018 SITE Conference for Psychology and Economics for comments and suggestions. We thank Kristy Kim, Christopher Lim, Maxim Massenkoff, Tashfeen Saeed, Jihong Song, and Ao Wang for outstanding research assistance. Our survey was approved by University of Chicago IRB, protocol IRB18-0144 and preregistered as trial AEARCTR-0002987 (DellaVigna and Pope 2019).

[†] Go to https://doi.org/10.1257/mic.20200129 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

wonders, "If we move beyond pure replication to conceptual replication, will the experimental results replicate? How do we even measure replication, if the units of measure in the replication differ from the original units?"

These three researchers are concerned about the generalizability of a set of experimental results as the design changes. This concern is labeled as being about *stability of results*, *conceptual replication*, or *external validity* (e.g., Rothwell 2005). A number of papers examine the stability of experimental results with respect to specific design choices, such as, for example, 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 (2020) 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 would 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 case as a road map for conceptual replication in experiments with multiple treatments arms (e.g., Gerber and Green 2000; Bertrand et al. 2010; Bhargava and Manoli 2015). Since some of the design changes produce results with different units of measurement, we propose rank-order correlation as a way to compare treatment effects. Further, since we are interested in not only how the results change, but also how researchers *expect* the results to change, as in the referee and researcher examples above, we collect forecasts about the stability of the 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 rerun 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; and (vi) (*consent form*) it may matter that subjects know that it is an experiment.

The initial task is a typing task documented in DellaVigna and Pope (2018a, b): subjects on Amazon Mechanical Turk (MTurk) have ten minutes to alternatively press the "a" and "b" buttons on their keyboards as quickly as possible. While the task is not meaningful per se, it lends itself to study motivation since the typing exercise becomes tiresome. In these previous papers, we compared effort for nearly 10,000 subjects across 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.

In this paper, we build on this previous experiment but consider several novel design variants, covering the six dimensions above, none of which are considered in our previous work. Specifically, we collect data on nearly 10,000 new MTurk subjects. In each variant we include 15 of the original treatments, following a preregistered design. First, we run a *pure replication* of the same experiment three 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 (WWII) conscription cards—and measure the number of cards coded within ten minutes.[1] 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.[2] 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 an experiment.

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 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 nonlinear 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.

[1] Nearly 400 subjects left positive comments about this task, such as "*What a fun hit! WW2 history* … Memorial Day …" and "INTERESTING WORK." In contrast, comments about the a-b task are typically about exhaustion.

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

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 experts in behavioral and experimental economics or experts on replication, yielding 55 responses. Each expert sees a description of the task and 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 ten 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 0.93 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 0.90 under full stability), (iv) the task and output changes will have a sizable impact (0.50 to 0.70 correlation), and (v) the disclosure of consent will have a modest impact (0.78 correlation, compared to 0.89 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.[3] 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 ten 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.

[3] This null effect of demographics was not obvious. For example, women tend to display more generous behavior and more reciprocity (Croson and Gneezy 2009), which would affect effort in the social preference treatments.

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 ten-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 ten-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 ten-minute task to the WWII extra-cards coding. The correlation between these tasks, which are both responsive to incentives, is quite high at 0.65 and 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, 0.53, than for just the output change, 0.61. 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 SE) under three piece rate treatments, indicating a flat and nonmonotonic response to incentives in the ten-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 compare the extra-cards WWII coding task with, and without, a consent form. The rank-order correlation is 0.84, close to the expert prediction (0.78) and to the full-stability measure (0.89). Thus, in our context, it does not matter 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. Eight out of 10 planned comparisons have a correlation above 0.60, and 6 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 prespecified. Indeed, our full-stability benchmark, which assumes full replication but allows for sampling noise, is an excellent predictor of the observed correlations.

Second, the experts have a mixed record in their ability to predict how much design changes affect the results, and they overestimate their own accuracy. This contrasts with evidence that experts 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 suggests that design choices and external validity judgments may be more tentative than we realize.

Turning to two specific results, our third takeaway 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. While we do not have direct evidence on this, a possibility is that 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's (1971) 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 open-science work on large-scale replication of experiments (Open Science Collaboration 2015; Camerer et al. 2016, 2018). Closer to our focus on conceptual replication are the "Many Labs" projects (Klein et al. 2014, 2018), which replicate dozens of psychological findings in different labs around the world, and Landy et al. (2020), which crowdsources 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. (2020) find larger heterogeneity in results when the design is left to the different researchers, leading presumably to larger design differences.

An example of conceptual replication is the comparison of experimental results across platforms, such as in the laboratory versus on MTurk (e.g., Horton, Rand, and Zeckhauser 2011 and Snowberg and Yariv 2021) or in the laboratory versus in the field (e.g, Falk and Heckman 2009). Our design does not include a platform comparison, since it is covered by previous work.

I. Design and Measure of Stability

A. *Experimental Design*

2015 Experiment and Model.— The starting point for the design is the real-effort task in DellaVigna and Pope (2018a, b) (itself based on the task in Ariely, Bracha, and Meier 2009), which we ran in May 2015 on 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 for experiments in marketing, psychology, and economics, with findings generally similar to the results in laboratory or field settings (Horton, Rand, and Zeckhauser 2011 and Snowberg and Yariv 2021), though with some evidence of a higher level of noise than in some student samples (Snowberg and Yariv 2021).

We recruited subjects on MTurk for a $1 payment 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 reads, "*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 ten-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 ten minutes, the subjects were presented with the total points and the payout, 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 sixteenth 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 4 treatments in Table 1 are piece rate treatments, with the piece rate varying from no piece rate to low piece rate ($0.01 per 100 points) to mid piece rate ($0.04 per 100 points) to high piece rate ($0.10 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 nonmonetary 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. (2018), 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

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

Table 1—Findings by Treatment: Effort in Different Versions

		Mean effort (SE)				
		Button pushing, 10 min		2018 WWII cards coding task		
Category (1)	Task: Treatment wording (2)	2015 exp. (3)	2018 exp. (4)	Ten-min (5)	Extra work (6)	Extra work, no consent (7)
Piece rate	"Your score [The number of [additional] cards you complete] will not affect your payment in any way."	1,521 (31)	1,367 (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]"	2,029 (27)	1,966 (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]."	2,132 (26)	2,119 (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]."	2,175 (24)	2,146 (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]."	1,883 (29)	1,801 (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]."	1,907 (27)	1,780 (50)	56.90 (1.80)	9.85 (0.84)	9.86 (0.71)
	"As a bonus, the Red Cross charitable fund will be given 1**0 cents** for every 100 points that you score [5 cents for every [additional] card you complete]."	1,918 (26)	1,839 (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]."	1,602 (30)	1,476 (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."	2,004 (27)	1,953 (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."	1,970 (29)	1,940 (53)	59.10 (1.83)	9.64 (0.76)	11.70 (0.82)

(*continued*)

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 enough, but we incorporate four piece rates to build in overidentification. We present the estimation details in Section IC.

Behavioral Treatments: The next treatments are motivated by behavioral research. In the paying-too-little treatment, we set a very low piece rate, \$0.01 for every 1,000 points, to test whether this crowds out intrinsic motivation. In the next 2 social preferences treatments, subjects earn a return for a charity by working (as in Imas 2014), with either a low return to the charity (\$0.01 per 100 points) or a high return (\$0.10 per 100 points). In the third social preference treatment, on gift exchange (as in Gneezy and List 2006), subjects receive an unconditional \$0.40 bonus.

We model these treatments as follows. For the paying-too-little treatment and 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 $\Delta s_{CO} = 0$.

Table 1—Findings by Treatment: Effort in Different Versions (Continued)

Category (1)	Task: Treatment wording (2)	Mean effort (SE)				
		Button pushing, 10 min		2018 WWII cards coding task		
		2015 exp. (3)	2018 exp. (4)	Ten-min (5)	Extra work (6)	Extra work, no consent (7)
Risk aversion and probability weighting	"As a bonus, you will have a **1 percent chance** of being paid an **extra $1** for every 100 points that you score [**extra 50 cents** for every [additional] card you complete]."	1,896 (28)	1,975 (47)	59.09 (1.68)	12.83 (0.76)	11.54 (0.79)
	"As a bonus, you will have a **50 percent chance** of being paid an **extra 2 cents** for every 100 points that you score [**extra 1 cents** for every [additional] card you complete]."	1,977 (25)	1,837 (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]]."	1,848 (32)	1,774 (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."	1,761 (31)	1,642 (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.**"	1,740 (29)	1,627 (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]."	—	2,056 (46)	56.18 (1.76)	10.81 (0.79)	13.3 (0.74)
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 treatments, 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.

For the two charity treatments, we allow for both a pure altruism parameter α and a "warm-glow" parameter a and model motivation as $s + \alpha p_{ch} + a \times 0.01$: the altruism parameter α multiplies the actual return to the charity p_{ch}, while the warm-glow term a multiplies the return to the charity for the low-return treatment ($0.01 per 100 presses for the a-b task). 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 2 treatments motivated by the research on present bias, the piece rate is $0.01 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 percent chance of being paid an extra $1 for every 100 points*," while in the second treatment, it is "*a 50 percent 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 $0.01 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 that subjects are not too risk averse, we expect higher effort in the 1-percent-of-$1 treatment than in the $0.01 treatment. The treatment with a 50 percent probability of a $0.02 piece rate provides evidence on the concavity of the value function, i.e., risk aversion, which we capture in reduced form as $\pi(0.5) < 0.5$.[4]

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.[5] Column 3 of Table 1 and online Appendix Figure 2 summarize the average effort in the 15 treatments.

2018 Experiment.—In May of 2018, we ran a new round of experiments on MTurk following a preanalysis plan, with design variants but otherwise as close as possible to the 2015 experiment. The data are available at DellaVigna and Pope (2021).

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 versions 2–4.

Within each version, the workers were randomized into 1 of 16 treatments with equal weights.[6] In addition to the 15 treatments from the earlier experiment, an additional sixteenth treatment combined a piece rate and a psychological manipulation.

[4] With just two probabilistic treatments, we cannot disentangle the curvature of the probability weighting from the curvature of the value function. In the estimates, we assume a linear utility function, thus loading all curvature on the function $\pi(P)$. In DellaVigna and Pope (2018a), we show that assuming a concave value function with the Tversky and Kahneman (1971) calibration has only a small impact on the estimate of $\pi(0.01)$.

[5] These three treatments turned out to be underpowered 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., $0.40 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.

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

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.[7]

Ten-Minute WWII Coding: The second version is also a ten-minute task, but subjects are assigned to code the occupation in World War II enrollment cards[8]: "*In this task you will be coding up conscription records about soldiers in World War II. You will have 10 minutes to complete as many cards as you can. Your job is to identify the occupation in field 7 of each record and to type it into the text box below each card. If you are unable to determine what the occupation is, or if field 7 is missing from the card, please type 'unclear.*'" We then show the subjects an example of a card and 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.[9]

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 except 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 $0.40.

Extra-Work WWII Coding: In versions 1 and 2, we measure productivity—the number of units produced within a given time—as in most real-effort experiments (e.g., Gneezy and List 2006). Yet, an alternative margin of effort is the extensive margin of how much *extra work* one is willing to do, as pioneered by Abeler et al. (2011).

In our third version, the subjects first code the occupation for 40 WWII cards (online Appendix Figure 1d) with no extra incentive. After they are done with the

[7] 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 payout page has slightly different wording. Arguably, in light of these changes, this may be a "conceptual replication." The forecasters could see the changes.

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

[9] We measure accuracy for worker i as the share of cards on which i's coding agrees with the modal coding of others for that card. We use this measure to exclude from the sample a small number of cheating workers (see "Sample").

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 ten-minute WWII card coding version. To partially compensate for this, the required number of cards, 40, is such that most subjects would finish earlier than in 10 minutes.[10]

No-Consent WWII Coding: While in all other versions, the workers see a consent form after clicking on the MTurk HIT, in this version, which is otherwise identical to the third 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 preanalysis plan, we set out to exclude subjects that (i) do not complete the task within 30 minutes of starting, (ii) exit and then reenter the task as a new subject (as these individuals might see multiple treatments), (iii) are not approved for any other reason (e.g., they did not have a valid MTurk ID), or (iv) 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: subjects that (v) in version 1 scored 4,000 or more a-b points, (vi) in version 2 coded 120 or more cards with accuracy below 50 percent, or (vii) in versions 3 and 4 completed the 40 required cards in less than 3 minutes with accuracy below 50 percent or completed the 20 additional cards in less than 1.5 minutes with accuracy below 50 percent. We set a target of 10,000 subjects completing the tasks after these restrictions.

We followed the preregistration sample rules. The experiment ran for 3weeks, at which point we had 12,983 subjects who started the task on Qualtrics. We removed 324 workers who had reentered 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 restrictions 4–7. Two final restrictions not included in the preregistration

[10] 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.

were excluding 21 MTurkers who appeared to have cheated on the card coding task in ways not covered above and 59 observations due to Qualtrics data "glitches."[11]

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.

B. *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 percent versus 54.4 percent), more older workers (55.4 percent above the age of 30, compared to 48.5 percent), and more college-educated workers (58.8 percent versus 54.8 percent). 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 4 versions.

Dimension 2. Demographics: We take advantage of the heterogeneity in the MTurk population and compare across three different demographic breakdowns, 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), and (iii) workers who are up to 30 years old (N = 5,259) versus workers who are older than 30 (N = 5,212).

Dimension 3. Geography/Culture: Using the latitude and longitude inferred from the IP address, we geocode the likely location of the workers (barring, say, the use of a virtual private network). Still pooling the 2015 and 2018 a-b task data, we compare workers in the United States (N = 8,803) versus workers in India (N = 1,225). 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.

[11] The 21 MTurkers eliminated for cheating had less than 10 percent 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 ten minutes in the first task (with a ten-second leeway for early timer starts), or (v) ten time stamps more than the total coded cards.

Table 2—Stability across Designs: Rank-Order Correlations, Forecasts versus Actual versus Full-Stability

Design comparison	Rank-ord. correl. full stability: Bootstrap from data (1)	Rank-ord. correl. full stability: Structural (2)	Average forecast of rank-order correlation: Faculty experts (3)	Average forecast of rank-order correlation: PhD students (4)	Average forecast of rank-order correlation: MTurkers (5)	Rank-. ord. correl actual (6)	*p*-value for difference: Experts versus full stability (7)	*p*-value for difference: Actual versus full stability (8)	*p*-value for difference: Actual versus experts (9)
Category: Pure repl.									
2015 AB task versus 2018 AB task (observations = 8,252; observations = 2,219)	0.93 (0.04)	0.94 (0.03)	0.82 (0.01)	0.87 (0.01)	0.75 (0.02)	0.91 (0.05)	0.005	0.647	0.060
Category: Demogr., typing task									
Male versus female (observations = 4,686; observations = 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	0.845	0.000
College versus no college (observations = 5,842; observations = 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	0.522	0.000
Young ($\leq$30) versus old (30+) (observations = 5,259; observations = 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	0.675	0.000
Category: Geogr./culture									
US versus India (observations = 8,803; observations = 1,225)	0.90 (0.05)	—	0.63 (0.02)	0.67 (0.03)	0.68 (0.02)	0.65 (0.11)	0.000	0.047	0.896
Category: task									
AB task versus 10-min card coding (observations = 10,471; observations = 2,537)	—	0.50 (0.19)	0.66 (0.02)	0.63 (0.03)	0.64 (0.02)	0.59 (0.14)	0.392	0.703	0.623
Category: Output									
10-min cards versus extra cards (observations = 2,537; observations = 2,188)	—	0.58 (0.17)	0.61 (0.02)	0.61 (0.03)	0.62 (0.02)	0.27 (0.17)	0.831	0.184	0.035
AB task versus extra cards (observations = 10,471; observations = 2,188)		0.85 (0.07)	0.53 (0.03)	0.56 (0.04)	0.58 (0.02)	0.65 (0.07)	0.000	0.050	0.099
AB task: First 5 min versus last 5 min (observations = 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	0.553	0.000
Category: Consent									
Cards: consent versus no consent (observations = 2,188; observations = 2,246)	0.89 (0.05)	0.84 (0.07)	0.78 (0.02)	0.81 (0.02)	0.70 (0.02)	0.84 (0.09)	0.056	0.632	0.536
Observations			55	33	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 number rank-o. corr within 0.1 of truth			3.99 (0.24)	4.95 (0.25)	4.66 (0.22)				
Average actual rumber rank-o. corr within 0.1 of truth			3.22 (0.23)	3.33 (0.24)	3.01 (0.15)				

Notes: The table lists the ten 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 realized effort. This benchmark is derived from a data-based bootstrap in column 1, while it uses structural 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 of the data-based bootstrap (column 1) when available. The structural estimates for the India sample do not converge due to the very noisy response to incentives in this subsample. We thus cannot compute the structural full-stability benchmark.

Dimension 4. Task: We compare the pooled 2015-18 results for the a-b task to the results for the ten-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 2015–2018 a-b coding task, we compare output in the first five minutes versus the last five 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).

C. *Measure of Stability*

In each of these dimensions, we want to compare the average effort for the 15 treatments in the 2 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, as typical with horse races of interventions (e.g., Bertrand et al. 2010 or Bhargava and Manoli 2015). In such cases, one is typically interested not only in how each treatment compares to a baseline group, but also in comparisons across treatment arms. To capture these multiple comparisons, we use the rank-order correlation between the treatment effectiveness in one version versus in another version.

As an example, consider the hypothetical results in online Appendix Figure 3a for a replication case: in the first panel, only 2 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 nonlinear transformations.

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 that 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, we (i) bootstrap from the 2015 sample (with replacement) 150 observations from each of the 15 treatments, mirroring the sample size for the 2018 experiment; (ii) compute the average effort in the 15 simulated cells; (iii) compute

the rank-order correlation of the 15 bootstrapped means with the actual 2015 results for the 15 treatments; and (iv) repeat this 1,000 times. The average rank-order correlation across the 1,000 iterations, 0.94, is the full-stability benchmark.[12]

Similarly, we compute a bootstrap for the demographic comparisons in the pooled 2015–2018 a-b task: in each of the 15 treatments, we (i) randomly assign a subject to demographics A or B, with the share assigned to group A matching the empirical one; (ii) compute the average effort in each of the $15*2$ cells and (iii) the rank-order correlation; and (iv) 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 that, as a null hypothesis, one may expect to be stable across versions (though, of course, they could differ by, say, culture). 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. We define two versions to have stable experimental findings if they have the same behavioral parameters, even if the incidental parameters vary. We discuss the details in Section IIIC.

II. Expert Forecasts of Stability

A. *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 their familiarity with the experiment, but we scaled back the sample given that our earlier results suggest that a couple dozen responses would provide sufficient statistical power.

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 and experimental economists with PhDs in 2015–2018 (not included in the earlier sample), drawing names from attenders and presenters at two behavioral conferences (Behavioral Economics Annual Meeting and Stanford Institute for Theoretical Economics Psychology and Economics). The 60 experts, by our coding, cover applied behavioral theory

[12] We hold the 2015 results as given at each iteration; the results are very similar if we bootstrap those as well.

(11), laboratory experiments (17), and behavioral field evidence (32). At least 37 experts have experience using MTurk or similar online samples. In addition, we identified ten 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 PhD students in economics at UC Berkeley and the University of Chicago, yielding 33 responses. Finally, we collected 109 forecasts on MTurk (for a $1 payment).[13]

Survey: The survey, which was expected to take 15–20 minutes, walked the forecasters through 4 steps. First, we briefly summarized the design in the 2015 experiment and displayed the 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) one forecast about pure replication; (ii) three forecasts about demographics (gender/education/age); (iii) one forecast about geography/culture comparing MTurkers in the United States to those in India; (iv) one forecast about task change; (v) three forecasts about output change, comparing first the ten-minute WWII coding to the extra-work WWII coding, then the a-b task to the extra-work WWII coding, and finally, within the a-b task, effort in the first five minutes to the last five minutes; and (vi) one 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 IC. 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 (0.90). We did not report a full-stability benchmark comparing across different tasks or output, given that this requires a full set of structural estimates.[14]

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.

B. *Forecasts of Correlation*

Figure 1, panels a–b and Table 2, columns 3–5 report the results from the forecasts. On average, the experts expected that the rank-order correlation for the pure

[13] 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.

[14] We did not report 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), as we wanted to remain as blinded to the 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.

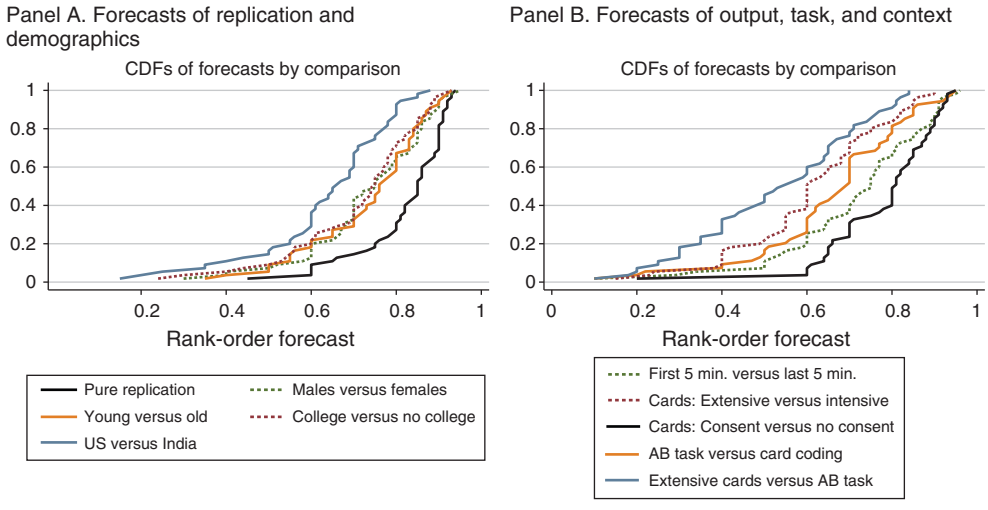

FIGURE 1. EXPERT FORECASTS, CDFS

Notes: Figure 1, panels A–B present the CDF of the forecasts by the 55 academic experts. Each expert made forecasts about rank-order correlation with respect to ten design changes. We split the ten forecasts into panel A and panel B.

replication would be quite high (0.82), though lower than the full-stability one (0.93), a difference that is statistically significant ($p = 0.005$, column 7). The CDF plot in Figure 1, panel A 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 3 demographic variables, with average forecasted rank-order correlation of 0.73 (gender), 0.71 (education), and 0.74 (age). As Figure 1, panel A 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 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, expect a correlation of 0.78, compared to the full-stability benchmark of 0.89.

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 IV.

The predictions of the PhD students track closely with the predictions of the experts; the forecasts of the MTurk subjects are, on average, somewhat lower, but they exhibit similar patterns. Thus, the expectations do not vary much with the population at hand; we present further splits in Section IV.

III. Stability of Experimental Results

A. *Main Results on Stability*

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

Pure Replication: As online Appendix Figure 4a–b shows, the distribution of effort across the 15 treatments for the a-b typing task is very similar in 2015 and 2018, if somewhat noisier in 2018 given the smaller sample size. Figure 2, panels A–B shows that effort also responds similarly to the piece rate incentives. As Figure 3 shows, the behavioral treatments for 2018 also line up very nicely with the 2015 results, only slightly below the 45-degree line (dashed line). Only the probability weighting treatment deviates by more than 100 points from the interpolating line (continuous line). The rank-order correlation of 0.91 is close to the full-stability benchmark of 0.93 and higher than the average forecast at 0.82 ($p = 0.060$ for the difference, column 9). Thus, our pure replication produces very similar results to the original ones.

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 4, panel A displays the treatment results separately for male and female subjects.[15] The data suggest 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.[16] 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 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 4, panel B, we compare subjects with a completed college degree and subjects without. The two groups differ in the level of effort: higher-education subjects exert less effort in any given

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

[16] In a meta-analysis of 17 studies, Bandiera et al. (2021) show that on average, women respond 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).

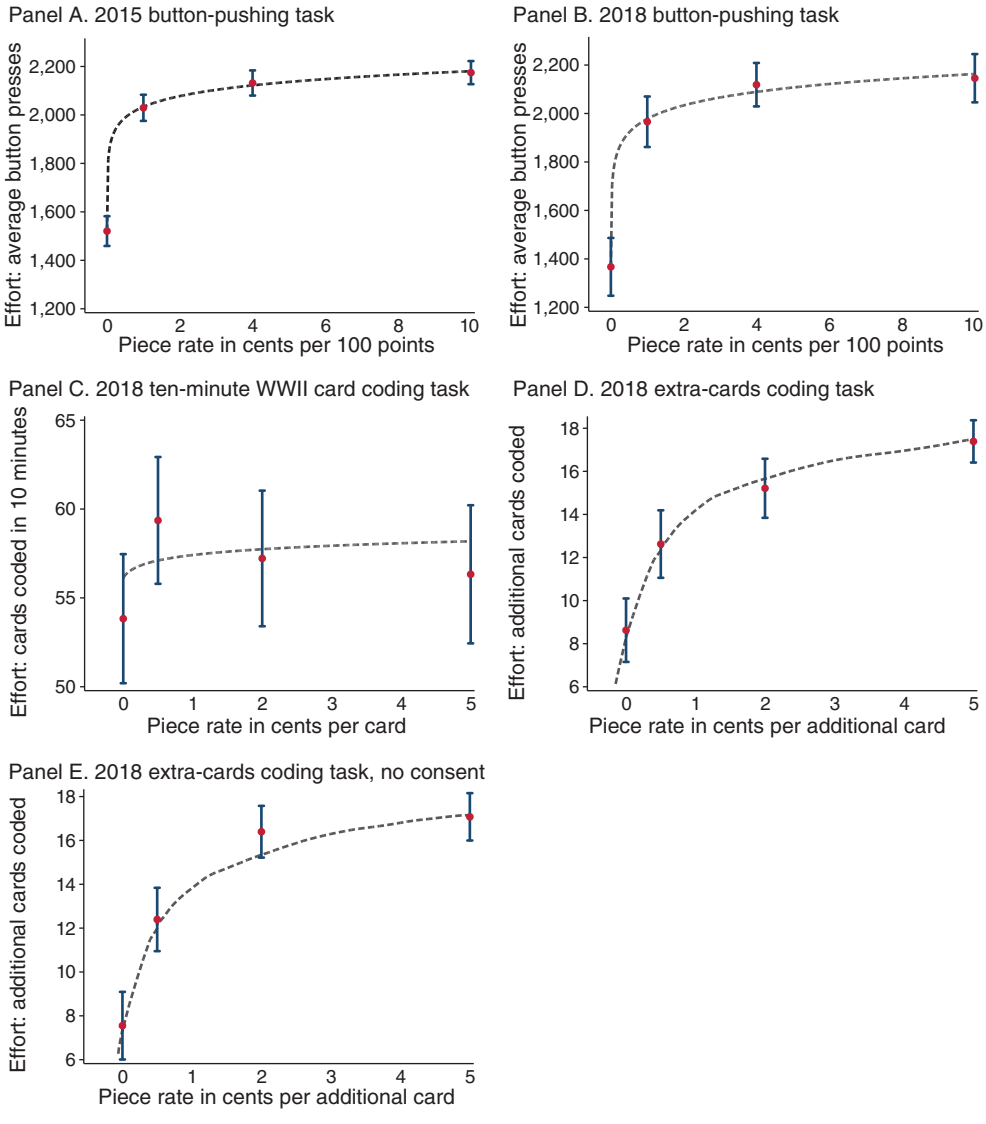

Figure 2. Average Effort in Piece Rate Treatments

Notes: Figure 2, panels A–E displays the average effort in four piece rate conditions (including the no-piece-rate baseline), separately in each of five experiments: the 2015 button press (panel A), the 2018 button press (panel B), the 2018 ten-minute card coding (panel C), the 2018 extra-card coding (panel D), and the 2018 extra-card coding with no consent form (panel E).The figures display a 95 percent confidence interval around the mean effort. The figure also displays with a dashed line the predicted effort from the structural estimates in Table 4, columns 1, 2, 9, 10, and 11.

treatment. But once we control for this difference, the treatment effects line up very nicely, yielding a rank-order correlation of 0.97, much larger than the average forecast of 0.71, a difference that again is statistically significant. Similarly, splitting the results by age in Figure 4, panel C, subjects younger than 30 years of age display higher effort than subjects that are older, but once again, the rank-order of the treatments is very high (0.98).

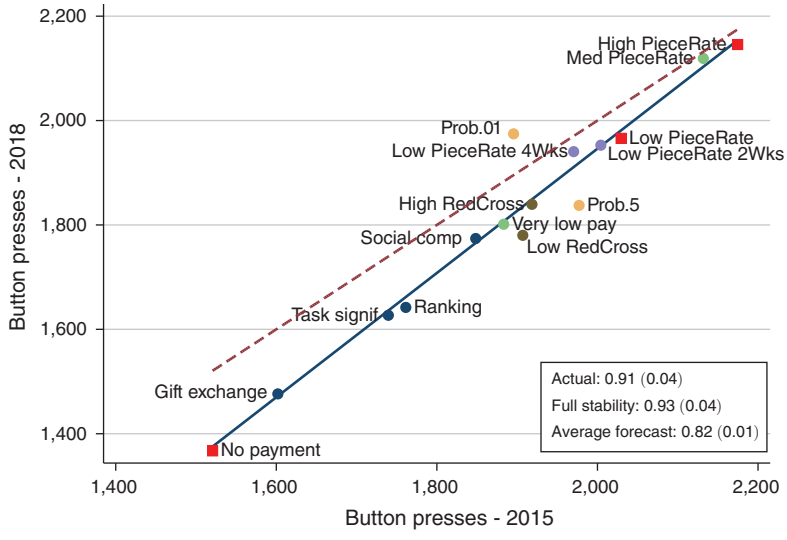

FIGURE 3. PURE REPLICATION, BUTTON-PUSHING TASK

Notes: Figure 3 displays, for each one of 15 treatments, the average effort across 2 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 dashed 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.

Geography/Culture: While the previous demographic features are self-reported, we now take advantage of the geolocation due to the IP address. We compare the average effort by treatment among the 12 percent of subjects with an IP in India versus the subjects with an IP in the United States. As Figure 5 shows, the subjects in India display lower average effort and lower elasticity. Adjusting for this difference, the behavioral and incentive treatments show a correlation of 0.65, statistically lower than the full-stability benchmark ($p = 0.047$) and nearly identical to the average forecast (0.62).

Task: We then compare the results in the ten-minute a-b typing task (pooling 2015 and 2018) to the results in a ten-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, is approximately normally distributed, with a median around 60 cards. Figure 2, panel C shows that the new task is unresponsive to the piece rate incentives.[17]

[17] 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.

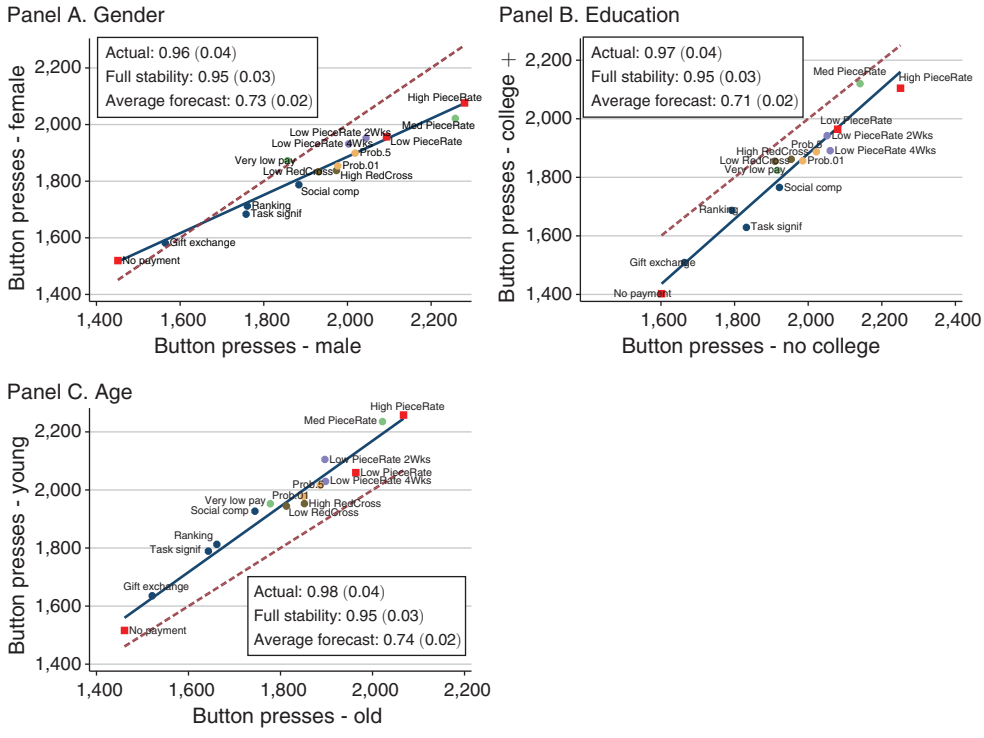

FIGURE 4. IMPACT OF DEMOGRAPHICS, BUTTON-PUSHING TASK

Notes: Figure 4, panels A–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 (panel A), education (panel B), and age (panel C). See notes to Figure 3 for more detail.

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 0.31.

Output Measure: In our fifth comparison, we consider how changes in measures of output for a given task affect the findings. First, we compare 2 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 majority of subjects code 0 extra cards or all 20 extra cards. Importantly, as Figure 2, panel D 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 design is well suited to capture variation in motivation.

Figure 7, panel A 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

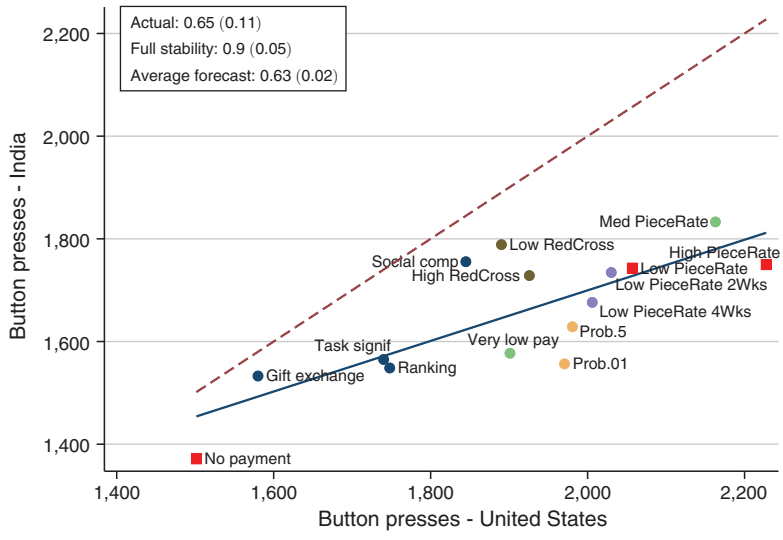

FIGURE 5. IMPACT OF GEOGRAPHY/CULTURE, BUTTON-PUSHING TASK

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 an S location (x-axis) or with a location in India (y-axis). See notes to Figure 3 for more detail.

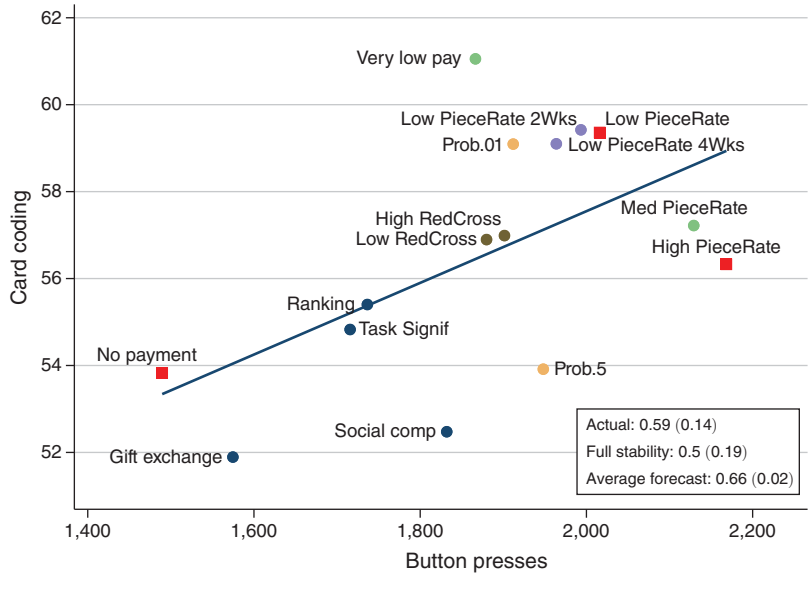

FIGURE 6. IMPACT OF TASK, BUTTON-PUSHING TASK VERSUS WWII CARD CODING TASK

Notes: Figure 6 displays, for each one of 15 treatments, the average effort across 2 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 ten-minute WWII card coding task. See notes to Figure 3 for more detail.

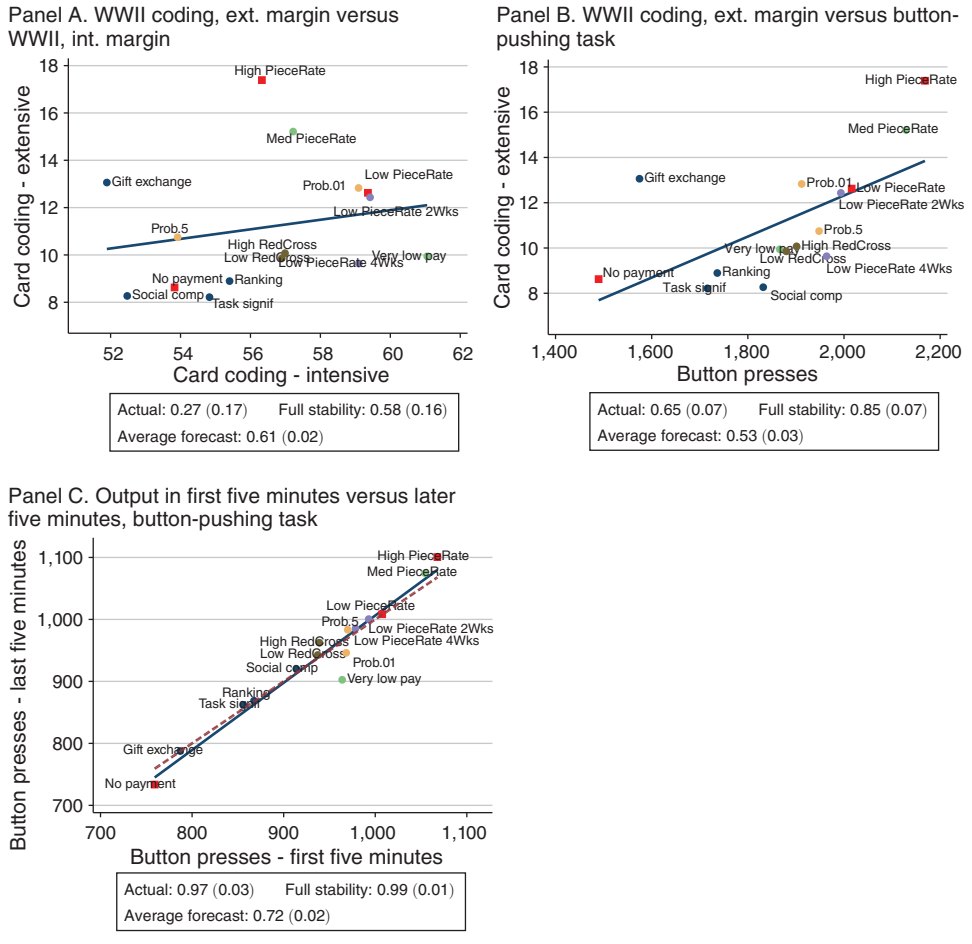

FIGURE 7. IMPACT OF OUTPUT

Notes: Figures 7, panels A–C display for each one of 15 treatments the average effort across 2 different output measures. In panel A, we compare the cards coded in the ten-minute WWII card coding task to the extra cards coded in the extra-work WWII card task. In panel B, we compare the a-b points in the ten-minute button-pushing task to the extra cards coded in the extra-work WWII card task. In panel C, we compare, within the button-pushing task (pooling 2015 and 2018), productivity in the first five minutes versus the next five minutes. See notes to Figure 3 for more detail.

(relative) instability has two possible explanations. First, changes in output may have truly changed the impact of behavioral motivators. Second, productivity in the ten-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.

As additional evidence, we do a combined output/task comparison of the a-b ten-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 7, panel B shows, the correlation for the joint task/output change, 0.65, is higher than it is for just the output change, 0.27. The experts instead expect the correlation to be higher for just the output change, 0.61, than for the task/output change, 0.53.

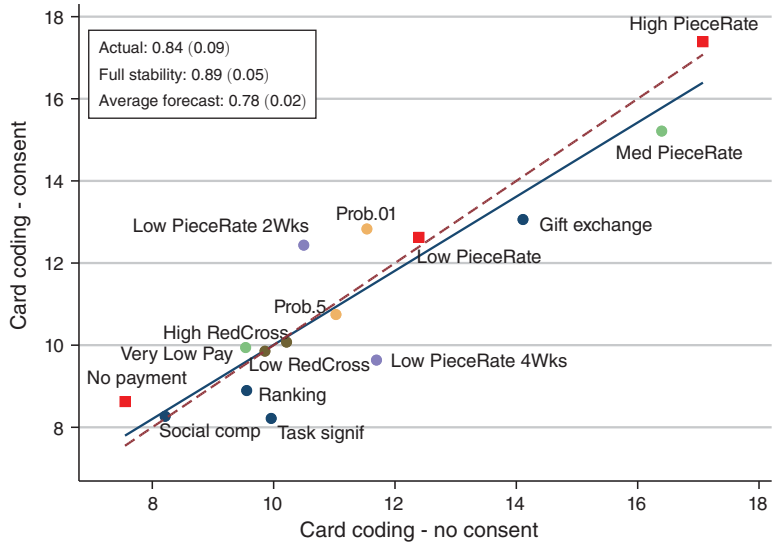

FIGURE 8. IMPACT OF CONSENT, WWII CODING TASK

Notes: Figure 8 displays, for each one of 15 treatments, the average effort for 2 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.

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 five minutes versus the next five minutes. As Figure 7, panel C 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 0.89 and the average forecast of 0.78.

B. *Robustness*

Alternative Measure of Stability: In online Appendix Table 4, we use alternative measures of stability, comparing the results to the rank-order measure reproduced in column 1. In column 2, we show that the results are very similar using the Pearson correlation measure. In columns 3–5, we consider all possible 105 binary comparisons of the 15 treatments and examine whether the version change affects which treatment is more effective. For example, 100 out of 105 treatment comparisons have the same sign for men and women, and 68 comparisons are statistically significant in the same direction in both samples, with 0 cases of significant comparisons going in opposite directions. This measure of stability is also highly correlated with the

Table 3—Stability across Designs, Additional Comparisons

Category	Version comparison	Rank-order correlations across designs: Full stability with noise (1)	Actual (2)	*p*-value for difference (3)
Demographics, ten-minute WWII coding task	Male versus female (observations = 1,014; observations = 1,523)	0.44 (0.18)	0.27 (0.21)	0.534
	College versus no college (observations = 1,478; observations = 1,059)	0.43 (0.18)	0.38 (0.21)	0.854
	Young versus old (observations = 1,128; observations = 1,409)	0.44 (0.17)	0.31 (0.22)	0.653
Geography/culture, extra-cards WWII coding	US versus India (observations = 3,668; observations = 492)	0.77 (0.10)	0.72 (0.14)	0.763
Geography/culture, AB typing task	Red states versus blue states (observations = 5,062; observations = 3,464)	0.94 (0.03)	0.96 (0.04)	0.760
Output, AB task	25th percentile versus 75th percentile (observations = 10,471)	0.94 (0.03)	0.95 (0.03)	0.724
Other selection, AB task	Enrollment in week 1 versus weeks 2–3 (observations = 6,359; observations = 4,112)	0.95 (0.03)	0.95 (0.04)	0.950
	Night versus day (observations = 4,556; observations = 5,195)	0.94 (0.03)	0.97 (0.04)	0.625
Other selection, ten-minute WWII coding task	Enrollment in week 1 versus weeks 2–3 (observations = 1,569; observations = 968)	0.43 (0.18)	0.58 (0.22)	0.591
	Night versus day (observations = 949; observations = 1,338)	0.37 (0.18)	−0.05 (0.21)	0.131
Other selection, WWII coding extra cards	Enrollment in week 1 versus weeks 2–3 (observations = 2,641; observations = 1,793)	0.88 (0.06)	0.83 (0.10)	0.639
	Night versus day (observations = 1,600; observations = 2,428)	0.88 (0.06)	0.82 (0.09)	0.542

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.

benchmark one. Finally, in columns 6 and 7, for each treatment, we compare the difference in effort in log points (column 6) or *z*-scores (column 7) relative to the baseline group. We then compute for each treatment the absolute difference in this effect across the 2 versions–say, between male subjects and female subjects—and then average across the 14 treatments. This measure also yields similar findings.

Alternative Comparisons of Designs: So far, we focused on 10 (prespecified) design changes. In Table 3, we consider 12 additional design comparisons.

The first three comparisons present the familiar demographic comparisons, but for the ten-minute WWII card coding task.[18] The rank-order correlations across the demographics are lower, given the noisiness of the WWII card estimates, but close to the full-stability benchmarks.

[18] We cannot make this comparison for the extra-cards 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.

We also revisit the geographic/culture comparisons. First, comparing between the India and United States sample, but for the extra-cards WWII card coding task,[19] we find a correlation of 0.72, close to the full-stability benchmark of 0.77. Second, returning to the a-b task, we compare 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: the twenty-fifth and seventy-fifth percentile of effort. Within the a-b task, in treatments where the twenty-fifth percentile worker exerts high effort, so does the, seventy-fifth percentile worker. In online Appendix Figures 6 and 7 and online Appendix Table 5, we replicate the key results using the twenty-fifth or seventy-fifth percentile of effort instead of the average effort.

Finally, we consider two further forms of sample selection that have been identified as potentially important for the productivity of MTurk workers (Case et al. 2017): (i) whether subjects sign up early on or later on in a experimental study, 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.

C. *Structural Estimates*

We now present estimates of the model in Section IA 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 sixteenth treatment, which we have not discussed so far; and (iv) to create a full-stability benchmark for design changes in which such a benchmark cannot be created with bootstraps from the data.

Estimation: We take the model in Section IA with an exponential cost of effort function, which conveniently implies a specification that expresses effort as a function of the motivation parameters. Building on DellaVigna et al. (2018), 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,

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

Equation (3) can be estimated with nonlinear least squares (NLS). The three parameters, $\hat{s}$, $\hat{k}$, and $\hat{\gamma}$, are overidentified given the four piece rates. We specify effort

[19] We pool across versions 3 and 4. We do not do such a comparison for the ten-minute WWII task given the noisiness of the estimates, since the Indian workers constitute only 12 percent of the sample.

e_j in the a-b button-pushing task as the number of button presses, in the ten-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.

In Table 4, we present estimates of the parameters using all 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 datasets, the elasticity of effort is precisely estimated to be about 0.04. Figure 2, panels A–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 four piece rates with three parameters.

The next rows show the estimates of the behavioral parameters. The estimate for the social comparison treatment $\widehat{\Delta s_{SC}} = 0.06$ indicates an impact equivalent to an incentive of \$0.06 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 $\widehat{\Delta s_{CO}} \approx 0$.

We estimate a precisely estimated zero effect for the altruism parameter, with point estimates $\hat{\alpha} = 0.003$ (standard error 0.010) for 2015 and $\hat{\alpha} = 0.010$ (standard error 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 $\pi(0.01) < 0.01$, 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).[20] 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 (standard error 0.003) for males but 0.019 (standard error 0.007) for females. The behavioral parameters, and especially the social preference parameters, are consistent across demographics.

[20] Online Appendix Table 6 reports the estimates on the pooled 2015 and 2018 sample, split also by education and age.

Table 4—Structural Estimates

	Button-pushing task, 10 min		Demographics, typing task, pooled 2015–2018		2018 WWII cards-coding task		
Parameters	2015 exp. (1)	2018 exp. (2)	Male (3)	Female (4)	Ten-min (5)	Extra work (6)	Extra work, no consent (7)
Category: Incidental parameters							
Curvature of cost of effort γ	0.016	0.012	0.012	0.019	2.000	0.045	0.055
	(0.004)	(0.005)	(0.003)	(0.007)		(0.014)	(0.014)
Implied elasticity	0.034	0.044	0.043	0.028	0.009	0.430	0.354
	(0.008)	(0.017)	(0.012)	(0.009)		(0.134)	(0.087)
Level of cost of effort k	−36.427	−29.183	−29.828	−42.500	−114.743	−3.469	−4.312
	(8.283)	(10.160)	(7.375)	(13.252)	(2.205)	(1.425)	(1.255)
Baseline motivation s	3.3e−04	5.1e−04	3.9e−04	2.2e−04	0.084	0.204	0.097
	(7.9e−04)	(0.002)	(0.001)	(7.3e−04)	(0.362)	(0.191)	(0.083)
Category: "Pay enough or don't pay"							
Δs_{CO}	−0.005	0.011	−0.051	0.104	1.6e+05	0.069	0.052
	(0.099)	(0.177)	(0.066)	(0.244)	(6.9e+05)	(0.102)	(0.076)
Category: Social pref. parameters							
Pure altruism alpha	0.003	0.010	0.009	8.7e−04	0.017	0.011	0.007
	(0.010)	(0.017)	(0.013)	(0.009)	(0.518)	(0.028)	(0.020)
Warm glow a	0.135	0.075	0.111	0.094	0.754	0.205	0.267
	(0.133)	(0.132)	(0.128)	(0.134)	(3.535)	(0.222)	(0.182)
Category: Social pref.: gift exch.							
Δs_{GE}	8.4e−04	0.001	0.001	5.1e−04	−0.083	0.831	0.908
	(0.002)	(0.004)	(0.002)	(0.001)	(0.360)	(0.348)	(0.356)
Category: Discounting							
Beta	1.075	0.855	0.769	1.422	228.893	5.396	0.215
	(1.122)	(1.306)	(0.931)	(1.861)	(2.1e+03)	(6.028)	(0.227)
Delta (weekly)	0.767	0.926	0.780	0.817	0.726	0.435	1.317
	(0.243)	(0.416)	(0.280)	(0.327)	(1.971)	(0.201)	(0.380)
Category: Social comparisons							
Δs_{SC}	0.055	0.079	0.068	0.039	−0.079	−0.018	0.024
	(0.065)	(0.132)	(0.086)	(0.066)	(0.357)	(0.071)	(0.044)
Category: Ranking							
Δs_{R}	0.014	0.015	0.015	0.009	1.864	0.001	0.103
	(0.020)	(0.033)	(0.025)	(0.019)	(7.787)	(0.071)	(0.075)
Category: Task significance							
Δs_{TS}	0.010	0.012	0.015	0.005	0.533	−0.007	0.143
	(0.015)	(0.028)	(0.024)	(0.011)	(2.569)	(0.070)	(0.092)
Category: Probability weighting parameters							
Pi (0.01)	0.001	0.010	0.002	0.001	0.626	0.013	0.005
	(0.001)	(0.008)	(0.002)	(0.002)	(2.518)	(0.006)	(0.003)
Pi (0.50)	0.207	0.087	0.171	0.169	1.5e−04	0.177	0.249
	(0.147)	(0.121)	(0.145)	(0.171)	(0.005)	(0.127)	(0.135)
Observations	8,252	2,219	4,686	5,785	2,537	2,188	2,246
Average effort	1,892.846	1,835.815	1,930.581	1,839.236	56.508	11.190	11.309
Root MSE	659.201	643.399	724.144	591.918	24.214	56.233	51.906
Category: Extra treat.: incentive + please try							
Out-of-sample pred.		1,979			57.03	12.91	13.62
		(47)			(1.089)		
Actual		2,056			56.18	10.81	13.30
		(46)			(1.756)	(0.785)	(0.738)

Notes: The table shows structural estimates of the incidental parameters (γ, k, and s) and psychological parameters estimated using all 15 treatments across 11 different samples. All models assume an exponential cost function. Columns 1–5 are estimated using NLS, while Columns 10–11 are estimated with maximum likelihood due to censoring. Column 1 refers to the 2015 typing task, Column 2 to the 2018 typing task. Columns 2–3 pool the 2015 and 2018 typing tasks but restrict the sample to a demographic subset. Columns 5–7 show estimates on the 2018 card coding treatments. Standard errors in parentheses.

Estimates, **Ten-Minute WWII Task:** For the 10-minute WWII task (column 5), we estimate an elasticity smaller than 0.01, really tiny, consistent with the very limited response to incentives.[21] Given the very small elasticity, the estimates of the key parameters are necessarily noisy.

Estimates, **Extra Work:** For the extra-work WWII coding (Columns 6 and 7), we estimate the model by maximum likelihood, accounting for censoring at 0 cards coded and 20 cards coded. The elasticity of effort to incentives, 0.43 in column 6, is among the highest in the literature (e.g., 0.1 for stuffing envelopes in DellaVigna et al. 2018 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. Figure 2, panels D–E show that the structural estimates capture well the 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 our sixteenth 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 use these estimates to compute a structural full-stability benchmark, as opposed to a bootstrap-based benchmark. We assume that the behavioral parameters remain constant across design changes but that the incidental parameters change. Consider, for example, the task change. We compare a simulated sample from the estimated a-b task parameters with a simulated sample that combines the behavioral 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. More precisely, (i) for the a-b task, we draw a sample of 700 observations per treatment given the a-b task structural estimates; (ii) 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 *behavioral* parameters for the a-b task; (iii) we compute the rank-order correlation of the sample means; and (iv) we repeat 1,000 times. As column 2 in Table 2 shows, the mean structural full-stability rank-order correlation is 0.50 (standard error 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.

For the output comparison, the structural full-stability benchmark is 0.58 (standard error 0.17), indicating that the observed correlation of 0.27 is largely (but not fully) explained by the noise. For the comparison between the a-b task and the

[21] 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.

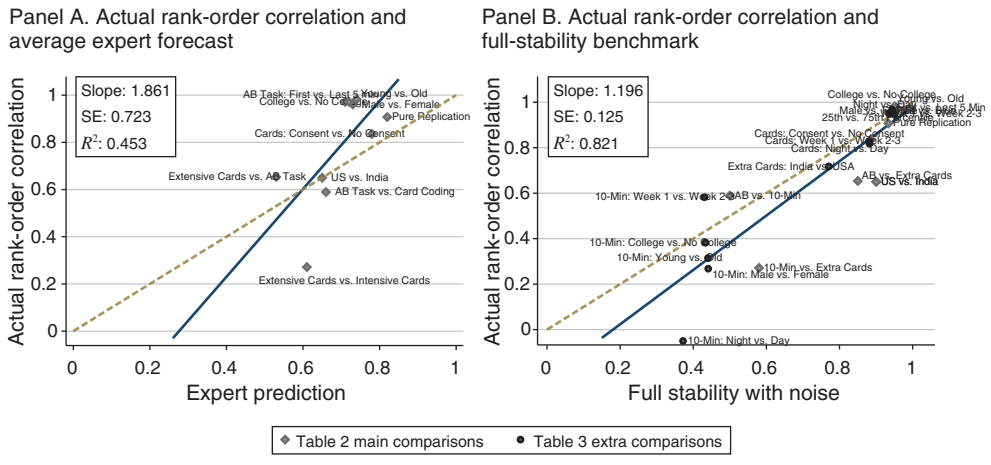

FIGURE 9. ACTUAL RANK-ORDER CORRELATION, AVERAGE FORECAST, AND FULL-STABILITY BENCHMARK

Notes: Panel A 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. Panel B presents the actual rank-order correlation versus the full-stability benchmark. For the full-stability benchmark, when both benchmarks are available, we use the data-based one. In panel B, we plot both the ten-benchmark version changes (Table 2) and the additional version changes (Table 3).

extra-work WWII task, the full-stability benchmark is much higher at 0.85 (standard error 0.07) given that the noise term is relatively small in both tasks.

Table 2 also shows that the structural full-stability benchmarks are very similar to the bootstrap-based ones for the demographic comparisons, validating the structural measure.[22]

D. *Summary: Predictors of Stability*

To summarize the results, we compare the predictive power of forecasts versus that of the full-stability benchmark. In Figure 9, panel A, we plot for each of the ten design changes the average expert forecast of correlation versus the actual correlation. In Figure 9, panel B, we plot the full-stability benchmark versus the actual correlation, including not only the ten 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.

[22] We do not produce a structural full-stability benchmark for the geographic comparison given that in the (relatively small) India sample, the response to incentives is so noisy that we cannot obtain reliable parameter estimates.

IV. 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 ten-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 ten-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 standard error) under the three key piece rate treatments, indicating a flat and nonmonotonic response to incentives in the ten-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, they do not appear to take sufficiently into account the noisiness of an experimental setup.

Forecaster Effort: As we document in DellaVigna and Pope (2018b), forecasters who appear to put in more effort by taking longer time and by clicking on links do a bit better in the their forecasts (at least in some conditions). In this setting, though, we do not find a difference splitting by the time taken to do the survey (columns 4 and 5).

Confidence: The forecasters indicated how many of their forecasts they expected to be within 0.1 of the truth. As the bottom row of Table 2 shows, faculty experts are overconfident, expecting on average to get 3.99 responses (standard error 0.24) close to the truth, while the average actual is 3.22 (standard error 0.23). PhD students and MTurk forecasters are even more overconfident.

Still, is there information in the confidence response? Figure 10 shows the number of forecasts within 0.1 of the truth for the forecasters making that forecast, pooling across faculty and PhD forecasters. 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.[23] 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

[23] Online Appendix Figure 8 displays the same evidence with respect to the absolute forecast error.

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

	Average forecast of rank-order correlation for the 15 treatments across designs						
		Version		Time spent on survey		Confidence	
Version comparison	Pooled experts and PhDs (1)	Info on piece rate (2)	No info on piece rate (3)	Long (18 mins+) (4)	Short (<18 mins) (5)	High (4+ corr. within 0.1) (6)	Low (<4 corr. within 0.1) (7)
Category: Pure repl.							
2015 AB task versus 2018 AB task	0.84 (0.01)	0.84 (0.01)	0.83 (0.02)	0.85 (0.01)	0.83 (0.02)	0.86 (0.01)	0.81 (0.02)
Category: Demogr., typing task							
Male versus female	0.75 (0.01)	0.76 (0.02)	0.74 (0.02)	0.76 (0.02)	0.73 (0.02)	0.77 (0.02)	0.71 (0.03)
College versus no college	0.72 (0.01)	0.72 (0.02)	0.73 (0.02)	0.75 (0.02)	0.70 (0.02)	0.76 (0.02)	0.67 (0.03)
Young (≤30) versus 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)	0.70 (0.02)
Category: Geogr./culture							
US versus India	0.65 (0.02)	0.65 (0.02)	0.65 (0.03)	0.67 (0.02)	0.62 (0.03)	0.69 (0.02)	0.58 (0.03)
Category: Task							
AB task versus card coding	0.65 (0.02)	0.62 (0.03)	0.69 (0.03)	0.66 (0.03)	0.64 (0.03)	0.67 (0.02)	0.62 (0.03)
Category: Output							
Ten-min cards versus extra cards	0.61 (0.02)	0.60 (0.03)	0.63 (0.03)	0.62 (0.03)	0.61 (0.03)	0.65 (0.02)	0.56 (0.03)
Extra cards versus AB task	0.54 (0.02)	0.54 (0.03)	0.54 (0.03)	0.57 (0.03)	0.51 (0.03)	0.60 (0.02)	0.45 (0.03)
AB task: first five min versus last five min	0.71 (0.02)	0.72 (0.02)	0.70 (0.03)	0.71 (0.03)	0.71 (0.03)	0.74 (0.02)	0.66 (0.03)
Category: Consent							
Cards: consent versus no consent	0.79 (0.01)	0.81 (0.02)	0.78 (0.02)	0.80 (0.02)	0.79 (0.02)	0.81 (0.01)	0.76 (0.03)
Observations	88	48	40	44	44	54	34
Average ind. abs. error	0.19 (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 (0.04)	0.17 (0.04)	0.15 (0.03)	0.17 (0.04)	0.15 (0.04)	0.19 (0.04)

Notes: The table considers the forecasts of subgroups. Column 1 presents the results for the overall group of academic experts and PhDs. In Columns 2 and 3, we split this group depending on whether the respondents were randomized to be provided information on the average effort by piece rate or not. In Columns 4 and 5, we split by the time taken to complete the survey. In Columns 6 and 7, we split by the expressed degree of confidence in the forecast. Bolded values are significantly different from one another at the 95 percent confidence level.

individual forecasts. We 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).

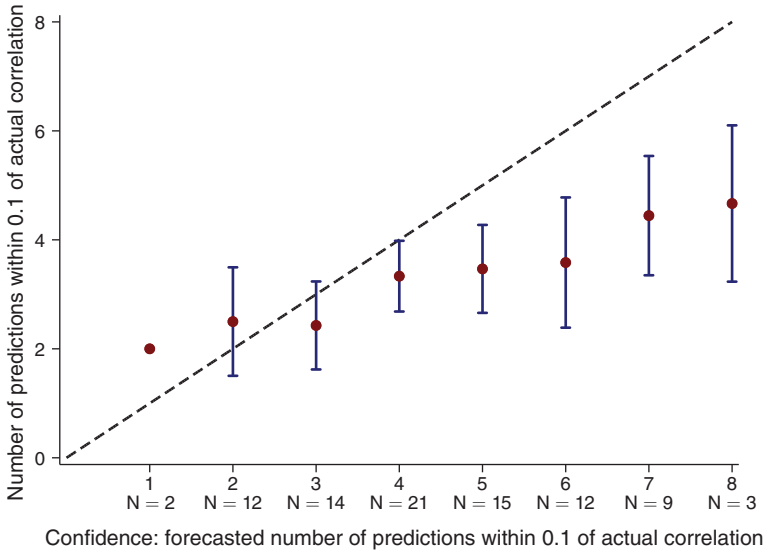

FIGURE 10. CONFIDENCE (IN THE FORECAST OF RANK-ORDER CORRELATION) AND ACCURACY

Notes: In the survey of forecasters, as a last question, we asked the expected number of forecasts of rank-order correlation that the forecasters expected to get within 0.1 of the correct answer. In Figure 10, we 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 dashed line is the 45-degree line indicating an unbiased estimate.

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. But does this ability as "superforecasters" (Tetlock and Gardner 2015; Mellers et al. 2015) translate across experiments? For the 35 individuals who made forecasts both in 2015 and in 2018, online Appendix Figure 9a provides no evidence of a correlation between their average absolute error (in terms of point) in 2015 and (in terms of rank-order correlation) in 2018. 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. While the wisdom-of-crowd forecast for a treatment is generally predictive of the average effort in that treatment, the average forecast underpredicts effort in the very-low-pay treatment and overpredicts effort for the probability weighting treatment and 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 whether 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 online Appendix Figure 9b 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 three, four, five, and six 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.

V. 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 the geographic/cultural mix of subjects, changes in the task and 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 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 setup 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 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 simple measure of stability with desirable properties.

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 an important role for noise in the experimental results: 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 role for noise, even when provided with diagnostic information.

What can explain the divergence between the replication results and the expectations of experts? While we do not have direct evidence, 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, 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* 101 (2): 470–92.

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 130 (3): 1117–65.

Araujo, Felipe A., Erin Carbone, Lynn Conell-Price, Marli W. Dunietz, Ania Jaroszewicz, Rachel Landsman, Diego Lamé, Lise Vesterlund, Stephanie W. Wang, and Alistair J. Wilson. 2016. "The Slider Task: An Example of Restricted Inference on Incentive Effects." *Journal of the Economic Science Association* 2 (1): 1–12.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review* 99 (1): 544–55.

Bandiera, Oriana, Greg Fischer, Andrea Prat, and Erina Ytsma. 2021. "Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments." *American Economic Review: Insights* 3 (4): 435-54.

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* 125 (1): 263–306.

Bhargava, Saurabh, and Dayanand Manoli. 2015. "Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment." *American Economic Review* 105 (11): 3489–3529.

Camerer, Colin, et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–36.

Camerer, Colin, et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015." *Nature Human Behavior* 2: 637–44.

Case, Logan S., Jesse Chandler, Adam Seth Levine, Andrew Proctor, and Dara Z. Strolovitch. 2017. "Intertemporal Differences among MTurk Workers: Time-Based Sample Variations and Implications for Online Data Collection." *Sage Open* 7 (2).

Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56 (3): 920–80.

Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature* 47 (2): 448–74.

DellaVigna, Stefano. 2018. "Structural Behavioral Economics." In *Handbook of Behavioral Economics*, Vol. 1, edited by Doug Bernheim, Stefano DellaVigna, and David Laibson. Amsterdan: Elsevier.

DellaVigna, Stefano, and Devin Pope. 2018a. "What Motivates Effort? Evidence and Expert Forecasts." *Review of Economic Studies* 85 (2): 1029–69.

DellaVigna, Stefano, and Devin Pope. 2018b. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy* 126 (6): 2410–56.

DellaVigna, Stefano, and Devin Pope. 2021. "Replication data for: Stability of Experimental Results: Forecasts and Evidence." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.38886/E135221V1.

de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and Bounding Experimenter Demand." *American Economic Review* 108 (11): 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* 112 (50): 15343–47.

Falk, Armin, and James J. Heckman. 2009. "Lab Experiments Are a Major Source of Knowledge in the Social Sciences." *Science* 326 (5952): 535–38.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–05.

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* 94 (3): 653–63.

Gneezy, Uri, and John A. List. 2006. "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments." *Econometrica* 74 (5): 1365–84.

Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature* 42 (4): 1009–55.

Horton, John J., David G. Rand, and Richard J. Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14 (3): 399–425.

Imas, Alex. 2014. "Working for the 'Warm Glow': On the Benefits and Limits of Prosocial Incentives." *Journal of Public Economics* 114: 14–18.

Klein, Richard A., et al. 2014. "Investigating Variation in Replicability: A 'Many Labs' Replication Project." *Social Psychology* 45 (3): 142–52.

Klein, Richard A., et al. 2018. "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings." *Advances in Methods and Practices in Psychological Science* 1 (4): 443–90.

Landy, Justin F., et al. 2020. "Crowdsourcing Hypothesis Test: Making Transparent How Design Choices Shape Research Results." *Psychological Bulletin* 146 (5): 451–79.

Mellers, Barbara, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar, and Philip Tetlock. 2015. "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions." *Perspectives on Psychological Science* 10 (3): 267–81.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): aac4716.

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

Rothwell, Peter M. 2005. "External Validity of Randomised Controlled Trials: 'To Whom Do the Results of This Trial Apply?'" *Lancet* 365 (9453): 82–93.

Snowberg, Erik, and Leeat Yariv. 2021. "Testing the Waters: Behavior across Participant Pools." *American Economic Review* 111 (2): 687–719.

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

Tversky, Amos, and Daniel Kahneman. 1971. "Belief in the Law of Small Numbers." *Psychological Bulletin* 76 (2): 105–10.

Vivalt, Eva. 2020. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economic Association* 18 (6): 3045–89.