Forecasting Social Science Research Results

Stefano DellaVigna, UC Berkeley and NBER

(based on collaborations with David Card, Devin Pope, Eva Vivalt)

UC Berkeley, SEEDEC

May 31, 2019

Motivation

- Sleep interventation in Chennai (Bessone, Rao, Schilbach, Schofield, Toma, 2019)
  - Eg: Provide sleep devices and examine impact on savings, productivity, etc.
  - Similar to many RCTs

- Six months ago: Researchers contact couple dozen "experts" and ask for forecast of results

- What did they do?

Motivation

- Step 1. Explain the study in brief

Overview of the Study

We are conducting a randomized controlled trial (RCT) examining the impact of reduced sleep deprivation on the lives of the poor in Chennai, India. We will enroll 450 low-income individuals for 28 work days each to perform data-entry work and to undertake a series of cognitive and decision-making tasks. The participants will also be able to save their earnings in our office.

To estimate the impact of improving participants' sleep, we randomly select some participants to receive one of the two of the following treatments designed to increase their sleep:

> **Treatment 1.** Participants receive devices intended to improve their sleep, such as a mattress, eyeshade, earplugs, and table fan, as well as encouragement to sleep more and information regarding the benefits of sleep.
>
> **Treatment 2.** Participants receive Treatment 1 plus monetary incentives to increase their sleep relative to their average sleep in the week-long baseline period[1].

We objectively measure the amount of sleep using state-of-the-art wristwatch-like devices called actigraphs.

Motivation

- Step 2. Ask for prediction, giving benchmark

Question 1. **Savings**

Please predict: What is the average impact of an additional hour of sleep on rupees saved per day (as measured by IV)? Control group participants earn on average Rs. 448 (USD 6.55) per day from the study and save Rs. 144 (USD 2.11) per day at the study office.

The numbers in the table below can be used as reference points for the magnitude of your answer.

Savings change by _____ *rupees per day*. (You can insert either zero, a positive, or a negative number).

Control Mean		Control SD
Rs. 144		Rs. 222
Answer	**% Change**	**SD Change**
±1	±1.0	±0.01
±14	±10.0	±0.06
±22	±15.4	±0.10
±44	±30.8	±0.20
±111	±77.1	±0.50
±167	±115.6	±0.75
±222	±154.2	±1.00

Motivation

- Step 3. Store forecasts and compare to results
  - How does average forecast compare to results?
  - Do results reject the null of wisdom-of-crowd forecast?
  - Do sleep experts agree with economists in forecast?
  - Is there much heterogeneity in prediction?

Kick-off

- Example of *Expert Forecasts about Research Results*
  1. Study (experiment or not) already run or to be run soon
  2. Results not known yet
  3. Contact forecasters (expert or not) to make forecasts of research results
  4. Store forecasts, so as to compare to results later

- So far, this is uncommon:
  - Prediction markets on replication: OSF, 2016; Camerer et al., 2016, 2018
  - Forecasts in (published) economics papers: Goh et al (2016); DellaVigna and Pope (2018a,b)

- It can be very useful! Highlight 5 motivations
- Work in progress on platform to scale this

Motivation 1: Updating on Science

- Example 1: Bertrand and Mullainathan (*AER* 2004)

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

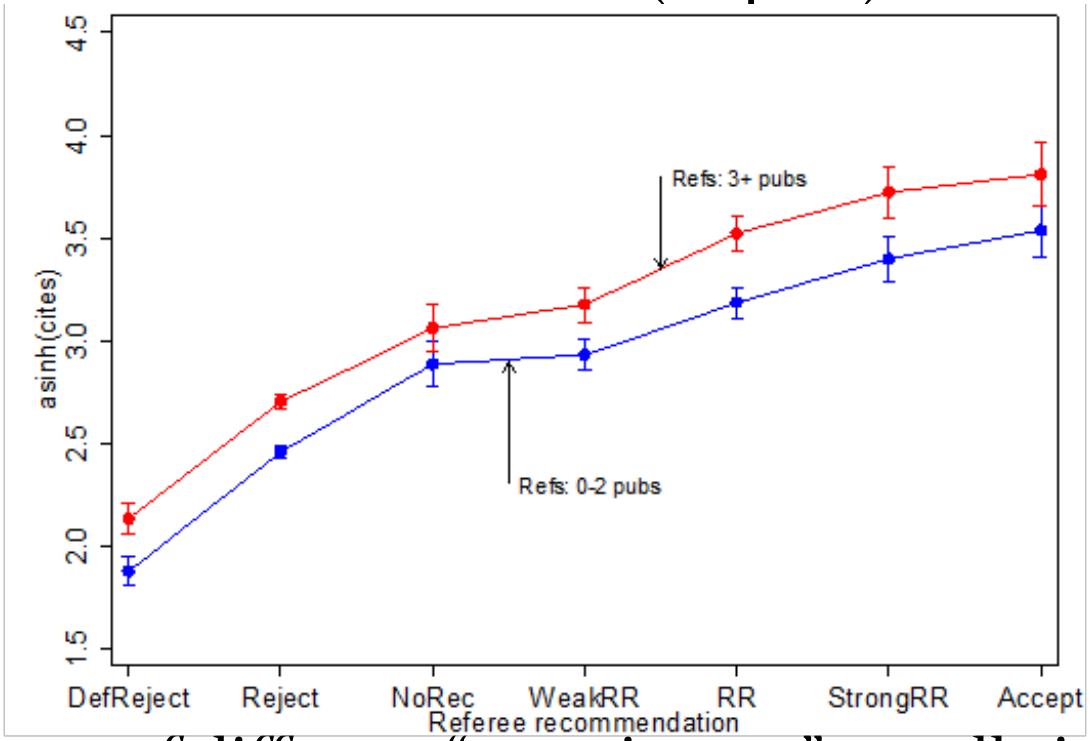
By Marianne Bertrand and Sendhil Mullainathan*

- Sendhil (2000): "We almost did not run this study"
- Conversation with MIT colleagues…

- Need to collect *priors ex ante* (hindsight bias)
- Motivation 1: <u>Capture updating on results</u>

Motivation 2: Testing Hypotheses

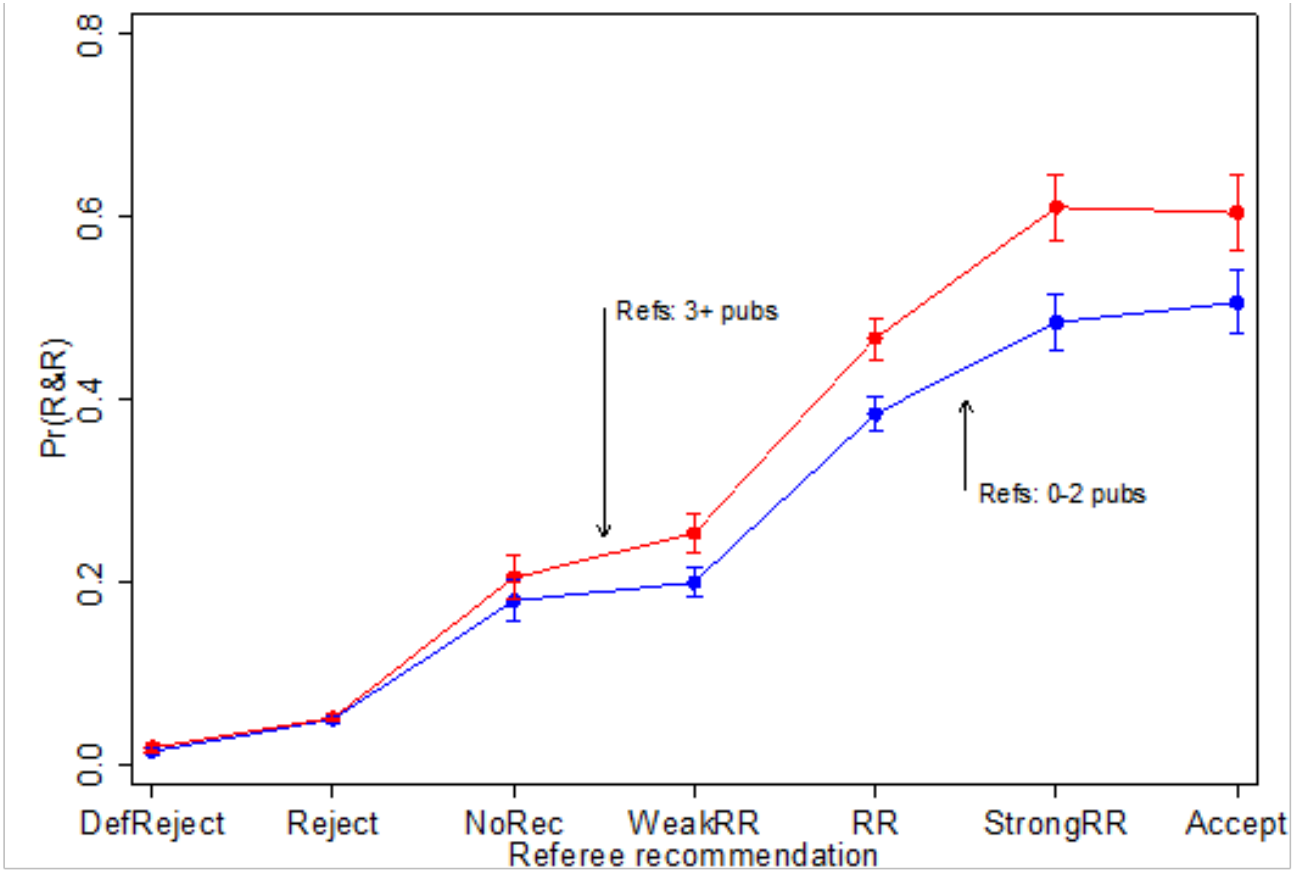
- Example 2: Card and DellaVigna (*REStat* forthc.)
  - Evidence on editorial process
  - Example: How informative are referees? Extent to which recommendations measure (ex post) citation



  - Referees of different "prominence" equally informative
  - Trust reports by junior researchers!

Motivation 2: Testing Hypotheses

- Example 2: Card and DellaVigna (*REStat* forthc.)
  - Do editors then put equal weight on refs?



- Editors overweight published referees by 25%

Motivation 2: Testing Hypotheses

- Two explanations:
  - Explanation 1 (*Ed pref*). Editors know refs equally informed, but prefer not to disagree with prominent refs
  - Explanation 2 (*Wrong beliefs*). Editors *think* prominent refs more informative
  - We asked editors to predict referee informativeness

Sample of Experts:	**REStud Meeeting**		**Univ. Zurich**
Question:	**Correct Answer (REStud Only)**	**Average Answer by Editors (N<=12)**	**Average Answer by Faculty (N<=13)**
Informativeness of Referee Recommendations			
How much higher is the percentile citation if a referee recommendation is positive versus if it is negative (for papers with 3 reports)?	*11.50*	17.50	14.8
What is the percentile citation increase for "prominent" referees?	*12.1*	24.3	22.2

  - It is wrong beliefs! May change with information out

Motivation 3: Publication Bias

- Example 3: Consider study of health intervention
  - Largely expected to be effective
  - Highly powered experiment
  - Finds a precise null effect → Hard to publish (Ioannidis 2008; Franco, Malhotra, and Simonovits 2014)

- If had collected expert forecasts, test against priors
- Results *reject* expert prior (mean forecast) → Paper is not a null result → More likely published

- Motivation 3: <u>Reduce publication bias</u>

Motivation 4: Measure & Interpret Expect.

- Example 4: Casey et al. (2018), Vivalt and Coville (2016)
  - What do experts expect about impact of intervention?
  - What do policy-markers expect?

  - If policy-makers consistently expect too much, bound for disappointment
  - Expectations informs policy in an case
  - Important to know who knows what (e.g. DellaVigna and Pope 2018b for behavioral treatments)

- Motivation 4: <u>Measure and Interpret Expectations</u>

Motivation 5: Optimal Exp. Design

- Example 5: UK Nudge unit does police retention study
  - Has 10 ideas on potential interventions
  - BUT can only run 3 arms with decent statistical power
  - Halperns calls Thaler to choose the 3 arms
  - Thaler: "*Don't ask me! Crowd-source it*"

- Collect systematic data set on accuracy of forecasts
  - *Who* has the more accurate forecasts?
  - How much sample size for *wisdom of crowds*?
  - What are *optimal weights* (eg, *Good Judgment Project)*?

- Can then use for optimal design
- Motivation 5: <u>Optimal experimental design</u>

Today's Conference

- Given five motivations, plan platform such that:
  1. Researcher posts summary of project
  2. Invite forecasts on project before results known
  3. Store forecasts, with characteristics of forecaster
  4. Yet protect anonymity

- Work with BITSS together with Eva Vivalt, in coordination with IPA
- Pilot platform: https://socialscienceprediction.org/

- Now more in detail about example of how this can be useful

Collaboration with Devin Pope

What Motivates Effort? Evidence and Expert Forecasts (*REStud,* 2018)

Predicting Experimental Results: Who Knows What? (*JPE*, 2018)

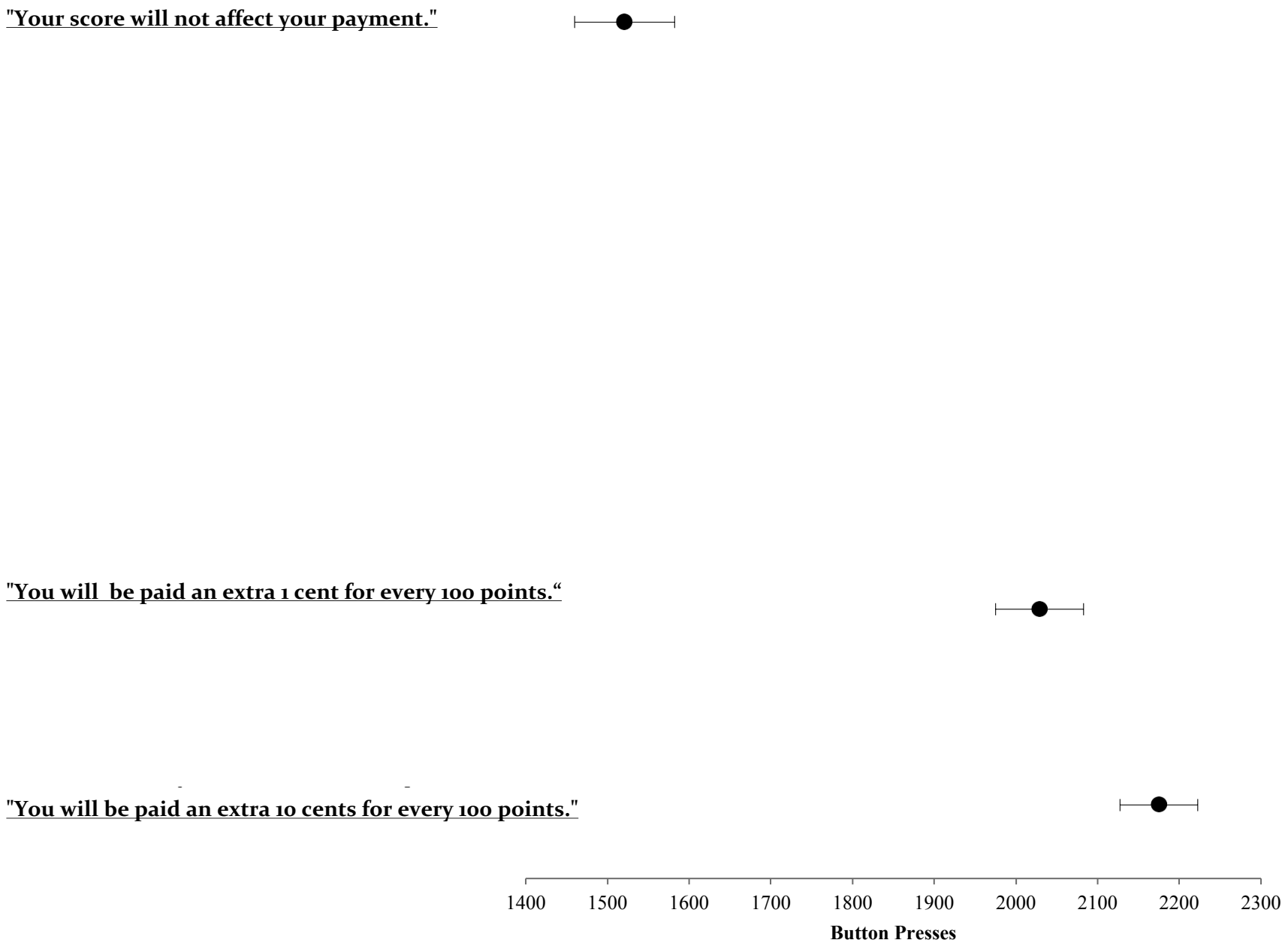
Stability of Experimental Results: Forecasts and Evidence (Working Paper, 2019)

Experimental design

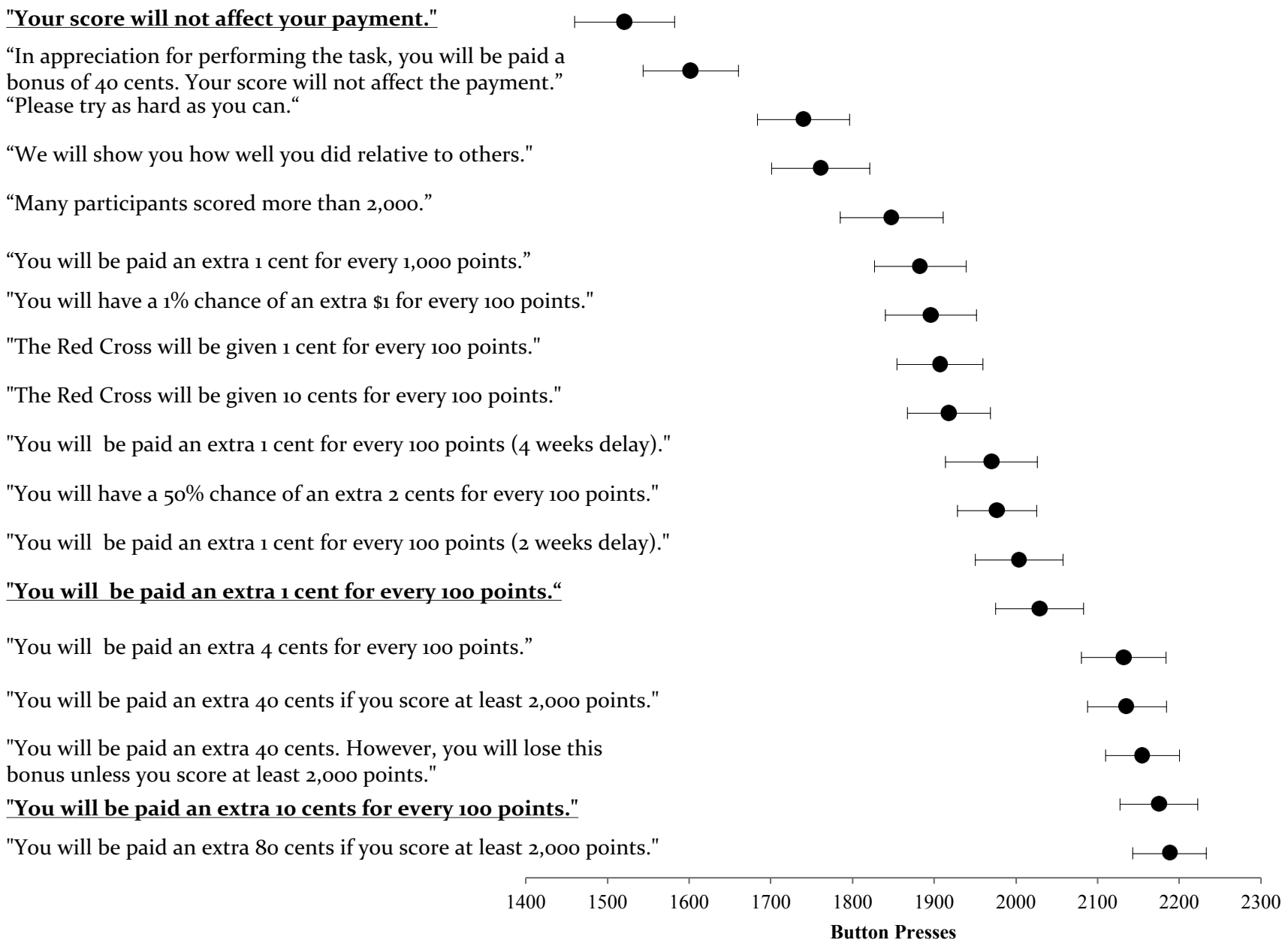
REAL-EFFORT TASK ON MTURK

- Recruited ~10,000 Mturk participants to do a 10-minute effort task
- Task: alternately press the “a” and “b” buttons as fast as they can
- Randomly assigned to one of 18 treatments intended to impact motivation
- Each treatment contained a unique sentence inspired by previous findings and designed to impact motivation.

Button Presses by Treatment (From Least to Most Effective) and Confidence Intervals

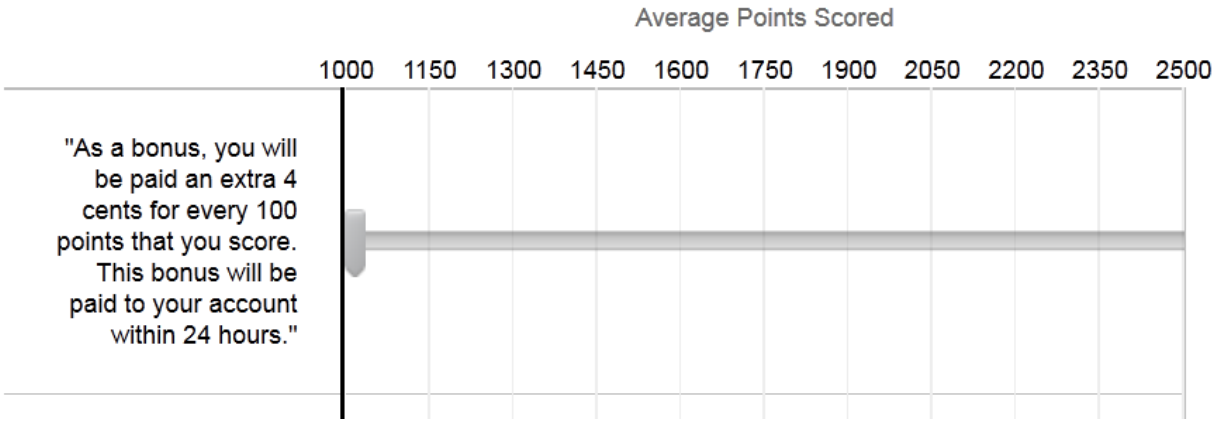


Button Presses by Treatment (From Least to Most Effective) and Confidence Intervals



Forecasts

- Prior to seeing the results of the pre-registered experiment, we contacted 312 academic experts in behavioral economics, decision making, and psychology. 208 completed the survey.
- Experts were shown instructions and could experience task for themselves
- The results of the three benchmark treatments were given to the forecasters
- Forecasters were asked to predict the results of the other 15 treatments using sliders



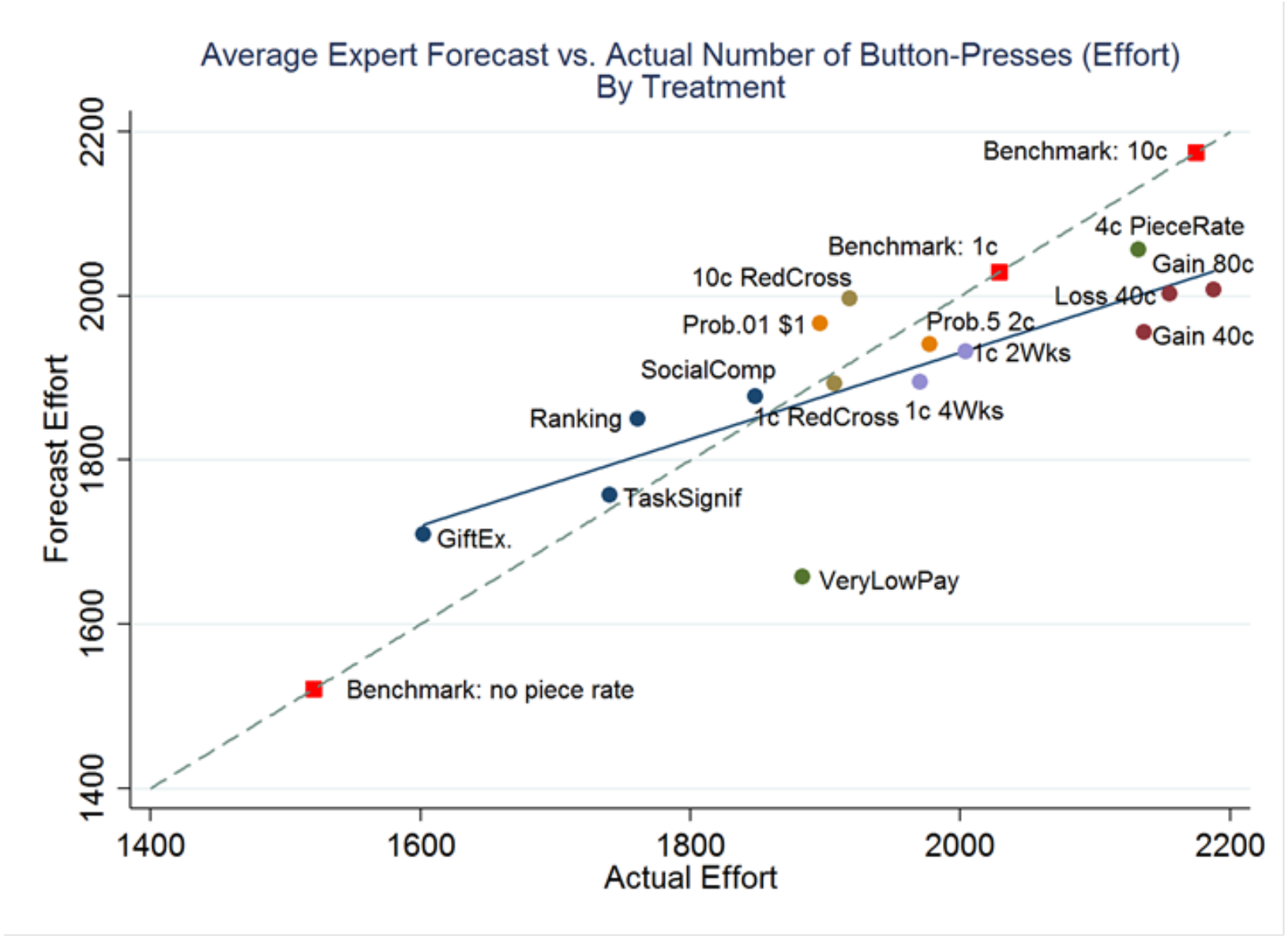
Additional forecasts

- 147 PhD students from top schools
- 158 Undergraduate students from Berkeley/Chicago
- 160 MBA students from Berkeley/Chicago
- 762 Mturkers

All survey takers were entered into a lottery where if chosen they would be paid an incentive-compatible reward based on the quality of their forecasts.

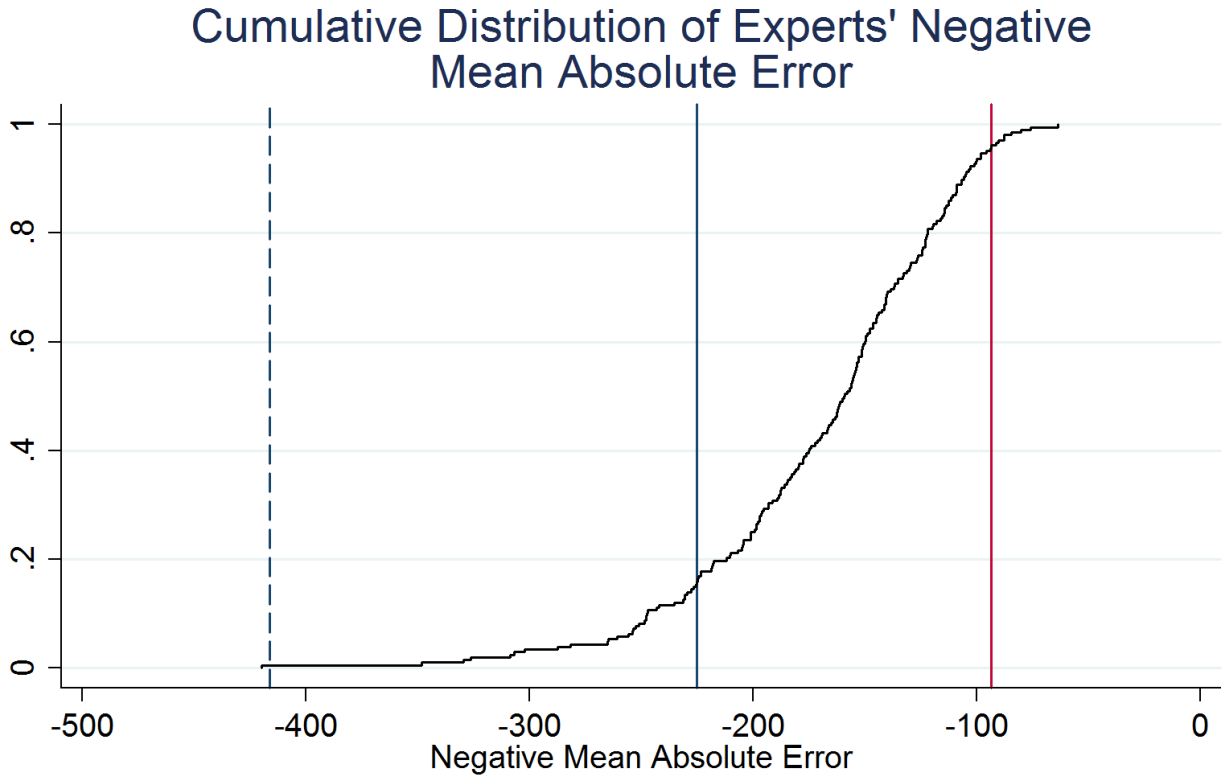
Academic experts were also promised anonymity and personalized feedback regarding the quality of their forecasts.

Accuracy of mean expert forecasts (Corr = .77)



II. Accuracy of Individual Forecast

- Quality of *individual* forecast?
  - 85 percent better than random choice b/w 1,500 and 2,200
  - BUT 97 percent worse than wisdom-of-crowds forecast



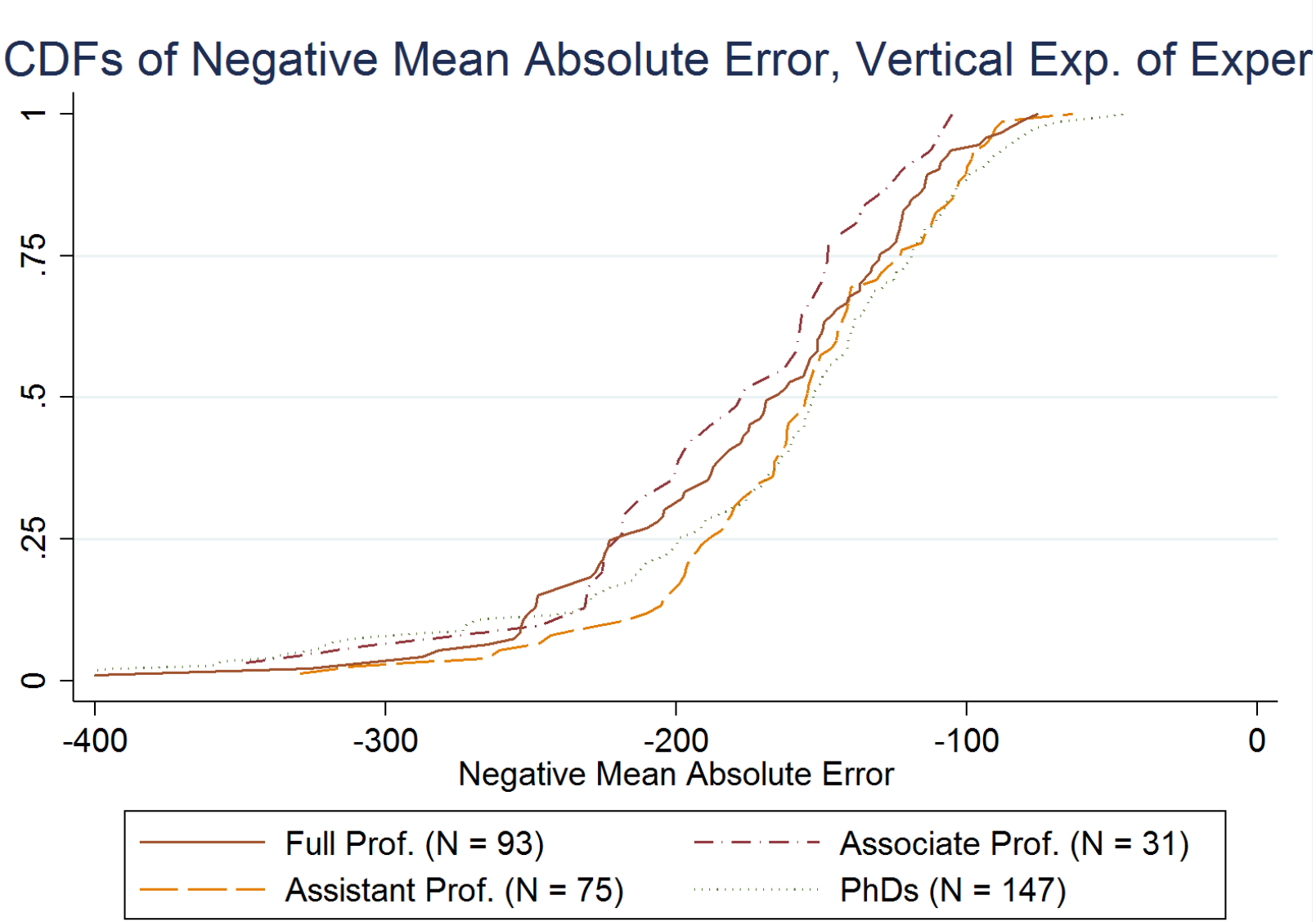
The vertical blue lines indicate the mean absolute error of uniformly distributed random forecasts on [1500,2200] (solid) and on [1000,2500] (dashed), the vertical red line denotes the mean absolute error of the mean forecast.

III. Forms of Expertise

- Which forms of expertise matter?
  1. *Vertical* Expertise
     - Citations
     - Seniority
  2. *Horizontal* Expertise
     - Having written a paper on a topic
     - Field of Research: Psychologists vs. Economist vs. Behavioral Economist
  3. *Contextual* Expertise
     - Experience with Mturk Platform

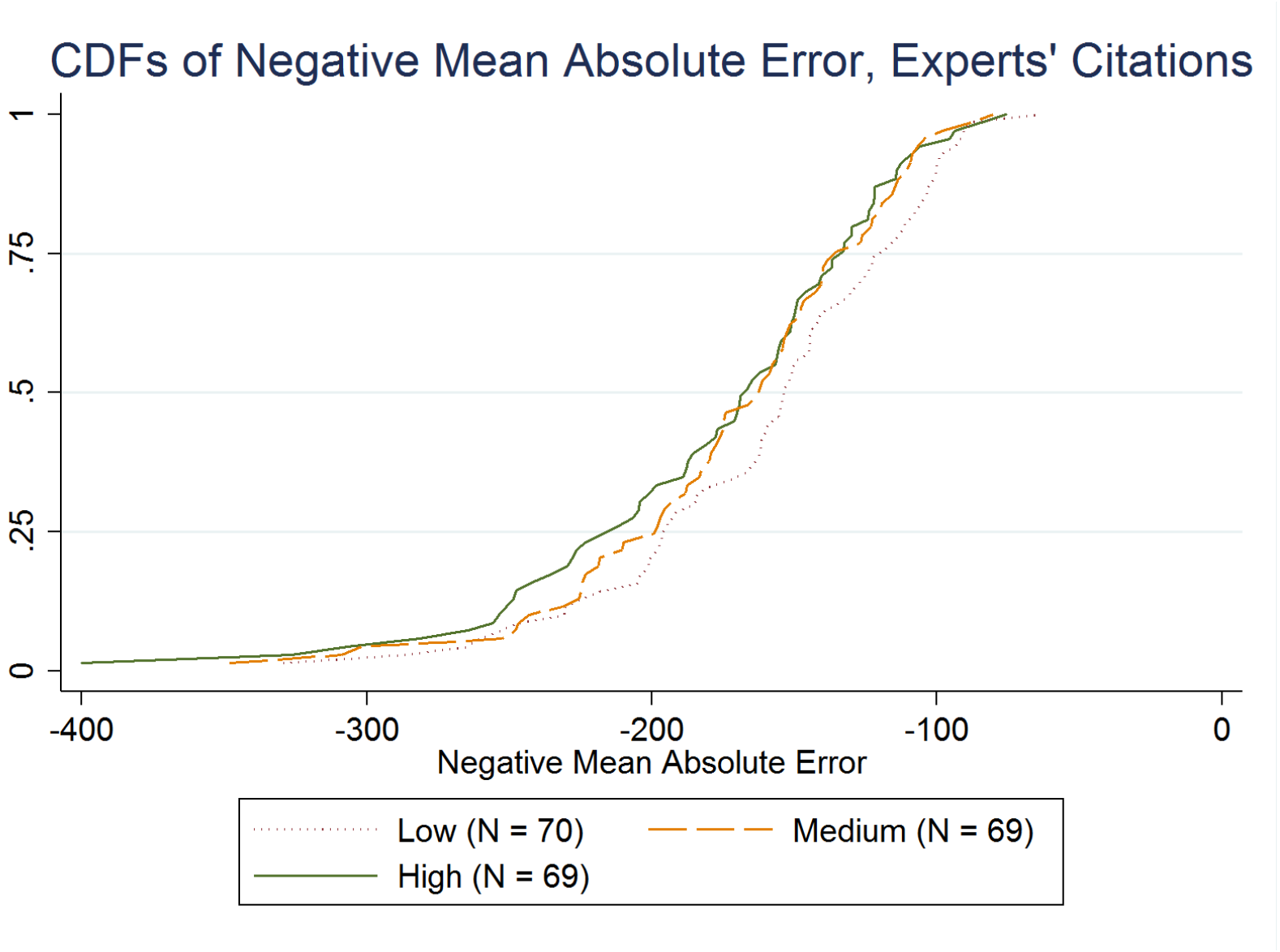
1. *Vertical* expertise? Not Academic Rank

- Compare Professors to Assistant, PhDs



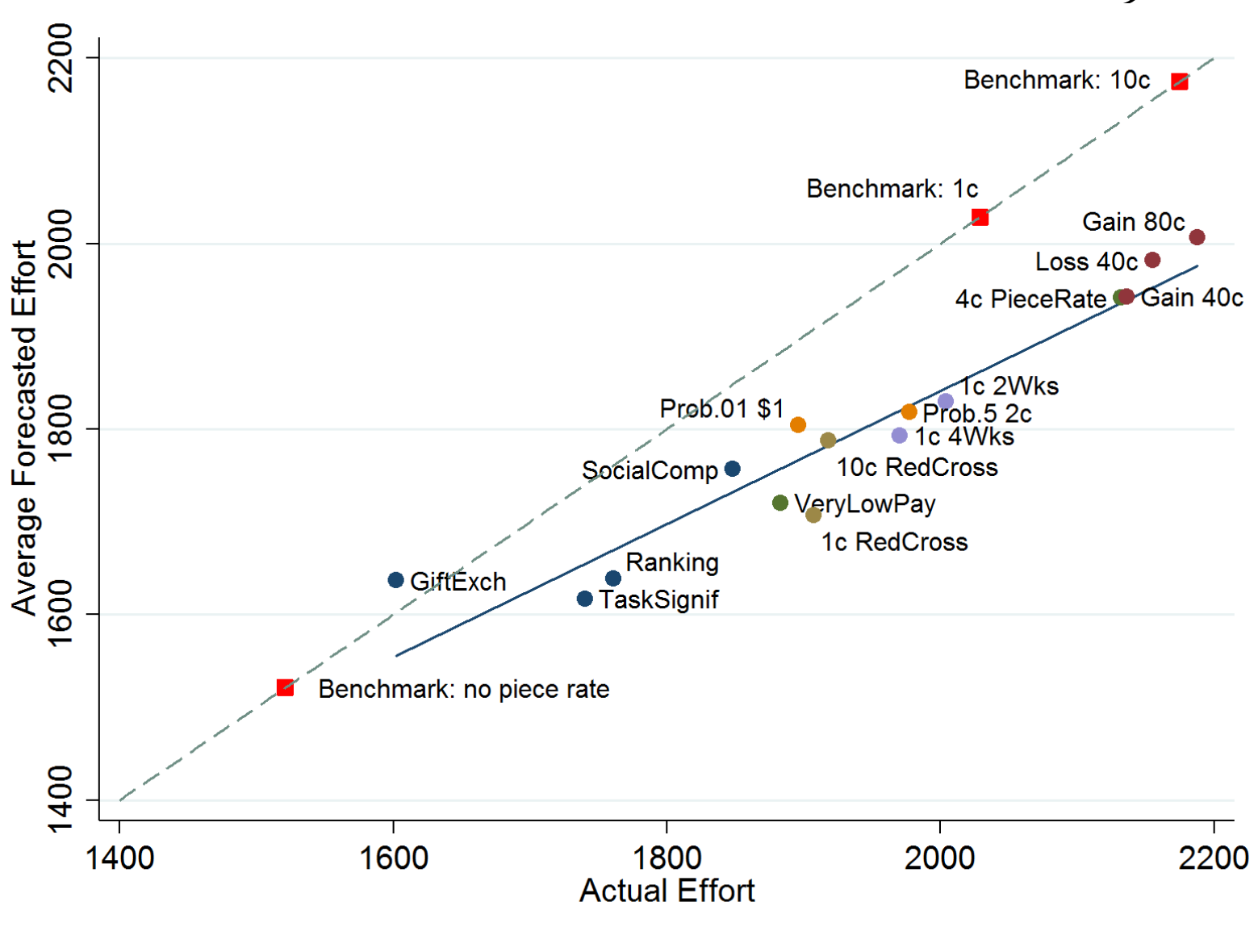
1. *Vertical* expertise? Not Citations

- Use Google Scholar Cites



IV. Experts vs. Non-Experts

- Wisdom-of-crowds rank-order corr. of Mturkers 0.92!



Additional findings

- Mean forecast (wisdom-of-crowd estimate) was better than 97% of individual forecasts (even pooling 4-5 forecasts has a dramatic effect on accuracy).
- No significant difference in accuracy across different types of academic experts (vertical, horizontal, or contextual expertise).
- Academic Experts = PhD Students > Undergrads > MBAs > Mturkers
- Switching measure of accuracy to rank-order accuracy, results in no difference in accuracy across any of the groups.
- Effects are not driven by amount of time taken on the survey or correlated with confidence (except for Mturkers).
- Superforecasters can be identified out of sample.
- Explore why academic experts make systematic errors for certain treatments.

Stability of experimental results: motivation

Situation 1. Economist designs lab experiment

-Primary features of design are ready, but what about things like sample, instructions, how to advertise. Do these secondary design choices matter?

Situation 2. Referee provides feedback to an editor

-Findings are interesting and have internal validity, but recommends rejection because worried about conceptual replication to other contexts

We are making important forecasting decisions all the time with research. Are we good at it?

We vary the design of our original a-b button pushing task and have people forecast the stability

1. Pure replication
2. Demographics
3. Geography/culture
4. Task
5. Measure of output
6. Consent/Natural experiment

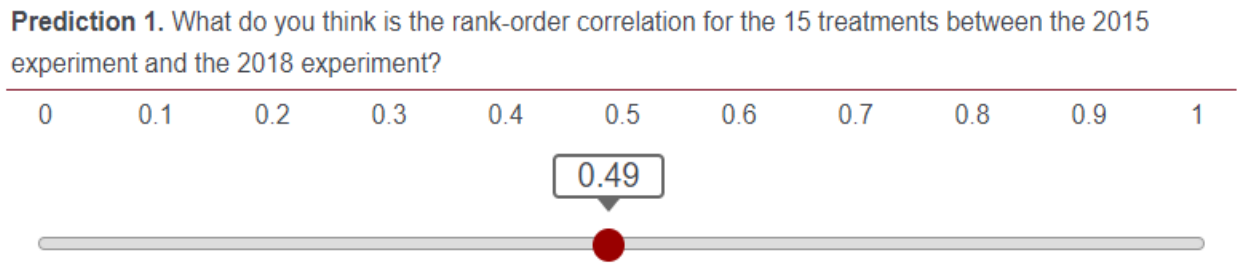
Measure of stability of results across design changes: rank-order correlation

Forecasters: We asked 60 behavioral economists (50 took the survey)

Expert Forecasts

- We elicit forecasts regarding rank-0. correlation
  - Contact 60 behavioral economists → N=50 responses
  - Contact 10 replication experts → N=5
  - Also contact PhD students at Berkeley and Chicago → N=33

- Experts make 10 rank-order forecasts
  - Show four examples of r-0 corr, then 10 forecasts made with slider



  - For A-B task, provide the rank-order correlation under full-stability
  - For WWII task, randomize
    - Baseline: see info on average effort and s.d.
    - Extra info: also informed on effort in 3 piece rate treatments, with s.e.

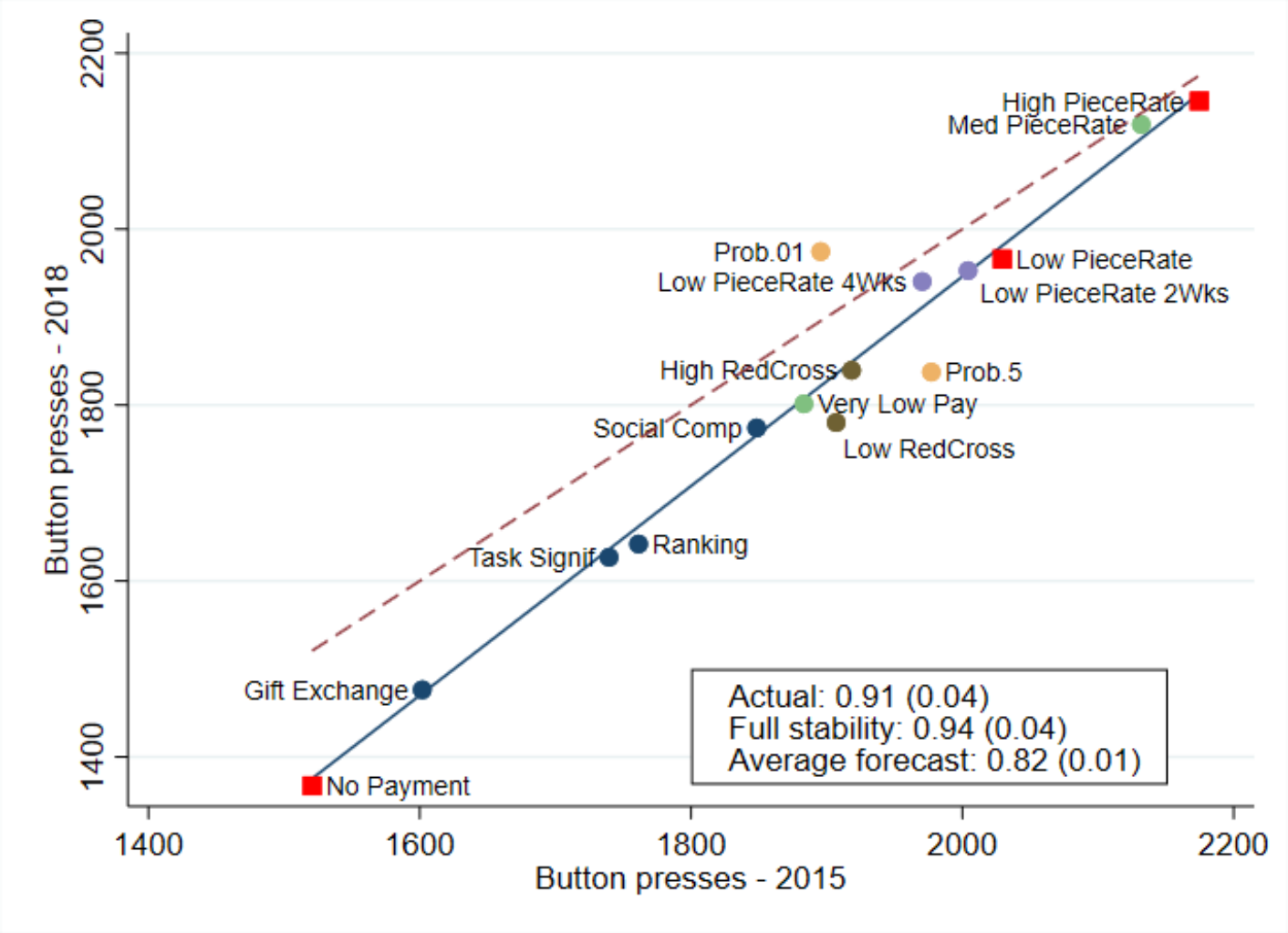
Expert Forecasts

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

Category	Design Comparison	Rank-Ord. Correlation	Average Forecast of Rank-Order Correlation			p-value
		Full Stability w/ Noise	Faculty Experts	PhD Students	Mturkers	Experts vs. Full Stability
		(1)	(2)	(3)	(4)	(6)
Pure Replication	2015 AB Task vs. 2018 AB Task (n=8,252; n=2,219)	0.94 (0.04)	0.82 (0.01)	0.87 (0.01)	0.75 (0.02)	0.004
Demogr., Typing Task	Male vs. Female (n=4,686; n=5,785)	0.95 (0.03)	0.73 (0.02)	0.77 (0.02)	0.73 (0.02)	0.000
	College vs. No College (n=5,842; n=4,629)	0.95 (0.03)	0.71 (0.02)	0.74 (0.02)	0.67 (0.02)	0.000
	Young (=<30) vs. Old (30+) (n=5,259; n=5,212)	0.95 (0.03)	0.74 (0.02)	0.76 (0.02)	0.66 (0.02)	0.000
Geography/ Culture	US vs. India (n=8,803; n=1,225)	0.89 (0.05)	0.63 (0.02)	0.67 (0.03)	0.68 (0.02)	0.000
Task	AB Task vs. 10-min Card Coding (n=10,471; n=2,537)	-	0.66 (0.02)	0.63 (0.03)	0.64 (0.02)	-
Output	10-min Cards vs. Extra Cards (n=2,537; n=2,188)	-	0.61 (0.02)	0.61 (0.03)	0.62 (0.02)	-
	Extra Cards vs. AB Task (n=2,188; n=10,471)	-	0.53 (0.03)	0.56 (0.04)	0.58 (0.02)	-
	AB Task: First 5 min vs. Last 5 min (n=10,471)	0.99 (0.01)	0.72 (0.02)	0.70 (0.03)	0.64 (0.02)	0.000
Consent	Cards: Consent vs. No Consent (n=2,188; n=2,246)	0.88 (0.05)	0.78 (0.02)	0.81 (0.02)	0.70 (0.02)	0.067

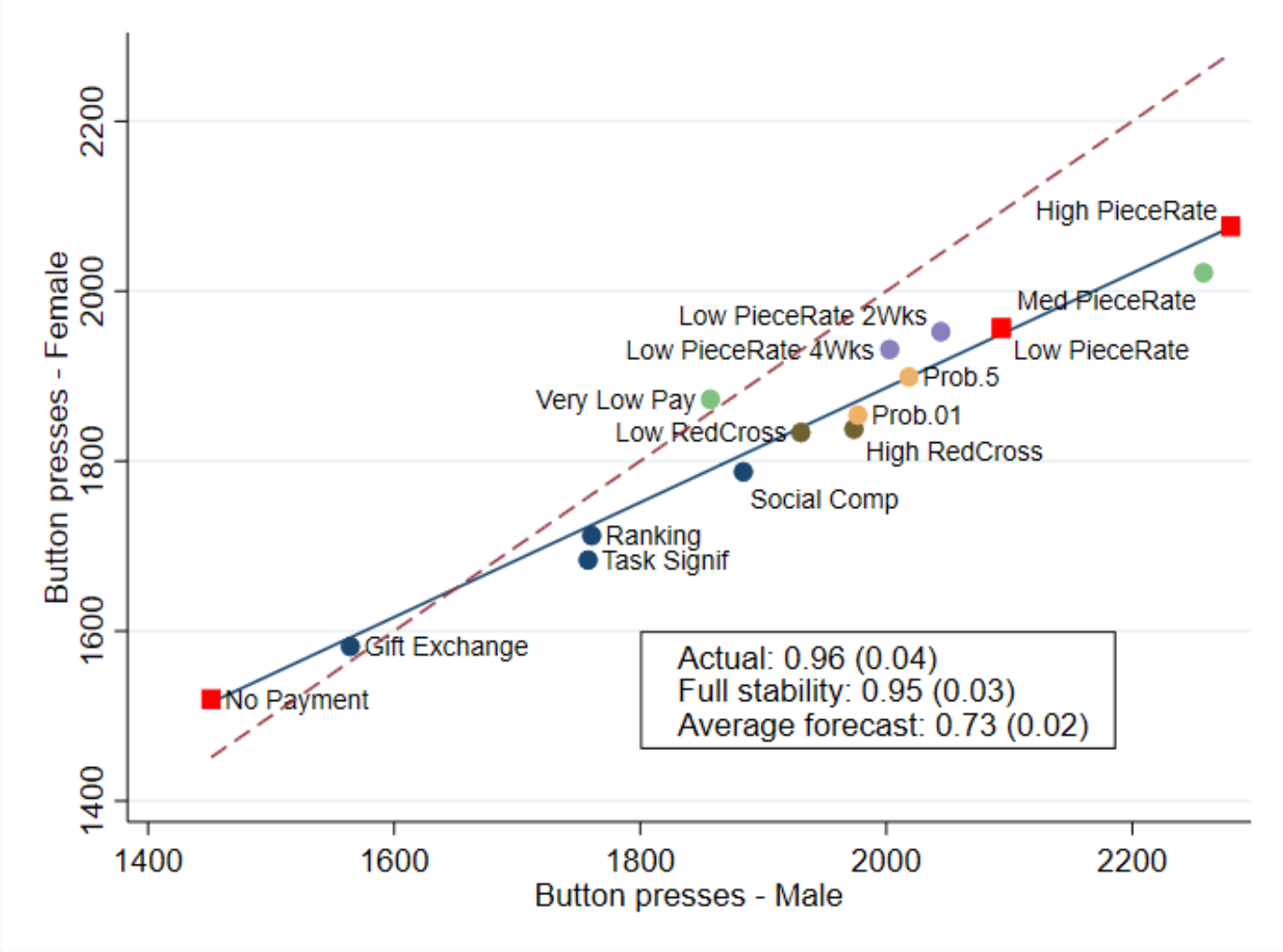
What Do We Find?

- Dimension 1. *Pure Replication.*
  - Results replicate very closely



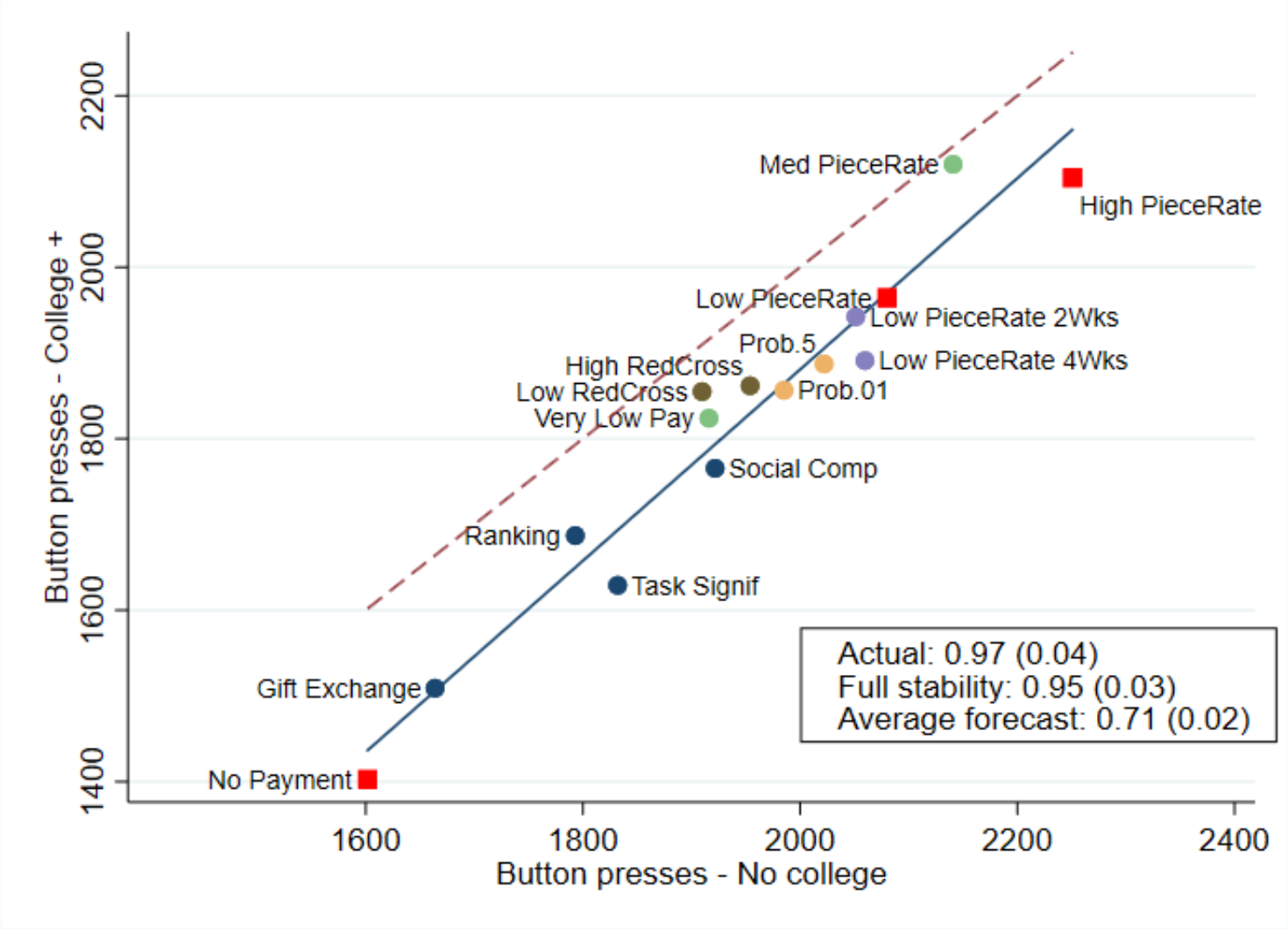
What Do We Find?

- *Dimension 2a. Demographics.* Correl. extremely high for gender



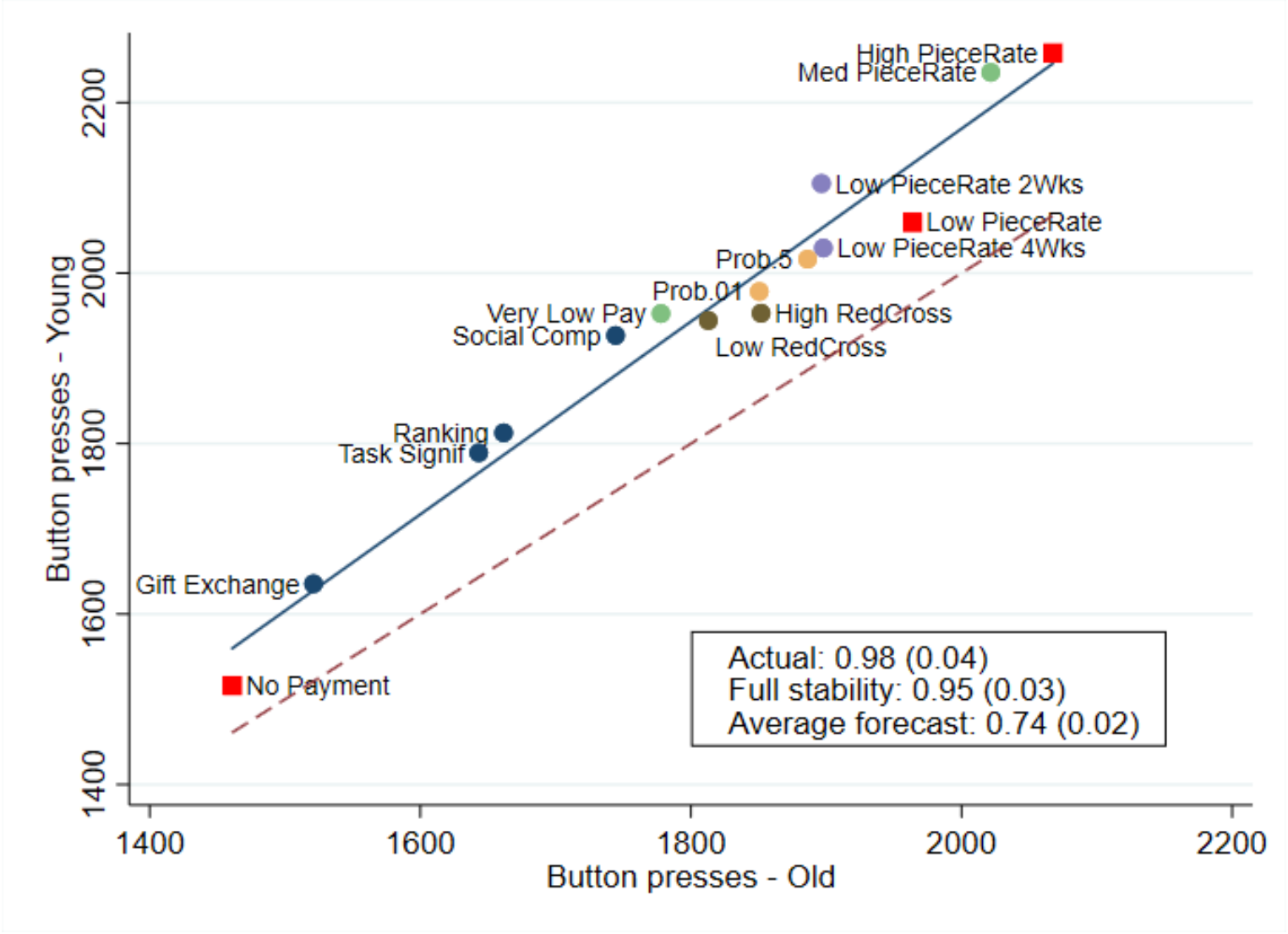
What Do We Find?

- *Dimension 2b. Demographics.* Correl. extremely high for educ



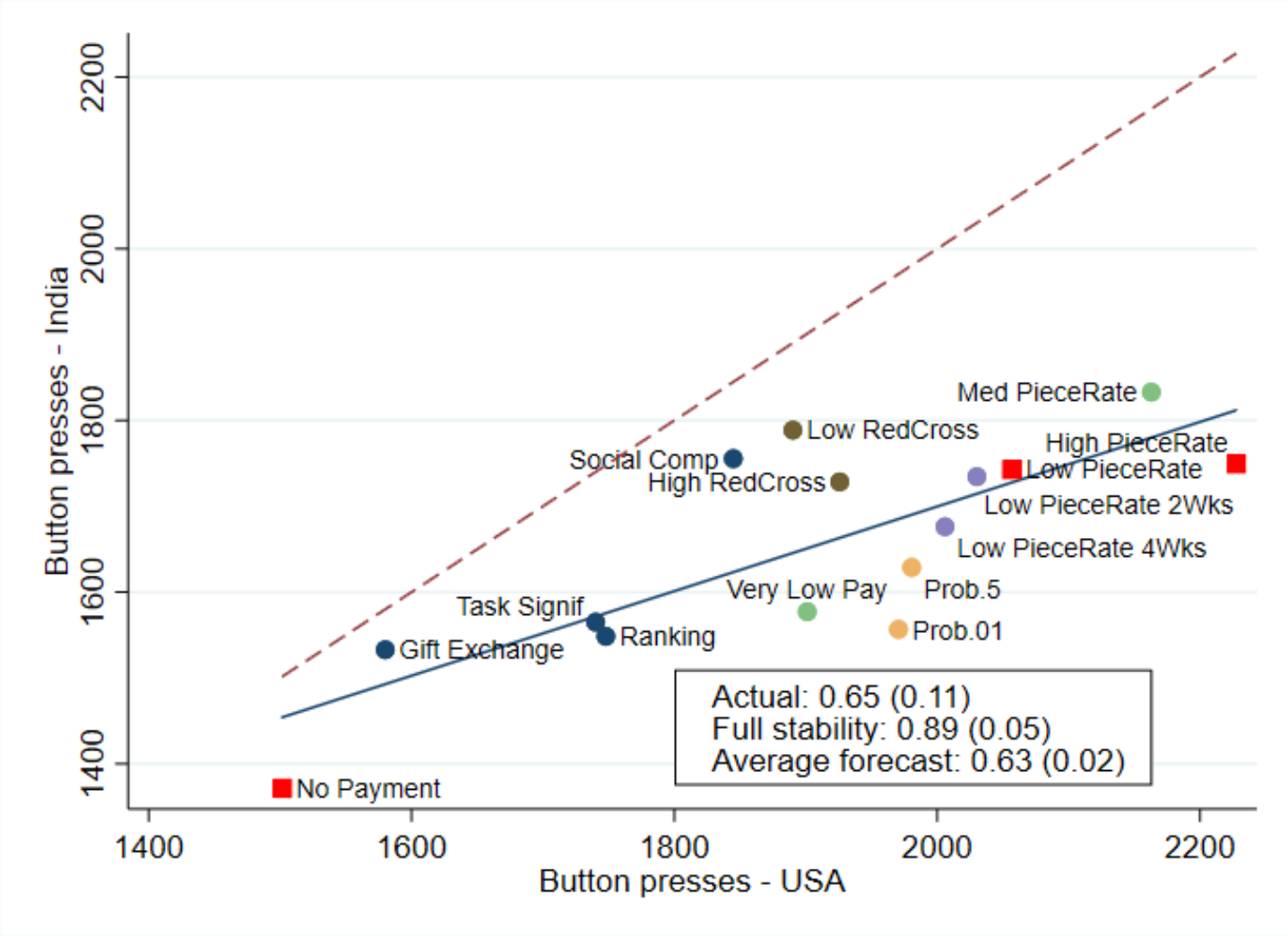
What Do We Find?

- *Dimension 2c. Demographics.* Correl. extremely high for age



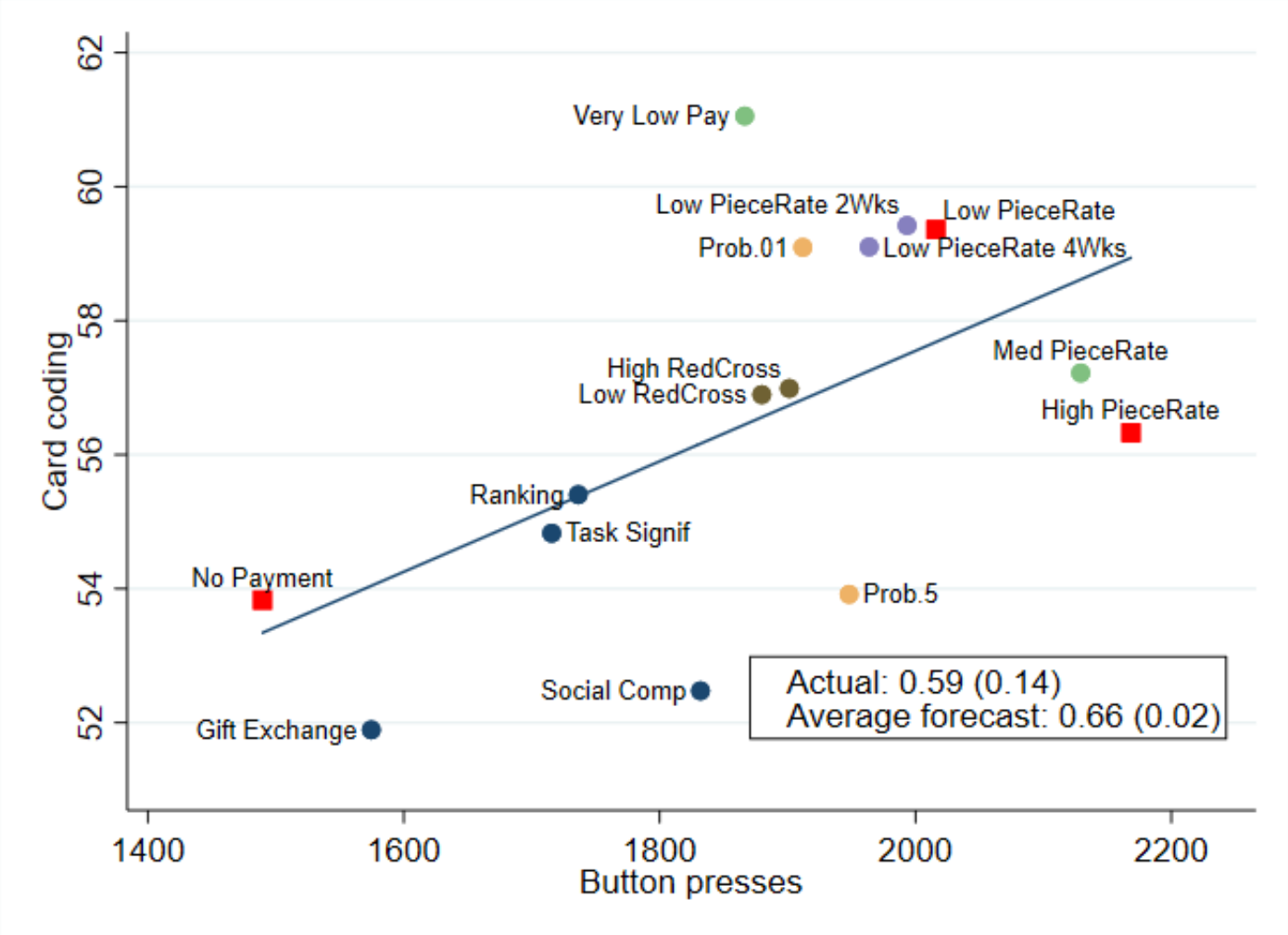
What Do We Find?

- *Dimension 3. Geography/Culture.*
  - Less stability, also taking into account sample size



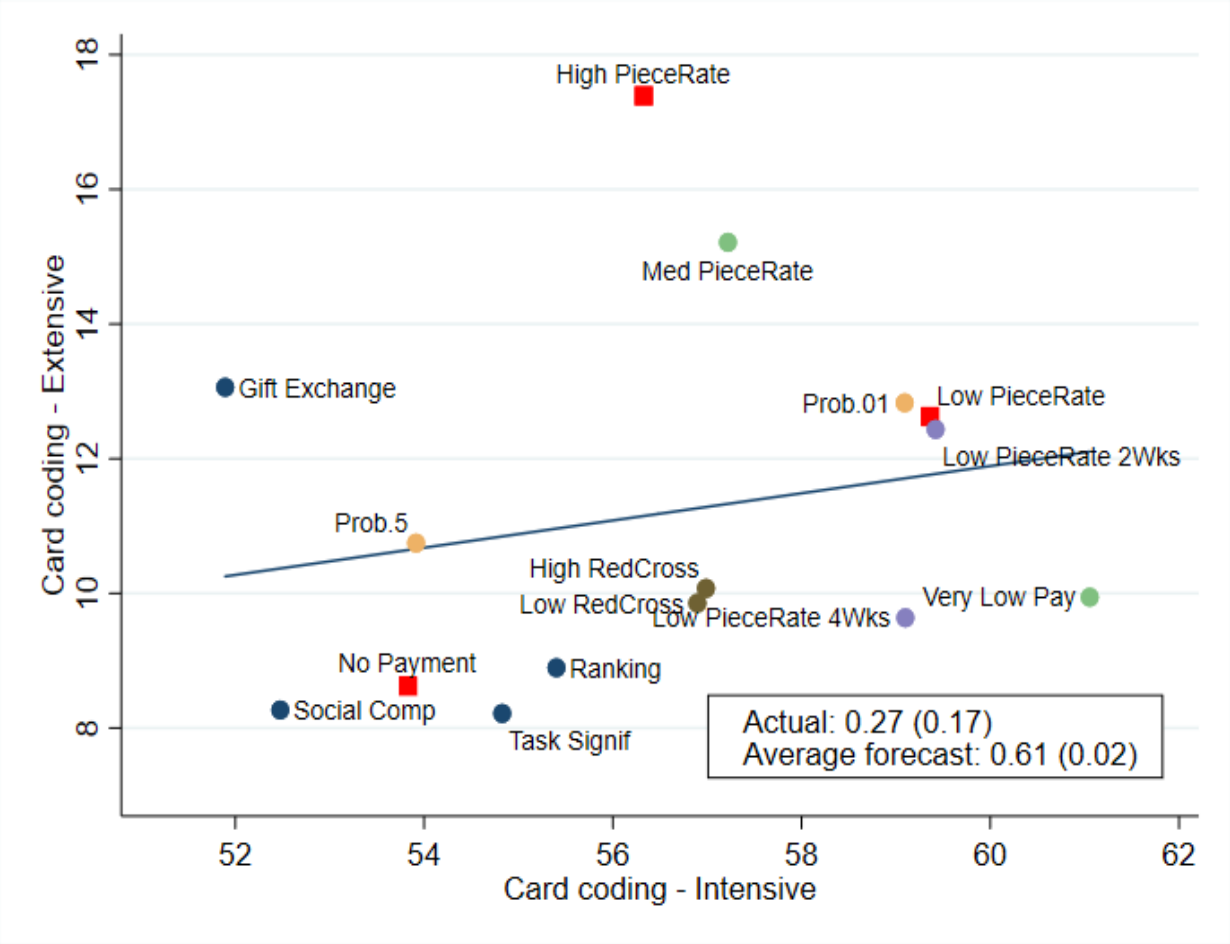
What Do We Find?

- Dimension 4. *Task.*
  - The results change quite a bit with more motivating task



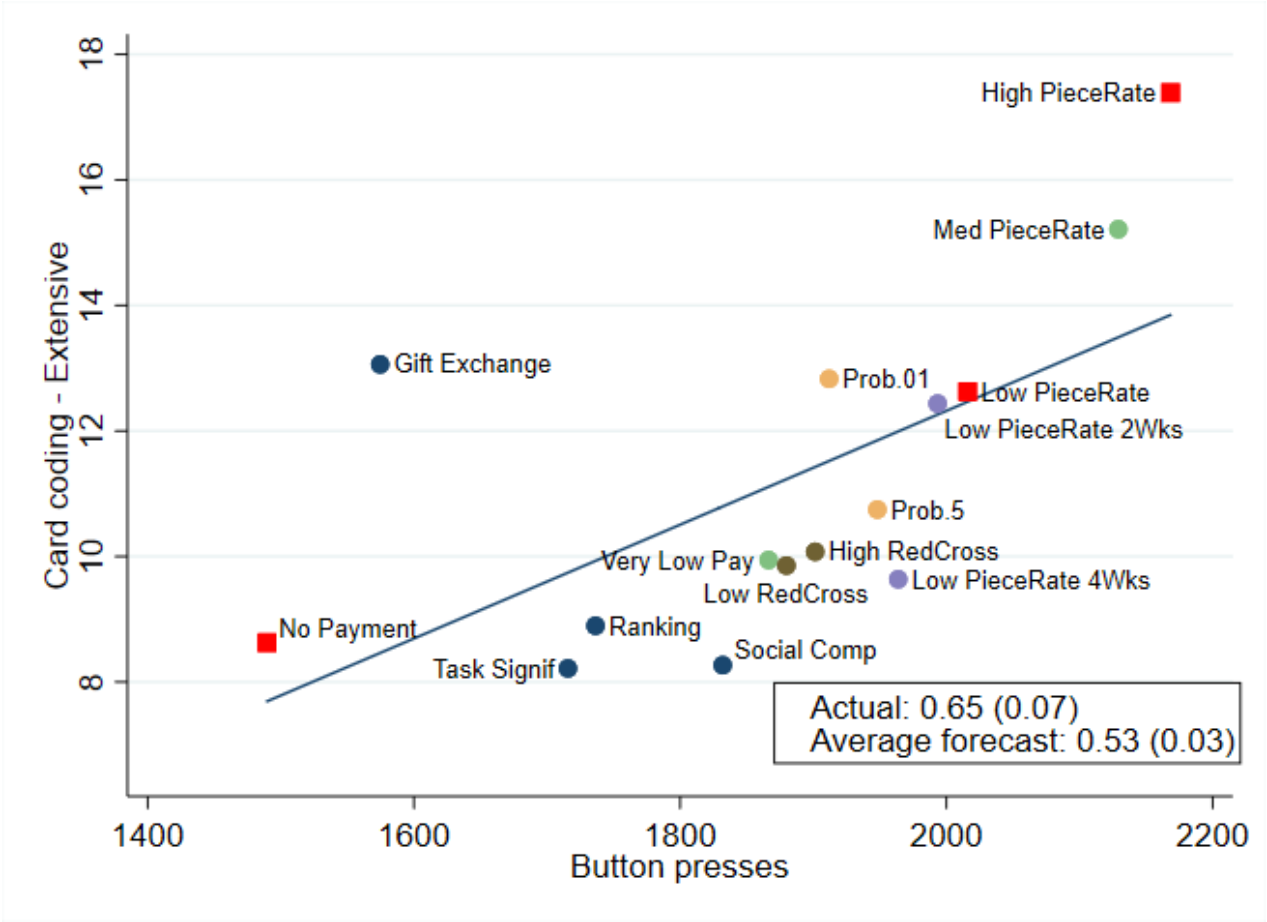
What Do We Find?

- Dimension 5a. *Output*
  - Extensive margin: *how long* will people work
  - Not much correlated with other card coding



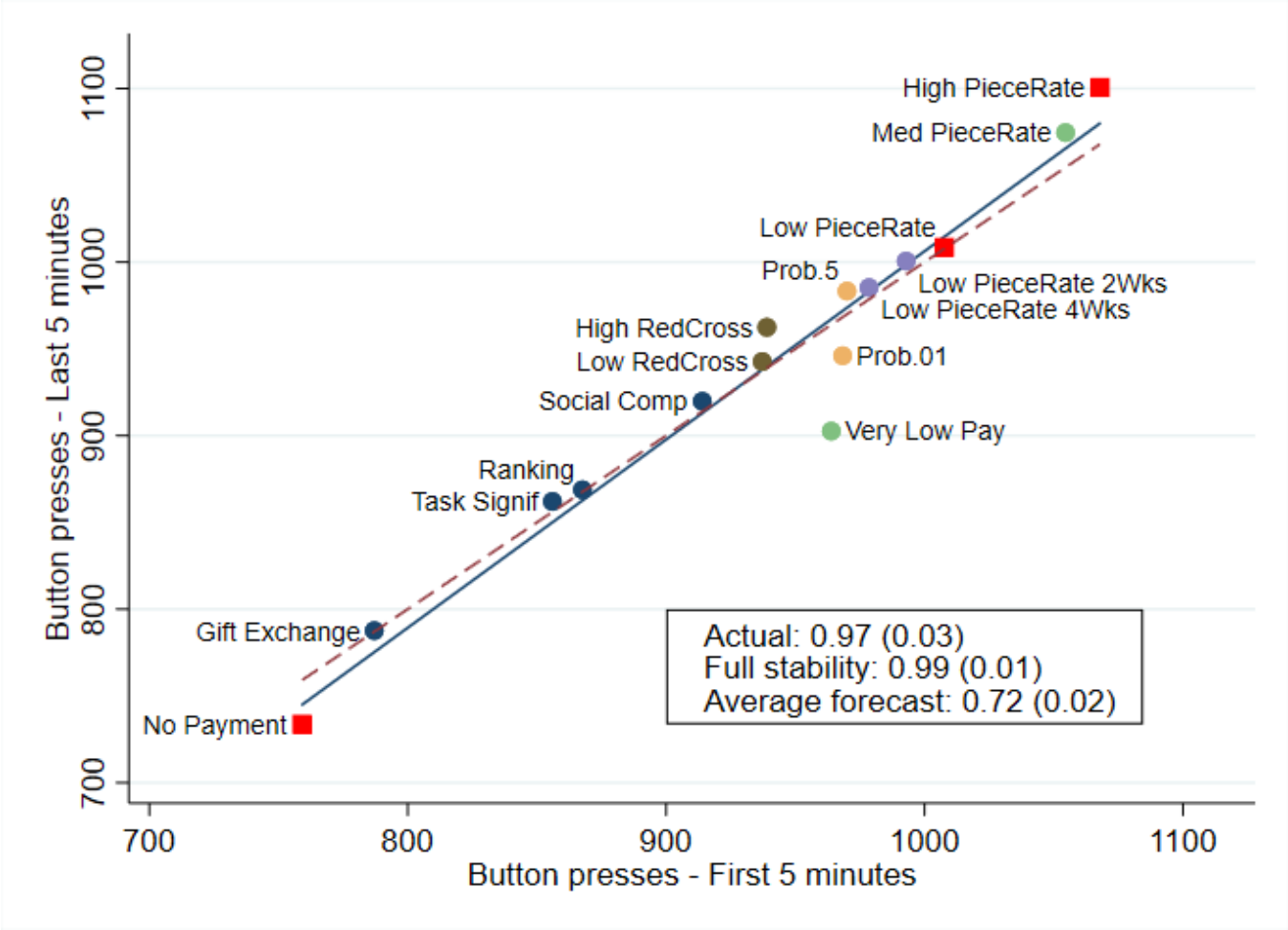
What Do We Find?

- Dimension 5b. *Output*
  - Extensive margin: *how long* will people work
  - More correlated with AB task (because less noise)



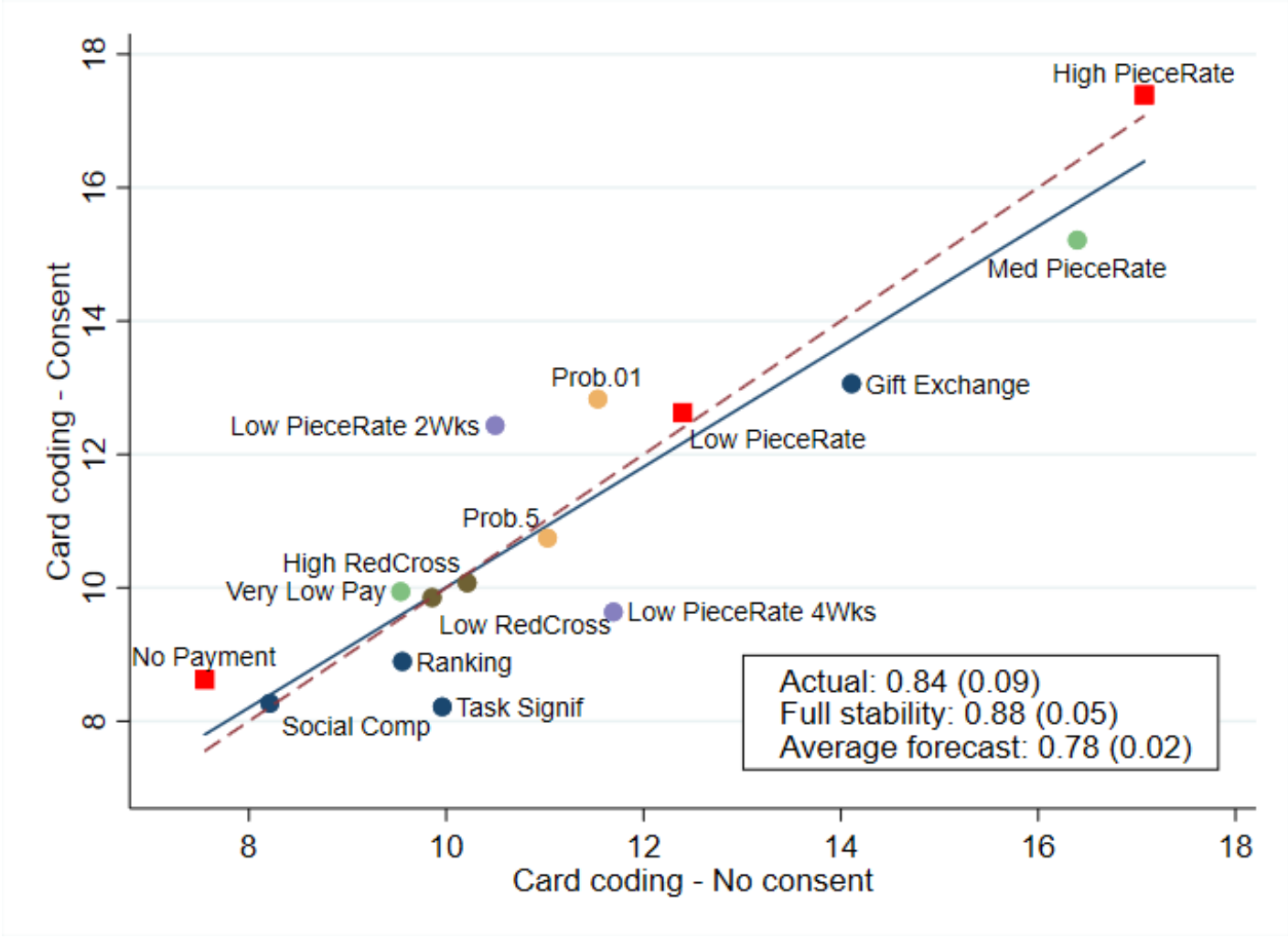
What Do We Find?

- Dimension 5c. *Output*
  - Output in first 5 minutes versus later 5 minutes
  - Very correlated



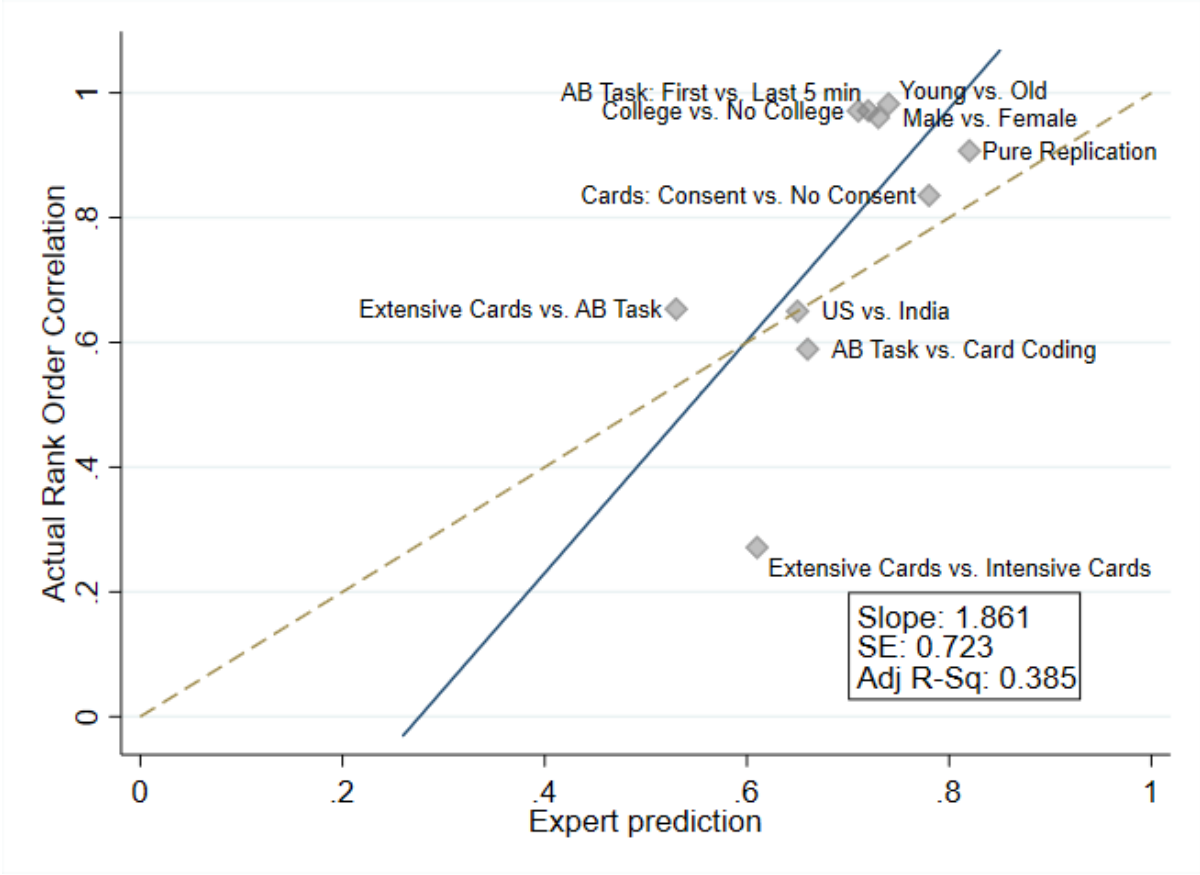
What Do We Find?

- Dimension 6. *Consent/Natural Experiment*
- Highly correlated



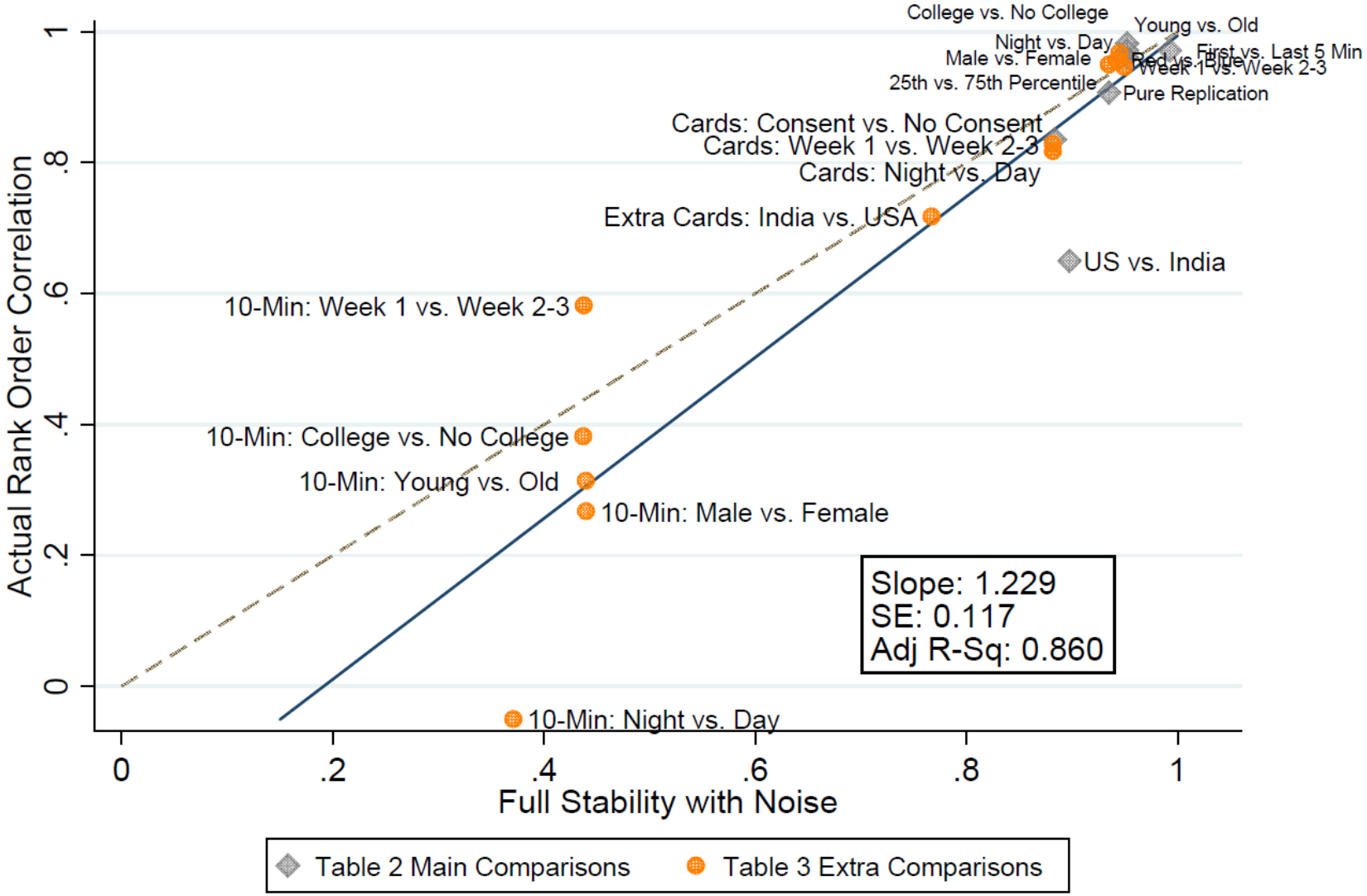
Summary Comparison

- Expert forecasts
  - overestimated the role of demographics
  - underestimate the role of noise
  - poorly correlated with actual correlation



Summary Comparison

- Benchmark of full stability much better predictor



Conclusion

- Given these motivations, platform such that:
  1. Researcher posts summary of project
  2. Invite forecasts on project before results known
  3. Store forecasts, with characteristics of forecaster
  4. Yet protect anonymity

- Work with BITSS together with Eva Vivalt, in coordination also with IPA
- Pilot platform: https://socialscienceprediction.org/