Online Appendix

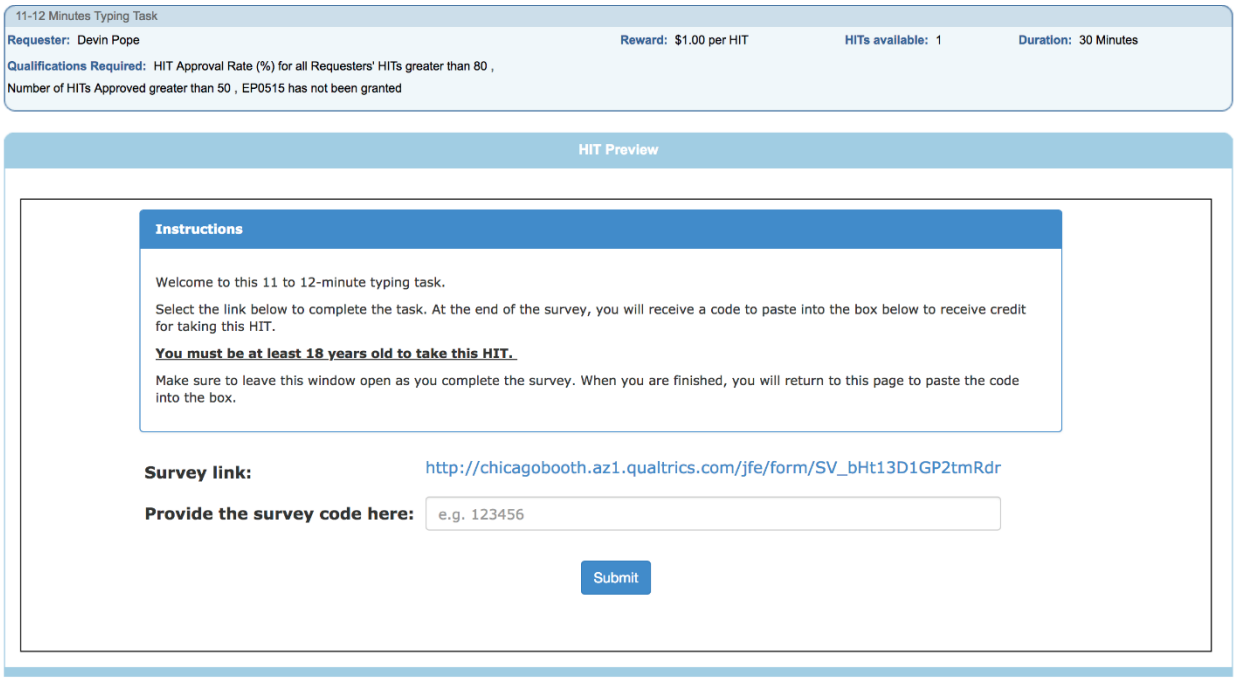
"Stability of Experimental Results: Forecasts and Evidence"

Stefano DellaVigna and Devin Pope

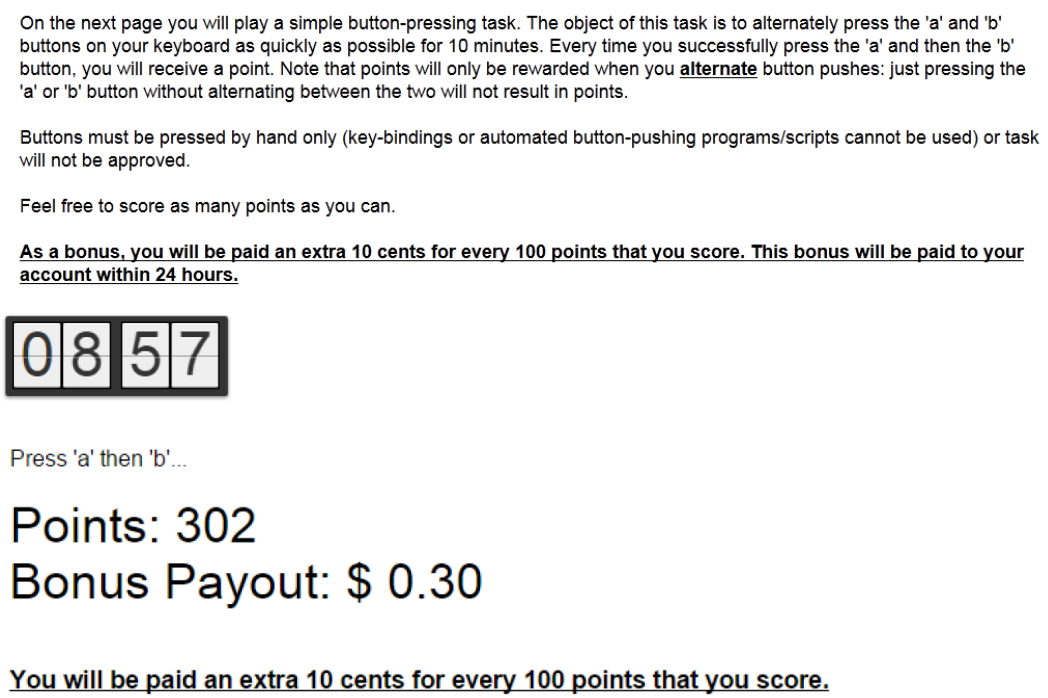
AEJ: Micro

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

Online Appendix Figure 1a. Recruitment Ad on MTurk



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



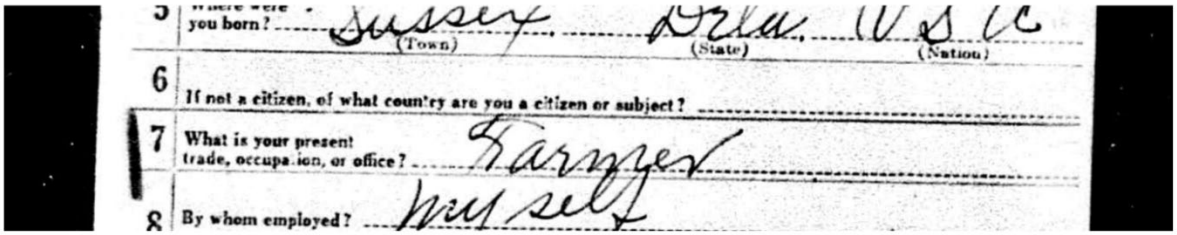
Online Appendix Figure 1c. Screenshot for WWII 10-minute Card Coding Task, Example

Time remaining: 9 Minutes, 55 Seconds

You have completed 4 cards.

Your current bonus is $0.02.

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



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

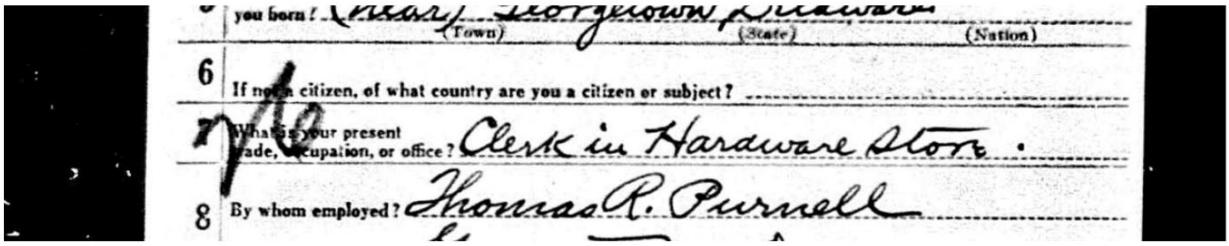
Type occupation here:

→

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

You have completed 3 of 40 required cards.

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



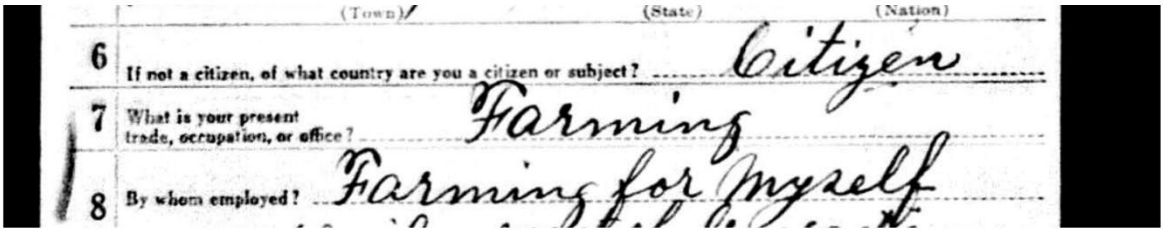
Type occupation here:

→

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

You have completed 1 additional cards.

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



The number of additional cards you complete will not affect your payment in any way.

Please click "I'm Finished" if you want to exit the survey, or click "Continue" if you want to work on more cards.

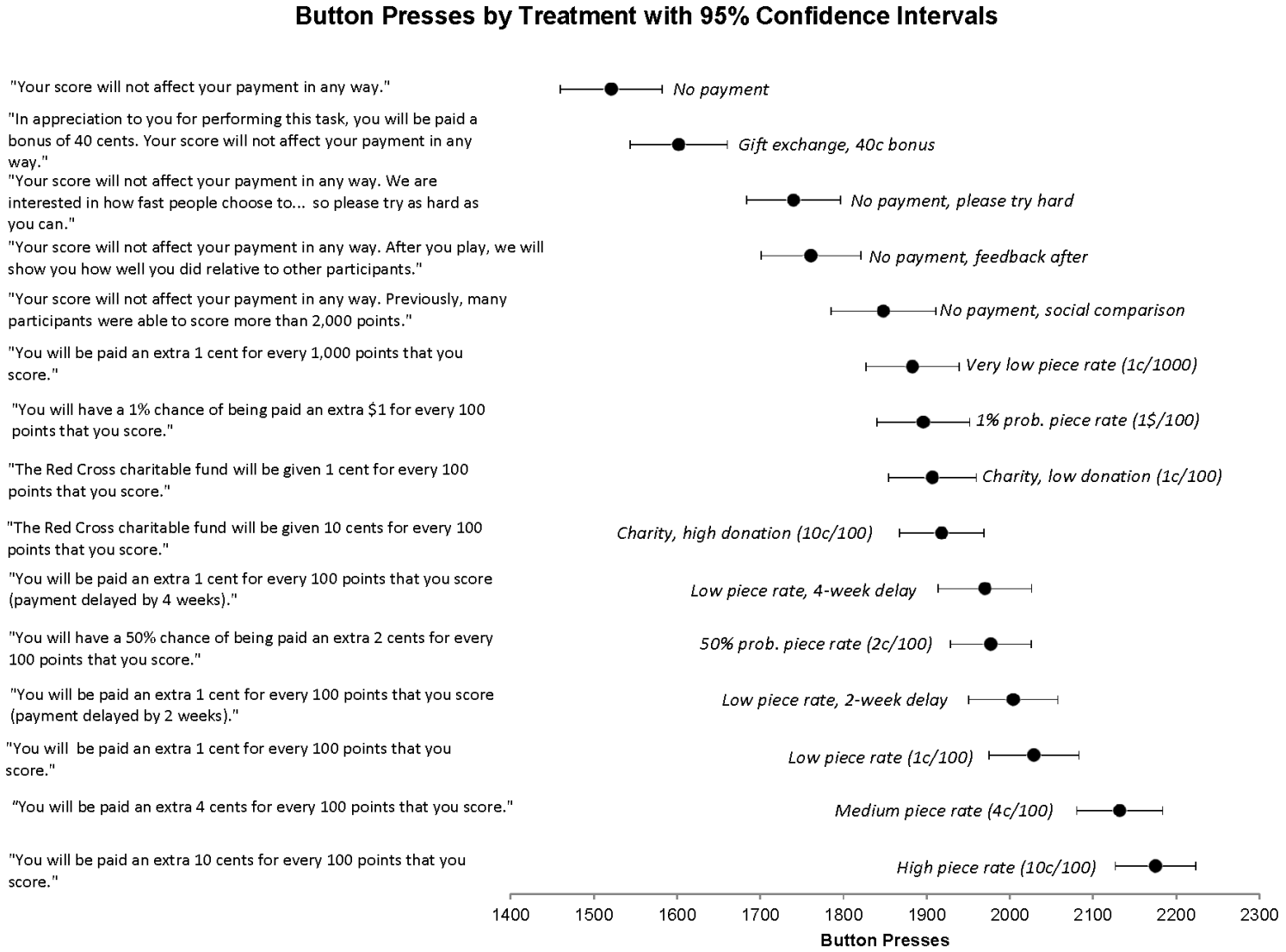
Type occupation here:

Continue

I'm Finished

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

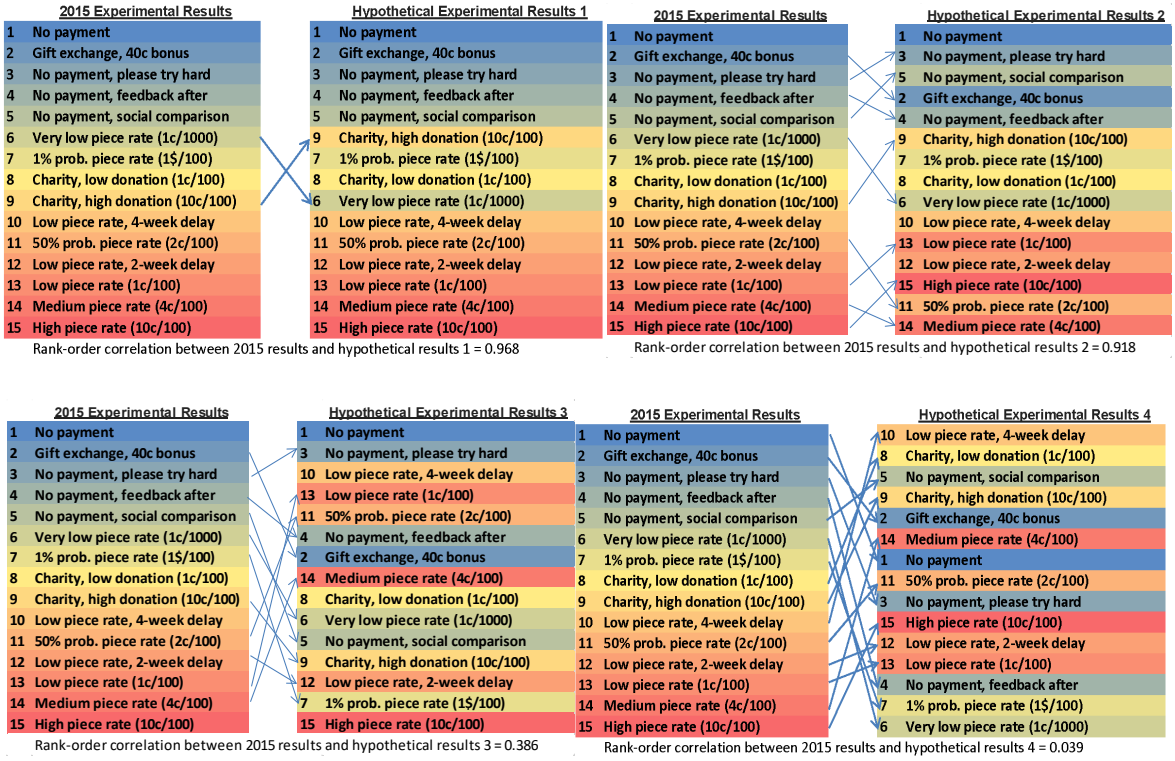
Online Appendix Figure 2. Summary of Treatments and Results from DellaVigna and Pope (2018)



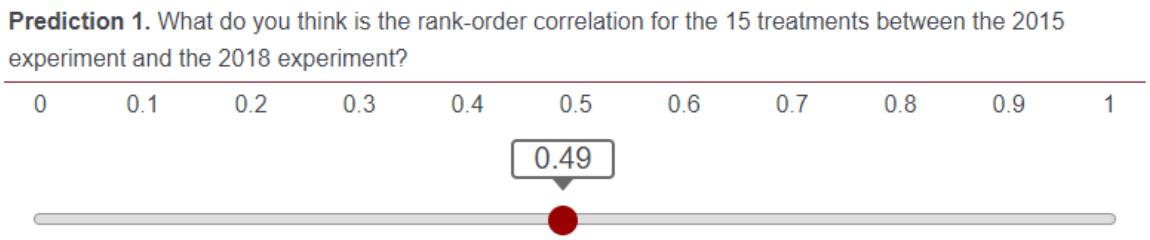
Notes: The figure summarizes the key wording as well as the average effort and standard error for the mean effort in the 2015 experimental results of DellaVigna and Pope (2018a) for the 15 treatments which we replicate. This image is as presented to the forecasters.

Online Appendix Figure 3. Expert Survey, Screenshots

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



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



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

Online Appendix Figure 4. Distribution of Effort Across All Treatments

Online Appendix Figure 4a. 2015 MTurk Button Pushing Task

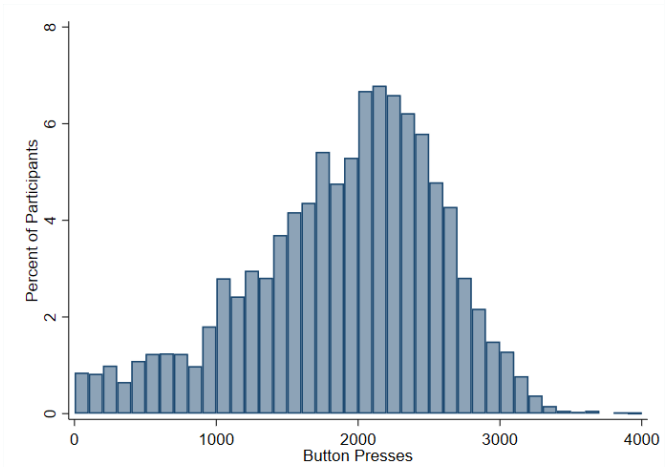


Figure 4b. 2018 MTurk Button Pushing Task

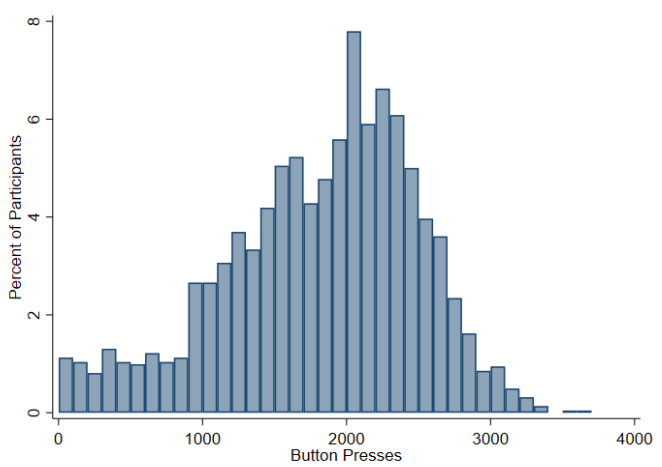


Figure 4c. 2018 10-Minute Card Coding Task

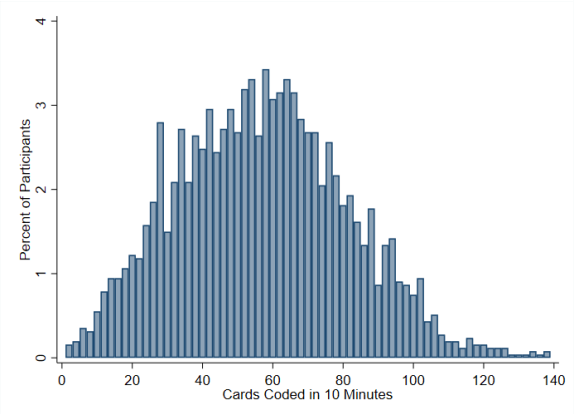


Figure 4d. 2018 Extra Card Coding Task

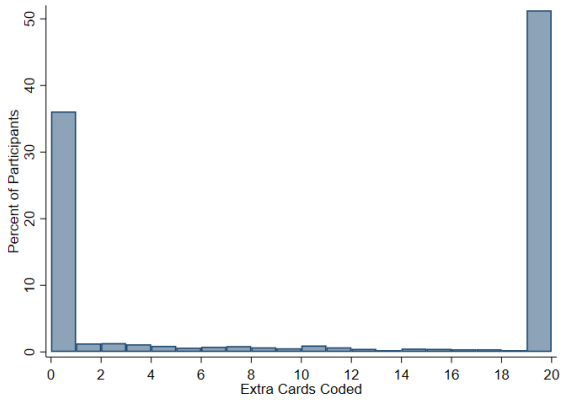
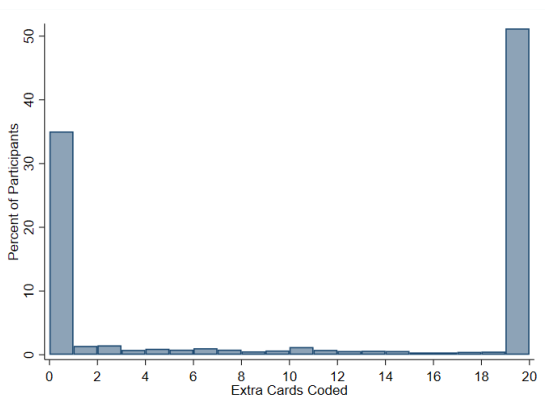
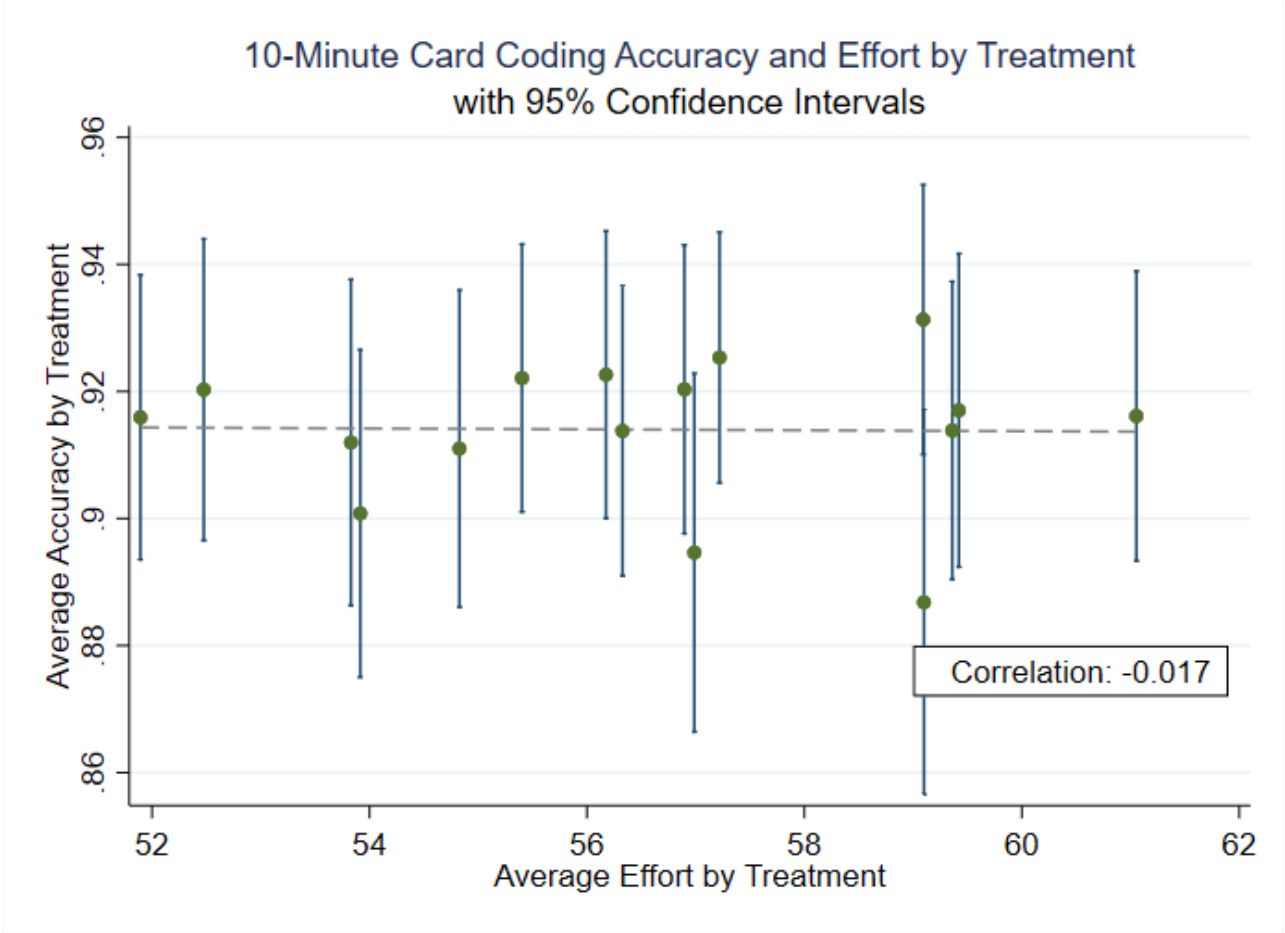


Figure 4e. 2018 Extra-Card Coding Task,No Consent



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

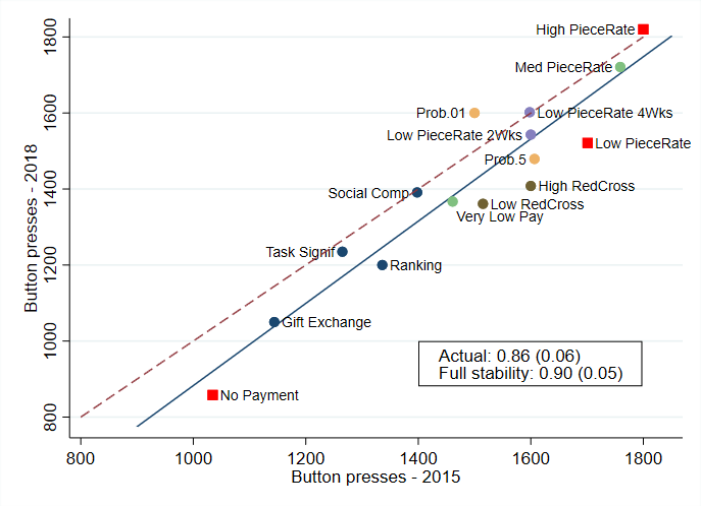
Online Appendix Figure 5. Average Accuracy and Effort by Treatment in the 10-Minute Card Coding Experiment



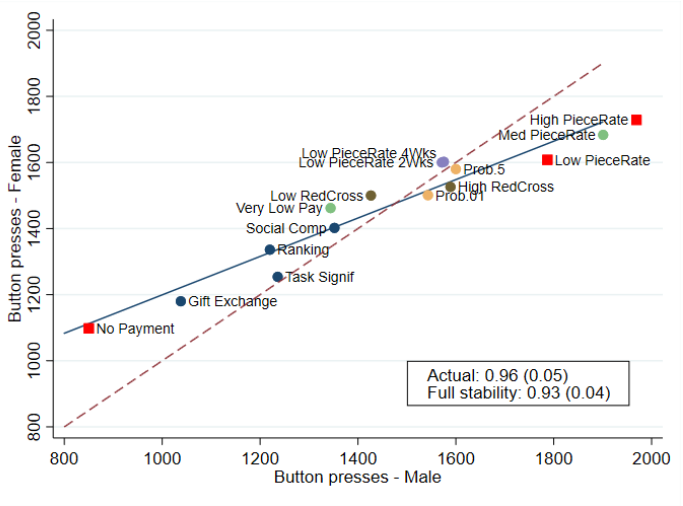
Notes: The figure displays evidence on accuracy for the 10-minute WWII coding task. The graph plots the average effort by treatment (on the x axis) against the average accuracy of coding (on the y axis). The measure of accuracy is the share of cards coded correctly, where we only considered cards for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

Online Appendix Figure 6. Comparison Across Versions, 25th Percentile of Effort

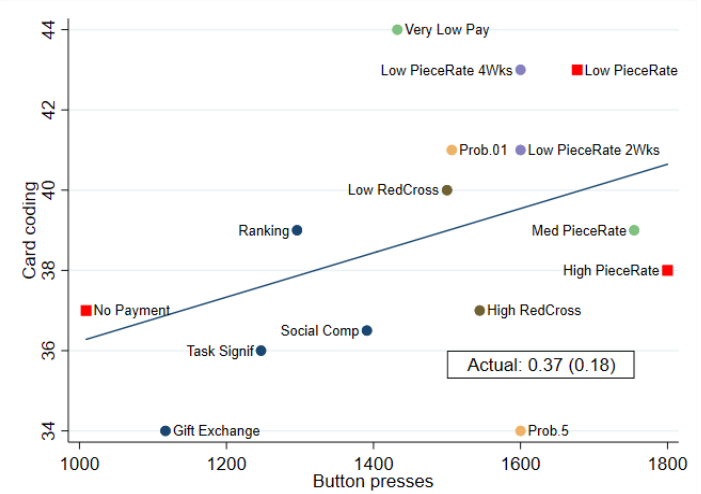
Onl. App. Figure 6a. Pure Replication, Button Pushing Task



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



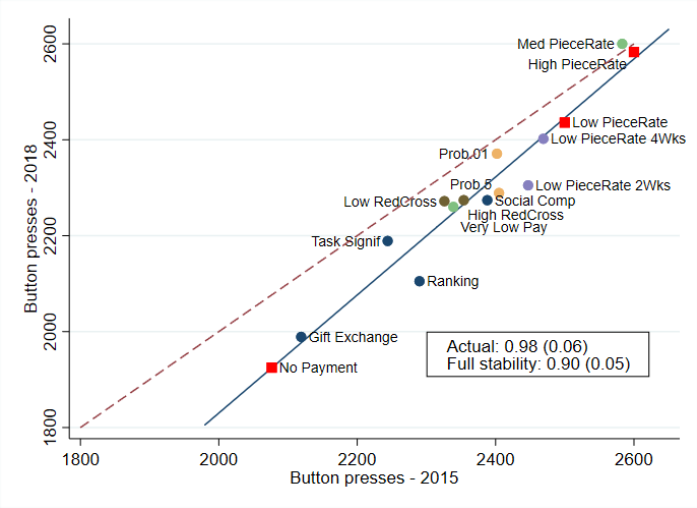
Onl. App. Figure 6c. Impact of Task



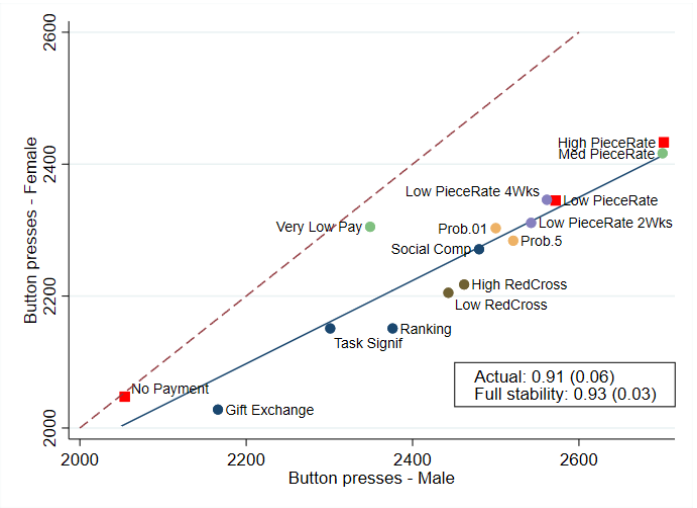
Notes: Online Appendix Figure 6a-c presents the equivalent material as in Figures 3, 4a, and 6, except that for each treatment we plot the 25th percentile of effort, instead of the mean effort as in the original figures. We do not plot these figures for comparisons involving the extra-work card task,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

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

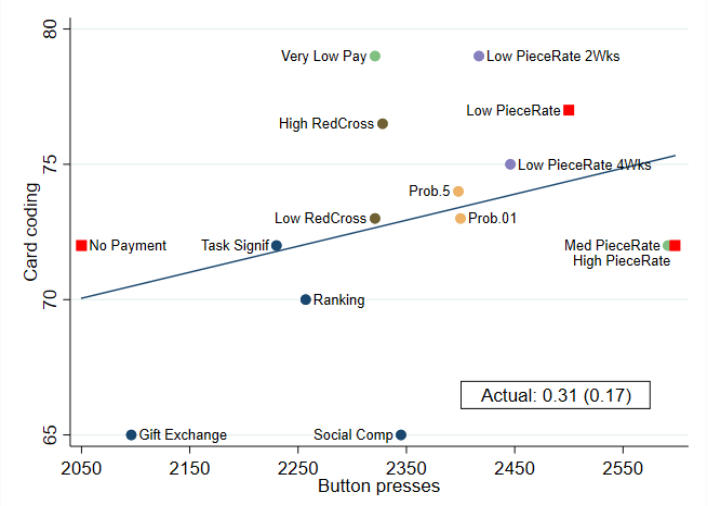
Onl. App. Figure 7a. Pure Replication, Button Pushing Task



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

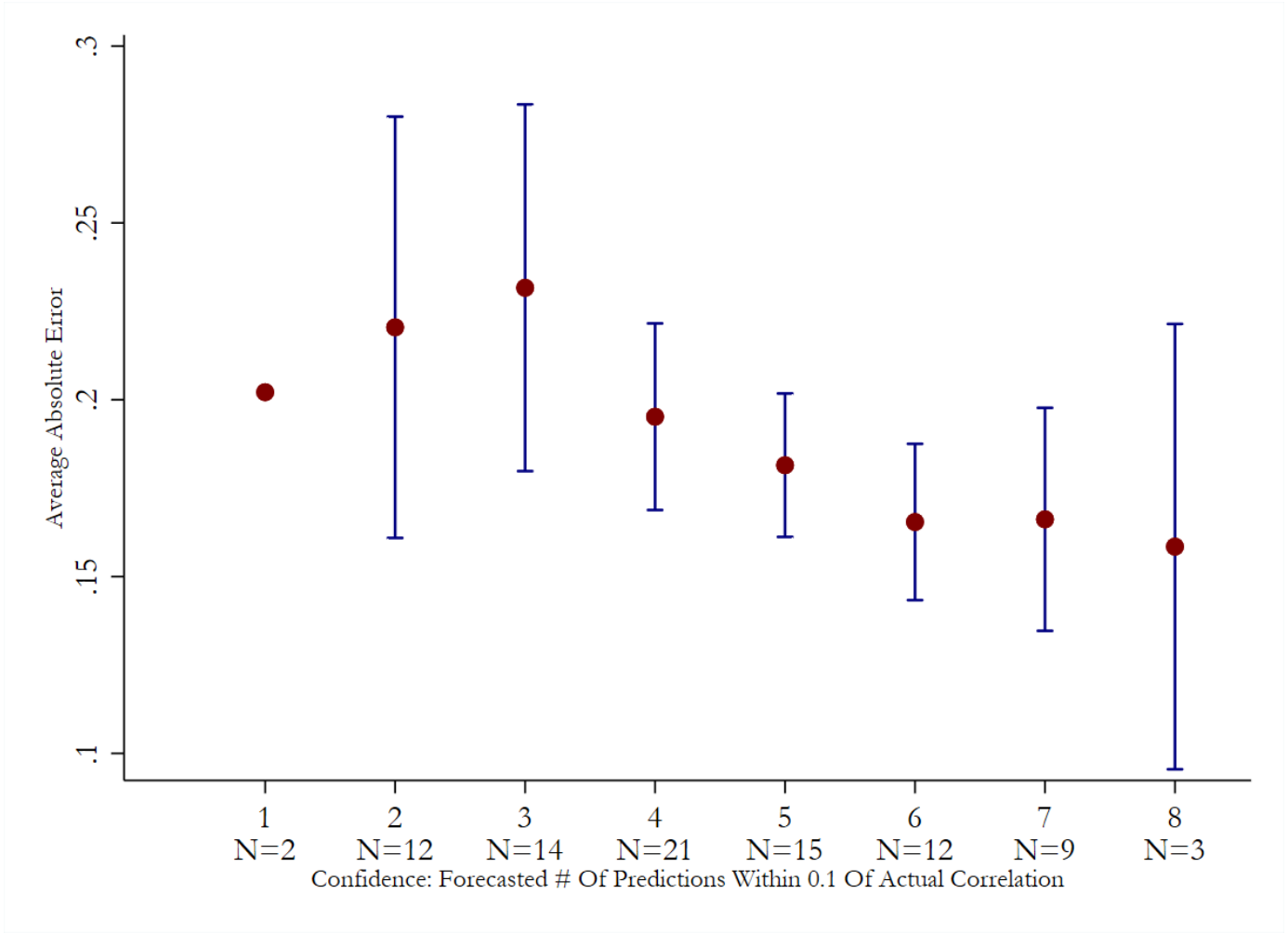


Onl. App. Figure 7c. Impact of Task



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

Online Appendix Figure 8. Confidence (in the Forecast of Rank-Order Correlation) and Average Absolute Error



Notes: In the survey of forecasters, as last question we asked the expected number of forecasts of rank-order correlation which the forecasters expected to get within 0.1 of the correct answer. In the figure we plot the average absolute error in the forecast, splitting by the measure of confidence, that is, the forecast (rounded to the closest round number) of the number of "correct" predictions. The sample includes academic experts, as well as PhDs.

Online Appendix Figure 9. Revisiting the 2015 Expert Forecasts

Figure 9a. Accuracy of 2015 Forecasts vs. 2018 Forecasts

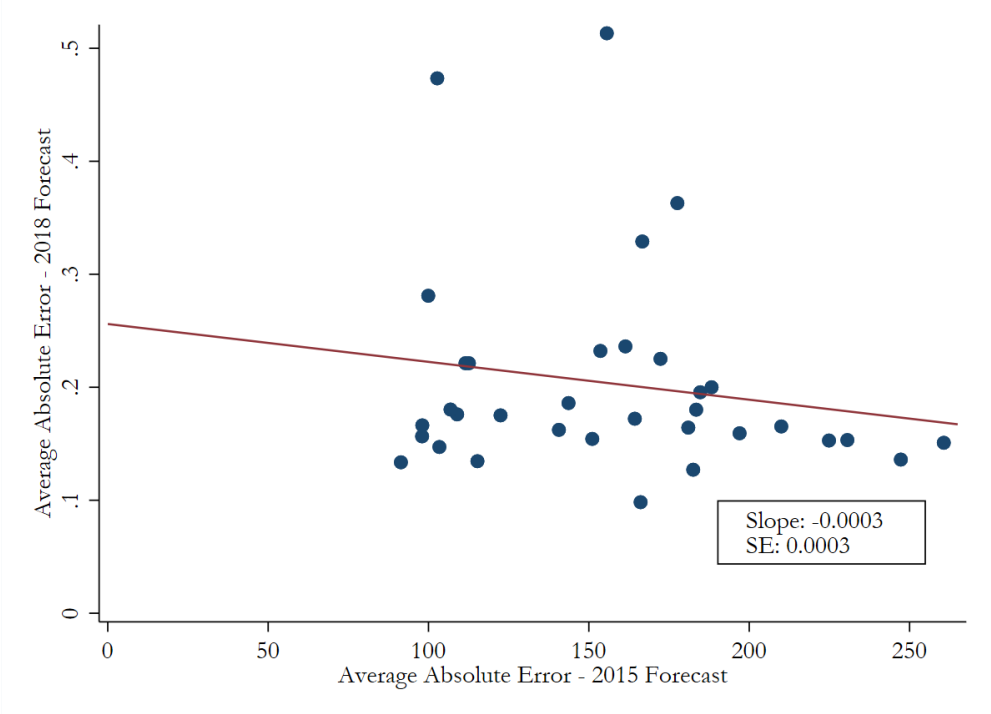
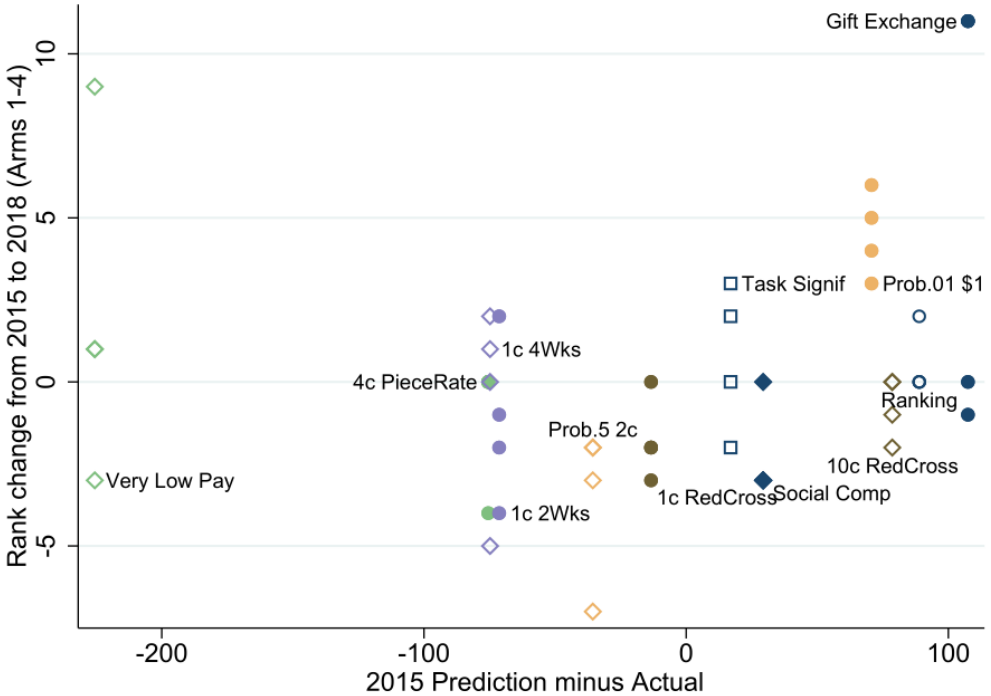


Figure 9b. Errors in 2015 Forecasts and Changes of Treatment Rank in 2018 Experiments



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

Online Appendix Table 1. Observation Counts by Treatment

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

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

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

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

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

Online Appendix Table 3. Accuracy in the 2018 Card-Coding Task

Category	Treatment Wording	10-Minute Card Coding	Required Cards, Pooled	Extra Cards, Pooled
		(1)	(2)	(3)
Piece Rate	No payment	0.912	0.928	0.920
		(0.013)	(0.009)	(0.018)
	Low piece rate	0.914	0.912	0.922
		(0.012)	(0.010)	(0.014)
	Medium piece rate	0.925	0.921	0.936
		(0.010)	(0.009)	(0.011)
	High piece rate	0.914	0.896	0.884
		(0.012)	(0.011)	(0.016)
Pay Enough or Don't Pay	Very low piece rate	0.916	0.919	0.898
		(0.012)	(0.009)	(0.02)
Social Preferences: Charity	Charity, low donation	0.920	0.932	0.906
		(0.012)	(0.008)	(0.017)
	Charity, high donation	0.895	0.920	0.929
		(0.014)	(0.009)	(0.015)
Social Pref: Gift Exchange	Gift exchange, 40c bonus	0.916	0.928	0.934
		(0.011)	(0.009)	(0.013)
Discounting	Low piece rate, 2-week delay	0.917	0.922	0.920
		(0.013)	(0.01)	(0.015)
	Low piece rate, 4-week delay	0.887	0.906	0.899
		(0.015)	(0.01)	(0.017)
Risk Aversion and Probability Weighting	1% prob. Piece rate	0.931	0.929	0.943
		(0.011)	(0.009)	(0.011)
	50% prob. Piece rate	0.901	0.914	0.920
		(0.013)	(0.01)	(0.015)
Social Comparisons	No payment, social comparison	0.920	0.909	0.896
		(0.012)	(0.010)	(0.019)
Ranking	No payment, feedback after	0.922	0.918	0.921
		(0.011)	(0.009)	(0.016)
Task Significance	No payment, please try hard	0.911	0.927	0.922
		(0.013)	(0.009)	(0.016)
Piece Rate + Task Significance	Low piece rate, please try hard	0.923	0.918	0.904
		(0.012)	(0.009)	(0.016)
	Number of Observations	2,706	4,723	3,026
	Average Accuracy	0.914	0.919	0.916
		(0.003)	(0.002)	(0.004)
	Prob > F	0.736	0.477	0.188

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

Online Appendix Table 4. Comparison Across Designs, Alternative Measures

Category	Version Comparison	Correlations Across Versions		Pairwise Comparisons of Treatments (out of 105 Possible Comparisons)			Average Difference From Baseline No-piece Rate Treatment	
		Rank-Order Correl.	Pearson Correl.	Same Direction	Same Direction and Stat. Sig.	Opposite Direction and Stat. Sig.	Log Points	Z-score
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Pure Replication	2015 AB Task vs. 2018 AB Task	0.91 (0.05)	0.97 (0.02)	96	67	1	0.07 (0.04)	0.13 (0.07)
Demographics	Male vs. Female	0.96 (0.04)	0.98 (0.02)	100	68	0	0.10 (0.04)	0.12 (0.06)
	College vs. No College	0.97 (0.04)	0.97 (0.02)	98	67	0	0.07 (0.03)	0.09 (0.05)
	Young vs. Old	0.98 (0.04)	0.98 (0.02)	101	70	0	0.04 (0.03)	0.08 (0.05)
Geography/ Culture	US vs. India	0.65 (0.11)	0.78 (0.09)	82	29	2	0.07 (0.03)	0.20 (0.06)
Task	AB Task vs. Card Coding	0.59 (0.14)	0.55 (0.13)	75	19	1	0.19 (0.04)	0.49 (0.09)
Output	Extensive Cards vs. Intensive Cards	0.27 (0.17)	0.21 (0.16)	62	7	3	0.21 (0.06)	0.23 (0.05)
	Extensive Cards vs. AB Task	0.65 (0.07)	0.63 (0.06)	82	47	8	0.16 (0.03)	0.37 (0.07)
	AB Task: First 5 min vs. Last 5 min	0.97 (0.03)	0.98 (0.01)	101	74	0	0.04 (0.02)	0.04 (0.02)
Ecological validity	Cards: Consent vs. No Consent	0.84 (0.09)	0.92 (0.05)	89	47	0	0.16 (0.09)	0.15 (0.07)
Correlation of Alternative Measure with Rank-Order Correl. Measure			0.98	0.99	0.95	-0.54	-0.80	-0.65

Notes: The Table presents alternative measures of stability of experimental results for the version comparisons of Table 2, comparing to the benchmark measure, rank-order correlation reproduced in Column 1. In Columns 3-5 we compare each of the 15 treatments to all other treatments, yelding 105 comparisons, Column 3 reports the treatment comparisons that are in the same direction across the versions being compared, Column 4 the comparisons that are not only in the same direction, but also statistically significantly different. Column 5 reports the comparisons that are in opposite directions and statistically significant in both versions. In Columns 6 and 7 we compute differences in log points (Column 6) or z-scores (Coumn 7) for each treatment compared to the baseline no-piece-rate treatment, and compare that measure across versions. The last row presents the correlation of the measures in Columns 2-7 with the benchmark measure, across the ten version comparisons.

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

Category	Design Comparison	25th Percentile		75th Percentile	
		Full Stability w/ Noise	Actual	Full Stability w/ Noise	Actual
		(1)	(2)	(3)	(4)
Pure Repl.	2015 AB Task vs. 2018 AB Task	0.90	0.86	0.90	0.98
	(n=8,252; n=2,219)	(0.05)	(0.07)	(0.05)	(0.06)
Demogr., Typing Task	Male vs. Female	0.93	0.96	0.94	0.91
	(n=4,686; n=5,785)	(0.04)	(0.05)	(0.03)	(0.06)
	College vs. No College	0.93	0.94	0.93	0.92
	(n=5,842; n=4,629)	(0.04)	(0.07)	(0.03)	(0.06)
	Young (=<30) vs. Old (30+)	0.93	0.93	0.93	0.89
	(n=5,259; n=5,212)	(0.04)	(0.06)	(0.03)	(0.06)
Geogr./ Culture	US vs. India	0.84	0.71	0.86	0.59
	(n=8,803; n=1,225)	(0.07)	(0.15)	(0.07)	(0.15)
Task	AB Task vs. 10-min Card Coding	-	0.37	-	0.31
	(n=10,471; n=2,537)		(0.19)		(0.16)
Output	10-min Cards vs. Extra Cards	-	-	-	-
	(n=2,537; n=2,188)				
	Extra Cards vs. AB Task	-	-	-	-
	(n=2,188; n=2,219)				
	AB Task: First 5 min vs. Last 5 min	0.95	0.93	0.96	0.97
	(n=10,471)	(0.03)	(0.05)	(0.02)	(0.04)
Consent	Cards: Consent vs. No Consent	-	-	-	-
	(n=2,188; n=2,246)				

Notes: The Table lists the 10 design changes to the experiment which constitute the focus of the paper. For example, in row 1 we compare the estimate of effort in the 15 treatments in the button pushing task, comparing the results in 2015 versus in 2018. We report the actual rank-order correlation, as well as the results under a full-stability benchmark (see Table 2). These results differ from the benchmark ones in Table 2 because we compute the effort estimate using the 25th and 75th percentile of effort instead of the mean effort. We do not report these measures of comparsons involving the extra-work task in which the 25th or 75th percentile effort is typically a corner solution (0 or 20).

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

		Button Pushing Task, 10 Min	Demographics, Typing, Pooled '15-'18						2018 WWII Cards Coding
Category	**Parameters**	**2015 + 2018 Pooled Exp.**	**College**	**No College**	**Young (=<30)**	**Old (30+)**	**USA**	**India**	**Extra Work, Pooled**
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Incidental Parameters	Curvature of Cost of Effort ɣ	0.015	0.015	0.014	0.011	0.022	0.013		0.051
		(0.003)	(0.005)	(0.004)	(0.003)	(0.010)	(0.003)	*	(0.010)
	Implied Elasticity	0.036	0.035	0.037	0.046	0.025	0.039		0.384
		(0.008)	(0.011)	(0.011)	(0.010)	(0.011)	(0.008)	*	(0.074)
	Level of Cost of Effort *k*	-34.591	-35.260	-33.447	-27.798	-48.420	-32.312		-3.954
		(6.619)	(9.628)	(8.703)	(5.595)	(19.659)	(5.879)	*	(0.934)
	Baseline Motivation *s*	3.7e-04	1.3e-04	0.001	0.002	1.4e-05	5.6e-04		0.137
		(7.5e-04)	(4.0e-04)	(0.003)	(0.004)	(7.6e-05)	(0.001)	*	(0.085)
Pay Enough or	Δs_{CO}	-1.2e-04	-0.015	0.017	0.156	-0.084	0.021		0.059
		(0.087)	(0.110)	(0.134)	(0.197)	(0.041)	(0.098)	*	(0.061)
Social Pref. Parameters	Pure Altruism *alpha*	0.005	0.002	0.010	0.003	0.005	0.007		0.008
		(0.008)	(0.011)	(0.014)	(0.015)	(0.011)	(0.008)	*	(0.017)
	Warm Glow *a*	0.117	0.136	0.097	0.231	0.030	0.097		0.241
		(0.100)	(0.158)	(0.122)	(0.184)	(0.071)	(0.085)	*	(0.139)
Social Pref.: Gift Exch.	Δs_{GE}	0.001	5.3e-04	0.002	0.005	3.9e-05	0.001		0.857
		(0.002)	(0.001)	(0.004)	(0.008)	(1.9e-04)	(0.002)	*	(0.245)
Discounting	*Beta*	1.021	1.171	0.692	3.320	0.222	0.954		0.995
		(0.888)	(1.435)	(0.852)	(3.112)	(0.450)	(0.818)	*	(0.674)
	Delta (Weekly)	0.803	0.673	1.061	0.654	1.016	0.849		0.789
		(0.210)	(0.260)	(0.372)	(0.202)	(0.514)	(0.217)	*	(0.169)
Social Comparisons	Δs_{SC}	0.060	0.034	0.127	0.192	0.008	0.056		0.007
		(0.058)	(0.053)	(0.146)	(0.161)	(0.022)	(0.054)	*	(0.037)
Ranking	Δs_{R}	0.014	0.010	0.020	0.052	0.001	0.015		0.056
		(0.018)	(0.019)	(0.031)	(0.056)	(0.004)	(0.018)	*	(0.047)
Task Significance	Δs_{TS}	0.010	0.004	0.035	0.040	7.9e-04	0.013		0.069
		(0.013)	(0.009)	(0.051)	(0.044)	(0.003)	(0.016)	*	(0.050)
Probability Weighting Parameters	Pi (0.01)	0.002	0.001	0.003	0.003	8.2e-04	0.003		0.008
		(0.001)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	*	(0.003)
	Pi (0.50)	0.168	0.113	0.259	0.263	0.091	0.177		0.213
		(0.107)	(0.113)	(0.201)	(0.164)	(0.138)	(0.107)	*	(0.092)
No. of Obs.		10471	5842	4629	5212	5259	8803	*	4434
Avg effort		1880.080	1823.903	1950.847	1954.960	1805.649	1910.458	*	11.250
Root MSE		656.589	666.189	637.689	686.057	618.030	653.489	*	54.078
Extra Treat.: Incentive + Please try	Out-of-Sample Pred.								13.13
	Actual								12.069 (0.543)

Notes: The Table shows structural estimates of the incidental parameters (ɣ, k, and s) and psychological parameters estimated using all 15 treatments across 11 different samples. All models assume an exponential cost function. Cols (1)-(6) are estimated using nonlinear least squares for the a-b yping task, while Col 8 is estimated on the extra-work task using maximum likelihood due to censoring. Standard errors in parantheses. The structural estimates for the India sample do not converge due to the very noisy response to incentives in this subsample.