J Econ Sci Assoc (2016) 2:1–12
DOI 10.1007/s40881-016-0025-7

ORIGINAL PAPER

The slider task: an example of restricted inference on incentive effects

Felipe A. Araujo[1] · **Erin Carbone**[2] · **Lynn Conell-Price**[3] · **Marli W. Dunietz**[1] · **Ania Jaroszewicz**[3] · **Rachel Landsman**[1] · **Diego Lamé**[1] · **Lise Vesterlund**[1] · **Stephanie W. Wang**[1] · **Alistair J. Wilson**[1]

Received: 30 October 2015 / Revised: 21 April 2016 / Accepted: 23 April 2016 /
Published online: 14 May 2016

Abstract Real-effort experiments are frequently used when examining a response to incentives. For a real-effort task to be well suited for such an exercise its measurable output must be sufficiently elastic over the incentives considered. The popular slider task in Gill and Prowse (Am Econ Rev 102(1):469–503, 2012) has been characterized as satisfying this requirement, and the task is increasingly used to investigate the response to incentives. However, a between-subject examination of the slider task's response to incentives has not been conducted. We provide such an examination with three different piece-rate incentives: half a cent, two cents, and eight cents per slider completed. We find only a small increase in performance: despite a 1500 % increase in the incentives, output only increases by 5 %. With such an inelastic response we caution that for typical experimental sample sizes and incentives the slider task is unlikely to demonstrate a meaningful and statistically significant performance response.

Keywords Real-effort · Slider task · Incentives

JEL Classification C90 · D01 · J30

This paper was written as a joint classroom project among faculty and students for a graduate course in experimental economics at the University of Pittsburgh.

✉ Lise Vesterlund
vester@pitt.edu

[1] Department of Economics, University of Pittsburgh, Pittsburgh, USA

[2] GSPIA, University of Pittsburgh, Pittsburgh, USA

[3] Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, USA

1 Introduction

Early economic experiments examining labor effort in the lab relied on the stated-effort design (for example Bull et al. 1987; Schotter and Weigelt 1992; Nalbantian and Schotter 1997; Fehr et al. 1993). Participants in the role of workers were given an endowment and asked to "purchase" a level of effort, which in turn benefited other participants in the role of principals. While stated-effort designs provided well-structured controls for participants' costs of effort and for the way in which that effort translated to output, the designs were seen as being abstract, overly distant from the types of labor effort the experiments were intended to capture. Scholars subsequently began to use real-effort designs, where participants are instead paid for performing an actual task in the lab.

Real-effort designs achieve less abstraction by trading off experimental control over the participants' effort costs and production function. This lack of control restricts the types of tasks that can be used to study a response to incentives. For example, take a simple decision-theoretic model of a real-effort task. In choosing her effort e between 0 and 1, participant i solves the following problem:

$$e_i^{\star}(w) = \arg\max_{e\in[0,1]} w \cdot f_i(e) - c_i(e),$$

where $w > 0$ denotes a piece-rate payment, while $f_i(e)$ and $c_i(e)$ represent the production and cost functions she brings into the lab. If we are to use a real-effort task to study the response to incentives in the laboratory, then it must be that the observed output $Y_i(w) = f_i\left(e_i^{\star}(w)\right)$ responds to the offered incentive w. That is, when comparing an increase in the piece-rate incentive from w_1 to w_2 we require an increase in the output so that $Y_i(w_2) - Y_i(w_1) > 0$.

As an extreme example, any task that individuals see as enjoyable will not produce a response to incentives. Suppose the cost of effort is strictly decreasing in effort (negative slope), then participants will exert full effort, and $e_i^{\star}(w) = 1$ for all $w \geq 0$. Tasks that are not enjoyable may, however, also be problematic. Persistent boundary solutions will result if the cost of maximal effort is too small, or if the task is so onerous that subjects disengage entirely. Furthermore, tasks will be unsuitable if the particular production and cost functions associated with the task lead to a perfectly inelastic response. For example, this will happen whenever the costs of the additional effort required to change output are large relative to the change in incentives.

Real-effort tasks for the laboratory must exhibit sufficiently elastic output at the offered incentives to secure responses that are not subordinate to noise and idiosyncratic variation in $c_i(\cdot)$ and $f_i(\cdot)$. To be well-powered at reasonable sample sizes, the task's average incentive effect should be large relative to observed variations at a fixed incentive level. An inelastic response will be seen when the production function is insensitive to the effort choice.

The experimental community has been quick to develop creative real-effort tasks. In considering easily implementable tasks that are short enough to be run repeatedly the "slider task"—which was introduced by Gill and Prowse (2012, hereafter abbreviated to G&P) to study disappointment aversion—has stood out. Participants are shown a screen with 48 sliders, where each slider has a range of positions from 0 to 100. Sliders are solved using the computer's mouse to move the slider's marker (initially placed at 0) to the midpoint of 50. Participants are given 2 min to solve as many sliders as possible, with the participant's chosen effort inferred by the number of sliders correctly positioned at 50 by the end of the 2 min. The task is normally repeated ten times and cumulative earnings across the entire experiment are given by $\sum_{t=1}^{10} w_t \cdot Y_{it}(w_t)$.

Initial evidence from the task indicated a positive and large response to incentives, and has led to the slider task being used frequently in papers measuring the incentive effects associated with various mechanisms and work environments. However, in contrast to the sensitivity to monetary incentives uncovered in the initial G&P study, more-recent slider-task studies find modest or non-existent treatment effects. Our paper's main result indicates that the slider task as currently operationalized has too inelastic a response to incentives to be recommended for future studies. The magnitude of the response uncovered with the slider task is negligible and lacking statistical significance. Our power calculations suggest recent null-results have a high likelihood of being type-II errors.

Where other studies have varied more complex elements of the payoff environment (strategic elements within a game, the nature of feedback, the frame, etc.) ours is a straighforward between-subject design, focused only on assessing whether the slider task's output responds to monetary incentives. In fact, we are the only paper to look at the slider task as a decision problem, with just monetary incentives varied between subjects so that experimenter-demand effects can not drive the response. Building on G&P's implementation of the slider task we conduct three treatments where we vary the piece-rate payment w that participants receive for each correctly positioned slider: a half cent at the low end, an intermediate two-cent treatment, and eight cents at the high end. This 16-fold increase (1500 %) in the piece rate corresponds to stark differences in participants' potential earnings, with maximum possible performance payments of $2.40, $9.60, and $38.40, respectively. However, despite substantial differences in the incentives offered, we uncover limited differences in average output: in order of increasing piece rates, we find that subjects complete 26.1, 26.6, and 27.3 sliders per 2-min round. This less than 5 % increase in response to a 1500 % increase in incentives is small both as an economic magnitude, but also relative to the size of learning effects and individual heterogeneity.

As a real-effort task, the slider task has many attractive characteristics. However, our paper shows that the task's output is too inelastic to be well suited for uncovering changes in output in response to changes in the incentives for standard laboratory samples sizes. This result implies similar caution is warranted for the inverse exercise for which the slider task has become frequently used. That is, a task that has an underpowered response will be as likely to produce type-II errors when used to detect changes in the underlying incentives through output.

2 Experimental design

Our experiments were conducted at the Pittsburgh Experimental Economics Laboratory, using subjects recruited from the student population, randomly assigned to one of three possible treatments.[1] Using a between-subject design, the piece rate is held constant throughout an experimental session so that each subject i receives a fixed payment per slider of $w_i \in \{0.5¢, 2.0¢, 8.0¢\}$.[2] After instructions on the nature of the task, each session began with a 2-min practice round for subjects to become familiar with the slider task. This was followed by ten paying rounds, each of which lasted 2 minute . In each round, subjects saw a single screen displaying 48 sliders of equal length and offset from one another, as per G&P.[3] At the end of each round there was a 10-s break during which subjects were reminded of how many rounds they had completed, the number of sliders completed (Y_{it}) and their corresponding earnings from that round $(w_i \cdot Y_{it})$.[4]

Once the ten paying rounds had concluded, subjects were asked to complete a survey.[5] Only after completing the survey were respondents informed of their total earnings for the session. Subjects were then privately paid, receiving a $10 participation payment on top of their earnings across the ten rounds $W_i = \sum_{t=1}^{10}(w_i \cdot Y_{it})$.[6]

In order to measure the extent to which the slider task responds to incentives, our paper's design adheres closely to that employed in G&P. There are four main differences: (1) The G&P design is within subject, where ours is between subject. (2) G&P examine a game between two randomly matched subjects competing over a variable prize; ours examines a decision problem, removing any externalities over payment. (3) The marginal incentives in G&P work through a probability of winning a prize, where each additional slider completed leads to a 1 % increase in the probability of winning a prize; in our experiment the marginal incentives work through a fixed piece rate per slider completed. (4) In G&P peer effects may be present, as subjects observe the other player's output at the end of each round; in our study there is no feedback on others' output levels.

[1] For consistency, one single member of the project read the instructions for all experimental sessions, and was assisted by another fixed experimenter. All data were collected over the course of 2 weeks in April of 2015, where all treatments were gender-balanced and interspersed across the data collection period. Initially, a total of three sessions were planned for each treatment; however, a computer error led to subjects' terminals freezing in one round in one session. Another session was, therefore, added to have three complete sessions for each treatment.

[2] The effective marginal incentives for a risk-neutral subject in G&P varied within a session between 0.15 and 6.2 ¢ with an average of 3.1 ¢.

[3] The experiment was programmed in z-Tree (Fischbacher 2007), and the program KeyTweak was used to disable all arrow keys on the keyboard, thereby ensuring that subjects only used the mouse to complete the slider tasks.

[4] Another difference between our design and the G&P design is that in their experiment subjects were given 2-min breaks while they waited for their opponent to complete the task.

[5] Data from the survey is available from the authors by request.

[6] Subjects in our 0.5 ¢ treatment had their final payoff W_i rounded up to the nearest whole cent.

Table 1 Results

(A) Summary statistics

Treatment	Output			N	Hourly rate
	Avg.	Min	Max		
0.5 ¢	26.1	6	44	42	$3.92
2 ¢	26.6	12	41	43	$15.95
8 ¢	27.3	10	46	63	$65.46
Total	26.7	10	46	148	

(B) Random-effect regressions

Estimate	Our data		G&P		G&P restricted	
	Linear	Log	Linear	Log	Linear	Log
Incentive effect, β	1.05 (0.65)	0.05 (0.03)	3.27[a] (0.75)	0.12 (0.04)	2.67[b] (0.65)	0.08 (0.02)
Initial output, $\hat{\eta}$	23.98 (0.48)	3.15 (0.02)	21.11 (0.89)	2.95 (0.06)	22.70 (0.72)	3.10 (0.03)
Learning effect, $\hat{\delta}_{10}$	4.34 (0.32)	0.17 (0.01)	4.35 (0.71)	0.19 (0.05)	4.24 (0.62)	0.16 (0.02)
Between SD, $\hat{\sigma}_u$ [c]	3.47 (0.29)	0.13 (0.01)	5.40 (0.68)	0.27 (0.06)	3.91 (0.33)	0.15 (0.01)
Within SD, $\hat{\sigma}_\epsilon$ [c]	2.77 (0.12)	0.12 (0.11)	3.87 (0.47)	0.29 (0.08)	3.22 (0.32)	0.13 (0.02)

Numbers in parentheses in Panel (B) are standard errors

[a] Marginal incentives for G&P first movers are calculated relative to our upper and lower incentive treatments. Because of this $\hat{\eta}$ has the interpretation of average output (average log of output) in round one at a 0.5 ¢ incentive in all regressions, and β has the interpretation as the estimated marginal effect of going from a 0.5 ¢ environment to an 8 ¢ environment in all regressions

[b] In G&P Restricted we excluded all of their participants whose performance was lower than our worst performing subject

[c] Standard errors for between and within standard deviations are drawn from a bootstrap of size 1000 that resamples across subjects, then subject-rounds

3 Results

Our experimental results are provided in Table 1. The first panel, Table 1(A), reports the average number of sliders completed per round, the minimum and maximum output, the total number of subjects N, and the effective average hourly wage rate (as the incentivized part lasts 20 min, this is simply $3 \cdot W_i$). On average, subjects across all of our experiments complete 26.7 sliders in each 2-min period. The lowest number of sliders solved by a subject in any round is ten, where the highest is 46 (two away from the 48 possible). Building on existing work we focus our analysis on the average number of sliders completed per round. Across treatments, we see that output increases with the piece rate: the average output is 26.1 for the lowest incentive of 0.5 ¢, somewhat higher at 26.6 for the middle incentive, and at its highest of 27.3 for the 8 ¢ incentive.[7]

Just from the averages in Table 1(A) it is apparent that the size of the incentive effect is small: going from a piece-rate of 0.5 ¢ to 2 ¢ leads to a 0.5 slider increase, and from 2 to 8 ¢ yields a 0.7 slider increase. Though the range of our incentives represents a substantial increase—from an effective hourly rate of about half the US federal minimum to just over $65 an hour[8]—this 1500 % increase in monetary incentives yields less that a 5 % increase in performance.

Across treatments and sessions, we observe substantial learning. Figure 1 presents the round averages for each of three treatments (where we have additionally provided bars indicating 95 % confidence intervals, given subject variation). In round one, the average output is 24.2 in both the 0.5 and 2 ¢ treatments, and 24.9 in the 8 ¢ treatment, though the variation across subjects is large. Across the session, output mostly increases so that the final output levels in round ten are 28.6 in the 0.5 ¢ treatments and 28.9 in both the 2 and 8 ¢ treatments. While the output in each treatment appears ordered according to incentives, it is noteworthy that the incentive order is only fully observed in six of the ten rounds.[9]

To quantify the effects from incentives while controlling for learning and subject-level variation, we run the following regression:

$$Y_{it} = \beta \cdot \left(\frac{w_i - 0.5}{8 - 0.5}\right) + \sum_{s=2}^{10} \delta_s \cdot 1_{s=t} + \eta + u_i + \epsilon_{it}, \tag{1}$$

where u_i is a subject-level random-effect, and ϵ_{it} an idiosyncratic error. The regressions include the treatment as a right-hand-side variable, rescaling the marginal incentive to run *linearly* from zero to one (0.5 ¢ at the low end, 8 ¢ at the high,

[7] If we drop the entire session with the programming error in the single round then the average output for the 8 ¢ treatment increases to 27.4.

[8] By way of comparison, the average lawyer makes an hourly wage of $64.17 according to the Bureau of Labor Statistics, while the average financial manager makes $62.61.

[9] Output in the high wage treatment appears to flatten out in the last few rounds more so than in the other two treatments. We can think of several reasons why this may be occurring. It is possible that higher wages might facilitate faster learning, or alternatively that they exert higher intial effort that produces subsequent fatigue. Alternatively this trend may be spurious. Unfortunately our design does not let us identify the cause of this trend. It may be of interest to more carefully study the cause of this difference in future work.

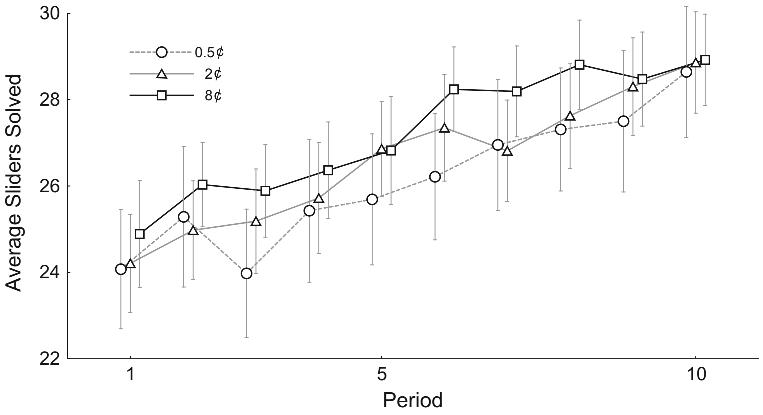

Fig. 1 Output across rounds

with the 2 ¢ marginal taking the intermediate value 0.2), and additionally adds nine period dummies as regressors, $\{\delta_t\}_{t=2}^{10}$, and a constant η. The first column in Table 1(B) reports the estimates for the incentive effect $\hat{\beta}$, the initial output level $\hat{\eta}$ at the beginning of the session, and the average amount of learning across the sessions $\hat{\delta}_{10}$. In addition, the table estimates the between-subject standard deviation, $\hat{\sigma}_u$, as 3.5 sliders, while the within-subject standard deviation, $\hat{\sigma}_\epsilon$, is estimated to be 2.8 sliders.

Unsurprisingly, given the overall averages in Table 1(A), the estimated value of β—where the coefficient represents the estimated marginal effect on sliders solved when moving from the 0.5 ¢ environment to the 8 ¢ environment—is close to one slider. Controlling for variation between and within subjects, as well as the across-session learning, the response to incentives is only marginally significant.[10] Interestingly, even our participants appear to be aware that their performance is not motivated by the payment they received. On the survey at the end of the experiment, we find that three-quarters of the participants do not think that there is any lower piece-rate payment at which they would decrease their performance.

Despite a 16-fold increase in the piece-rate, the increase in performance of only one slider is small relative to other variations within the task. In terms of heterogeneity in natural ability, a one-slider increase represents under a third of a between-subject standard deviation. In terms of idiosyncratic variation, it represents slightly over a third of a standard deviation. Across the entire session subjects seem to learn to complete more than four additional sliders, relative to their output in round one. So the observed incentive effect represents less than a quarter of the average learning effect.[11]

[10] Including attempted sliders in place of completed sliders, we find an incentive effect of 0.82 sliders ($p = 0.201$).

[11] Allowing rounds to enter into our estimating equation linearly and including a round-treatment interaction term, we fail to reject the null of no differences in learning effects between treatments ($p = 0.64$).

Table 2 Power: pairwise treatment comparisons

Treatment N	$0.005 to $0.02			$0.02 to $0.08			$0.005 to $0.08		
	p_{Crit}			p_{Crit}			p_{Crit}		
	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
20	0.150	0.083	0.020	0.238	0.150	0.047	0.383	0.270	0.104
30	0.170	0.099	0.026	0.300	0.199	0.070	0.496	0.373	0.171
40	0.191	0.114	0.033	0.360	0.249	0.097	0.597	0.472	0.243
50	0.212	0.130	0.039	0.418	0.300	0.126	0.681	0.561	0.320
60	0.234	0.146	0.045	0.471	0.349	0.156	0.748	0.639	0.397
70	0.254	0.163	0.054	0.520	0.395	0.190	0.802	0.706	0.470
80	0.275	0.178	0.062	0.569	0.442	0.223	0.847	0.763	0.540
90	0.294	0.194	0.069	0.610	0.486	0.257	0.882	0.809	0.604
100	0.314	0.208	0.078	0.651	0.528	0.293	0.910	0.848	0.661
150	0.406	0.290	0.121	0.801	0.702	0.467	0.978	0.955	0.861
200	0.491	0.367	0.170	0.891	0.821	0.618	0.995	0.988	0.950

We resample our subject-average data with replacement and for each pairwise treatment comparison regress the average subject output across the ten rounds on a dummy for the incentives. Figures indicate the fraction of null rejections where $p < p_{\text{Crit}}$ using a t-test from 100,000 simulations

The second column modifies the regression to use logarithms of the main dependent variable (log of completed sliders) and shifts the right-hand-side incentive variable to measure it in logs.[12] The interpretation of the β estimate in the log regressions is the percentage increase in output as we increase the incentives by 1500 %. Though marginally significant in a one-sided test ($p = 0.066$) the estimate of the incentive effect remains very low. Similar to the linear regressions, the 5 % estimate of the incentive effect is low relative to the 17 % increase attributable to learning, and to the 12–13 % effect from a within- or between-subject standard deviations.

Even if we disregard the small economic magnitudes and only focus on significance, the slider task is underpowered for uncovering a response to incentives with a typical experimental sample size. To demonstrate this, Table 2 provides power calculations for a response to incentives (see Slonim and Roth 1998 for a similar exercise). For each subject we generate their average output across the ten rounds of the experiment. Resampling N subject averages (with replacement) from each treatment we run 100,000 simulated regressions examining pairwise comparisons of our treatments—so the total counterfactual experiment sizes are $2N$. The figures in the table indicate the fraction of simulations where we reject the null of no response to incentives at the relevant p-value (using a t-test).

Fixing the required confidence level at 95 %, a fourfold increase from a half-cent to a two-cent incentive the experiment will lead to a type-II error approximately

[12] More exactly, the RHS incentive variable in our log regressions is rescaled and renormalized so that the incentive runs linearly from zero to one with the 2 ¢ marginal incentive taking the value of 0.5 (as our wage rates are 2^{-1}, 2^{1} and 2^{3}), where our linear regression had 2 ¢ representing just 20 % of the overall shift in incentives.

two-thirds of the time with 200 subjects per treatment, while at the same sample size the fourfold increase from a two-cent to eight-cent incentive will fail to reject one-fifth of the time. Turning to the most-extreme comparison, the 1500 % increase from a half-cent to an eight-cent incentive, the table indicates 90 total subjects per treatment are necessary to have eighty-percent power. Given this sample size requirement to reach 80 % power—and ignoring the fact that the incentive shift we are considering is economically very large—the overwhelming majority of slider-task experiments are underpowered.

4 Discussion

With an output elasticity of 0.025 (using the more conservative midpoint method of calculation) our between-subject design finds that the slider task is very inelastic. This finding is surprising given the initial G&P finding that output in the task was sensitive to incentives. We now examine how our results compare to G&P.

In G&P, two players i (a first mover) and j (a second mover) are randomly matched and compete to win a common prize of size $100 \cdot w_{it}$ cents, drawn randomly from an interval. The probability of player i winning the prize is given by $\frac{1}{100}\left(50 + Y_{it} - Y_{jt}\right)$, so for a risk-neutral participant the expected marginal incentive is w_{it}.[13] The sequencing of the game is such that the first mover's output (Y_{it}) is observed by the second mover j, and the second mover's response is the main focus in G&P. In looking at the response to incentives, we follow Gill and Prowse (2012) and look only at the first movers.

As noted earlier, the first mover's task in G&P is different from that in our study: (1) Their sessions have within-subject variation over the incentive w_{it} that may generate demand effects; (2) the tournament structure has own output inflicting a negative externality on the other player; (3) payment is incentivized only probabilistically; and (4) there is feedback on other participants' output levels. Changes in levels between G&P and our own study may come from any of these differences, and future research might help isolate each of these channels. However, it is still of interest to compare the magnitudes of the incentive effects.

Paralleling the regression results from our data in the first pair of columns in Table 1(B), the next two pairs of columns provide similar random-effects regressions from the G&P data. The first pair of G&P regressions provide results under the linear and log specification for the $N = 60$ first-movers.[14] The coefficient $\tilde{\beta}$ reflects the estimate from the G&P data for the incentive effect in our experiment, showing that the G&P data predict a significant 3.26 sliders increase as the marginal incentive is raised from 0.5 to 8 ¢. Our incentive estimate $\hat{\beta}$ from Table 1(B) is much smaller and is significantly different from the G&P level estimate ($p = 0.000$).

[13] The raw prizes in G&P are drawn uniformly over $\{£0.10, £0.20, \ldots, £3.90\}$. We transform these to expected marginal incentives for a risk-neutral agent, and then convert to US cents at a conversion rate of £0.65 = 100 ¢.

[14] To distinguish between estimates on our data and G&P's we will use the notation $\hat{\beta}$, $\hat{\eta}$, etc., for estimates from our data, and $\tilde{\beta}$, $\tilde{\eta}$, etc., for estimates from the G&P data.

The high incentive effect stems in part from a number of first-mover subjects who have very low output levels in the G&P data. There could be several reasons for producing low output. One possibility that exists in G&P but not in our study is that subjects might be trying to pick the efficient outcome (both exerting zero effort and equally splitting the chance to win the prize).[15] As a partial control for this, we re-run the same random-effects regressions excluding the G&P first-movers whose average output across the ten rounds is *lower than the lowest subject average* in our between-subject data (18.5 sliders, from the 0.5 ¢ treatment). This excludes six subjects, representing 10 % of the G&P first mover subjects.[16]

The regression results for the G&P subsample are given in the final pair of columns in Table 1(B). Though the estimated incentive effect is lower than the full sample—decreasing to 2.67 sliders—our estimate is still significantly different ($p = 0.012$). Moreover, despite the large differences in the estimated incentive effects, the other regression coefficients are remarkably similar.

Looking at the results in the linear specification with $N = 54$ (where we remove subjects in the left tail of the distribution), and comparing them to our results in the first column in Table 1(B), we find many commonalities. First, subjects on average increase performance across the session by approximately four sliders ($\hat{\delta}_{10}$ and $\tilde{\delta}_{10}$ are not significantly different).[17] Second, though the initial output level estimates of η are significantly higher in our sessions at 24 sliders in comparison to 22.7 in G&P, the size of the difference is quantitatively small.[18] Third, between- and within-subject standard deviations for output after controlling for the incentive effects (σ_u and σ_ϵ, respectively) are very similar, though in both cases the estimated variation in our experiments is smaller than in G&P.

Comparing our results to those of G&P, it is hard not to attribute the majority of the observed incentive effect to some combination of a within-subject effect (demand or peer effects) and a strategic or social effect (with the negative externality pushing subjects to exert low effort). While we leave it to future research to disentangle which of these factors are driving the additional incentive effects, it is clear that the incentive effect observed in our data can at best be described as marginal.

5 Conclusion

Using a between-subject design, we examine how performance in the slider task responds to changes in monetary incentives. Despite a 1500 % increase in incentives we find only a 5 % increase in output. With such an inelastic output

[15] Gill and Prowse (2012) note that 2 subjects (whom we will also exclude) appear to have difficulty positioning sliders at exactly 50 until a few rounds into the session.

[16] Note that only subjects with low *average* performance are eliminated from the data. Data from subjects with particular rounds with less than 19 sliders completed are still included in the analysis, provided that the subject's average across the session is above 18.5 sliders.

[17] All three of our treatments, as well as both movers in G&P show fairly consistent increases in average output across the session.

[18] A joint regression across both sets of data indicates no significant difference over the two constants ($p = 0.123$).

response we argue that the slider task is poorly suited for studying the response to incentives.

While our experiment documents the insensitivity of the slider task to incentives, it does not allow us to identify why a limited response is seen. Certainly the high level of performance is not consistent with the insensitivity resulting from individuals not being incentivized to perform. What is unclear is why they did not respond to the changes in incentives. If individuals are intrinsically motivated to exert full effort absent incentives, the insensitivity may result from a ceiling effect. The satisfaction of securing a high performance may on its own motivate individuals to exert maximal effort. An alternative explanation is that while incentives change effort levels, the tasks production function is insensitive to such changes.

Though we cannot identify what causes the insensitivity, our own experience in performing the task has provided some insights. In preparing the study we (all ten authors) took part in an un-incentivized ten-round trial run of the experiment. Reflecting on that experience, each one of us tried our hardest, with the aim of beating our personal best. The desire to beat previous outcomes dominated the effort costs of concentrating on the task. This was particularly true as the alternative to solving sliders was to do nothing at all. Eager to increase performance we were frustrated by our inability to increase output. While we each tried to increase effort, it did not result in increased output. Of course it is not possible to determine whether this stemmed from a ceiling effect or from the production function's insensitivity to effort.

A more sensitive output response may be expected if we were to extend the duration of the task. While sustaining concentration and maximum effort is not too costly for a short period of time, such concentration may become more costly when the task lasts for a longer period of time. A more elastic response may also be seen if there was an alternative activity to not performing on the task.

Three recent studies point to techniques that might offer more-constructive results for real-effort tasks in the lab. Gächter et al. (2015) introduce a ball-catching task where the cost of effort is directly manipulated by the experimenter. With suitable parameterizations, interior solutions can, therefore, be ensured. Less directly, Corgnet et al. (2014) and Eckartz (2014) examine a variety of real-effort tasks and find that the presence of outside leisure activities and paid outside options, respectively, lead to stronger incentive effects. These different approaches—the one with greater experimental control, the other with greater flexibility extended to subjects—suggest possible solutions for researchers wishing to use the slider task in the lab.

While there are several reasons that the incentive effect might be larger in the G&P data, our paper motivates future research on the potential greater sensitivity in within-subject designs.[19] One explanation for stronger results in within-subject designs is that they allow for better controls for the large variation in individual-level ability of the slider task.[20] An alternative, but undesirable explanation, is that

[19] In mirroring responses to incentives in labor markets one may wish to think of within-subject designs as capturing short-term effects and between-subject designs as capturing long-run responses.

[20] If this is the main channel, one way to reduce noise from individual heterogeneity is to measure baseline ability via a common task with a fixed incentive level at the start of each treatment, à la Lilley and Slonim (2014), with subsequent tasks chosen with the desired between-subject variation.

the additional response is an experimenter-demand effect. Future research is needed to identify the cause of these differences.

Whatever the cause, a reasonable criterion when using *any* real-effort task to study the incentives is a demonstrated response to explicit monetary incentives between subject. Statistical significance aside, desirable tasks should be able to demonstrate an incentive effect which is large relative to uncontrolled variation within the task (individual ability, learning, etc.). With respect to this above criterion, our paper sounds a cautious note for the slider task. While the task has many appealing properties its highly inelastic response makes it a poor candidate for uncovering a measurable and statistically significant response with typical experimental sample sizes and incentives.

References

Bull, C., Schotter, A., Weigelt, K. (1987). Tournaments and piece rates: An experimental study. *Journal of Political Economy*, *95*, 1–33.

Corgnet, B., Hernán-González, R, Schniter, E. (2014). Why real leisure really matters: Incentive effects on real effort in the laboratory. *Experimental Economics*, *18*, 284–301.

Eckartz, K., Task enjoyment and opportunity costs in the lab: The effect of financial incentives on performance in real effort tasks, 2014. Jena Economic Research Papers, 2014-005.

Fehr, E., Kirchsteiger, G., Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics*, *108*, 437–459.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, *10*(2), 171–178.

Gächter, S., Huang, L., & Sefton, M. (2015). Combining "real effort" with induced effort costs: the ball-catching task. *Experimental Economics*. doi:10.1007/s10683-015-9465-9

Gill, D., & Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, *102*(1), 469–503.

Lilley, A., & Slonim, R. (2014). The price of warm glow. *Journal of Public Economics*, *114*, 58–74.

Nalbantian, H., & Schotter, A. (1997). Productivity under group incentives: An experimental study. *American Economic Review*, *87*(3), 314–341.

Schotter, A., & Weigelt, K. (1992). Asymmetric tournaments, equal opportunity laws, and affirmative action: Some experimental results. *Quarterly Journal of Economics*, *107*(2), 511–539.

Slonim, R., & Roth, A.E. (1998) Learning in high stakes ultimatum games: An experiment in the Slovak Republic. *Econometrica*, *63*(3), 569–596.