Experimental Economics (2024) 27:973–1000
https://doi.org/10.1007/s10683-024-09841-1

The roles of selection and practice in mitigating negative responses to high-powered incentives

Rosario Macera[1]

Received: 4 March 2021 / Revised: 24 August 2024 / Accepted: 26 August 2024 /
Published online: 15 October 2024

Abstract

Despite substantial evidence for the effectiveness of monetary incentives, some experiments have shown that high-powered incentives might lead to lower performance than lesser incentives. This study explores whether firms have means to counter these potential negative effects. Building on a standard experimental design identifying the drawbacks of large-stake rewards, it shows that when workers either self-select into the task or have prior practice, high-powered incentives lead to higher average performance than a smaller reward. This effect is driven mainly by selection and practice increasing the share of workers who respond positively to high-powered incentives. These results suggest that firms have natural instruments to deal with the potential adverse effects of high-powered incentives.

Keywords High-powered incentives · Choking under pressure

JEL classification D03 · D86 · D90 · J33

I specially thank CONICYT Fondecyt Regular #1190305 and Instituto de Sistemas Complejos de Ingeniería (ISCI), ANID PIA/PUENTE AFB230002 for funding. Hugo Correa, Carla Guadalupi, Alejandro Guin-Po, Alejandro Hirmas, Guillermo Irarrázabal, Fernanda y Sofía Lozano, Maria Cristina Riquelme and Pablo Sánchez provided excellent research assistance. I thank Raicho Bojilov, Edgar Kausel, Carlos Noton, Joaquín Poblete, Zoe Rahwan and Mike Waldman for useful comments. I further thank participants at the 9th Maastricht Behavioral and Experimental Economics Symposium, the 7th International Conference of the French Association of Experimental Economics, and EEA-ESEM Geneva 2016. Seminars participants at the Universidad de Chile, Universidad Alberto Hurtado, Universidad de Los Andes, and Centro de Encuestas UC also provided useful feedback. AER registry ID AEARCTR-0005745. Data, code and experimental instructions in the repository https://doi.org/10.60525/04teye511/UUCN1B.

Rosario Macera
rosario.macera@uc.cl

[1] School of Management, Pontificia Universidad Católica de Chile, Santiago, Chile

1 Introduction

Significant evidence supports the efficacy of monetary incentives. A wide array of research, encompassing laboratory, field, and natural experiments, shows that well-designed incentives significantly increase productivity and effort levels across economic settings. For instance, in a comprehensive review of 17 field and laboratory studies, Bandiera et al. (2021) found that implementing performance-based pay leads to an average increase in output of 0.36 standard deviations..[1]

Despite the evidence supporting the efficacy of monetary incentives, some studies suggest that high-powered incentives can adversely affect performance. Pokorny (2008) used two real-effort tasks to evaluate performance under various reward levels —very low, low, and high—compared to a group with no incentives. Their findings indicate that a low reward is more effective than a high one. Similarly, Ariely et al. (2009) randomly assigned subjects into a low, medium, or high reward group in six real-effort cognitive tasks. The results show that performance is greater in the low versus high-reward condition. In the field, Azmat et al. (2016) used random variation in the stakes of tests to show that the positive gap between female and male students' grades decreases with the stake of the test.[2]

The evidence of high-powered incentives' detrimental effect leads to a natural economic question: Can firms mitigate these negative impacts, should they exist? Even though the field of psychology has extensively studied the mechanisms that can lead high-powered incentives to affect productivity adversely, it is still unclear whether firms possess effective strategies for counteracting these detrimental effects should they need to.,[3] Shedding light on this question is important as high-powered incentives are common in real-world compensation schemes.[4]

[1] Classical references on the positive effects of monetary incentives are Jensen & Murphy (1990); Prendergast (1999); Lazear (2000); Lavy (2009); see Cole et al. (2015) for evidence on the positive effects of high-powered incentives in particular. For a more recent review of the theoretical underpinnings of the positive effects of monetary incentives, see (Lazear, 2018) For more recent evidence on the effectiveness of monetary incentives, see Englmaier et al. (yyyy); Bripi & Grieco (2023), and citations therein.

[2] Most studies using quasi-random variation in observable data to explore the adverse effects of high-stake incentives belong to the realm of sports. See Dohmen (2008); Cao et al. (2011); González-Díaz et al. (2012); Feri et al. (2013); Teeselink et al. (2020).

[3] Since the seminal works of Baumeister (1984) and Baumeister & Showers (1986) were published, psychologists have documented that high-stake situations, such as high-powered incentives, can create performance-impairing cognitive pressure. Following (DeCaro et al., 2011) there are two channels through which cognitive pressure might impair performance: first, "distraction theories, where impairment occurs because the attention needed to perform a task is affected by task-irrelevant thoughts and worries (e.g., Mobbs et al. (2009), DeCaro et al. (2011)); second, "explicit monitoring theories, where pressure leads subjects to attend so closely to the process of performing the task that it disrupts the attention needed for its execution. For reviews, see Markman et al. (2006); DeCaro et al. (2011). Related literature has studied the neurological mechanisms underpinning underperformance under monetary incentives (e.g., Aarts et al. (2014); Yu (2015)).

[4] For example, from 2011 to 2014, the highest-paid CEOs in the US received an average annual cash bonus of 4.5 million US dollars, with their total average yearly earnings of around 24 million, according to data from The New York Times in collaboration with Equilar (available on the NYT website). High-powered incentives are not limited to top executive roles but are also prevalent in more common occupations with measurable outcomes. Shelef & Nguyen-Chyung (2015) analyzed a dataset of 40,000 real estate agents who earned about 70,000 US dollars annually, on average. Their research found that the most

This study introduces a real-effort experiment designed to assess if two prevalent personnel practices in organizations—selection into the task and practice—can help mitigate the potential adverse impacts of large-stake incentives. This research question departs from previous literature, which primarily aimed to identify the existence of adverse effects of high-powered incentives.[5] This paper, in contrast, investigates the possibility of reducing these effects, should they exist, through the routine operations of firms employing standard personnel practices: selection in hiring and training.

In the experiment, participants from three universities solved a real-effort math task. After an unpaid round to familiarize themselves with the task, they solve it under two compensation schemes: a low and a high reward, the latter being ten times the former. Subjects have a negative response to high-powered incentives if productivity decreases with the stake of the incentive and a positive response if productivity increases. In the *Baseline* treatment, subjects responded to a generic campus flyer advertising a paid study and enrolled without knowledge of the task. Once in the laboratory, a random subset of the subjects was assigned to the *Practice* treatment, where they had the opportunity to extensively practice during the unpaid round before executing the task under the high and low payments (in random order). Conversely, in the *Selection* treatment, subjects were recruited through a campus flyer that explicitly advertised a paid study based on math skills, and they enrolled after receiving detailed information about the task. Subjects in this treatment, as well as those in the *Baseline*, did not have the opportunity to practice extensively during the unpaid round. At recruitment, all participants knew that their earnings would be performance-based, but the details of the incentive stakes were only revealed during the study itself.

The experimental results show that selection and practice decrease the negative effects of high-powered incentives. Both personnel practices improve the average productivity response to high-powered incentives: from a 6% insignificant decline in productivity from the low to the high payment in the *Baseline* treatment to 12% and 9% significant increases in the *Selection* and *Practice* treatments, respectively. This average productivity increase is driven mainly by the extensive margin: the share of subjects negatively responding to high-powered incentives decreases significantly by 17 and 19 percentage points in the *Selection* and *Practice* treatments, respectively. The share of subjects who increase their performance under high versus low incentives also significantly increases under both personnel practices: 15 percentage points in the *Selection* and 10 percentage points in the *Practice* treatment.

What mechanisms enable selection and practice to improve the response to high-powered incentives? Exploratory analysis suggests that subjects in the *Practice* treatment solved the math problems more quickly when faced with the high payment

Footnote 4 continued
typical contract in that field offered no fixed salary with a 50-50% commission split or a 100% commission-based payment.

[5] Another strand of the literature has studied whether these negative effects are present in highly qualified subjects. For instance, Teeselink et al. (2020) found that professional dart players appeared less susceptible to "choking under pressure relative to amateur players, while Bühren et al. (2024) show that high-stakes pressure remains evident even in World Champions and Olympic Alpine skiers.

than the low one. Participants in the *Selection* treatment, in turn, correctly solved more math problems that were, by random chance, easier and worked on the task for longer when facing high-powered incentives. Therefore, selection and practice help the subjects develop different task-solving strategies.

Showing that personnel practices can minimize the potential negative consequences of high-powered incentives is important for at least two reasons. The first is to improve our understanding of the extent to which monetary incentives can produce undesirable outcomes in actual labor markets. Based on the evidence showing that monetary incentives might sometimes harm performance, industry practitioners and mainstream media have advised against using monetary rewards, treating their potentially detrimental impacts as a norm rather than an exception.[6] This paper's results highlight that it is important to consider the firms' ability to deal with psychological reactions to monetary incentives before advising them to move away from monetary incentives to motivate the workforce.

Second, showing that selection and practice ameliorate the negative effects of high-powered incentives highlights the importance of firms' orchestration of their hiring, training, and compensation practices. This paper's results support the view in the management literature on "High Performance Work Systems, which has acknowledged the benefits of bundling these practices for firm success (Way (2002); Combs et al. (2006); Shin & Konrad (2017)). This paper offers causal evidence of one more reason whereby coordinating human resources practices can improve firms' results: ensuring a healthy response to monetary incentives by hiring best-fit workers and investing in training to ensure the best use of compensation resources.[7]

This paper contributes to the literature showing that the perverse effects of monetary incentives might be traced to deficient designs that fail to account for all relevant elements of the economic environment in which the incentive is applied. For example, Ederer and Manso (2013) demonstrate that monetary rewards do not negatively impact innovation if the incentives are deferred, allowing workers to experiment without immediate risk. In a field experiment in a retail company, Brahm & Poblete (2018) showed that a zero-average response to changes in sales targets is due to the disregarding of the fact that supervisors are heterogeneously adjusting targets based on current sales performance. Cole et al. (2015) found that the beneficial effects of high-powered incentives on loan officers' risk assessments can be muted due to delayed compensation and limited liability. In this paper, if a firm observes a negative response to high-powered incentives, it can trace it to hiring and

[6] For instance, in a 30-million-viewers TED talk, industry practitioner Dan Pink claims that "when a task calls for even rudimentary cognitive skills, a larger reward leads to poorer [rather than better] performance so "if we really want high performance [...] the solution is not to do more of the wrong thing. Beyond Pink's conclusion, quotes from practitioners such as "once basic needs are covered the psychological benefits of money are questionable or "why do organizations continue to spend countless dollars trying to motivate and engage their staff with monetary incentives if we know it doesn't work? abound in the popular press.

[7] The management literature has identified several channels whereby high-performance work systems increase firm performance such as improved employee motivation (e.g., Takeuchi et al. (2009), Jiang et al. (2012)) and better relationships and coordination (Evans & Davis (2005); Gittell et al. (2010)).

training practices that do not adequately account for the fact that workers will be paid using high-powered incentives.[8]

This study also adds to a body of research indicating that negative psychological reactions to monetary incentives are heterogeneous across subjects. For instance, in the literature on the crowding out of intrinsic motivation by extrinsic incentives, Huffman & Bognanno (2018) use a within-subjects design in a real-world labor setting to investigate how worker motivation changes after removing a monetary incentive. They find that 53% reported a decreased enjoyment of the task following the removal of the incentive. At the same time, 35% of workers indicated that their task enjoyment actually increased once the monetary reward was no longer in place.[9] In the same vein, Schlosser et al. (2019) document that around 25% of subjects outperform their real scores in the GRE test (a high-stake situation) whenever they repeat it in a voluntary non-incentivized experimental session (a low-stake situation).[10] In this paper, the finding that around 40% of the participants in the *Baseline* treatment had a positive reaction to high-powered incentives reinforces the idea that average results can mask substantial heterogeneity in the response to such incentives, and aligns with the intuition that an average negative reaction to high-powered incentives "does not mean that any person hired for a particular job would *choke* while doing it (Kamenica (2012)).[11]

The rest of this paper is organized as follows. Section 2 presents the experimental design. Section 3 shows the productivity response to high-powered incentives in the intensive and extensive margins. Section 4 discusses potential confounds and extensions, and Sect. 5 provides a general discussion of the findings.

2 The experimental design

(1) *Subject pool and procedures*. Two hundred and ninety-three students from three universities in Chile participated in four experimental sessions spanning from Fall 2015 to Spring 2020. Flyers posted across campuses advertised the study, and, except for the first two sessions, the flyers were also posted on each university's social

[8] Other unintended consequences of monetary incentives that can be traced to designs that forgo important aspects of the economic environment are: employee gaming (Oyer (1998); Larkin (2014); Pierce et al. (2022)), outcome manipulation (Jacob & Levitt (2003); Fisman & Wang (2017)), and distorted career concerns (Acemoglu et al. (2020)).

[9] In the crowding out of intrinsic incentives by extrinsic incentives, monetary incentives are detrimental to performance as they deplete the consumption value of the task arising from social preferences (e.g., Englmaier & Leider (2012); Siemens (2013)) or the workers' intrinsic valuation of the task (e.g., Deci (1971); Deci & Ryan (1985)). See (Kőszegi, 2014) for a review.

[10] See their Fig. 3. They further document that the grade gap between the high and low-stake results of the GRE test depends on demographic factors such as ethnicity and gender.

[11] The heterogeneity in the negative response to large-stake incentives is also in line with the growing literature emphasizing that behavioral traits such as social preferences (e.g., Kranton & Sanders (2017); Cappelen et al. (2020)), loss aversion (e.g., Goette et al. (2019); Gächter et al. (2022)), and intertemporal preferences (e.g., Huffman et al. (2019); Aycinena et al. (2022)) are also heterogeneous across subjects.

media. The experiment took place in each university's computer laboratory through a private web page designed in JavaScript for an improved graphical interface.[12]

(2) *The task*. The subjects had to find the two numbers adding up to 10 in a 4×3 table containing two-decimal random numbers. Subjects had to click on their two chosen numbers and then click a "next button. Once they clicked "next, they could not return to previous tables. They had to solve as many tables as possible in a four-minute period, with a maximum of 20 tables. At the top of the page was a counter showing the number of correctly solved tables and a chronometer displaying the elapsed time. There was no penalty for incorrect answers.[13]

This task, introduced by Ariely et al. (2009), has several advantages. First and foremost, because it is simple and fast, it can be administered repeatedly to compare performance under different incentive schemes. Second, its outcome (number of correct tables) is observable and has no prosocial component. Performance, therefore, should increase with the incentive stake.[14] Third, it is a cognitive task, i.e., a task that requires mental effort or "cognitive resources—including perception, memory, and judgment (Russo & Dosher (1983); Cooper-Martin (1994)). Using a cognitive task is important since pressure from high-powered incentives does not impair performance in non-cognitive tasks (Ariely et al., 2009).

(3) *Payments (within subjects)*. Once in the laboratory, the platform instructed the subjects that they would perform the task three times: in an unpaid round to get familiar with it and then in two paid rounds. The platform further instructed the subjects that the exact compensation would be described before each round.

The payment structure followed that in Ariely et al. (2009) for comparability. The low payment offered 13.3 US dollars for correctly solving 10 tables and a piece rate of 1.6 US dollars for each correct table above 10. The high payment was 10 times the low payment: 133 dollars for solving 10 correct tables and a piece rate of 16 dollars for each correct extra table.[15] The order of the high and low payments was randomized for each subject. Table 1 summarizes the payments.

[12] The experiment was registered on the AEA RCT registry site in April 2020 before the last two experimental sessions were implemented (33% of the sample). Ninety-eight percent were attending three prominent Chilean universities, while 2% belonged to other smaller institutions. See Table 2 for a description of the data collection.

[13] There was no button for skipping tables. However, subjects could submit a random answer and proceed to the next table without penalty. See the full experimental platform in the Appendix. Further, the external validity of tasks like this has improved over time. For instance, managers increasingly use similar computerized assignments to screen for cognitive skills during hiring and recruitment processes. In response, firms are starting to sell products such as the "General Mental Ability test, which practitioners recommend as superior to unstructured hiring tools (Tarki & Sanandaji (2020)).

[14] Whenever outcomes are not observable, low-powered incentives can be optimal even if subjects are not affected by performance-impairing pressure (e.g., Lazear (1986), Holmström & Milgrom (1991)). The absence of a prosocial component is important because monetary incentives—and especially high-powered ones—can harm performance, e.g., they can spoil cooperation (Falk & Fehr, 1999) or the signal sent by task execution that the worker is of a prosocial type (Bénabou & Tirole, 2006).

[15] The design did not include a very low-powered incentive, as this paper's goal is to address the role of firms in mitigating potential negative effects of large-stake rewards. For findings on the negative effects of very small payments on productivity, see (Gneezy and Rustichini, 2000); Gneezy & Rey-Biel (2014). For evidence showing no effect of a small payment on performance, see (Pokorny, 2008); DellaVigna && Pope (2018). Similarly, the compensation was not calibrated to be cost-effective. Rather, its structure

Table 1 Monetary Incentives in the Low and High Payments

Number of correct tables	Order randomized at subject level	
	Low payment (US dollars)	High payment (US dollars)
$\leqslant 9$	0	0
$= 10$	13.3	133
+ 1 above 10	1.6 each	16 each

Payments were round and easy to understand in their Chilean pesos (CLP) equivalent: In the low (high) payment, subjects received 8,000 (80,000) CLP for reaching the 10-table threshold and 1,000 (10,000) CLP for each correct table after that. The exchange rate varied between 600-800 CLP per US dollar during implementation, while inflation remained low

The high payment was large for this subject population. The standard hourly wage for an undergraduate in the hosting universities was around 6.6 US dollars, while the average monthly tuition was 600-850 US dollars.[16] Since the maximum earning under the high payment was approximately 300 US dollars, a student could earn half of his monthly tuition or make a whole month (45 h) of a research assistant's salary for a study taking less than an hour.

(4) *Treatments (between subjects)*. To study the effects of selection and practice on the response to high-powered incentives, the treatments varied in the amount of information provided about the task at recruitment and the opportunity to practice it before the paid rounds.

(4.1) *Baseline treatment* (N=144). Subjects in this treatment received no information about the task at the recruitment stage and had no chance to practice it.

4.1.1) *Recruitment in the Baseline treatment*. Flyers invited students to participate in a brief study without any reference to the task. See the flyer in Fig. 1, panel (a). The response email to interested subjects advertised a "study on productivity and contained only logistic information such as date, location, and a minimum payment of approximately 3.3 US dollars. In the email, they were informed that the payment could be greater if their "involvement in the study was good.

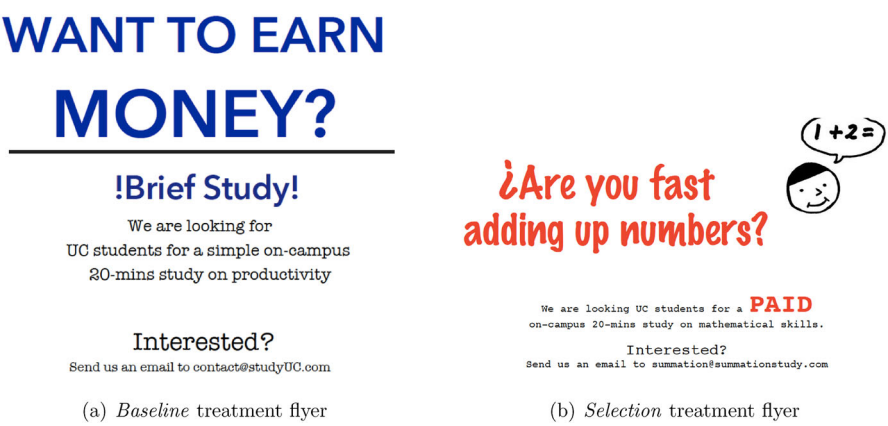
4.1.2) *Task execution in the Baseline treatment*. Once in the laboratory, the subjects executed the task three times, once in the unpaid round and then with a high and low payment (in random order), as described before.

(4.2) *Selection treatment* (N=84). In this treatment, the recruitment flyer and the reply email offered details about the task.

4.2.1) *Recruitment in the Selection treatment*. Contrarily to the generic flyer used in the *Baseline* treatment, the flyer in the *Selection* treatment explicitly asked "Are you fast at adding up numbers? and advertised a "study on mathematical skills. See

Footnote 15 continued
followed that in Ariely et al. (2009) for comparability. An interesting (but out of scope) research question would be to estimate the compensation structure and its corresponding elasticity of effort to determine the optimal level of high-powered incentives.

[16] Approximate amounts considering the students' majors and an average from 2015 to 2018. Inflation in Chile was stable and low during the experimental period.



(a) *Baseline* treatment flyer (b) *Selection* treatment flyer

Fig. 1 Recruitment Flyers in the *Baseline* and *Selection* Treatments.
Notes. Flyers were independently distributed across campuses, each featuring a distinct contact email (contact@studyUC.com and summation@summationstudy.com). The same flyers were used during social media advertising. Subjects who simultaneously applied to both studies were assigned to a waiting list and then rejected from both studies

Fig. 1, panel (b).[17] The response email sent to interested subjects contained the same logistic information offered in the *Baseline* treatment plus the following paragraph with the task description:[18]

> *We will present you with a series of tables with 12 numbers. In each table, you will have to find the two numbers that add up to 10. We will offer you a total of 20 tables and a time limit for solving them.*

4.2.2) *Task execution in the Selection treatment*. Once in the laboratory, the study was the same as that in the *Baseline* treatment.

(4.3) *Practice treatment* (N=65). In this treatment, subjects were allowed to practice the task extensively before executing the task in return for payment.

4.3.1) *Recruitment in the Practice treatment*. Subjects in this treatment were randomly selected from those enrolled in the *Baseline* treatment.

4.3.2) *Task execution in the Practice treatment*. Once in the laboratory, subjects in this treatment were offered the opportunity to rehearse the task. To this end, the unpaid round was not constrained to a four-minute round; rather, subjects could practice for up to 20 min (they could stop practicing at any time before the 20 min elapsed). Except for the duration of the unpaid round, the rest of the experiment was the same as that used for the subjects in the *Baseline* and *Selection* treatments.

[17] Two different research assistants designed the flyers to minimize the possibility that students would link them. Further, the flyers for the *Baseline* and *Selection* treatments were posted separately whenever possible. Finally, students could participate only in one of the two studies. In the rare case that students replied to both flyers, they were assigned to a waiting list and rejected from both studies once enrollment closed.

[18] See the complete response email to interested subjects in the online Appendix.

Table 2 Data collection

	Treatment	Subjects per treatment	Shifts per session	Number of Universities
Session 2015	*Baseline*	88	6	1
(University 1)	*Selection*	36		
	Practice	0		
Session 2018	*Baseline*	8	3	2
(University 2)	*Selection*	9	(within shift	
	Practice	32	randomization)	
(University 3)	*Baseline*	9	4	
	Selection	17	(within shift	
	Practice	1	randomization)	
Session 2020	*Baseline*	39	27	1
(University 1)	*Selection*	22	(within shift	
	Practice	32	randomization)	
Total	Baseline	144	37	3
	Selection	84		
	Practice	65		

The 2015 session did not collect data for the *Practice* treatment. This session collected data for a variation of the *Practice* treatment where a subsample of the subjects in the *Baseline* were invited to repeat the study (see a description of this treatment in Sect. 4). Within each shift, only the *Practice* treatment was randomized (the *Selection* treatment is not randomized by design). The experiment was registered in the AEA RCT registry site in April 2020 (ID AEARCTR-0005745) before the last two experimental sessions were implemented (33% of the sample). Preregistration considered all treatments. The outcome variable, measured as the number of correctly solved tables, was registered

Table 3 Descriptive statistics

	Woman (%)	Math degree (%)	Reaches 10 tables under low payment (%)	Reaches 10 tables under high payment (%)	*Baseline* productivity
Baseline	61.81	40.28	15.97	12.50	1.29
Selection	41.67	46.43	29.76	33.33	1.80
Practice	64.62	36.92	20.00	23.08	1.50

Math degree is a dummy for subjects in majors such as business, engineering, or statistics. *Baseline productivity* is the number of correct tables per minute in the unpaid round (inputing a zero to a few observations that did not solve any table during the unpaid round). The high payment is 10 times the low payment. See Table 1

(5) *Data collection*. Table 2 provides details of the data collection across sessions. Table 3 presents the summary statistics by treatment. Regressions in the upcoming Sect. 3 show that no results change when controlling for session and university fixed effects.

3 Results

3.1 Effects of high-powered incentives are heterogeneous

Result 1 compares the average productivity across payments for subjects in the *Baseline* treatment where subjects have no prior knowledge of the task, nor can they practice it beforehand. Across all results, productivity is defined as the number of correctly solved tables.[19]

Result 1 On average, subjects in the Baseline treatment solve 0.361 fewer correct tables under the high than the low payment. This decrease, however, is not statistically significant.

Table 4, column (1), examines this result in a controlled regression framework. It shows the OLS estimates of the number of tables correctly solved regressed on a dummy for the high payment. Following the within-subject design, standard errors are clustered at the subject level. The average number of correct tables under the low payment is 6.18. The high payment reduces this average by 0.36 tables, a decrease that is not statistically significant (p-value = 0.176).[20] Column (2) adds university and session fixed effects to show that this result is robust across experimental sessions: the point estimate remains unchanged, and so does the p-value (p-value = 0.182). Columns (3) and (4) include demographics (dummies for gender and math degree) and baseline productivity measured as the number of correct tables per minute in the initial unpaid round. In both columns, the point estimate and its p-value remain similar (p-value = 0.184 and 0.169, respectively). Following (Malmendier & Schmidt, 2017); DellaVigna et al. (2019), the regression in Table 4, bottom panel, also clusters standard errors at the shift level using the wild-cluster bootstrap of Cameron et al. (2008) to account for the small number of clusters. Clustering does not change the non-significance of the high-payment dummy: the wild-cluster bootstrap p-value ranges between 0.1621 and 0.1784.

Beyond the average treatment effect, the within-subject design identifies the response to high-powered incentives at the individual level. Result 2 shows that the small negative average response to high-powered incentives hides substantial heterogeneity.

Result 2 The small, non-significant average behavior in the *Baseline* treatment is not driven by most subjects not responding to incentives. On the contrary, 39.58% of the subjects increase their productivity under the high relative to the low payment, while 50.00% decrease their productivity.

Figure 2 displays a histogram of the change in the number of correct tables between the high versus the low payment and its associated density. Around 40% of the subjects in the *Baseline* treatment increase their productivity when the

[19] In this task, the number of correct tables is the natural productivity measure as it defines the subjects' payoffs. Other studies showing that stressors harm performance use similar measures (e.g., Pokorny (2008); Essl & Jaussi (2017)). However, this definition of productivity differs from that in Ariely et al. (2009) who use the participant's earnings as a fraction of total possible income.

[20] This non-significance contrasts with the statistically significant negative average result in Pokorny (2008); Ariely et al. (2009).

Table 4 Response to High-Powered Incentives in the *Baseline* Treatment

	OLS estimates of dependent variable:			
	Number of correct tables			
	(1)	(2)	(3)	(4)
High payment	−0.361	−0.361	−0.381	−0.385
	(0.266)	(0.270)	(0.272)	(0.278)
Woman			−0.545	−0.390
			(0.439)	(0.441)
Math degree			0.574	0.356
			(0.466)	(0.476)
Baseline productivity			1.461	1.719
			(0.311)***	(0.309)***
Constant	6.181	7.109	4.488	7.089
	(0.284)***	(1.272)***	(0.549)***	(1.302)***
University & session fixed effects		✓		✓
Wild-cluster bootstrap-t at the shift level:				
High payment	0.1712	0.1784	0.1621	0.1621
R^2	0.00	0.05	0.21	0.25
N	288	288	288	288

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Standard errors clustered at the subject level. Th sample includes all subjects in the *Baseline* treatment (N=144 with two observations per subject). *High payment* takes value one for the high and zero for the low payment. *Math degree* is a dummy for subjects in majors such as business, engineering, or statistics. *Baseline productivity* is the number of correct tables per minute in the unpaid round. The wild-cluster bootstrap uses the boottest with 9999 replications, a seed of 4500 for replicability, and Rademacher or Webb weights (Roodman et al., 2019)

reward size increases, solving an average of 2.74 *more* correct tables under the high payment. In turn, 50% decrease their productivity, solving an average of 2.89 *fewer* correct tables under the high relative to the low payment. The histogram further reveals that, on the left tail, 34.72% of the subjects solve two or fewer correct tables, while 25.00% solve three or fewer. On the right tail, 27.08% of the subjects solve two or more correct tables, and 18.06% solve three or more. Therefore, the insignificant average response in Result 1 stems from two groups responding to incentives with similar average magnitudes but in opposite directions.

3.2 Selection and practice improve the response to high-powered incentives

Figure 3 and Result 3 show that both personnel practices improve the average productivity change between the low and high payments relative to the *Baseline* treatment.[21]

[21] Subjects in the *Practice* treatment had 20 min available to practice in the unpaid round, during which they practiced for an average of 12.29 min and solved 19.28 tables (15.63 correct ones). Participants in the *Baseline* and *Selection* treatments, whose unpaid round was constrained to the standard 4 min, practiced for 3.36 and 3.64 min on average and solved 6.51 and 7.75 tables (4.88 and 6.57 correct ones), respectively.

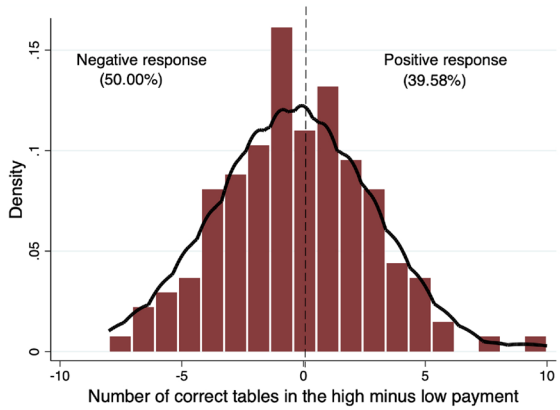


Fig. 2 Productivity Change Between the High and Low Payments in the *Baseline* Treatment.
Notes. The high payment is 10 times the low reward. Productivity is the number of tables correctly solved under each payment. Productivity change is productivity under the high payment minus that under the low payment. A subject has a "Negative response ("positive response) to high-powered incentives if the number of correct tables is lower (higher) in the high versus the low payment. Other subjects solve the same number of correct tables across payments

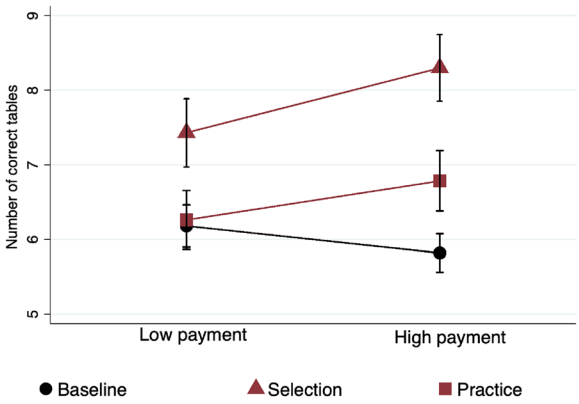


Fig. 3 Average Productivity From the Low to the High Payment Across Treatments.
Notes. The capped bars represent the standard error of the mean. The high payment is 10 times the low reward. Productivity is the number of tables correctly solved under each payment. The order of the high and low payments is randomized at the subject level. In the *Selection* treatment, subjects receive information about the task before enrollment. In the *Practice* treatment, subjects had the opportunity to rehearse the task for up to 20 min before executing it for payment. In the *Baseline*, there is no selection nor practice

Result 3 Contrary to the 6% productivity decline in the *Baseline*, in the *Selection* treatment, productivity increases by 12%, from 7.43 correct tables in the low payment to 8.30 in the high payment. In the *Practice* treatment, productivity increases by 9%, from 6.26 to 6.79. Both increases are statistically significant.

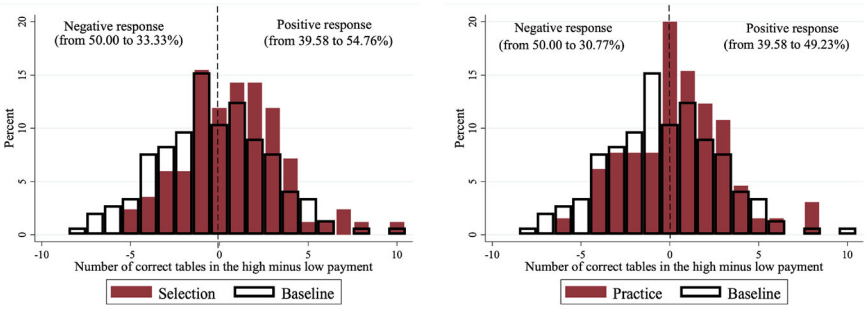
Table 5 Response to High-Powered Incentives Across Treatments

	OLS estimates of dependent variable:			
	Number of correct tables			
	(1)	(2)	(3)	(4)
High payment	−0.361	−0.361	−0.370	−0.370
	(0.266)	(0.268)	(0.269)	(0.271)
Selection	1.248	0.436	0.557	0.079
	(0.537)**	(0.524)	(0.464)	(0.478)
High payment x *Selection*	1.230	1.230	1.284	1.286
	(0.410)***	(0.413)***	(0.417)***	(0.420)***
Practice	0.081	0.180	−1.846	−2.185
	(0.486)	(0.600)	(0.642)***	(0.721)***
High payment x *Practice*	0.884	0.884	0.931	0.932
	(0.435)**	(0.438)**	(0.469)**	(0.474)*
Woman			−1.331	−1.195
			(0.345)***	(0.332)***
Math degree			1.020	0.777
			(0.343)***	(0.346)**
Baseline productivity			0.611	0.633
			(0.126)***	(0.127)***
Constant	6.181	7.109	5.839	7.241
	(0.284)***	(1.264)***	(0.410)***	(1.167)***
University & session fixed effects		✔		✔
Wild-cluster bootstrap-t *p*-value at shift level:				
Selection	0.0049	0.2922	0.1211	0.8327
High payment x *Selection*	0.0010	0.0010	0.0009	0.0008
Practice	0.8676	0.6828	0.2469	0.0084
High payment x *Practice*	0.0233	0.0233	0.0179	0.0177
R^2	0.06	0.14	0.23	0.28
N	586	586	586	586

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Standard errors clustered at the subject level. The sample includes subjects in all treatments: 144 in *Baseline*, 84 in *Selection*, and 65 in *Practice* (293 subjects with two observations per subject). *High payment* is a dummy taking value one for the high payment and zero for the low payment. *Math degree* is a dummy for subjects in mayors such as business, engineering, or statistics. *Baseline productivity* is the number of correct tables per minute in the unpaid round. The wild-cluster bootstrap uses the boottest with 9999 replications, a seed of 4500 for replicability, and Rademacher or Webb weights (Roodman et al., 2019)

Table 5 explores the statistical significance of the average treatment effects on productivity pictured in Fig. 3. It shows the OLS estimates of the number of tables correctly solved regressed on a dummy for the high payment, dummies for



(a) Mann-Whitney p-value = 0.0048. (b) Mann-Whitney p-value = 0.0529.

Fig. 4 Productivity Change Between the High and Low Payments Across Treatments.
Notes. Productivity change is productivity under the high payment minus that in the low payment. A subject has a "Negative response ("positive response) to high-powered incentives if the number of correct tables is lower (higher) in the high versus the low payment. Other subjects solve the same number of correct tables across payments. The null hypothesis in the Mann–Whitney test is that both samples are drawn from the same distribution

both treatments, and their interaction. Standard errors are clustered at the subject level.

Column (1) starts by replicating the insignificant average decrease of −0.36 correct tables in the *Baseline* treatment, captured by the coefficient of the high-payment dummy. In contrast, the productivity increase in response to the high payment is significant for the *Selection* and *Practice* treatments: the point estimate of the interaction between the high payment and *Selection* dummies is 1.23 (p-value = 0.003), while that for the *Practice* treatment is 0.88 and significant (p-value = 0.043). These productivity responses to high versus low incentives are economically important. For the *Selection* treatment, they imply that productivity increases by 12% from the low to the high payment (an increase of 1.23−0.36 tables from a productivity of 6.18+1.25 tables under the low payment), while in the *Practice* treatment it corresponds to a 9% productivity increase (an increase of 0.88−0.36 from 6.18+0.08). Columns (2) and (3) show that these effects are robust to adding university and session fixed effects plus demographic controls and baseline productivity.

Appendix Table A1 shows that the effects reported in Result 3 are consistent across the experimental sessions. For the 2015, 2018, and 2020 sessions, the parameter of the interaction between the dummies for the *Selection* treatment and the high payment ranges from 1.00 to 1.54 more correct tables. Similarly, the interaction with the high payment in the *Practice* treatment ranges from 0.86 to 1.31 more correct tables. Additionally, the high-payment dummy has a negative parameter estimate across all sessions, ranging from −0.26 to −0.65, indicating that the result reported in Result 1 is also robust.

Next, I explore the role of the extensive and intensive margins in the average productivity improvements in the *Selection* and *Practice* treatments. To this end, Fig. 4 starts by showing that the distribution of the productivity change between the

high and low payments in the *Selection* and *Practice* treatments is shifted to the right relative to that for the *Baseline*. This is important as it shows that the average productivity increases caused by selection and practice are not simply due to improvements in the responses of a few superstar subjects.

Result 4 *Selection* and *Practice* decrease the share of subjects whose productivity declines in the high relative to the low payment: from 50% in the *Baseline* treatment to 33% and 31% in the *Selection* and *Practice* treatments, respectively.

In further detail, in the *Selection* treatment, the high payment leads to a 14.40 percentage point increase in the share of subjects solving two or more correct tables and a 15.77 percentage point increase in the share solving three or more (relative to the *Baseline*). Conversely, the percentage of subjects solving two or fewer correct tables decreases by 16.68 percentage points, while the percentage solving three or fewer tables decreases by 13.1 percentage points. A similar pattern emerges in the *Practice* relative to the *Baseline* treatment, where the high payment leads to a 6.77 percentage point increase in the share solving two or more tables and a 15.77 percentage point rise in the share solving three or more. The percentage of subjects solving two or fewer tables decreases by 11.64 percentage points, while the percentage solving three or fewer tables decreases by 9.62 percentage points.

Table 6 explores, in a regression framework, the magnitude and significance of the average treatment effects of *Selection* and *Practice* on the share of subjects with a negative and positive response to high-powered incentives, as described in Result 4. It shows the OLS estimates of a dummy taking the value one if productivity strictly decreased under the high versus the low payment (columns (1) and (2)) or a dummy taking the value one if productivity strictly increased (columns (3) and (4)), regressed on an intercept and dummy variables for the *Selection* and *Practice* treatments. Standard errors are robust.[22]

Column (1) shows that the share of adversely affected subjects significantly decreases by 17 percentage points. (p-value = 0.013) in the *Selection* treatment and by 19 percentage points. in the *Practice* treatment (p-value = 0.007) relative to the 50% of subjects with a negative response in the *Baseline* treatment. Column (2) shows these results do not change when controlling for demographics (gender and math degree), baseline productivity, and university and session fixed effects. As before, the wild-cluster bootstrap at the shift level leads to the same significance level (p-values of 0.0235 and 0.0283, respectively).

Column (3) shows that the share of subjects with a positive response to high-powered incentives also increases with selection and practice. The *Selection* treatment increases the share of positive responses by 15 percentage points. (p-value = 0.027), while *Practice* increases the share by a non-significant 10 percentage points. (p-value = 0.197). However, column (4) shows that this increase does reach statistical significance when adding the standard set of controls used in all previous

[22] OLS estimates are used to ease the interpretability of the parameters of interest. Robust standard errors account for the natural heteroscedasticity arising in OLS with a binary dependent variable. Using logit/probit does not change the significance of the results.

Table 6 Change in the Share of Subjects With a Negative and Positive Response

	OLS estimates of dependent variable:			
	Dummy for negative response		Dummy for positive response	
	(1 if High < Low; 0 otherwise)		(1 if High > Low; 0 otherwise)	
	(1)	(2)	(3)	(4)
Selection	−0.167	−0.172	0.152	0.146
	(0.067)**	(0.078)**	(0.068)**	(0.079)*
Practice	−0.192	−0.183	0.096	0.182
	(0.071)***	(0.088)**	(0.075)	(0.091)**
Woman		0.068		−0.051
		(0.063)		(0.064)
Math degree		0.047		−0.063
		(0.067)		(0.067)
Baseline productivity		−0.019		0.008
		(0.035)		(0.038)
Constant	0.500	0.499	0.396	0.503
	(0.042)***	(0.203)**	(0.041)***	(0.203)**
University & session fixed effects		✓		✓
Wild-cluster bootstrap-t p-value at shift level:				
Selection	0.0167	0.0235	0.0074	0.0346
Practice	0.0097	0.0283	0.1560	0.0611
R^2	0.03	0.06	0.02	0.04
N	293	293	293	293

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Standard errors are robust. The dependent variable is a dummy taking value one if the subject had a negative response to high-powered incentives, i.e., if solved strictly more tables in the low payment (columns (1) and (2)), or a dummy taking value one if the subject had a positive response to high-powered incentives, i.e., if solved strictly less tables in the low payment (columns (3) and (4)). The sample includes all 293 subjects: 144 in *Baseline*, 84 in *Selection*, and 65 in *Practice*, with one observation per subject. *Math degree* is a dummy for subjects in mayors such as business, engineering, or statistics. *Baseline productivity* is the number of correct tables per minute in the unpaid round. The wild-cluster bootstrap uses the boottest with 9999 replications, a seed of 4500 for replicability, and Rademacher or Webb weights (Roodman et al., 2019)

regressions. Clustering standard errors at the shift level led to the same result (p-values of 0.0335 and 0.0534, respectively).

Finally, Table 7 shows that only the *Selection* treatment has a statistically significant effect on the intensive margin, particularly among those with a negative response to high-powered incentives. Columns (1) and (2) show that for the subsample of subjects whose performance is greater under the low payment, only the interaction between the high payment and the *Selection* treatment dummies is

positive and significant. For this subgroup, there is no effect for subjects in the *Practice* treatment. Columns (3) and (4) show that there are no significant effects for either treatment for the subsample of subjects whose performance was greater under the high payment.

3.3 Potential mechanisms

How do *Selection* and *Practice* improve the response to high-powered incentives? Exploratory analysis suggests that they affect the subjects' strategies to solve the task under high versus low payments. I use three proxies for subjects' solving strategies: first, *speed*, measured as the average number of seconds spent in the initial tables, correct or incorrect; second, *difficulty*, measured as the sum of the percentages of correct tables that have one correct number in the first column and the first row of the table;[23] third, *persistence* on the task as time elapses, measured by the number of seconds between the last submitted table and the round's endpoint when the four minutes have elapsed. These measures were not registered.[24]

Result 5 When facing high-powered incentives, subjects in the *Practice* treatment become differentially faster than when facing the low payment, while those in the *Selection* treatment pick easier tables and keep solving tables right before the four-minute round elapses.

Table 8 presents Result 5 in a regression framework. It shows OLS regressions for each of the three proxies for subjects' solving strategies regressed on a dummy for the high payment and its interaction with the treatment dummies. Standard errors are clustered at the subject level.

Column (1) shows that high-powered incentives induce subjects in the *Practice* treatment to become faster at solving the first three tables. Under the low payment, subjects in the *Baseline* treatment take, on average, 33.48 s to solve these tables, and this average does not change under the high payment (0.84 s longer; p-value = 0.681). Instead, those who practice the task take 8.69 s less under the high payment to solve a table (p-value = 0.017). Subjects in the *Selection* treatment do not seem to become differentially faster: They decrease the average time per solved table under the high payment by three seconds, but this point estimate is not significant (p-value = 0.224). These results are robust when adding the standard set of controls (column (2)).

[23] This proxy for difficulty emanates from the top-to-bottom and left-to-right reading convention. As a result of this convention, I assume that participants tend to begin their search for the correct numbers in this column or row, potentially reducing search time. If they consider a table is too difficult, they can submit a random answer and move to the next table. Finally, recall that the positions of the correct numbers (and the correct numbers themselves) are randomly drawn in each table for each subject.

[24] Despite not being registered, the data for these proxies were recorded in all sessions on the proprietary web page where the study was implemented, except for the first eight subjects who participated in the first shift of the study. See Table 8 notes.

[25] Recall that subjects can choose a random answer to proceed to the next table if they consider that solving the current table is taking too long.

Table 7 Response to High-Powered Incentives Split by Subjects With a Negative and Positive Response

	OLS estimates of dependent variable:			
	Number of correct tables			
	Sample: Subjects with negative response		Sample: Subjects with positive response	
	(1)	(2)	(3)	(4)
High payment	−2.889	−2.936	2.737	2.752
	(0.219)***	(0.226)***	(0.251)***	(0.259)***
Selection	1.587	0.433	2.149	0.907
	(0.880)*	(0.824)	(0.639)***	(0.589)
High payment x *Selection*	0.746	0.789	0.154	0.205
	(0.332)**	(0.344)**	(0.390)	(0.404)
Practice	1.044	−1.569	0.899	−1.140
	(0.701)	(1.093)	(0.618)	(1.127)
High payment x *Practice*	0.289	0.149	−0.049	0.072
	(0.371)	(0.401)	(0.417)	(0.496)
Woman		−1.493		−1.176
		(0.551)***		(0.451)**
Math degree		1.053		0.862
		(0.539)*		(0.497)*
Baseline productivity		0.856		0.506
		(0.275)***		(0.162)***
Constant	7.556	10.425	4.351	3.000
	(0.394)***	(1.152)***	(0.351)***	(0.712)***
University & session fixed effects		✓		✓
Wild-cluster bootstrap-t *p*-value at shift level:				
Selection	0.1319	0.6571	0.0026	0.1127
Practice	0.2506	0.2884	0.0853	0.0826
R^2	0.20	0.45	0.22	0.42
N	240	240	270	270

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Standard errors are robust. The table shows the treatment effects of *Selection* and *Practice* (relative to *Baseline*) in the number of correctly solved tables for two groups of subjects: those with a negative response to high-powered incentives (columns (1) and (2); 120 subjects) and those with a positive response (columns (3) and (4); 135 subjects). *Math degree* is a dummy for subjects in mayors such as business, engineering, or statistics. *Baseline productivity* is the number of correct tables per minute in the unpaid round. The wild-cluster bootstrap uses the boottest with 9999 replications, a seed of 4500 for replicability, and Rademacher or Webb weights (Roodman et al., 2019)

Column (3) shows that subjects in the *Selection* treatment seem to pick easier tables under high-powered incentives.[25] Under the low payment, 33% of the correctly solved tables by subjects in the *Baseline* treatment are "easy tables, i.e.,

Table 8 Task Solving Strategies

	OLS estimates of dependent variables:					
	Speed (average seconds per table solved)		*Difficulty* (% of easier correct tables)		*Persistence* (remaining seconds at the end)	
	(1)	(2)	(3)	(4)	(5)	(6)
High payment	0.837	0.886	−0.012	−0.012	2.873	2.834
	(2.036)	(2.057)	(0.018)	(0.018)	(2.057)	(2.045)
Selection	−6.172	−0.675	0.060	−0.015	9.342	7.184
	(2.529) **	(2.495)	(0.031) *	(0.028)	(4.315) **	(4.340) *
High payment x *Selection*	−3.022	−3.373	0.062	0.065	−13.360	−13.123
	(2.480)	(2.511)	(0.027) **	(0.028) **	(4.539) ***	(4.545) ***
Practice	3.549	17.431	−0.004	-0.128	8.305	−3.543
	(4.275)	(6.757) **	(0.030)	(0.039) ***	(4.368) *	(5.248)
High payment x *Practice*	−8.686	−8.944	0.041	0.044	−2.593	−2.389
	(3.625) **	(3.701) **	(0.032)	(0.034)	(4.633)	(4.360)
Constant	33.481	34.407	0.331	0.327	16.786	9.092
	(2.131) ***	(2.823) ***	(0.017) ***	(0.025) ***	(1.424) ***	(2.922) ***
University & session fixed effects, baseline productivity, gender, and math degree		✓		✓		✓
Wild-cluster bootstrap-t *p*-value at shift level:						
Selection	0.0337	0.5937	0.0336	0.4718	0.0213	0.0923
High payment x *Selection*	0.1325	0.7901	0.0030	0.0031	0.0054	0.0059
Practice	0.4807	0.0004	0.9269	0.0127	0.0286	0.5701
High payment x *Practice*	0.0056	0.0047	0.2589	0.2173	0.4222	0.4375
R^2	0.04	0.17	0.04	0.25	0.03	0.09
N	571	571	572	572	572	572

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

Notes. Standard errors clustered at the subject level. The sample includes all subjects in the *Baseline*, *Selection* and *Practice* treatment, except for the eight first subjects for whom the study webpage did not record the time stamps and the seven first for whom it did not record the position of the correct numbers. The dependent variables are *Speed*, measured as the average number of seconds in the initial three tables, considering correct or incorrect tables; *Difficulty*, measured as the percentage among correct tables that had one of the correct numbers in the first column plus the percentage with one correct number in the first row; and *Persistence*, measured as the seconds between the last table is submitted (correct or incorrect) and 240 s (when the four minutes window elapses). The wild-cluster bootstrap uses the boottest with 9999 replications, a seed of 4500 for replicability, and Rademacher or Webb weights (Roodman et al., 2019)

tables where one of the correct numbers is in the first row or column. This percentage barely changes under the high payment (1% more easy tables; p-value = 0.513). In contrast, subjects who self-selected into the task solve 6% more easy tables under the low payment (p-value = 0.055) plus an extra 6% under the high payment (p-value = 0.022). Column (4) shows that these results are robust when adding controls. Subjects in the *Practice* treatment do not show any significant effect.

Finally, column (5) shows that subjects in the *Selection* treatment persist longer at the task under large-stake rewards. Relative to the low payment case, the time difference between the last submitted table and the 240-second mark (the end of the four-minute round) is 13.36 s less (p-value = 0.004). This effect is absent in the *Practice* treatment: under the high payment, the time difference between the last submitted table and the time elapsing is short and insignificant (−2.59 s; p-value = 0.576). Column (6) shows that these results are robust when adding the standard set of controls.[26]

4 Robustness

This section discusses four aspects of the experimental design that could muddy the interpretation of the results. The data does not offer support for any of these confounds.

(1) Order effects. The (random) order of the high and low payments might affect the productivity response to high-powered incentives. For instance, if the effort cost function is non-separable across payment rounds, facing the low payment first implies that subjects will be operating in the steepest part of the cost function when solving the task under the high payment. Appendix Table A2 shows that the response to high-powered incentives in the *Selection* and *Practice* treatments are similar when one divides the sample by the payment order. Columns (1) and (2) show the OLS estimates (with and without controls) of the number of correct tables for the subgroup of subjects who randomly received the high payment first. Columns (3) and (4) show the same regressions for the subgroup that received the low payment before the high payment. For the *Selection* treatment, the interaction between the high-payment dummy and the treatment dummy is positive across all regressions, ranging from 0.99 to 1.46 (relative to the 1.23 estimate in the pooled sample). For the *Practice* treatment, the interaction between the high-payment dummy and the treatment dummy ranges from 0.66 to 1.19 (relative to the 0.88 estimate in the pooled sample). This suggests that the average productivity response to high-powered incentives is independent of the payment order.

(2) Income target. If subjects have an income target for the payment rounds (as in Goette et al. (2004) and Fehr & Goette (2007)), the high and low payments could differently affect their incentives to exert effort above the 10-tables threshold. Under

[26] These results suggest that subjects in the *Selection* treatment can stay focused for longer, even under the pressure of the four minutes being about to elapse. This might reflect a higher orientation towards motivated goals in this selected group (Payne et al., 2007). Alternatively, they could be more able to use memory and attention in parallel. This ability called "working memory, has been shown fundamental for cognitive tasks (Baddeley & Hitch (1974); Baumeister (1984); Baumeister & Showers (1986)).

the high payment, incentives to solve additional tables after 10 have been successfully solved drop as the bonus, at this point, will probably exceed the income target. On the contrary, under the low payment, reaching the 10-table threshold will not necessarily meet the income target. Thus, incentives to keep solving tables are preserved. This delivers a testable prediction: If subjects have an income target, an adverse reaction to high-powered incentives could arise due to subjects solving fewer tables above 10 in the high versus the low payment.

Appendix Table A3 shows that, above the 10-table threshold, the productivity differences between the high and low payments are small and statistically insignificant. In the *Baseline* treatment, the difference is 0.25 more correct tables in the high versus the low payment, and the p-values of this difference are large using a test of means and a t-test with clustered standard errors (at the individual and shift levels). For the *Selection* treatment, the difference is equally small: 0.22 tables and not statistically significant under any computation of the standard errors. In the *Practice* treatment, the difference is larger and negative (0.70 more correct tables under the low payment) but not significant. Further, Appendix Table A3 shows that, for the *Selection* and *Practice* treatments, the percentage of subjects who reach the 10-table threshold is higher under the high payment. For the *Baseline* treatment, the reverse holds.

(3) Does selection confound practice? Since subjects in the *Selection* treatment knew the task characteristics in advance, it could be the case that they (somehow) practiced it independently before the study. To explore this possibility, 31 new subjects were recruited as in the *Selection* treatment and were offered 20 min of practice as in the *Practice* treatment. If the effects of selection are only due to independent practice, then the productivity response to high-powered incentives in this new group should be similar to that in the *Practice* treatment.

Appendix Table A4 shows that the subjects recruited as in the *Selection* treatment but who were also offered practice behave differently from those in the *Practice* treatment. The table shows the OLS estimate of the number of tables correctly solved, regressed on dummies for the high payment, the treatments, and all their interactions. The parameter of interest is the triple interaction between the high payment, selection, and practice, as this parameter captures the added effect of practice on top of selection. With or without the standard controls, the parameter has the opposite sign to the interaction between the high payment and practice. A Wald test comparing these two parameters rejects the null hypothesis of equality (p-value = 0.0208). This result suggests there is more to selection than just informal practice before the study occurs.[27]

(4) Could knowledge of the compensation stakes decrease the effects of practice? In actual firms, the extent of training is motivated by the stakes of its future returns (Becker, 1962). It is possible, however, that knowledge of the prospective high-

[27] Further evidence suggesting that the effect of selection is not due to unobserved practice comes from the findings suggesting that the subjects in the *Selection* and *Practice* treatments rely on different task-solving strategies to improve the effectiveness of the high-powered incentives. As shown in Table 8 above, the data suggest that, under high-powered incentives, practice makes the subjects faster at solving the tables, while self-selection prompts the subjects to pick easier tables and work longer on the task.

powered incentives induces cognitive pressure, damaging the positive effects of practice. To test this hypothesis, a group of 80 subjects from the *Baseline* treatment repeated the study after having the opportunity to rehearse the task online for six days.[28] If practicing the task with knowledge of the payment stakes worsens the negative effects of high-powered incentives, practice should have decreased these subjects' productivity response to large-stake rewards relative to their first-time participation.

Appendix Table A2 shows that the response to high-powered incentives does not decrease, but actually improves when subjects repeat the study after practicing the task from home. Standard errors are clustered at the subject level. An OLS regression of the number of tables correctly solved, regressed on a dummy for the high-payment case, a dummy for the study's repetition, and their interaction, shows that the subjects solved 1.20 more tables in the high versus the low payment, relative to their first-time participation (column (1), p-value = 0.013). Column (2) shows that controls do not change the results.

Columns (3) and (4) show the estimates when the sample is split according to subjects' first-time participation response to high-powered incentives. Those who previously had a negative response now solve 2.17 more correct tables under the high payment (column (3); p-value = 0.001); those who previously had a positive response still display the same response, as the parameter of the interaction between the high-payment dummy and the dummy for the study's repetition is negative but small and non-significant (column (4); p-value = 0.274). Clustering standard errors at the shift level does not change any results. These results show that practice with knowledge of the payment stakes improves productivity for those previously adversely affected by high-powered incentives. In contrast, it does not affect those who previously had a positive response.[29]

5 Discussion

This paper shows that selection and practice, key aspects of the employment relationship in actual firms, can mitigate the potential negative effects of high-powered incentives. Both selection and practice substantially increase the average productivity response to high-powered incentives. In the *Selection* treatment, the

[28] Just as standard economic theory would predict, knowledge of the compensation stakes boosted practice: Subjects who agreed to participate trained intensively and faster than those in the *Practice* treatment. They practice for an average of 47 min, with a maximum of more than four hours. Further, while subjects in the *Practice* treatment solved tables at the same speed as those in the *Baseline* in their unpaid round, 1.84 tables per minute, subjects in the repeated study solved 2.45 tables per minute, a 33% speed increase.

[29] As in any two-stage experiment, a possible confound is selective attrition. If subjects with a negative response to high-powered incentives are less likely to agree to participate in the repeated study, then the effect could overestimate the positive effect of practice. The data suggest that this is not the case. In an OLS regression of whether the invited subject showed up or not to the repeated study, the point estimate of the dummy taking the value one if the subject had a negative response to incentives in her first-time participation is never significant. The decision not to participate again was instead driven by the ease of accessing the campus where the repeated study took place and whether the subject could reap payments beyond the showing-up fee (by reaching the 10-table threshold in the low-payment case).

average increase is due to extensive and intensive margin improvements: high-powered incentives harm fewer people, and those harmed have a smaller negative response than the control. In the *Practice* treatment, the average productivity increase emanates from a smaller share of subjects negatively reacting to high-powered incentives. These results suggest that firms with adequate recruitment, selection, and training practices can safely rely on high-powered incentives to motivate their workforces.

The result that selection into the task improves the effectiveness of the high-powered incentives relates to other research exploring the effects of selection into a task on the response to incentives. Notably, in the context of a non-routine task, Englmaier et al. (forthcoming a) used a large field experiment to show that incentives improve performance by the same magnitude for teams that self-select into the task and those that participate without prior knowledge of the task.[30] Since, in their paper, selection was induced by intrinsic motivation, their results align with those in Ashraf et al. (2020), who showed that offering career benefits at the recruitment stage only crowds out prosocial traits of low-skill applicants, as the marginal applicants are more talented and equally prosocial.[31] In this paper's experiment, selection into the task can also relate to an intrinsic taste for it, as subjects enrolling in the *Selection* treatment identified themselves as being skilled in the advertised task, thus presumably enjoying it.

Further, the benefits of allowing subjects to self-select into the task relate to the broader evidence that workers also self-select into payment structures. Lazear (2000) first showed that changes in the structure of pay-for-performance affect workforce composition, while Dohmen & Falk (2011) presented laboratory evidence that differences in productivity, risk attitudes, and self-assessment of skills drive this sorting. Bellemare & Shearer (2010) showed field evidence that workers employed in a job with substantial daily income risk were significantly less averse to risk than the broader population. Dohmen & Falk (2010) presented field and laboratory evidence that workers can self-select based on personality traits and social preferences.[32]

The result that practice improves the effectiveness of high-powered incentives relates to the evidence in psychology that practicing a task under mild pressure ameliorates the negative effects of stressors. Evidence spans from math problems (Beilock et al., 2007) to sports such as golf putting and dart throwing (Oudejans & Pijpers (2010; 2009)). This evidence is further related to the literature showing that the adverse effects of pressure decrease with experience (e.g., Teeselink et al.

[30] In their paper, however, incentives do decrease the willingness to explore new solutions for the sample of subjects who did not self-select into the task.

[31] This result contrasts with those in Deserranno (2019), who showed that higher pay discouraged people with strong prosocial preferences from applying for a health-promoter position in Uganda. This result aligns with the standard "crowding out of intrinsic motivation by intrinsic incentives" hypothesis found in Deci (1971); Deci et al. (1999).

[32] More broadly, there is plenty of evidence that subjects strategically sort in and out, not only in the labor market but in other economic environments, such as prosocial situations (Dana et al. (2007); DellaVigna et al. (2012); Lazear et al. (2012); Andreoni et al. (2017)).

[33] Grip & Sauermann (2012) showed that call-center executives' productivity increased by around 10% because of a randomly assigned training program. Using administrative data, they ruled out that this estimate was affected by worker selection into training. Using a control-function approach on panel data to

(2020)), even though experience confounds practice with selection. It also relates to the literature showing a causal impact of on-the-job training on performance.[33]

I speculate that this paper's positive effects of selection and practice on the effectiveness of high-powered incentives are lower bounds than those we should observe in the real labor market. First, real-world firms screen applicants. By actively eliciting the relevant traits that drive a positive response to high-powered incentives, firms can strengthen the sorting induced by the self-selection into the task used in this experimental design. Second, even if firms cannot screen, they can improve upon the basic selection into the task by providing candidates with richer information. For instance, information about the firm's culture and other compensation package details can also enhance selection. Finally, in real-world firms, selection occurs repeatedly through dismissals and resignations. Even though costly, dismissals might be optimal if the performance decrease due to high-powered incentives is as large.

Despite their ameliorating effect, in real-world firms, the power of selection and practice can be restricted by the difficulty of substituting workers. For example, in highly competitive markets with very specific human capital, firms might prefer to keep workers adversely affected by pressure as they outperform in other scant and thus expensive skills. This might explain why field evidence of "choking under pressure is prevalent in sports (e.g., Dohmen (2008); Hickman & Metz (2015)). In other cases where the adverse effects of cognitive pressure have been identified, such as school performance (e.g., Ramirez & Beilock (2011); Azmat et al. (2016)), substitution can be unfeasible or undesirable.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10683-024-09841-1.

References

Aarts, E., Wallace, D. L., Dang, L. C., Jagust, W. J., Cools, R., & D'Esposito, M. (2014). Dopamine and the cognitive downside of a promised bonus. *Psychological Science, 25*(4), 1003–1009.

Acemoglu, D., Fergusson, L., Robinson, J., Romero, D., & Vargas, J. F. (2020). The perils of high-powered incentives: Evidence from Colombia's false positives. *American Economic Journal: Economic Policy, 12*(3), 1–43.

Andreoni, J., Rao, J. M., & Trachtman, H. (2017). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy, 125*(3), 625–653.

Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (2009). Large stakes and big mistakes. *Review of Economic Studies, 76*(2), 451–469.

Ashraf, N., Bandiera, O., Davenport, E., & Lee, S. S. (2020). Losing prosociality in the quest for talent? Sorting, selection, and productivity in the delivery of public services. *American Economic Review, 110*(5), 1355–1394.

Aycinena, D., Blazsek, S., Rentschler, L., & Sprenger, C. (2022). Intertemporal choice experiments and large-stakes behavior. *Journal of Economic Behavior & Organization, 196*, 484–500.

Footnote 33 continued
correct for the endogeneity, Konings and Vanormelingen (2015) found a sizable effect on productivity, which was larger than the resulting wage premium. Dearden et al. (2006) find a similar result.

Azmat, G., Calsamiglia, C., & Iriberri, N. (2016). Gender difference in response to big stakes. *Journal of the European Economic Association, 14*(6), 1372–1400.
Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Amsterdam: Elsevier.
Bandiera, O., Fischer, G., Prat, A., & Ytsma, E. (2021). Do women respond less to performance pay? Building evidence from multiple experiments. *American Economic Review: Insights, 3*(4), 435–454.
Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology, 46*(3), 610–620.
Baumeister, R. F., & Showers, C. J. (1986). A review of paradoxical performance effects: Choking under pressure in sports and mental tests. *European Journal of Social Psychology, 16*(4), 361–383.
Becker, G. (1962). Investment in human capital: A theoretical analysis. *The Journal of Political Economy, 70*(5), 9–49.
Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General, 136*(2), 256–276.
Bellemare, C., & Shearer, B. (2010). Sorting, incentives and risk preferences: Evidence from a field experiment. *Economics Letters, 108*(3), 345–348.
Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review, 96*(5), 1652–1678.
Brahm, F., & Poblete, J. (2018). Incentives and ratcheting in a multiproduct firm: A field experiment. *Management Science, 64*(10), 4471–4965.
Bripi, F., & Grieco, D. (2023). Participatory incentives. *Experimental Economics, 26*, 1–37.
Bühren, C., Gschwend, M., & Krumer, A. (2024). Expectations, gender, and choking under pressure: Evidence from alpine skiing. *Journal of Economic Psychology, 100*, 102692.
Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics, 90*(3), 414–427.
Cao, Z., Price, J., & Stone, D. F. (2011). Performance under pressure in the NBA. *Journal of Sports Economics, 12*(3), 231–252.
Cappelen, A., List, J., Samek, A., & Tungodden, B. (2020). The effect of early-childhood education on social preferences. *Journal of Political Economy, 128*(7), 2739–2758.
Cole, S., Kanz, M., & Klapper, L. (2015). Incentivizing calculated risk-taking: Evidence from an experiment with commercial bank loan officers. *The Journal of Finance, 70*(2), 537–575.
Combs, J., Liu, Y., Hall, A., & Ketchen, D. (2006). How much do high-performance work practices matter? A meta-analysis of their effects on organizational performance. *Personnel Psychology, 59*(3), 501–528.
Cooper-Martin, E. (1994). Measures of cognitive effort. *Marketing Letters, 5*(1), 43–56.
Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory, 33*(1), 67–80.
De Grip, A., & Sauermann, J. (2012). The effects of training on own and co-worker productivity: Evidence from a field experiment. *The Economic Journal, 122*(560), 376–399.
Dearden, L., Reed, H., & Van Reenen, J. (2006). The impact of training on productivity and wages: Evidence from british panel data. *Oxford Bulletin of Economics and Statistics, 68*(4), 397–421.
DeCaro, M. S., Thomas, R. D., Albert, N. B., & Beilock, S. L. (2011). Choking under pressure: Multiple routes to skill failure. *Journal of Experimental Psychology: General, 140*(3), 390.
Deci, E. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology, 18*(1), 105–115.
Deci, E., Koestner, R., & Ryan, R. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*, 627–668.
Deci, E., & Ryan, R. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer.
DellaVigna, S., List, J., & Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *Quarterly Journal of Economics, 127*(1), 1–56.
DellaVigna, S., List, J., Malmendier, U., & Rao, G. (2019). Estimating social preferences and gift exchange with a piece-rate design, Working paper.
DellaVigna, S., & Pope, D. G. (2018). What motivates effort? Evidence and expert forecasts. *Review of Economic Studies, 85*(2), 1029–1069.
Deserranno, E. (2019). Financial incentives as signals: Experimental evidence from the recruitment of village promoters in Uganda. *American Economic Journal: Applied Economics, 11*(1), 277–317.

Dohmen, T., & Falk, A. (2010). You get what you pay for: Incentives and selection in the education system. *The Economic Journal, 120*(546), F256–F271.
Dohmen, T., & Falk, A. (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review, 101*(2), 556–590.
Dohmen, T. J. (2008). Do professionals choke under pressure? *Journal of Economic Behavior & Organization, 65*(3), 636–653.
Ederer, F., & Manso, G. (2013). Is pay for performance detrimental to innovation? *Management Science, 59*(7), 1496–1513.
Englmaier, F., Grimm, S., Grothe, D., Schindler, D., & Schudy S. (forthcoming a) The effect of incentives in non-routine analytical team tasks. *Journal of Political Economy.*
Englmaier, F., S. Grimm, D. Grothe, D. Schindler, & S. Schudy (forthcoming b) The efficacy of tournaments for non-routine team tasks. *Journal of Labor Economics*.
Englmaier, F., & Leider, S. (2012). Contractual and organizational structure with reciprocal agents. *American Economic Journal: Microeconomics, 4*(2), 146–183.
Essl, A., & Jaussi, S. (2017). Choking under time pressure: The influence of deadline-dependent bonus and malus incentive schemes on performance. *Journal of Economic Behavior & Organization, 133*, 127–137.
Evans, W. R., & Davis, W. D. (2005). High-performance work systems and organizational performance: The mediating role of internal social structure. *Journal of Management, 31*(5), 758–775.
Falk, A., & Fehr, E. (1999). Wage rigidity in a competitive incomplete contract market. *Journal of Political Economy, 107*(1), 106–134.
Fehr, E., & Goette, L. (2007). Do workers work more if wages are high? Evidence from a randomized field experiment. *American Economic Review, 97*(1), 298–317.
Feri, F., Innocenti, A., & Pin, P. (2013). Is there psychological pressure in competitive environments? *Journal of Economic Psychology, 39*, 249–256.
Fisman, R., & Wang, Y. (2017). The distortionary effects of incentives in government: Evidence from China's"Death Ceiling" program. *American Economic Journal: Applied Economics, 9*(2), 202–18.
Gächter, S., Johnson, E. J., & Herrmann, A. (2022). Individual-level loss aversion in riskless and risky choices. *Theory and Decision, 92*(3), 599–624.
Gittell, J. H., Seidner, R., & Wimbush, J. (2010). A relational model of how high-performance work systems work. *Organization Science, 21*(2), 90–506.
Gneezy, U., & Rey-Biel, P. (2014). On the relative efficiency of performance pay and non-contingent incentives. *Journal of the European Economic Association, 12*(1), 62–72.
Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *Quarterly Journal of Economics, 115*(3), 791–810.
Goette, L., Graeber, T., Kellogg, A., & Sprenger, C. (2019). Heterogeneity of gain-loss attitudes and expectations-based reference points, *SSRN 3589906*.
Goette, L., Huffman, D., & Fehr, E. (2004). Loss aversion and labor supply. *Journal of the European Economic Association, 2*(2–3), 216–228.
González-Díaz, J., Gossner, O., & Rogers, B. W. (2012). Performing best when it matters most: Evidence from professional tennis. *Journal of Economic Behavior & Organization, 84*(3), 767–781.
Hickman, D. C., & Metz, N. E. (2015). The impact of pressure on performance: Evidence from the PGA tour. *Journal of Economic Behavior & Organization, 116*, 319–330.
Holmström, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization, 7*(2), 24–52.
Huffman, D., & Bognanno, M. (2018). High-powered performance pay and crowding out of nonmonetary motives. *Management Science, 64*(10), 4669–4680.
Huffman, D., Maurer, R., & Mitchell, O. S. (2019). Time discounting and economic decision-making in the older population. *The Journal of the Economics of Ageing, 14*, 100121.
Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics, 118*(3), 843–877.
Jensen, M., & Murphy, K. (1990). Performance pay and top-management incentives. *Journal of Political Economy, 98*(2), 225–264.
Jiang, K., Lepak, D. P., Hu, J., & Baer, J. C. (2012). How does human resource management influence organizational outcomes? A meta-analytic investigation of mediating mechanisms. *Academy of Management Journal, 55*(6), 1264–1294.
Kamenica, E. (2012). Behavioral economics and psychology of incentives. *Annual Review of Economics, 4*(1), 427–452.

Kőszegi, B. (2014). Behavioral contract theory. *Journal of Economic Perspectives, 52*(4), 1075–1118.
Konings, J., & Vanormelingen, S. (2015). the impact of training on productivity and wages: Firm-level evidence. *Review of Economics and Statistics, 97*(2), 485–497.
Kranton, R. E., & Sanders, S. G. (2017). Groupy versus non-groupy social preferences: Personality, region, and political party. *American Economic Review, 107*(5), 65–69.
Larkin, I. (2014). The cost of high-powered incentives: Employee gaming in enterprise software sales. *Journal of Labor Economics, 32*(2), 199–227.
Lavy, V. (2009). Performance pay and teachers effort, productivity, and grading ethics. *American Economic Review, 99*(5), 1979–2011.
Lazear, E. (2000). Performance pay and productivity. *American Economic Review, 90*(5), 1346–1361.
Lazear, E. P. (1986). Salaries and piece rates. *Journal of Business, 59*(3), 405–431.
Lazear, E. P. (2018). Compensation and incentives in the workplace. *Journal of Economic Perspectives, 32*(3), 195–214.
Lazear, E. P., Malmendier, U., & Weber, R. A. (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics, 4*(1), 136–163.
Malmendier, U., & Schmidt, K. M. (2017). You owe me. *American Economic Review, 107*(2), 493–526.
Markman, A. B., Maddox, W. T., & Worthy, D. A. (2006). Choking and excelling under pressure. *Psychological Science, 17*(11), 944–948.
Mobbs, D., Hassabis, D., Seymour, B., Marchant, J. L., Weiskopf, N., Dolan, R. J., & Frith, C. D. (2009). Choking on the money: Reward-based performance decrements are associated with midbrain activity. *Psychological Science, 20*(8), 955–962.
Oudejans, R. R., & Pijpers, J. R. (2009). Training with anxiety has a positive effect on expert perceptual-motor performance under pressure. *Quarterly Journal of Experimental Psychology, 62*(8), 1631–1647.
Oudejans, R. R., & Pijpers, J. R. (2010). Training with mild anxiety may prevent choking under higher levels of anxiety. *Psychology of Sport and Exercise, 11*(1), 44–50.
Oyer, P. (1998). Fiscal year ends and non linearincentive contracts: The effect on business seasonality. *Quarterly Journal of Economics, 113*(1), 149–185.
Payne, S. C., Youngcourt, S. S., & Beaubien, J. M. (2007). A meta-analytic examination of the goal orientation nomological net. *Journal of Applied Psychology, 92*(1), 128–150.
Pierce, L., Rees-Jones, A., & Blank, C. (2022). The negative consequences of loss-framed performance incentives, Discussion paper, NBER working paper 26619.
Pokorny, K. (2008). Pay but do not pay too much: An experimental study on the impact of incentives. *Journal of Economic Behavior & Organization, 66*(2), 251–264.
Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature, 37*(1), 7–63.
Ramirez, G., & Beilock, S. L. (2011). Writing about testing worries boosts exam performance in the classroom. *Science, 331*(6014), 211–213.
Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., & Webb, M. D. (2019). Fast and wild: Bootstrap inference in stata using boottest. *The Stata Journal, 19*(1), 4–60.
Russo, J. E., & Dosher, B. A. (1983). Strategies for multiattribute binary choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(4), 676–696.
Schlosser, A., Neeman, Z., & Attali, Y. (2019). Differential performance in high versus low stakes tests: Evidence from the GRE test. *The Economic Journal, 129*(623), 2916–2948.
Shelef, O., & Nguyen-Chyung, A. (2015). Selecting among high-powered incentives: Evidence from real estate agent careers, in *Academy of Management Proceedings*, 2015(1).
Shin, D., & Konrad, A. M. (2017). Causality between high-performance work systems and organizational performance. *Journal of Management, 43*(4), 973–997.
Takeuchi, R., Chen, G., & Lepak, D. P. (2009). Through the loooking galss of a social system: Cross-level effects of high-performance work systems on employees attitudes. *Personnel Psychology, 62*(1), 1–29.
Tarki, A., & Sanandaji , T. (2020). What top consulting firms get wrong about hiring. *Harvard Business Review.* https://hbr.org/2020/01/what-top-consulting-firms-gets-wrong-about-hiring.
Teeselink, B. K., van Loon, R. J. P., van den Assem, M. J., & van Dolder, D. (2020). Incentives, performance and choking in darts. *Journal of Economic Behavior & Organization, 169*, 38–52.
von Siemens, F. A. (2013). Intention-based reciprocity and the hidden costs of control. *Journal of Economic Behavior & Organization, 92*, 55–65.
Way, S. A. (2002). High performance work systems and intermediate indicators of firm performance within the US small business sector. *Journal of Management, 28*(6), 765–785.

Yu, R. (2015). Choking under pressure: The neuropsychological mechanisms of incentive-induced performance decrements. *Frontiers in Behavioral Neuroscience, 9*, 19.