An economic approach to test-taking[*]

by Lex Borghans, Huub Meijers, Benedikt Vogt,[†] and Bas ter Weel

June 2018

Abstract

When taking a test, individuals often face a trade-off between time investment and time pressure. Therefore test scores will depend on cognitive ability and how people deal with the time pressure. We set up a laboratory experiment in which we vary incentives and track what people would have answered before and after they submit their answer. Our results show that the probability of finding the right answer is an increasing function of the time invested and the functional form does not change if we vary incentives. Individuals respond to incentives by adjusting time investments in accordance with maximizing output.

Key words: Choice process; time pressure; monetary incentives; cognitive test

JEL codes: D03; J24

[*] We would like to thank Andrew Caplin, Thomas Dohmen, Robert Dur, Eva Feron, Andries de Grip, Franziska Tausch, Paul Schempp, Trudie Schils, Martin Strobel, Dinand Webbink, and Ulf Zölitz, as well as seminar participants in Aarhus, Brussels, Maastricht, and Reus for helpful comments. We would like to thank Eva Feron, Franziska Tausch, and Ulf Zölitz further for their help with conducting the experiment.
The programs used to conduct the experiment, as well as the data and code used to derive the results presented in this paper, are available upon request.

[†] Corresponding author is Benedikt Vogt. Borghans: Department of Economics and ROA, Maastricht University; lex.borghans@maastrichtuniversity.nl. Meijers: Department of Economics and UNU-MERIT, Maastricht University; huub.meijers@maastrichtuniversity.nl. Vogt: CPB Netherlands Bureau for Economic Policy Analysis b.vogt@cpb.nl. Weel: SEO Amsterdam Economics and Department of Economics, University of Amsterdam; b.terweel@seo.nl.

1. Introduction

Testing is an important element at school, college, and in the workplace. People are sorted and selected based on tests that aim to measure their abilities. The deliberation process when taking a test can be viewed as an economic problem. Under the time pressure of a test, people face a trade-off between time investment, which improves the probability of finding a solution, and the cost of time. As a consequence, how people respond to these incentives determines their performance on a test. To investigate this trade-off between time investment and costs, information is needed about which solutions people would have come up with at different points in time.

The aim of this study is to investigate how individuals respond to changes in incentives when they take tests. In particular, this research answers the following four questions: What is the relationship between time investment and the probability of solving a test? Does this production function shift if incentives change? Do people adjust their time investment if incentives change? To what extent is the adjustment optimal?

We developed a novel experimental approach to study the deliberation process and performance on test-taking in a laboratory. We provided subjects with two independent monetary incentives. First, the decision to commit to a solution was incentivized by rewarding early submissions more than late submissions. By varying the incentives for providing the right answer under time pressure, we were able to identify how respondents change their strategy. Second, we rewarded subjects for the number of seconds they took to pick the right solution during the deliberation process, which took 60 seconds at most. This way, subjects revealed their immediate solution, after which they had time to rethink and adjust their solution as many times as they wished. This allowed us to monitor their behavior and problem-solving ability. The reward per second was not varied during the experiment. The information that we acquired with the second incentive scheme allowed us to identify what a respondent would have earned under counterfactual strategies. In this way, we can answer the four aforementioned research questions about the shape of the production function, whether time pressure changes the production function, how time-pressure affects the strategy of people, and to what extent these adjustments are optimal.

Subjects had to solve 45 Raven matrices, which were randomly assigned to different treatments.[1] To check whether Raven matrices lead to specific outcomes that are not generalizable, we replicated the approach using a test in which participants had to solve numerical problems.[2] Monetary incentives and time pressure were varied across three different treatments.[3]

The results can be summarized as follows. First, the probability of finding the right answer is an increasing function of the time invested.[4] Second, the relationship and functional form do not significantly change if we increase time pressure or the level of incentives. Third, people respond to incentives by adjusting time investments in accordance with maximizing output. Fourth, however, these changes are small. If we triple the average reward for a correct solution, subjects invest 60 percent too little time compared with optimal behavior. These results are robust to different subsamples and to taking into account possible effects of aggregation bias. We find that smarter people are most effective at adapting to a changing context (monetary incentives and time pressure) and thus are more likely to maximize output. When we investigate subsamples, we observe that those subjects who we define as high performers are also best at making efficient choices.

With this study, we contribute to three strands of literature in economics and psychology. The first strand is the literature in economics that aims at measuring deliberation processes in economic decision making. Agranov et al. (2015), for instance, make use of a probabilistic

[1] Raven matrices are a well-established measure of problem solving ability (e.g., Carpenter et al. 1990). We make use of these matrices because they are the best available proxy for problem solving, independent of culture and educational background. The design of the test is such that subjects have to choose from a set of eight alternatives, of which one is the right solution to solving a 3×3 matrix of figures. For each problem, a limited amount of time is available to come up with a solution. We discuss the more salient details of these matrices below.

[2] These numerical problems are similar to the ones used by Caplin et al. (2011) to examine decision-making processes in a search-theoretic choice experiment. Our application is different in the sense that we are interested in success rates and monetary rewards during a test and not in the distinction between optimal and satisficing behavior during search processes.

[3] Borghans et al. (2008b) and Borghans et al. (2013) use a similar strategy to provide incentives. Time pressure is measured in a different way in this study compared with our previous work, as we will discuss below in more detail. There are numerous studies, both in psychology and economics, which find that incentives affect cognitive test scores (e.g., Duckworth et al. 2011; Edlund 1972; Lloyd & Zylla 1988; Segal 2012). Almlund et al. (2011) provide an excellent overview on the results of these studies.

[4] The finding that people respond to incentives seems to be consistent with recent work that reports positive short-term effects of financial incentives on achievement test scores and graduation rates (e.g., Rodriguez-Planas 2012). Angrist and Lavy (2009) and Angrist et al. (2009) only find effects for girls, while other studies find no effects (Fryer 2011) or only effects on math test scores (Bettinger, 2011). In addition, Borghans et al. (2008b), Duckworth et al. (2011), and Duckworth and Seligman (2005) document that certain personality traits seem to explain a substantial part of differences in performance on a cognitive test.

payment scheme to measure the deliberation process in guessing games. The study that comes closest to this paper is that of Caplin et al. (2011), who make use of a probabilistic payment scheme to test whether search behavior can be explained with utility maximizing or is more in line with satisficing behavior. Their key finding is that individual search behavior can be best explained by a model of satisficing (Simon, 1955). Their subjects stop searching when they reach a satisficing level of utility rather than when they maximize utility.[5] Our contribution to this literature is threefold. First, we aim at observing choice process data before *and* after an individual made a decision. Second, we investigate whether the choice process and behavior changes when we vary the stakes and the time pressure for coming up with a correct solution. Third, we can investigate if behavior is optimal in terms of expected payoff maximization.

The second strand is the literature that investigates decision-making under time pressure. Kocher and Sutter (2006), for instance, investigate decision-making under time pressure in an experimental beauty-contest game. They find that time-dependent payoffs lead to faster decisions without a loss of quality. Our findings are in line with their findings, since incentives affect behavior but do not change the number of correctly solved matrices. Lindner and Sutter (2013) investigate level-k reasoning under time pressure in a laboratory experiment. Their findings suggest that choices converge to equilibrium behavior compared with a situation in which there is no time pressure.[6] We see our evidence as complementary to their work since we are interested in the effects of time pressure on productivity and behavior.

The third strand is the literature in economics and psychology that investigates the measurement of skills and the interaction of cognitive ability with economic decision making (see Rustichini (2015) for a recent summary). Gill and Prowse (2016), for instance, find a positive relationship between level-k reasoning and cognitive ability. Higher ability types are more likely to play payoff maximizing strategies in a multi-player environment. Hence, individuals with higher cognitive abilities are also more likely to act more in a way economic theory predicts. Several papers in economics and psychology show that performance is a function of incentives, preferences, personality traits, and cognitive ability. This can be

[5] Similar studies were conducted by Gabaix et al. (2006), Manzini and Mariotti (2007), and Reutskaja et al. (2011), who make use of choice-process data to test consumer choice models.

[6] Kocher et al. (2013) also provide evidence that subjects tend to become more loss averse and more gain seeking when they are set under time pressure.

performance on a test, but also in different types of problem solving, in school, the labor market, etc. (see e.g. Almlund et al. (2011), Borghans et al. (2008), and Kautz et al. (2014) for reviews of the progress in this field of research). We contribute to this literature in two ways. To the best of our knowledge, we are the first to plot successfully a production function of a cognitive test. Second, we show that the outcome of a test depends on the behavior during this test, and also the individual's ability to adapt to the test environment. This means that cognitive tests always measure at least two things: the cognitive ability and an individual's ability to adapt to the test environment.

The setup of this paper is as follows. Section 2 presents the theoretical background and our translation to a laboratory experiment. Section 3 presents our main results. Section 4 analyzes why behavior is relatively inelastic and to what extent behavior is optimal. Section 5 reports on the robustness checks, and Section 6 concludes the paper.

2. Framework

2.1. Theoretical Background

We model the behavior of a risk-neutral agent who has to find the answer to a question.[7] More skilled agents have a higher probability to find the solution immediately at all levels of complexity. In addition, a more complex problem reduces the quality of the immediate choice for all subjects. Finally, we assume that subjects only maximize their income. This is because in our experimental setup, monetary incentives are very salient. Moreover, the decrease in monetary rewards when answering later is so high that it is unlikely that the joy of thinking (i.e. intrinsic motivation) about an answer would substantially change the optimal timing.

The search phase is costly because of the disutility of time or the cost of providing effort. Agents face the option of staying with their immediate choice or searching for a possible better solution. Searching for the right answer has two possible outcomes. First, agents could find out that the immediate choice is the right choice. In this case, they will not change their choice, but they do incur costs. Second, agents find a better choice and switch to this alternative. The search process continues until they run out of time or until the disutility of time becomes larger than the probability of finding a better alternative. This is equivalent to

[7] The qualitative predictions do not change if our agent is risk averse, or risk neutral.

the approaches developed in Caplin and Dean (2011), Gabaix et al. (2006), and Simon (1955) for various types of choice processes.

Finally, economic agents have an incentive to limit the length of the search process and reveal their final choice as soon as possible. The reason for this is that there exists a trade-off between searching and the disutility of time. The probability of improving the outcome is decreasing with contemplation time, while the utility constantly decreases with contemplation time.[8]

The utility function of agent i solving problem q in context τ can be written as follows:

$$U_{i,q,\tau}(t,\gamma,F) = p_{i,q,\tau}(t)(F_\tau - \gamma_\tau t) \tag{1}$$

In equation (1), $p_{q,i,\tau}(t)$ is the probability function of knowing the solution at point t, which is measured in seconds. F_τ is a fixed payment and γ_τ the amount deducted per second. We assume that the probability function is upward sloping with diminishing returns: $\frac{\partial p_{q,i,\tau}(t)}{\partial t}$>0 and $\frac{\partial^2 p_{q,i,\tau}(t)}{\partial t^2}$<0.

Agent i optimizes his time, after which he submits a solution in each respective context. Maximizing (1) with respect to t yields

$$\frac{\partial U_{i,q,\tau}(t)}{\partial t} = \frac{\partial\, p_{i,q,\tau}(t)}{\partial t}F_\tau - \frac{\partial p_{i,q,\tau}(t)}{\partial t}\gamma_\tau t - p_{i,q,\tau}(t)\gamma_\tau = 0 \tag{2}$$

To predict differences in behavior between different levels of incentives and time pressure, we present comparative statics with respect to the time pressure parameter γ_τ and the incentive parameter F_τ. We apply the envelope theorem and derive from (2) our first prediction:

[8] More formally, the probability function of individual i for having the correct solution on problem q in incentive environment τ has the following properties: $0 \leq p_{q,i,\tau}(t) \leq 1$, $\partial p_{q,i,\tau}(t)/\,\partial \mathrm{t} > 0$, $\partial^2 p_{q,i,\tau}(t)/\partial^2 t < 0$ for $\forall\, t > 0$. The variable $t \,\epsilon\, (0,+\infty]$ is the time investment made to solve a matrix.

Prediction 1: Optimal time investment increases with greater incentives:

$$\frac{\partial t^{*}(F_{\tau},p_{i,q,\tau},\gamma_{\tau})}{F_{\tau}}=\frac{-\frac{\partial p_{i,q,\tau}}{\partial t}}{\frac{\partial^{2}p_{i,q,\tau}}{\partial t^{2}}(F_{\tau}-\gamma_{\tau}t)-\frac{\partial p_{i,q,\tau}}{\partial t}2\gamma_{\tau}}>0 \qquad (3)$$

This implies that, in order to behave optimally, subjects should finalize their solutions to the problems later when incentives are higher. In a similar way, we arrive at our second prediction.

Prediction 2: Optimal time investment decreases with increasing time pressure:

$$\frac{\partial t^{*}\left(F_{\tau},p_{i,q,\tau},\gamma_{\tau}\right)}{\gamma_{\tau}}=\frac{-\frac{\partial p_{i,q,\tau}}{\partial t}t-p_{i,q,\tau}}{-\frac{\partial^{2}p_{i,q,\tau}}{\partial t^{2}}(F_{\tau}-\gamma_{\tau})+\frac{\partial p_{i,q,\tau}}{\partial t}2\gamma_{\tau}}<0 \qquad (4)$$

This implies that, in order to behave optimally, agents should finalize their solutions to the problems earlier when they face higher time pressure.

2.2. Experimental Design

We conducted an experiment at the Behavioural and Experimental Economics Laboratory (BEElab) at the School of Business and Economics of Maastricht University. The experiment was conducted using the software package z-Tree (Fischbacher, 2007). We used the recruiting software ORSEE to recruit subjects from the student population at Maastricht University (Greiner, 2003), and conducted the analysis on a set of *n*=128.[9]

2.2.1. Tasks

To measure the process of answering questions, we selected Raven matrices (Raven, 1962), which measure fluid intelligence and have been used to measure skills in psychology (e.g. Carpenter et al. 1990; Roberts et al. 2005).[10] The results of these tests have also been applied

[9] Online Appendix A provides an overview of the recruitment procedure and presents background information about the subjects. Appendix A also provides detailed information about the instructions, the intensive trial phase to make subjects understand the matrices and numerical problems, and the setup of the experiment.

[10] Fluid intelligence reflects the ability to learn independent of the actual knowledge and cultural background of the individual. Carpenter et al. (1990) document that Raven matrices measure the skill to encode and induce regularities in problems. In addition, well-performing test takers are able to “induce abstract relations and [exhibit] the ability to dynamically manage a large set of problem solving goals in working memory” (p. 404). In addition, there is a distinction between two types of cognitive processes: those executed quickly with little

in the economic literature as measures of skill (Almlund et al. 2011; Kautz et al. 2014). Subjects had to select the correct missing figure, which completes the sequence of nine consecutive figures designed as a 3×3 matrix. They needed to select the missing figure from a set of eight figures. All figures were connected in a logical manner and no prior knowledge was needed in order to be able to solve these matrices. We used a set of 45 Raven matrices. All subjects had to solve the matrices in three sessions of 15 problems. The order in which subjects had to solve the matrices was randomized. Panel A of Figure 1 presents an example of a Raven matrix.

Figure 1 about here

To ensure that the results were not mainly driven by the type of problem, we developed a second type of task. Subjects had to solve numerical problems in which they faced a set of eight different addition and subtraction problems. Each of these problems consisted of three terms and each of the terms consisted of a number between zero and one hundred. The number was either depicted in Arabic numerals or written out in words on the screen. Subjects had to find the solution, which summed up to the highest amount in the set of eight problems. Panel B of Figure 1 shows an example of a numerical problem.[11]

Before subjects could start the experiment, they had to go through a trial phase, which made them familiar with the problems and the payment schemes. In the trial phase, subjects had to solve five easy problems. While solving the first two problems, we introduced payment schemes. During the last three problems, we introduced the payment schemes simultaneously. After each problem in the trial phase, subjects had to calculate their payoff and could only continue if they passed this test.[12]

conscious deliberation and those that are slower and more reflective. Raven matrices refer to the former (e.g. Epstein 1994).

[11] These numerical problems are similar to the ones used by Caplin et al. (2011) to study consumer-choice behavior. We have developed an equivalent of their numerical problems, with a maximum complexity level equal to their level 3. In our experiment, subjects only had 60 seconds to solve the numerical problems.

[12] Online Appendix A.2 presents more information about the Raven matrices and the numerical problems and their rank order. Table B.1 in the online Appendix shows the descriptive statistics of the number of tries in the trial phase for each of the five problems. We can learn two lessons from the data. The first one is that the median number of tries is one in any of the trial problems. Secondly, are more detailed look at the data shows that in the last trial phase more than 80% of our subjects immediately gave the correct solution to the trial

2.2.2. Monitoring the Decision-Making Process

We study the decision-making process and performance of subjects on each Raven matrix and numerical problem separately.[13] Figure 2 shows a screenshot of the decision screen for a typical Raven matrix and for a numerical task. The problem set was displayed in the middle of the screen. Each problem had a time limit of 60 seconds, which was not varied during the experiment.[14] The remaining time was indicated with the green bar on the left part of the screen. Subjects had to wait until the time elapsed to proceed to the next problem.

Figure 2 about here

We paid subjects depending on the time it took for them to select the right solution during the deliberation process, which took 60 seconds at most. This payment scheme yielded a monetary reward of 0.5 cents for every second the correct solution was selected during the 60-second time period. Hence, the initial choice that rational subjects tended to make before solving the problem was to choose an immediate random first solution.[15] If a subject picked the correct solution immediately and did not change the selection, he or she earned 30 cents from the payment scheme. The information about the payment scheme was displayed on the right part of the screen. We call this the blue payment scheme.

The second role of the blue system is that it allowed us to monitor the decision-making process for each problem. After the initial choice, subjects could change their solution as

problem and all but one subject came up with the correct solution after the second try. We see this as convincing evidence that the trial problems made people understand our experimental design.

[13] This is different from how a usual cognitive test is conducted and evaluated. However, we are not primarily interested in the measurement of cognitive ability, but in the decision-making process during a problem-solving task.

[14] We restricted the time for each problem to 60 seconds because in a previous experiment we found that the average time spent on a problem in a cognitive test is 49 (st. dev. 4) seconds (e.g., Borghans et al. 2008b).

[15] To induce an immediate choice, there was a popup message if subjects did not select a solution within the first five seconds. The payment scheme in the blue system is the same in all treatments and could yield a maximum payoff of €45 (90 problems). This incentive scheme is consistent with the scheme used by Caplin et al. (2011). They used a probabilistic incentive scheme to reveal the actual preferred choice of an individual. More specifically, a random point in time during solving a problem was selected and the option chosen at that moment was evaluated and rewarded. We implemented a non-probabilistic payment version of their approach.

often as they liked. This allowed us to identify how the provisional choices evolved with contemplation time. It also allowed us to investigate time investments to solve the problem.[16]

At the same time, a second payment scheme was active. Subjects were incentivized to submit their solutions as soon as possible by pressing the "submit" button. The purpose of the payment scheme was to induce time pressure and provide incentives to solve the problem as soon as possible. In the baseline treatment, we linearly decreased the amount of money for a correctly submitted solution from 25 to 5 cents. Subjects were able to stop the decline by pressing the button. The actual payment was displayed on the right side of the screen. Note that after pressing the submit button, the blue system was still active. This means that after submitting the solution, subjects still had an incentive to keep on thinking about their submission because they received 0.5 cents per second for selecting the correct solution. To help the respondents to distinguish the payment schemes, we called this the red payment scheme.

2.2.3. Changing Incentives and Time Pressure

Next to the baseline treatment, we varied the context by changing both incentives and time pressure. First, we changed the level of the incentives, but not the slope, in the red system, meaning that it ran from 55 to 35 cents during the 60-second time period in which the Raven matrix had to be solved. We call this the HL treatment, which stands for High incentives and Low time pressure. The comparison of this treatment to the baseline treatment (Low stakes and Low time pressure (LL)) yields information about the effect of incentives on decision-making processes. From an economic point of view, we expect choice behavior to be different across the two treatments. Since in HL the payment for a solution is higher at any point in time, subjects should submit their final choice later (cf. the first prediction of the model).

Second, we also changed the slope of the incentives to change the context and increase time pressure. This HH treatment (High stakes and High time pressure) ran from 55 to 5 cents. We expected subjects to submit their solution earlier compared with HL, because time pressure

[16] This way of approaching decision-making is closely related to choice data gathered for understanding consumer search under time pressure (e.g., Caplin et al. 2011, Reutskaja et al. 2011). It differs from the approaches in consumer choice experiments because in our setting the solution is either right or wrong. By contrast, consumers can decide to stop searching for alternatives when they have reached a satisficing level of utility (e.g., Simon 1955, Stigler 1961). In our case this would be equivalent to stop changing the solution.

was higher and incentives were lower at any point in time (cf. the second prediction of the model).

Note that the model is ambiguous about the effects of changing time pressure *and* incentives at the same time. This is related to the comparison of the LL treatment with the HH treatment. Time pressure can decrease time investments, whereas greater incentives can increase optimal time investment.

3. Results

In this section, we present our main results. First, we analyze the relationship between time investment and the probability of solving a problem correctly (the production function). Second, we analyze whether this probability alters if we change the incentives. Third, we analyze whether individuals adjust their time investment if incentives change. Fourth, we analyze to what extent this adjustment is optimal.

3.1. Production Function and Time Investment

3.1.1. Production Function

Table 1 shows the success rates in LL. Panel A reports the average probability of solving the problem over time in terms of the fraction of solutions, and Panel B reports the average cumulative earnings in the blue system. We document the probabilities and earnings for both types of problems. The Raven matrices are split into three levels of complexity.[17] The unit of observation in Table 1 is the matrix or the numerical problem. The total number of observations equals $1{,}920$, which is the result of 128 subjects solving 15 Raven matrices or numerical problems.[18] At six points we measure the probability and earnings up to that point in time. That is, at time $t = 10$ we document probabilities and average earnings from the first to the tenth second.

[17] We define the degree of difficulty by the number of the respective Raven matrices in the test manual. This provides us with an exogenous definition of difficulty. In the actual intelligence test, the Raven matrices are ordered according to their degree of difficulty. They start with the easiest and end up with the most difficult. Hence a low number in the actual test manual indicates an easy matrix. We ended up with 15 easy problems, 16 moderate problems, and 14 difficult problems. More information on the items can be found in Appendix A.2.

[18] We experienced minor computer problems for seven questions. As such, we conducted the whole analysis excluding these questions. The results did not change.

Table 1 about here

Panel A shows that the probability of solving a problem and earnings rise over time. The overall performance suggests that about half (0.519) of the matrices were solved correctly within 60 seconds. The rate of improvement is highest in the first 30 seconds (262.2 percent improvement). There are differences in performance across different levels of complexity. For relatively easy matrices, the probability equals about 70 percent (0.722), whereas the probability is only 28 percent for the most difficult matrices. Also, in terms of improvement, there exists heterogeneity across different levels of difficulty. In the last 30 seconds the rates of improvement are 12.1, 12.7, and 7.6 percent for the easy, medium, and difficult matrices, respectively.

Statistically significant differences between consecutive points in time and within columns are denoted by an asterisk ($*$). Overall, the probabilities are different across ten-second time intervals. The only exception is the increase in the probability in the final ten seconds. Across different levels of complexity this pattern is confirmed. For difficult matrices, the probability does not seem to differ statistically from the 30th second onwards. Statistical differences between complexity levels are denoted by a plus ($+$). We observe statistically significant differences of moderate and difficult Raven matrices relative to the easy matrices and relative to one another. We observe that the probabilities are always different at the one percent level when comparing different levels of complexity.

The average cumulative earnings (Panel B) reveal a somewhat convex pattern that differs across levels of complexity. This pattern is consistent with the concave pattern in Panel A because after a certain point there is not much improvement in performance. Overall, the earnings for solving Raven matrices equal 10.7 cents, which is about a third of the maximum payoff of 30 cents.[19] Payoffs are higher for easy matrices. The patterns of statistically significant differences confirm the findings of Panel A.

The final column in Panels A and B shows the performance when solving numerical problems. The pattern in this column is comparable to the easy Raven matrices. In the end,

[19] This maximum payoff implies that subjects choose the right solution at the very first second and never change their choice during the 60-second time period.

the probability of finding the solution equals 75.5 percent, with average cumulative earnings of 16.7 cents. Statistically significant differences are obtained between different time intervals.

3.1.2. Time Investment

Table 2 presents the performance in the red system in three panels. Panel A reports the cumulative fraction of submitted solutions at different points in time during the 60-second time period in which a problem has to be solved. Panel B reports the cumulative fraction of correctly submitted solutions at the same points in time. Finally, Panel C reports the average earnings of solutions that were correct and submitted within the different intervals.

Table 2 about here

Panel A shows that solutions to the majority of Raven matrices were submitted after 60 seconds (96 percent). Subjects failed more often to submit their solutions to the most difficult matrices. Solutions to almost all numerical problems have been submitted (99.3 percent). The pattern suggests that on average half of the solutions to the Raven matrices have been submitted after 30 seconds. Only for the most difficult matrices was this number lower after 30 seconds (38.8 percent). In terms of statistically significant differences, we observe that all comparisons are different at the one percent level.

The numbers in Panel B of Table 2 suggest that correct submissions increase with time and that the pattern is concave. This is generally true for all types of problems. Again, most differences are statistically significant, with the exception of the last row in Panel B. The number of correctly submitted solutions does not seem to differ from the 50th to the final second.

The ratio between the fraction of correct submissions and the fraction of all submissions in Panel B and Panel A provides information about the success rate of the submitted answers. For all types, the ratio peaks after 30 seconds and declines afterwards. This suggests that those who submit their solutions between the 20th and 30th second are more likely to submit the right solution relative to those submitting earlier or later.

Finally, Panel C of Table 2 documents average earnings in cents within different time intervals of ten seconds (note that these earnings are different from the ones in Panel B in

Table 1, where we document cumulative earnings from the blue system). The low earnings up to the 10th second are the result of a low number of submissions and a low rate of correct submissions. In the second interval of ten seconds, average earnings from the red system are highest for all types of problems. This suggests that earnings peak earlier than the success rate. In terms of statistically significant differences, we observe that all comparisons are different at the one percent level.

3.2. The Impact of Incentives and Time Pressure on the Production Function

Figure 3 shows the probability of solving the problem over the 60 seconds for all three treatments: LL, HH, and HL. The figure is a graphical equivalent to Panel A of Table 1 for all different treatments. Panels A and B document the probability of solving the problem over time for the Raven matrices and numerical tasks. We construct these functions by taking the mean of the correctly selected solutions over all problems and subjects in a respective treatment at each second. The grey areas indicate the 95 percent confidence intervals.

Figure 3 about here

In Panel A we show the curve for the baseline treatment and the curves of the probabilities of HL and HH of the Raven matrices. The figure indicates that the confidence intervals of the HL and HH overlap with each other and with the confidence interval of LL. These graphs show that the speed of solving a problem does not vary with our variation of monetary stakes and time pressure in the red system. We observe the same pattern for the different treatments in the numerical tasks in Panel B. All confidence intervals overlap, which suggests that the speed of thinking does not differ across the different incentive environments.

A comparison between Panel A and Panel B shows that Raven matrices are on average more difficult to solve than numerical problems. The probability over time increases faster and reaches a higher level after 60 seconds for the numerical tasks compared with the Raven matrices. The finding that the probability of solving a problem does not differ between the three different treatments suggests that the adaptation to incentives seems to stem from a change in submission behavior.

3.3. The Impact of Incentives and Time Pressure on Time Investment

In Panels C and D of Figure 4, we compare submission behavior between different treatments. Panel C shows the cumulative fraction of submitted solutions of the Raven matrices and Panel D shows the fraction of submissions for the numerical tasks. The pattern in Panel C suggests that subjects change submission behavior across treatments. In the first 20 seconds the curves of all treatments overlap, but afterwards solutions in LL have been submitted significantly faster than in HL and HH. Even though subjects submitted fastest in LL, we do not observe statistically significant differences between LL and HH. Panel D shows that behavior is similar for tasks. However, solutions to the numerical tasks were submitted more quickly overall than to the Raven matrices.

Figure 4 about here

We also analyze the fraction of correctly submitted solutions over time. Panels E and F of Figure 4 show the fraction of correctly submitted solutions for both tasks in all treatments. The picture that emerges from these graphs is that the fraction of correctly submitted solutions is different across treatments. Panels E and F reveal that in HL, compared with LL, subjects wait longer until they submit a correct solution. This is in line with what we expect from the predictions of our model. In contrast to our theoretical prediction, we observe no significant differences in submission behavior between HL and HH.

Table 3 about here

Finally, Table 3 reports panel regressions with problem and individual fixed effects and the submission time in seconds as the dependent variable. We include treatment dummies for HL and HH. Columns (1) to (3) report the results for the Raven matrices, whereas columns (4) to (6) report the numerical tasks. The estimated coefficients suggest that subjects change submission behavior significantly in HL compared with LL when they are confronted with the Raven matrices. We do not observe a significant difference in the timing of submission if we compare HH and LL. The coefficient of HH and HL are significantly different from each other in all specifications. In the numerical task, every treatment yields significantly different average submission times. However, the actual change in timing of the solution is small. The maximum change we observe is that subjects wait on average three seconds longer until they submit their solution in HL compared with LL. This change in timing is equally small for both tasks.

3.4. Is Adaptation Optimal?

Figure 5 shows the expected earnings of the Raven matrices in all treatments. We calculate the expected earnings by multiplying the fraction of correct solutions in each treatment with the payoff from the red payment system at each point in time. Panel A shows the results in LL, Panels B and C present the results for HH and HL, and Panels D, E, and F show the results for the numerical tasks. The vertical straight line indicates the average submission time in each treatment and the dashed grey lines indicate the respective 95 percent confidence bounds. The dotted grey line indicates the time when the expected earnings reach the maximum.[20]

Figure 5 about here

All treatments reveal different times, which maximize the expected earnings. In LL expected earnings peak at the 27th second for the Raven matrices and the 22nd second for the numerical tasks. In HH the optimal submission time is the 27th second for the Raven matrices and 25th second for the numerical tasks. The difference between the treatments is the strongest for HL. Earnings for the Raven matrices peak at the 40th second and for the numerical task at the 39th second. All observed average submission times deviate statistically significantly from the optimal submission times and all panels reveal concave patterns. Subjects submit their solutions 4.1 (3.8) seconds later in LL when solving Raven matrices (numerical tasks), compared with the optimal submission time (two-tailed t-tests, p-value<0.001). In HH, submissions are 4.0 (1.8) seconds later (two-tailed t-tests, p-value<0.001), and in HL the submit button is pressed 6.7 (10.0) seconds earlier compared with the point that maximizes expected payoff time (two-tailed t-tests, p-value<0.001).

From an economic point of view, it is also important to compare the expected payoffs at the point of submission and at the maximum. Therefore, we compare the expected payoffs at the

[20] We define the maximum in a non-parametric way by searching for the highest value in the time span of 60 seconds.

average point of submission with the expected earnings at the optimal time of submission.[21] In LL this difference is not significantly different from zero for both tasks. The mean difference is 0.07 (0.15) cents for the Raven matrices (numerical tasks) and the p-value is 0.34 (0.13).[22] In HL, subjects could have earned on average 1.0 cents (0.7 cents) more for each problem if they had thought longer before submitting their solution. The difference is significant at the one percent level (p-value<0.005 for the numerical tasks). In HH, expected earnings are slightly higher in the optimum than at the actual submission time. However, the difference of 0.26 cents is only significant at the 10 percent level for the Raven matrices. For the numerical tasks, expected earnings could have been 0.7 cents higher in the optimum (p-value <0.010).

The incentives that we provide induce only a small change in behavior during problem solving. If we triple the average reward for correctly solving Raven matrices, subjects only invest 12 percent more time. Optimal behavior, however, which maximizes payoffs, would lead to an increase in time investment of 32 percent. We give three potential reasons why we observe this behavior.

First, since the overall improvements in expected earnings are either small or insignificant, if one compares the expected earnings at the actual submission times with the expected earnings at the optimal submission times, these small adjustments in timing could be explained by the small gains subjects make by adjusting their behavior.

A second reason why we do not observe strong changes in submission behavior could be due to heterogeneity in our data. We will address this point in the next section.

A third reason for the small adaptations could be that subjects are intrinsically motivated in solving problems. The intrinsic motivation to solve problems could overrule the motivation triggered by the monetary incentives. At a certain level, monetary incentives only seem to play a minor role in influencing problem-solving behavior.

[21] To control for the correlation of these two points in time at the individual level, we take the difference and test whether this is significantly different from zero.

[22] All p-values reported are obtained from a two-tailed t test.

4. Why Do Subjects Adapt So Little? Heterogeneity in Questions and Performance

The probability of solving a problem hardly changes across contexts but submission behavior changes in the way our model predicts. However, the adjustment in the timing of submission is small, as are the gains in expected earnings. We investigate this pattern further by looking at different subsamples.

4.1. Production Function and Time Investment

We first explore heterogeneity in problems and analyze if the choice process and submission behavior varies across different levels of difficulty. We then analyze differences across skill types. We only provide evidence for Raven matrices because results for the numerical tasks are equivalent.

4.1.1. Easy and Difficult Problems

Panels A, B, and C of Figure 6 show the probability of solving a problem over time for three levels of difficulty. We define the level of difficulty by the number of the respective Raven matrices in the test manual. This provides us with an exogenous definition of difficulty. The order of easy, moderate, and difficult problems was randomized across treatments. The grey areas indicate the 95 percent confidence intervals. Similar to the aggregate level presented in Figure 4, the confidence intervals overlap in all treatments and for all degrees of difficulty.

Figure 6 about here

Panels D, E, and F of Figure 6 show submission behavior for different levels of difficulty. These panels show the cumulative fraction of easy, moderate, and difficult problems in all treatments. The picture that emerges from these three panels is that submission behavior is heterogeneous across different levels of difficulty. Panel D shows that there is no significant difference in submission behavior between the treatments for easy Raven matrices. Panels E and F reveal that the submission behavior varies between treatments for moderate and difficult problems. Subjects submit solutions earliest in LL and latest in HL. Similar to our results at the aggregate level, we cannot identify significant differences between HH and the HL treatment.

Since submission behavior seems to be heterogeneous for different levels of difficulty across different treatments, we also investigate whether this holds for submission behavior of a correct solution (Panels G, H, and I of Figure 6). The panels suggest that the fraction of correctly submitted solutions is only different for two scenarios. Easy and moderate problems are solved more rapidly in the LL than in the HL treatment. We do not observe significant differences between all other treatments. Moreover, we do not obtain significant differences between the treatments in the fraction of correctly submitted solutions for difficult problems.

We observe strong changes between treatments in terms of submission behavior, but very few changes in the submission of a correct answer. Interestingly, the treatment effect on submissions becomes stronger with the difficulty of the problems. Subjects take more time to submit an answer if the problems become more difficult and incentives are higher. However, this does not influence the quality of the submitted solutions. Our analysis of the data shows that subjects submit a correct solution faster only if they face lower incentives for easy and moderate problems. Most interestingly, the fraction of correctly submitted solutions in the 60th second does not differ across treatments.

4.1.2. High and Low Performance

Panels A, B, and C of Figure 7 show the probability of solving the problem over time for different levels of performance. We take the fraction of correctly selected solutions in the 30th second as a performance measure and split the sample into three groups of equal size: high, moderate, and low performance types.[23]

Figure 7 about here

Panels D, E, and F in the middle of Figure 7 document the cumulative fraction of submitted solutions for high, moderate, and low performers in all treatments, respectively. The panels show that the reaction to a change in the red payment system is heterogeneous across different levels of performance. First, high-performance subjects adapt their submission behavior strongest compared with moderate and low-performance types. Individuals with a high problem-solving score submit their solutions significantly later in HL and HH compared

[23] The results do not change if we take performance at the 20th, 40th, 50th, or 60th second as the criterion.

with LL. Subjects with a lower performance also submit their solutions in HL and HH later than in LL, but the difference is smaller than for high-performance types.

We also analyze submission behavior of a correct solution for different levels of performance. Panels G, H, and I of Figure 7 show the fraction of correctly submitted solutions across all three treatments. The pictures show that subjects with a higher performance also adapt their submission behavior of correct solutions more strongly to the context relative to subjects with a moderate or low performance. The best performing problem solvers wait longer until they submit correct solutions in a context with higher incentives, compared with other subjects. These results are in line with what we find for submission behavior in Panels D, E, and F.[24]

It seems that the best performers show the strongest adjustment of behavior between different treatments. The next question is whether submission behavior is optimal in terms of payoff maximization.

4.2. Adaption Behavior

4.2.1. Easy and Difficult Problems

Up to this point, we constructed the expected payoff function by aggregating the fraction of correctly submitted solutions over all subjects and problem types. We now analyze heterogeneity in earnings. Figure 8 shows the expected payoffs for the Raven matrices in all three treatments for easy, moderate, and difficulty problems. Panels A, B, and C show the results for easy problems across all three treatments, Panels D, E, and F for moderate problems, and G, H, and I for difficult problems. The vertical black line indicates the average submission time and the dotted grey line the optimal submission time for each level and each treatment.

Figure 8 about here

The pattern of submission behavior is heterogeneous for different levels of difficulty. If subjects face an easy problem, they submit their solutions too early compared with the

[24] We also investigate submission behavior of a correct solution for the numerical tasks. The results are consistent with what we find for the Raven matrices. An equivalent to Figure 7 is Figure B.2, which can be found in Appendix B.

optimal submission time in LL and HL. In HH, the actual submission time is not significantly different from the optimum. All average submission times for moderate problems deviate significantly from the point where expected earnings peak. In LL and HH, solutions are always submitted too late. Subjects do not wait long enough to submit their solutions in HL. The picture is different for difficult problems. We do not observe a significant difference between actual and optimal behavior with respect to expected payoff maximization in LL and HL for difficult problems.

The picture that emerges from the deviations in expected payoffs is mixed. In the baseline treatment, expected earnings for easy problems are not significantly higher in the optimum compared with the actual average submission time.[25] In HH, earnings are not significantly different from the optimum,[26] whereas subjects could have earned 2.5 cents more per problem by submitting their solutions to easy problems later in the HL treatment (p-value<0.01, t-test). The picture is equivalent for moderate problems. Earnings in the optimum are slightly higher compared with the actual submission time in LL.[27] We do not observe significant differences in HH[28] but expected earnings are 1.95 cents higher at the optimum in HL (p-value=0.0167, two-tailed t-test). Difficult problems seem to be solved in a payoff maximizing way. We do not observe differences in expected payoffs for difficult problems in all treatments, compared with the maximal expected payoff.[29] As on the aggregate level, potential improvements of expected earnings are low.

4.2.2. High and Low Performance

Figure 9 shows the expected payoff functions for different performance levels. Panels A, B, and C show the expected payoffs for subjects with the highest performance in the blue system after 30 seconds for all three treatments. Panels D, E, and F and G, H, and I show the expected payoffs and submission times for moderate and low performance types, respectively. Our findings are consistent over different levels of performance. The solutions

[25] Difference is 0.25 cents, p-value=0.45, two-tailed t-test.
[26] Difference is 0.11 cents, p-value=.88.
[27] Difference .19 cents, p-value=.51
[28] Difference .47 cents, p-value=.37.
[29] LL treatment: 0.05 cents, p-value .82; HH treatment 0.28 cents, p-value 0.54; HL treatment 0.13 cents, p-value 0.86.

are submitted too late in LL and HH. Moreover, subjects submit their solutions too early in HL.[30]

Figure 9 about here

The deviations in expected payoffs are consistent with what we find on the aggregate level. We find no differences in expected payoffs at the actual submission time and the optimal submission time for high performers and moderate performers in all treatments. Only low performers can improve their earnings in the HL treatment by 2.3 cents (p-value<0.01) if they wait longer to submit their answer. An equivalent of Figure 9 for the numerical tasks can be found in Figure B.1 in the online Appendix.

Overall, we find only small adaptations in behavior for different performance levels. Most interestingly, the total deviation from the optimal submission times between all treatments differs systematically across performance types. In total, low performers deviate by 20 seconds, moderate performers by 10 seconds, and high performers by 9 seconds from the optimal submission time. This implies that adaptation to a change in incentives and time pressure is associated with performance level.

5. Robustness Checks

In this section we provide robustness checks, first showing that the behavior in the blue system is independent of the treatment variation. Next, we show that it is unlikely that our results are driven by aggregation bias.

5.1. Elicitation of the Choice Process

To estimate the probability of a correct solution over time, we need to be able to identify an agent's thoughts over time. Since it is impossible to obtain a pure measure of the currently preferred solution, we take behavior in the blue payment scheme as a proxy measure. The blue payment scheme provides an incentive to reveal immediately the solution to a problem independent of the actual submitted solution. The data of the selection times show that subjects used the blue payment scheme to indicate their preferred choice in the course of

[30] Except for two cases, all deviations are highly significant ($p<0.01$). High performance subjects do not deviate significantly from the optimal submission time in LL and low performance types do not deviate significantly from the optimum in HH.

solving a problem. It is important to document whether behavior in the blue payment system was indeed independent from behavior in the red payment system. To show this, we conduct panel regressions with problem and individual fixed effects, as well as the time of the first choice as the dependent variable. Table 4 shows the results, with columns (1) to (3) indicating the regressions for Raven matrices. The first column only contains problem fixed effects. We run a specification only with individual fixed effects in column (2). The specification in column (3) contains problem and individual fixed effects. Columns (4) to (6) show the equivalent regressions for the numerical tasks. There is no significant difference in the time of first selection between the treatments for Raven matrices and numerical tasks.

Table 4 about here

A second check for independence of the blue payment scheme across treatments is the number of choices subjects make. Table 5 shows regression results with the number of choices per problem as a dependent variable. We again conduct panel regressions with problem fixed effects and subject fixed effects. For the Raven matrices, none of the treatment dummies are significantly different from zero. The only exception is for numerical tasks in which subjects tend to make fewer choices in HH compared with LL. However, the coefficient is only significant at the 10 percent level and the point estimate of 0.14 clicks seems to be economically small.

Table 5 about here

5.2. Aggregation Bias

Our empirical method bears the danger of aggregation bias. We assume that the aggregation of choices in the blue system is a valid representation of an individual’s probability function and an individual’s expected payoff function. Since we would only observe jumps from zero to one on an individual basis, we have to aggregate the data to learn more about the decision-making process. We test the robustness of our results by disaggregating the data into various subsamples.

Table 6 about here

Panel A of Table 6 shows the results for the Raven matrices and Panel B for the numerical tasks. The first column shows the average submission times in the three treatments. The second column documents the optimal submission times from Figure 5.

In the following columns, we split the data into various subsamples. First, we split the sample into different subgroups of performance types and different degrees of problem difficulty. Next, we create subgroups for performance and different degrees of difficulty. Each cell documents the mean and the standard error of the respective state of disaggregation. Columns (4) to (14) show that the means are generally not different from the aggregated mean. This indicates that our results do not seem to be driven by aggregation bias.

6. Interpretation and Conclusion

In this paper, we have analyzed individual performance on a cognitive test from an economic point of view. We propose a new method to shed light on the trade-offs that people face in the decision-making process. By designing an experimental set up in the laboratory, we were able to plot the result of the trade-off between time investment and the cost of time and the probability of finding the right answer to a question during a cognitive ability test.

Our main findings are consistent with an economic model in which the success rate depends on time investment. We observe heterogeneity in outcomes and behavior in our sample of relatively homogenous subjects (all are students from one university). This heterogeneity in behavior reflects different choices and different responses to monetary incentives and time pressure. For relatively low performing subjects there is scope for improvement if they lengthen their deliberation process. Understanding the way in which people try to find answers to questions provides teachers, employers, and policymakers, for example, with information as to how time pressure and monetary incentives change the quality of a decision and the behavior of different types when being put to a test.

Our main contribution to the literature can be summarized by the following key results. We plot the returns to time investments during a cognitive test. Our results suggest that the probability of finding a correct solution to a problem is a concave function of the time invested. In our experiment, the incentive environment does not seem to influence (statistically) significantly this probability. Hence, as of a certain level of incentives, subjects do not come up faster with a correct solution. However, submission behavior differs across incentive environments as predicted by economic theory. Subjects invest more time when monetary incentives are higher and less when time pressure is increased. Overall, the changes in the timing are relatively small. We explore three potential reasons. One potential reason for our finding could be that in our setup the differences between actual expected earnings and optimal expected earnings are economically small. Another explanation can be heterogeneity across subjects. We observe heterogeneity in behavior if we split the sample into different levels of problem difficulty and different performance types. Behavior is adjusted the most for the more difficult problems. This suggests that individuals adapt their behavior to the incentive environment. The adaptation in behavior does not improve outcomes. Next, higher performing individuals adjust their behavior relatively optimally when monetary incentives are changed. Further controlling for aggregation bias does not change our results.

Another contribution of this research is that we show that subjects adjust their behavior to the environment. This suggests that incentives and time pressure could help improve performance but could also lead to declines in performance. In particular, we observe that in treatments with high time pressure or low stakes, people seem to wait too long to submit their answers, while in the treatment with high stakes people submit their answers too soon. This implies that time pressure and monetary incentives do not necessarily maximize performance but that the optimal response depends on skills and context. One first conclusion for teachers and employees is that behavior in different incentive environments is heterogeneous across different types of individuals. Low-performing subjects in particular adjust the least to new environments, whereas high-performing subjects seem to be able to adjust their behavior relatively optimally to different environments. This could also indicate that low-performing subjects have little knowledge about their production function.

Next, our results indicate that as of a certain level of incentives the speed of thinking does not increase anymore. From an economic point of view this suggests that there is a certain range in which incentives seem to be able to improve outcomes. Moreover, we show that subjects fail to maximize their expected payoffs. One possible reason for this result could be that non-financial motives play a role too, while working on a cognitive ability test.

Finally, the experimental method applied in this paper offers an opportunity to track an individual’s thoughts before final decisions are made. We believe that such a setup is useful in other settings too, in order to learn more about decision-making processes.

For policymakers and applied researchers, these results suggest that behavior and performance during cognitive ability tests are not only a pure measure of ability, but also of the way in which people adapt behavior to different contexts. Our results indicate that the context influences behavior on a task but not necessarily the speed of thinking. An important avenue for future research is to investigate further the heterogeneity across different types by, for example, taking into account the impact of differences in personality traits and economic preferences on behavior during a test.

7. References

Agranov, Marina, Andrew Caplin, and Chloe Tergiman. 2015. “Naive Play and the Process of Choice in Guessing Games.” *Journal of the Economic Science Association* 1 (2): 146–57.

Almlund, Mathilde, Angela L. Duckworth, James J. Heckman, and Tim D. Kautz. 2011. “Personality Psychology and Economics.” In *Handbooks of the Economics of Education*, 1–158. Amsterdam, North Holland: Elsevier.

Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. 2009. “Incentives and Services for College Achievement: Evidence from a Randomized Trial.” *American Economic Journal: Applied Economics* 1 (1): 136–63.

Angrist, Joshua, and Victor Lavy. 2009. “The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial.” *The American Economic Review* 99 (4): 1384–1414.

Bettinger, Eric P. 2011. “Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores.” *Review of Economics and Statistics* 94 (3): 686–698.

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas ter Weel. 2008. “The Economics and Psychology of Personality Traits.” *Journal of Human Resources* 43.

Borghans, Lex, Huub Meijers, and Bas Ter Weel. 2008. “The Role of Noncognitive Skills in Explaining Cognitive Test Scores.” *Economic Inquiry* 46 (1): 2–12.

Borghans, Lex, Huub Meijers, and Bas ter Weel. 2013. “The Importance of Intrinsic and Extrinsic Motivation for Measuring IQ.” *Economics of Education Review* 34: 17–28.

Caplin, Andrew, Mark Dean, and Daniel Martin. 2011. “Search and Satisficing.” *The American Economic Review* 101 (7): 2899–2922.

Caplin, Andrew, and Martin Dean. 2011. “Search, Choice, and Revealed Preferences.” *Theoretical Economics* 1 (6): 19–48.

Carpenter, Patricia A., Marcel Adam Just, and Peter Shell. 1990. “What One Intelligence Test Measures: A Theoretical Account of the Processing in the Raven Progressive Matrices Test.” *Psychological Review* 97: 404–431.

Duckworth, Angela L., and Martin E. P. Seligman. 2005. “Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents.” *Psychological Science* 16 (12): 939–944.

Duckworth, Angela Lee, Patrick D. Quinn, Donald R. Lynam, Rolf Loeber, and Magda Stouthamer-Loeber. 2011. “Role of Test Motivation in Intelligence Testing.” *Proceedings of the National Academy of Sciences*, April.

Edlund, Calvin V. 1972. “The Effect on the Behavior of Children, as Reflected in the IQ Scores, When Reinforced After Each Correct Response.” *Journal of Applied Behavior Analysis* 5 (3): 317–319.

Epstein, Seymour. 1994. "Integration of the Cognitive and the Psychodynamic Unconscious." *American Psychologist* 49 (8): 709–724.

Fischbacher, Urs. 2007. "Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10 (2): 171–178.

Fryer, Roland G. 2011. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." *Quarterly Journal of Economics* 126 (4): 1755–1798.

Gabaix, Xavier, David Laibson, Guillermo Moloche, and Stephen Weinberg. 2006. "Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model." *The American Economic Review*, 1043–1068.

Gill, D., & Prowse, V. (2016). Cognitive ability, character skills, and learning to play equilibrium: A level-k analysis. *Journal of Political Economy*, *124*(6), 1619-1676.

Greiner, Ben. 2003. *An Online Recruitment System for Economic Experiments*. Vol. GWDG Bericht 63. Forschung Und Wissenschaftliches Rechnen 2003. Kurt Kremer, Volker Macho.

Kautz, Tim, James J. Heckman, Ron Diris, Bas ter Weel, and Lex Borghans. 2014. "Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success." Working Paper 20749. National Bureau of Economic Research. http://www.nber.org/papers/w20749.

Kocher, Martin G., Julius Pahlke, and Stefan T. Trautmann. 2013. "Tempus Fugit: Time Pressure in Risky Decisions." *Management Science* 59 (10): 2380–91.

Kocher, Martin G., and Matthias Sutter. 2006. "Time Is Money - Time Pressure, Incentives, and the Quality of Decision-Making." *Journal of Economic Behavior & Organization* 61 (3): 375–392.

Lindner, Florian, and Matthias Sutter. 2013. "Level-Reasoning and Time Pressure in the 11–20 Money Request Game." *Economics Letters* 120 (3): 542–45.

Lloyd, Margaret E., and Therese M. Zylla. 1988. "Effect of Incentives Delivered for Correctly Answered Items on the Measured IQs of Children of Low and High IQ." *Psychological Reports* 63 (2): 555–561.

Manzini, Paola, and Marco Mariotti. 2007. "Sequentially Rationalizable Choice." *The American Economic Review* 97 (5): 1824–1839.

Raven, John C. 1962. *Advanced Progressive Matrices*. London: H. K. Lewis & Co. Ltd.

Reutskaja, Elena, Rosemarie Chariklia Nagel, Colin F. Camerer, and Antonio Rangel. 2011. "Search Dynamics in Consumer Choice under Time Pressure: An Eye-Tracking Study." *American Economic Review* 101 (2): 900–926.

Roberts, Richard D., Pippa M. Markham, Gerald Matthews, and Zeidner Moshe. 2005. "Assessing Intelligence: Past, Present, and Future." In *Handbook of Understanding and Measuring Intelligence*, 333–360. Thousand Oaks, CA: Sage Publications Inc.

Rodriguez-Planas, Nuria. 2012. "Longer-Term Impacts of Mentoring, Educational Services, and Learning Incentives: Evidence from a Randomized Trial in the United States." *American Economic Journal: Applied Economics* 4 (4): 121–139.

Rustichini, Aldo. 2015. "The Role of Intelligence in Economic Decision Making." *Current Opinion in Behavioral Sciences*, Neuroeconomics, 5 (October): 32–36.

Segal, Carmit. 2012. "Working When No One Is Watching: Motivation, Test Scores, and Economic Success." *Management Science* 58 (8): 1438–1457.

Simon, Herbert A. 1955. "A Behavioral Model of Rational Choice." *The Quarterly Journal of Economics* 69 (1): 99–118.

Stigler, George J. 1961. "The Economics of Information." *The Journal of Political Economy*, 213–225.

Table 1. Performance on Raven Matrices and Numerical Tasks

Panel A. Success Rate

	Raven Matrices							Numerical Tasks
Time elapsed (sec.)	All	Easy vs. Difficult	Easy	Easy vs. Moderate	Moderate	Moderate vs. Difficult	Difficult	All
	0.156		0.216		0.146		0.099	0.316
10	(0.008)	+++	(0.016)	++	(0.014)	+++	(0.012)	(0.011)
0 vs. 10	***		***		***		***	***
	0.311		0.464		0.300		0.152	0.526
20	(0.011)	+++	(0.019)	+++	(0.018)	+++	(0.015)	(0.011)
10 vs. 20	***		***		***		***	***
	0.409		0.601		0.396		0.208	0.628
30	(0.011)	+++	(0.019)	+++	(0.019)	+++	(0.017)	(0.011)
20 vs. 30	***		***		***		**	***
	0.456		0.663		0.440		0.242	0.690
40	(0.011)	+++	(0.018)	+++	(0.019)	+++	(0.018)	(0.011)
30 vs. 40	***		**		ns		ns	***
	0.499		0.699		0.498		0.274	0.736
50	(0.011)	+++	(0.018)	+++	(0.019)	+++	(0.018)	(0.010)
40 vs. 50	***		ns		**		ns	***
	0.519		0.722		0.523		0.284	0.755
60	(0.011)	+++	(0.017)	+++	(0.019)	+++	(0.019)	(0.010)
50 vs. 60	ns		ns		ns		ns	ns
Observations	1920		662		671		587	1920

Panel B. Average Cumulative Earnings per Question (in Cents)

	Raven Matrices							Numerical Tasks
Time elapsed (sec.)	All	Easy vs. Difficult	Easy	Easy vs. Moderate	Moderate	Moderate vs. Difficult	Difficult	All
10	0.512 (0.031)	+++	0.600 (0.053)	ns	0.531 (0.054)	+	0.390 (0.050)	0.955 (0.037)
0 vs. 10	***		***		***		***	***
20	1.752 (0.068)	+++	2.420 (0.122)	+++	1.702 (0.116)	+++	1.055 (0.106)	3.164 (0.080)
10 vs. 20	***		***		***		***	***
30	3.601 (0.108)	+++	5.137 (0.192)	+++	3.515 (0.181)	+++	1.965 (0.164)	6.092 (0.122)
20 vs. 30	***		***		***		***	***
40	5.794 (0.151)	+++	8.335 (0.261)	+++	5.617 (0.250)	+++	3.131 (0.227)	9.414 (0.161)
30 vs. 40	***		***		***		***	***
50	8.213 (0.194)	+++	11.796 (0.326)	+++	7.968 (0.320)	+++	4.451 (0.291)	12.998 (0.199)
40 vs. 50	***		***		***		***	***
60	10.770 (0.239)	+++	15.360 (0.393)	+++	10.542 (0.395)	+++	5.855 (0.359)	16.742 (0.236)
50 vs. 60	***		***		***		***	***
Observations	1920		662		671		587	1920

Note. Panel A shows the fraction of correctly selected answers in steps of 10 seconds. The first column reports the numbers for all Raven matrices. Columns (2) – (4) report the numbers for different degrees of difficulty. The last column reports the number for the calculation task. Panel B shows the cumulative earnings from the blue payment system in steps of 10 seconds. Results from t-tests performed in order to compare differences between rows (between columns) are reported with asterisks (pluses). *** (+++) p<0.01, **(++) p<0.05, *(+) p<0.1. "ns" indicates that the differences are not statistically significant. Standard errors are reported in parentheses.

Table 2. Submission Behavior on Raven Matrices and Numerical Tasks

Panel A. Submissions (Percentage)

	Raven Matrices							Numerical Tasks
Time elapsed (sec.)	All	Easy vs. Difficult	Easy	Easy vs. Moderate	Moderate	Moderate vs. Difficult	Difficult	All
10	0.070 (0.006)	+++	0.110 (0.012)	ns	0.054 (0.009)	ns	0.044 (0.008)	0.068 (0.006)
0 vs. 10	***		***		***		***	***
20	0.291 (0.010)	+++	0.432 (0.019)	+++	0.253 (0.017)	+++	0.177 (0.016)	0.382 (0.011)
10 vs. 20	***		***		***		***	***
30	0.555 (0.011)	+++	0.708 (0.018)	+++	0.535 (0.019)	+++	0.405 (0.020)	0.690 (0.011)
20 vs. 30	***		***		***		***	***
40	0.734 (0.010)	+++	0.856 (0.014)	+++	0.726 (0.017)	+++	0.606(0.020)	0.870 (0.008)
30 vs. 40	***		***		***		***	***
50	0.876 (0.007)	+++	0.946 (0.009)	+++	0.878 (0.013)	+++	0.794 (0.017)	0.964 (0.004)
40 vs. 50	***		***		***		***	***
60	0.960 (0.004)	+++	0.985 (0.005)	ns	0.954 (0.008)	+++	0.939 (0.010)	0.993 (0.002)
50 vs. 60	***		***		***		***	***
Observations	1920		662		671		587	1920

Panel B. Correct Submissions (Percentage)

	Raven Matrices							Numerical Tasks
Time elapsed (sec.)	All	Easy vs. Difficult	Easy	Easy vs. Moderate	Moderate	Moderate vs. Difficult	Difficult	All
	0.019		0.042		0.012		0.002	
10	(0.003)	+++	(0.008)	+++	(0.004)	++	(0.002)	0.036 (0.004)
0 vs. 10	***		***		**		ns	***
	0.148		0.266		0.121		0.046	
20	(0.008)	+++	(0.017)	+++	(0.013)	+++	(0.009)	0.244 (0.010)
10 vs. 20	***		***		***		***	***
	0.286		0.447		0.280		0.112	
30	(0.010)	+++	(0.019)	+++	(0.017)	+++	(0.013)	0.472 (0.011)
20 vs. 30	***		***		***		***	***
	0.365		0.536		0.365		0.170	
40	(0.011)	+++	(0.019)	+++	(0.019)	+++	(0.016)	0.598 (0.011)
30 vs. 40	***		***		***		***	***
	0.414		0.577		0.423		0.220	
50	(0.011)	+++	(0.019)	+++	(0.019)	+++	(0.017)	0.654 (0.011)
40 vs. 50	***		ns		**		**	***
	0.441		0.595		0.459		0.247	
60	(0.011)	+++	(0.019)	+++	(0.019)	+++	(0.018)	0.672 (0.011)
50 vs. 60	*		ns		ns		ns	Ns
Observations	1920		662		671		587	1920

Panel C: Earnings (Cents)

	Raven Matrices							Numerical Tasks
Time elapsed (sec.)	All	Easy vs. Difficult	Easy	Easy vs. Moderate	Moderate	Moderate vs. Difficult	Difficult	All
10	5.901 (0.108)		8.009 (0.159)		4.882 (0.198)		1.391 (0.134)	11.564 (0.127)
		+++		+++		+++		
20	8.315 (0.065)		9.864 (0.094)		7.945 (0.113)		4.738 (0.133)	9.924 (0.053)
10 vs. 20	***	+++	***	+++	***	+++	***	***
30	4.994 (0.046)		6.167 (0.081)		5.337 (0.077)		2.956 (0.074)	6.678 (0.045)
20 vs. 30	***	+++	***	+++	***	+++	***	***
40	2.261 (0.037)		2.995 (0.075)		2.360 (0.063)		1.518 (0.053)	3.693 (0.044)
30 vs. 40	***	+++	***	+++	***	+++	***	***
50	0.844 (0.023)		1.104 (0.056)		0.856 (0.037)		0.700 (0.033)	1.563 (0.038)
40 vs. 50	***	+++	***	+++	***	+++	***	***
60	0.187 (0.012)		0.288 (0.036)		0.267 (0.026)		0.107 (0.013)	0.353 (0.028)
50 vs. 60	***	+++	***	+++	***	+++	***	***

Note. Panel A shows the fraction of people who submit their answer after the 10th, 20th second and so forth. Panel B shows the fraction of correctly submitted answers in the same categories as in the previous tables. Panel C shows the payoffs from the red payment scheme. We calculate the earnings in intervals of 10 seconds. Asterisks indicate significant differences between rows (*** p<0.01, ** p<0.05, * p<0.1). Pluses indicate significant differences between columns (+++ p<0.01, ++ p<0.05, + p<0.1). "ns" indicates that the differences are not statistically significant. Standard errors are reported in parentheses.

Table 3. Determinants of Submission Time

	(1)	(2)	(3)	(4)	(5)	(6)
		Raven Matrices			Numerical Tasks	
HH treatment	0.598	0.916*	0.598	0.919**	1.032**	0.919**
	(0.434)	(0.513)	(0.459)	(0.341)	(0.460)	(0.419)
HL treatment	2.838***	3.171***	2.838***	3.105***	3.144***	3.105***
	(0.380)	(0.540)	(0.490)	(0.323)	(0.526)	(0.473)
Constant	30.42***	30.20***	22.80***	25.76***	25.71***	23.29***
	(0.235)	(0.289)	(0.985)	(0.177)	(0.274)	(0.737)
Observations	5,760	5,760	5,760	5,760	5,760	5,760
R-squared	0.008	0.010	0.304	0.013	0.018	0.268
Question FE	YES	NO	YES	YES	NO	YES
Individual FE	NO	YES	YES	NO	YES	YES
p-value LL vs. HH	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Note. The table shows the results of linear panel regressions with the submission time in seconds for Raven matrices and numerical tasks as the dependent variable. The baseline payment scheme serves as the reference group. The last row reports the p-value of the F-test, which tests the equality of both treatment dummies. We control for potential confounding factors which could be driven due to certain individuals or questions by making use of question fixed effects in columns (1) and (4), individual fixed effects in columns (2) and (5) and controlling for both in columns (3) and (6). Robust standard errors are reported in parentheses. Standard errors are clustered at the question level in columns (1) and (4) and at the subject level in the remaining columns. *** p<0.01, ** p<0.05, * p<0.1

Table 4. Time of First Choice

	(1)	(2)	(3)	(4)	(5)	(6)
		Raven Matrices			Numerical Tasks	
HH treatment	0.0158	0.0380	0.0158	0.166	0.173	0.166
	(0.378)	(0.279)	(0.278)	(0.211)	(0.213)	(0.215)
HL treatment	-0.157	-0.117	-0.157	0.407*	0.394	0.407
	(0.370)	(0.257)	(0.254)	(0.228)	(0.303)	(0.304)
Constant	6.071***	6.051***	5.904***	3.750***	3.752***	3.352***
	(0.231)	(0.157)	(0.602)	(0.122)	(0.164)	(0.227)
Observations	5,760	5,760	5,760	5,760	5,760	5,760
R-squared	0.000	0.000	0.019	0.001	0.002	0.022
Question FE	YES	NO	YES	YES	NO	YES
Individual FE	NO	YES	YES	NO	YES	YES
p-value HL vs. HH	.601	.512	.462	.326	.209	.171

Note. The table shows panel regression with question fixed and subject fixed effects. The dependent variable is the time of the first choice per question. HH treatment and HL treatment are dummies for the respective red payment scheme. The baseline payment scheme serves as the reference group. The last row reports the p-value of the F-test, which tests the difference between the coefficients of HH and HL. Standard errors are clustered at the question level in columns (1) and (4) and at the subject level in the remaining columns. *** p<0.01, ** p<0.05, * p<0.1

Table 5. Number of Choices

	(1)	(2)	(3)	(4)	(5)	(6)
		Raven Matrices			Numerical Tasks	
HH treatment	0.0751	0.0776	0.0751	-0.141***	-0.128*	-0.141*
	(0.0873)	(0.0882)	(0.0853)	(0.0483)	(0.0752)	(0.0738)
HL treatment	0.00636	0.00885	0.00636	-0.0793	-0.0661	-0.0793
	(0.0828)	(0.0906)	(0.0879)	(0.0507)	(0.0547)	(0.0535)
Constant	2.808***	2.806***	2.624***	2.718***	2.709***	2.472***
	(0.0462)	(0.0546)	(0.162)	(0.0282)	(0.0351)	(0.101)
Observations	5,760	5,760	5,760	5,760	5,760	5,760
R-squared	0.000	0.000	0.022	0.001	0.002	0.073
Question FE	YES	NO	YES	YES	NO	YES
Individual FE	NO	YES	YES	NO	YES	YES
p-value HL vs. HH	.49	.34	.343	.237	.431	.418

Note. The table shows panel regression with question fixed and subject fixed effects. The dependent variable is the number of choices per question. HH treatment and HL treatment are dummies for the respective red payment scheme. The baseline payment scheme serves as the reference group. The last row reports the p-value of the F-test, which tests the difference between the coefficients of HH and HL. Standard errors are clustered at the question level in columns (1) and (4) and at the subject level in the remaining columns. *** p<0.01, ** p<0.05, * p<0.1

Table 6. Controlling for Aggregation Bias

Panel A. Raven Matrices

Treatment	Actual submission time	Aggregated optimum	Mean of disaggregated optima										
			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Baseline	28.953	27	29.125	26.711	26.113	26.214	26.299	26.72	28.965	27.748	26.202	25.564	23.797
	(0.042)		(0.004)	(0.008)	(0.002)	(0.011)	(0.009)	(0.023)	(0.038)	(0.035)	(0.034)	(0.036)	(0.035)
HH	29.94	27	27.32	27.648	29.23	27.485	26.972	27.359	27.309	26.986	26.396	25.681	25.898
	(0.042)		(0.001)	(0.005)	(0.006)	(0.009)	(0.011)	(0.02)	(0.035)	(0.033)	(0.035)	(0.035)	(0.034)
HL	32.106	40	39.906	43.875	39.946	44.502	42.764	42.704	34.928	38.972	37.789	35.594	36.817
	(0.044)		(0.004)	(0.021)	(0.005)	(0.017)	(0.018)	(0.028)	(0.042)	(0.04)	(0.037)	(0.043)	(0.042)

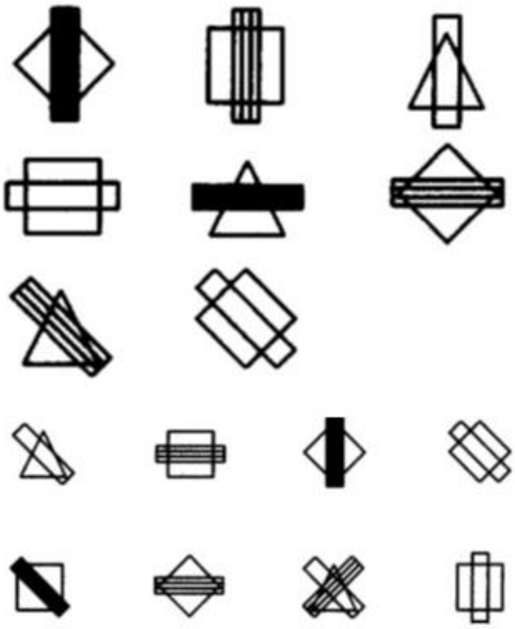
Panel B. Numerical Tasks

Treatment	Actual submission time	Aggregated optimum	Mean of disaggregated optima										
			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Baseline	25.46	22	23.969	23.914	24.053	24.524	22.587	23.712	24.436	24.306	24.456	23.671	23.261
	(0.034)		(0.008)	(0.01)	(0.007)	(0.01)	(0.008)	(0.014)	(0.027)	(0.026)	(0.028)	(0.028)	(0.027)
HH	26.552	25	21.734	23.414	25.869	22.575	22.328	21.661	25.053	24.856	24.54	23.201	22.484
	(0.034)		(0.011)	(0.011)	(0.004)	(0.011)	(0.011)	(0.016)	(0.026)	(0.027)	(0.026)	(0.028)	(0.027)
HL	28.676	39	39.016	37.055	39.778	40.245	40.956	37.621	29.805	31.242	31.427	31.167	31.154
	(0.037)		(0.012)	(0.01)	(0.008)	(0.022)	(0.018)	(0.025)	(0.036)	(0.038)	(0.035)	(0.037)	(0.039)

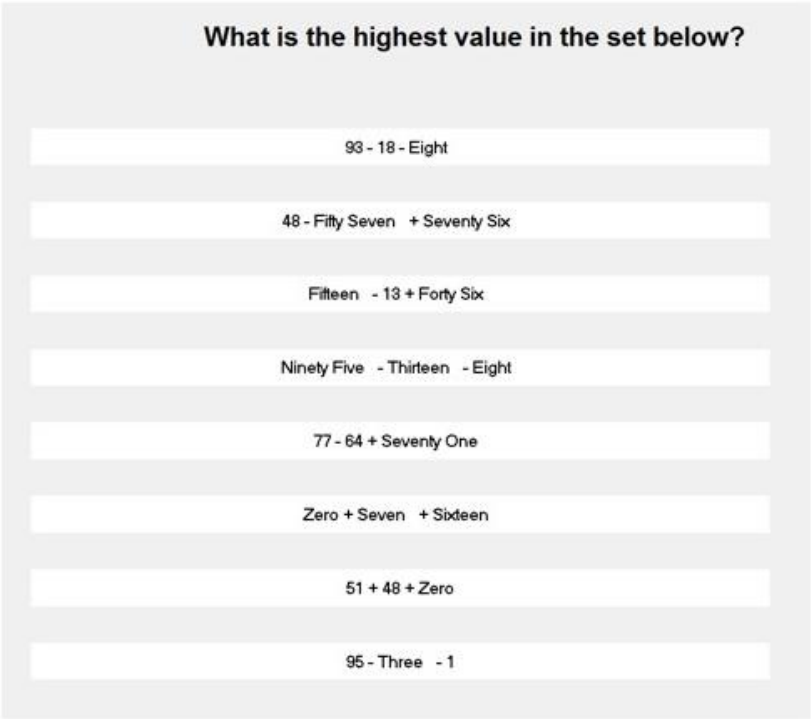
Note. The table shows the actual submission time for the three different treatments in the first column. The second column shows the optimal submission time in the fully aggregated state. The last 11 columns show the means of optima in different states of disaggregation. Standard errors are reported in the row below the means. The 11 different states of disaggregation for each column are as follows: 1: Three different performance types; 2: Six different performance types; 3: Three different degrees of difficulty; 4: Six different degrees of difficulty; 5; Three different performance types and three different degrees of difficulty; 6: Six different performance types and six different degrees of difficulty; 7: By 12 different submission times; 8: By 12 different submission times and three different types of performance; 9: By 12 different submission times and three different degrees of difficulty; 10: By 12 different submission times, three different degrees of difficulty, and three different performance types; 11: By 60 different submission times.

Figure 1. Examples of the Tasks

Panel A. Raven Matrix



Panel B. Numerical Task



What is the highest value in the set below?

93 - 18 - Eight

48 - Fifty Seven + Seventy Six

Fifteen - 13 + Forty Six

Ninety Five - Thirteen - Eight

77 - 64 + Seventy One

Zero + Seven + Sixteen

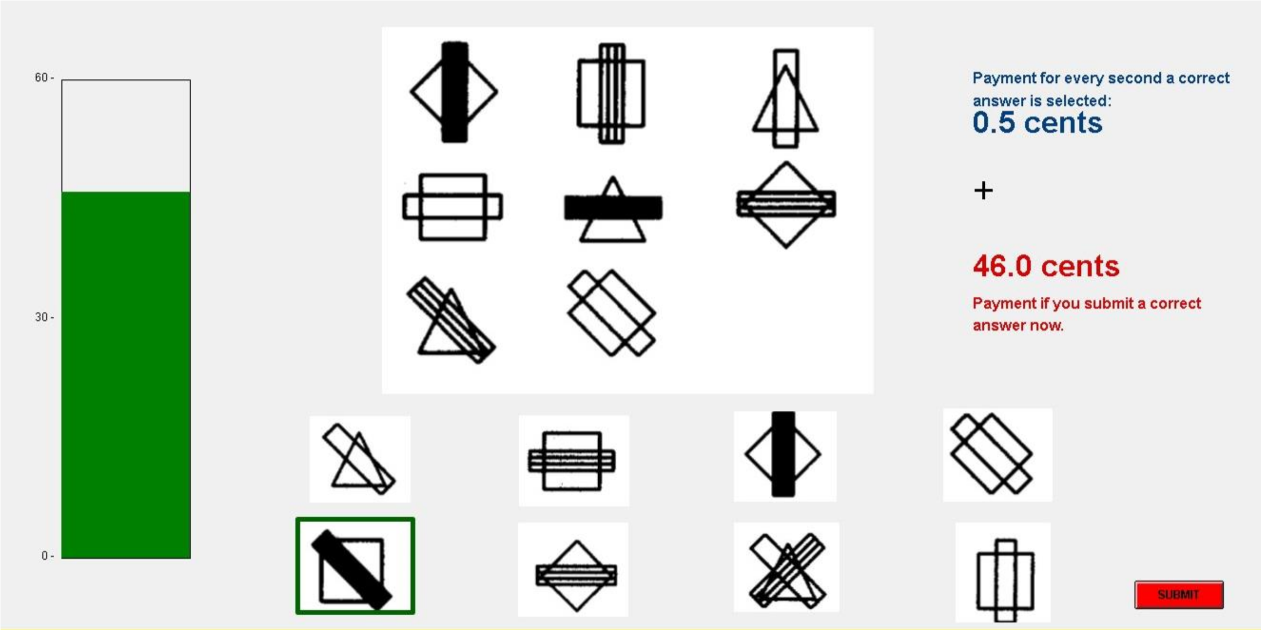
51 + 48 + Zero

95 - Three - 1

Note. Panel A shows a typical Raven matrix, taken from Carpenter et al. (1990, p. 407). None of our experimental Raven matrices is shown since the psychological tests are meant to be kept confidentially. Panel B shows a typical numerical task.

Figure 2. Screenshot of the Decision Screen

Panel A. Raven Matrix



Payment for every second a correct answer is selected:

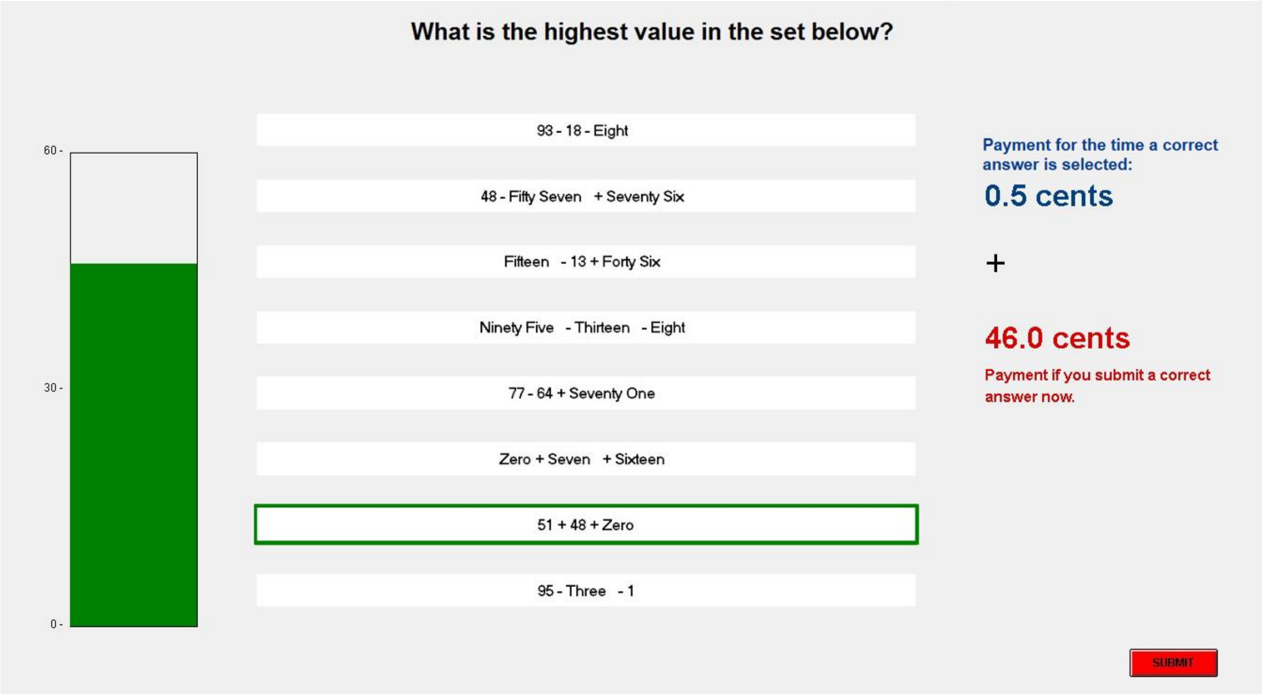
0.5 cents

+

46.0 cents

Payment if you submit a correct answer now.

Panel B. Numerical Task



What is the highest value in the set below?

93 - 18 - Eight

48 - Fifty Seven + Seventy Six

Fifteen - 13 + Forty Six

Ninety Five - Thirteen - Eight

77 - 64 + Seventy One

Zero + Seven + Sixteen

51 + 48 + Zero

95 - Three - 1

Payment for the time a correct answer is selected:

0.5 cents

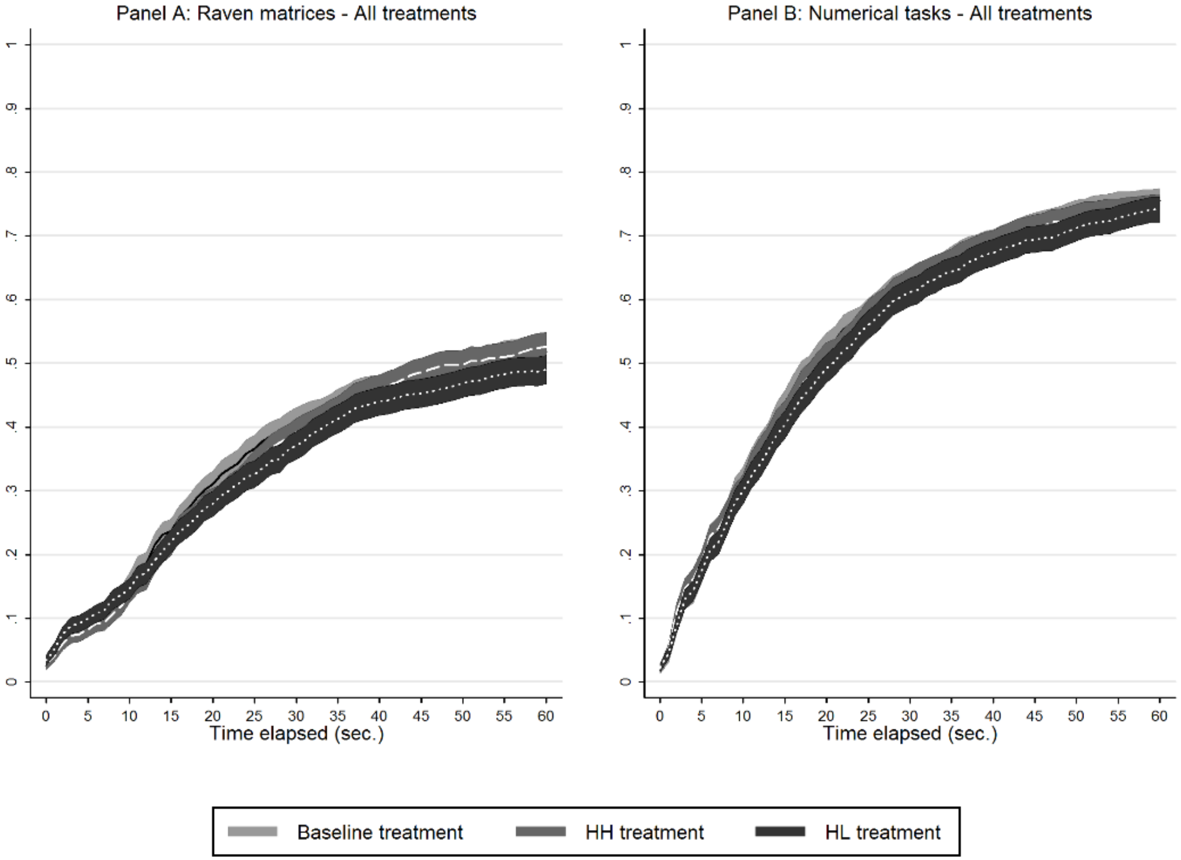
+

46.0 cents

Payment if you submit a correct answer now.

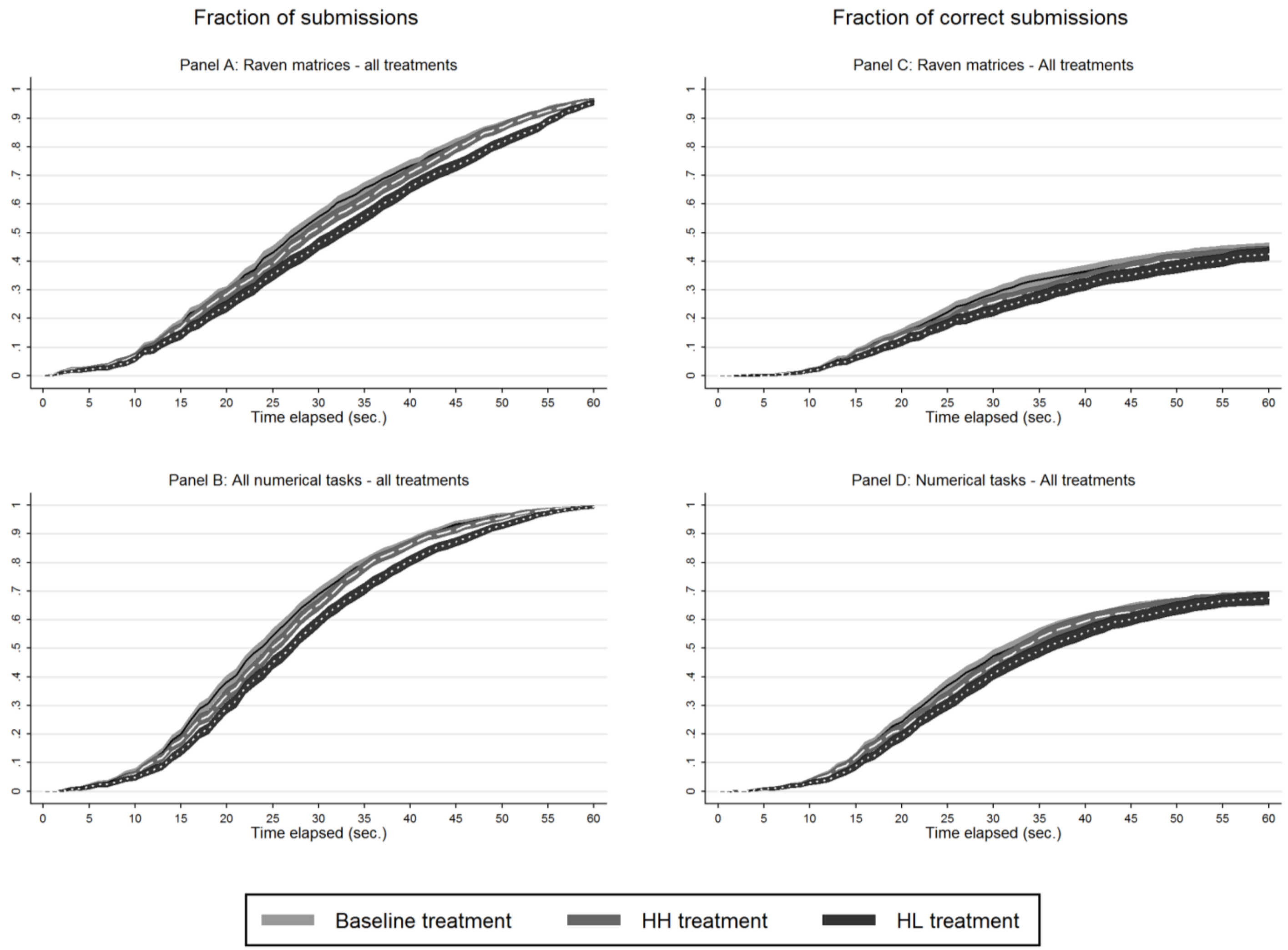
Note. Panel A shows a screenshot of a decision screen of a typical Raven matrix, taken from Carpenter et al. (1990, p. 407). None of our experimental Raven matrices is taken since the psychological tests are meant to be kept confidentially. The correct solution is the option with the green frame. Panel B shows a screenshot of a decision screen of a typical numerical task. The correct solution is the option with the green frame.

Figure 3. The Production Function across Different Incentive Schemes and Tasks



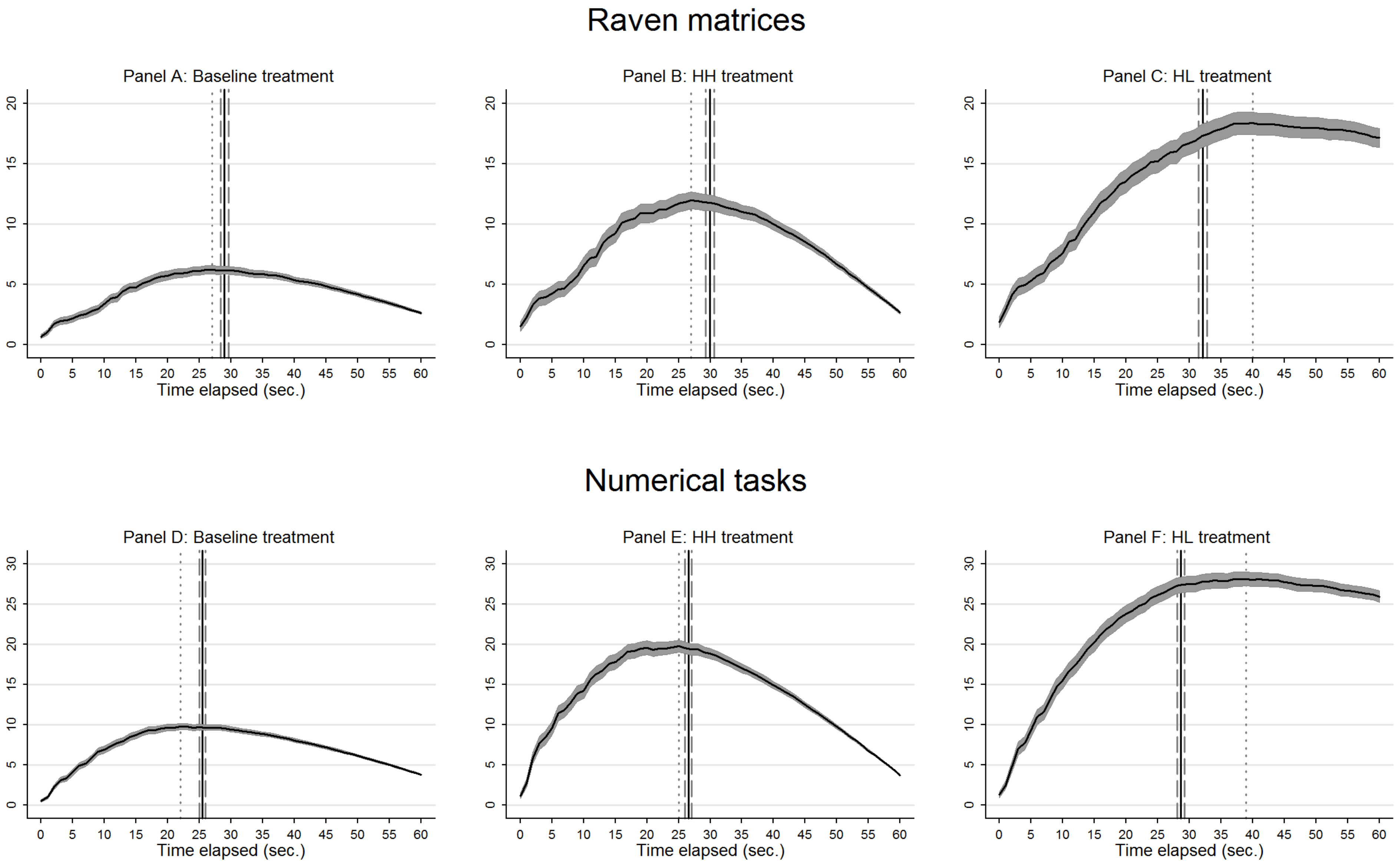
Note. Panels A and B report the probability of knowing the correct answer over time for the Raven matrices and the numerical tasks.

Figure 4. Performance on Raven Matrices and Numerical Tasks



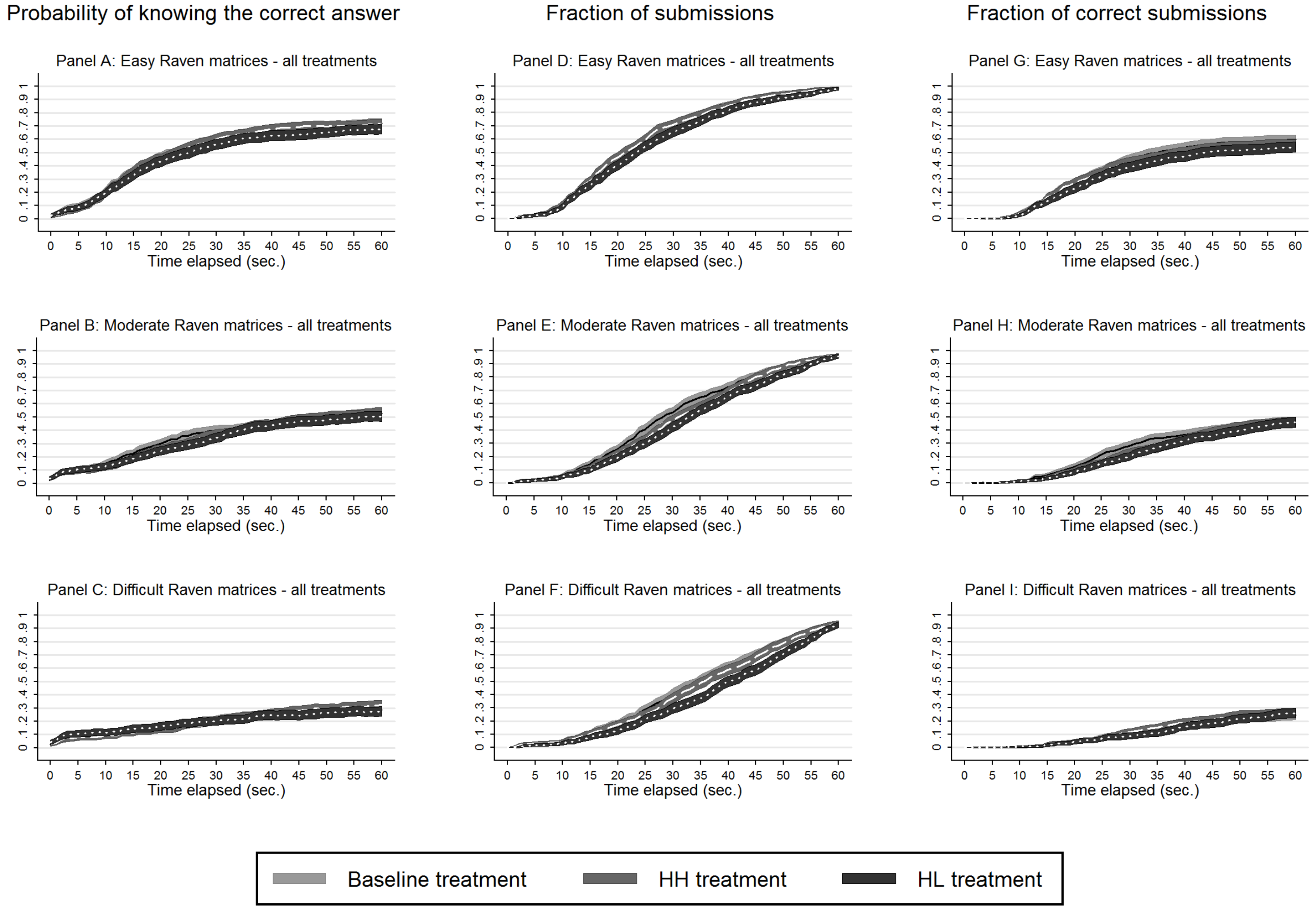
Note. Panels C and D show the cumulative distribution of submissions over time across all three treatments for both tasks. Panels E and F show the cumulative distribution of correct submissions. The grey areas indicate 95% confidence bounds.

Figure 5. Expected Earnings and Submission Times of Raven Matrices and Numerical Tasks



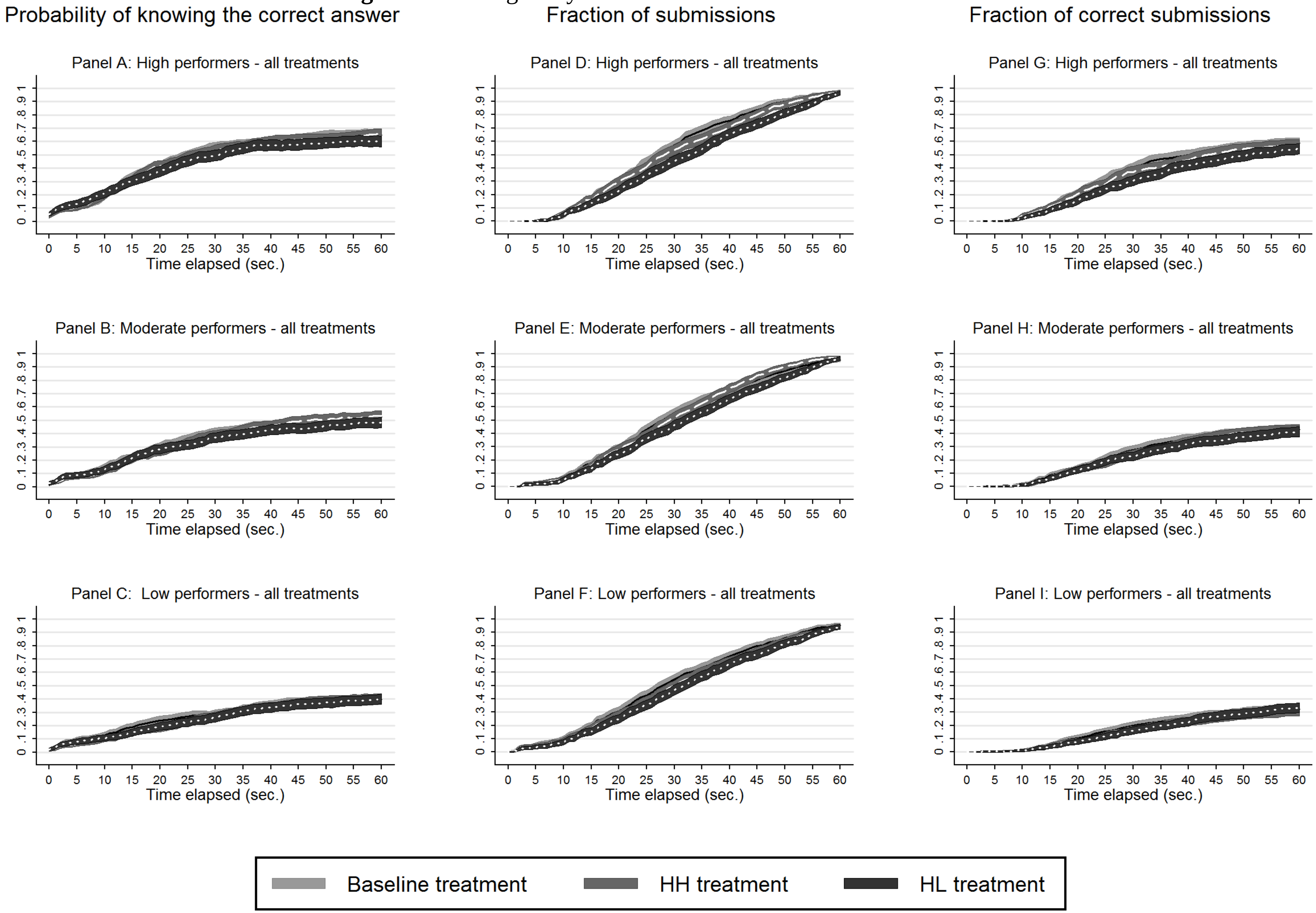
Note. The figure shows the expected earnings over time for Raven matrices (Panels A-C) and numerical tasks (Panels D-F) for each treatment. The grey areas indicate the 95% confidence bounds.

Figure 6. Heterogeneity in Questions between Raven Matrices



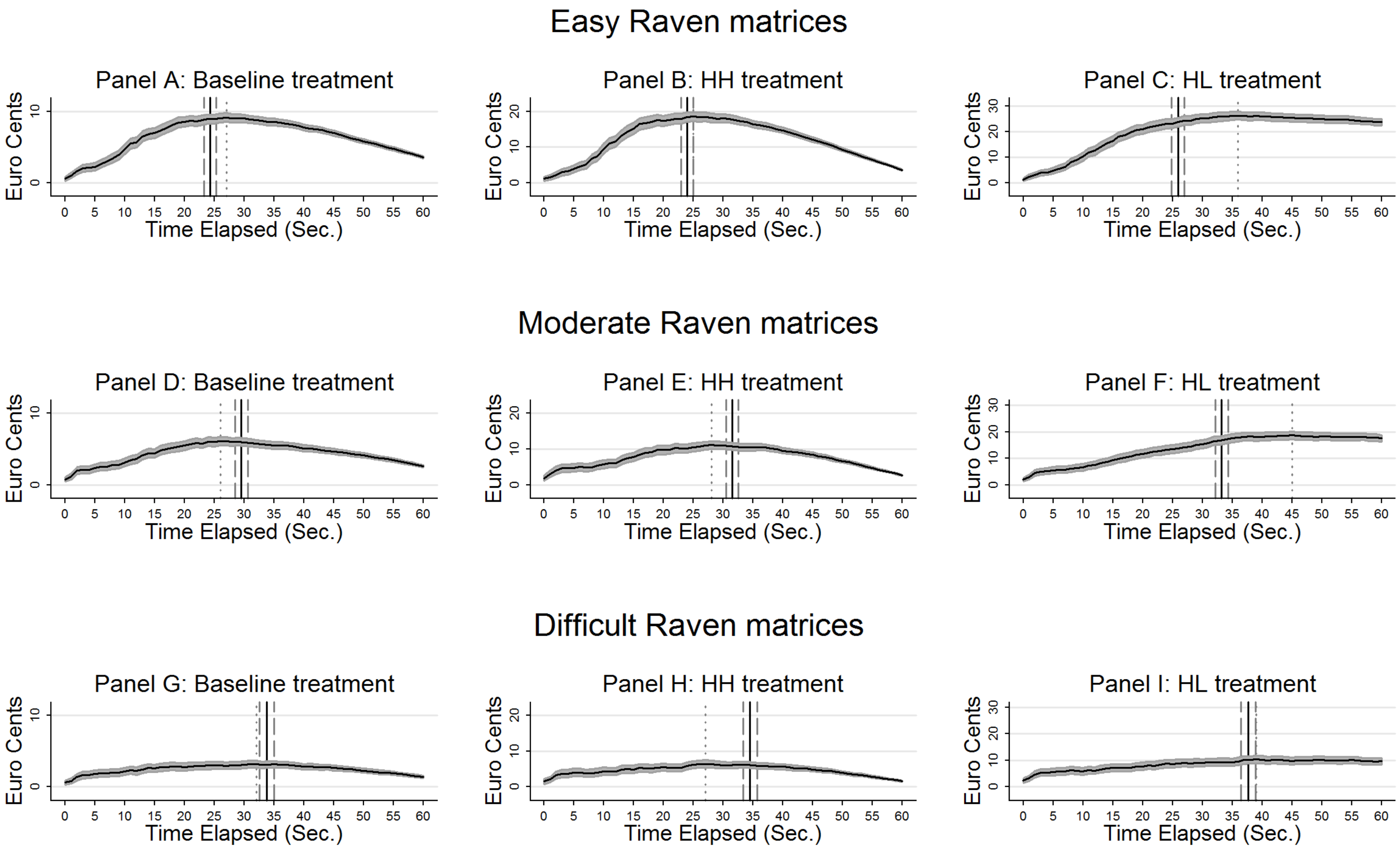
Note. The figure shows the probability of knowing the correct answer over time, the cumulative distribution of submissions, and the cumulative distribution of correct submissions of the Raven matrices for all three treatments separately. We split our data into three degrees of difficulties (easy, moderate, and difficult). The grey areas indicate the 95% confidence bounds.

Figure 7. Heterogeneity in Performance between Raven Matrices



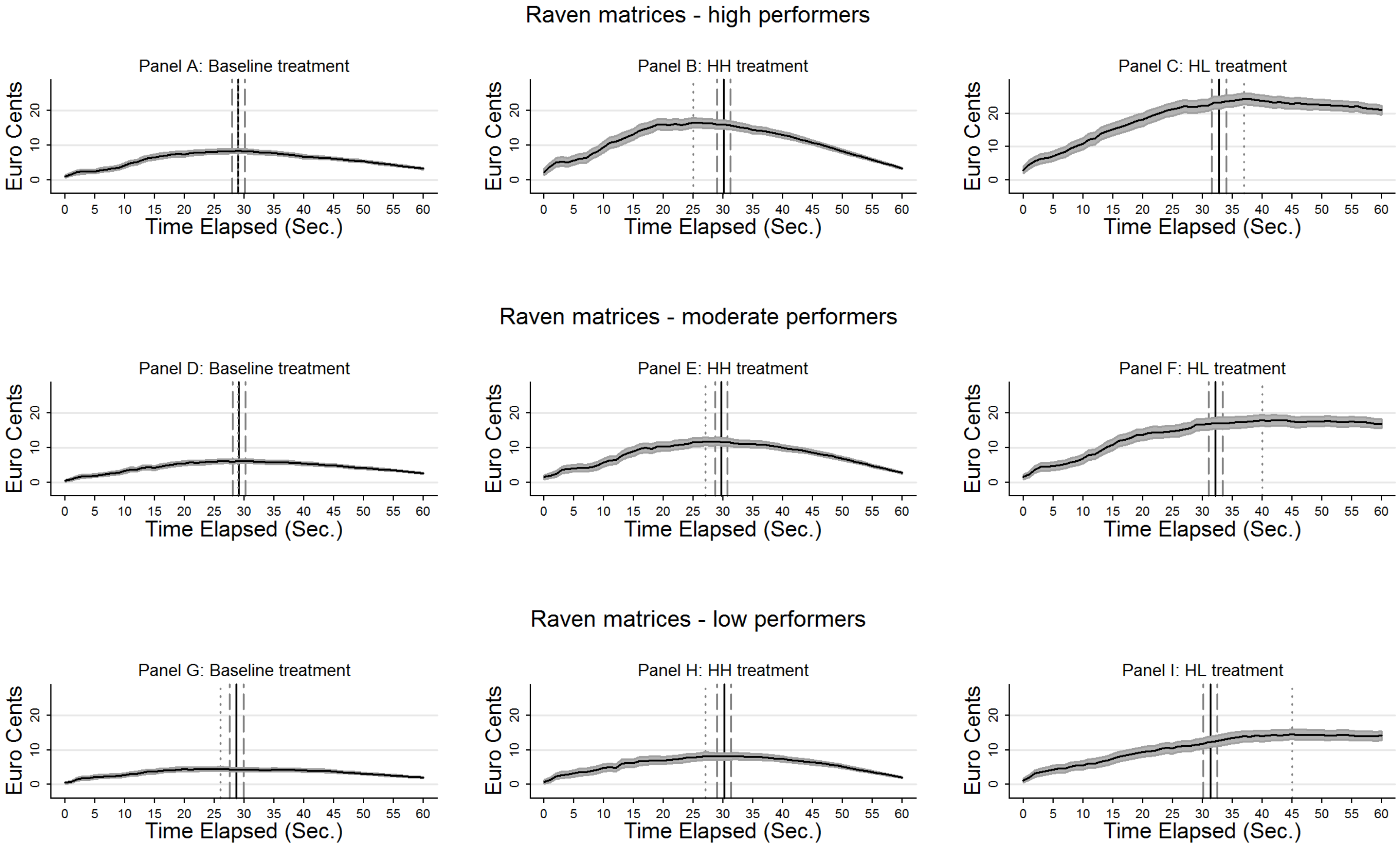
Note. The figure shows the probability of knowing the correct answer over time, the cumulative distribution of submissions, and the cumulative distribution of correct submissions for three performance types and all treatments. The grey areas indicate the 95% confidence bounds.

Figure 8. Expected Earnings and Submission Times of Raven Matrices by Degree of Difficulty



Note. The figure shows the expected earnings over time for Raven matrices. We split the sample into three degrees of difficulty. The grey areas indicate the 95% confidence bounds.

Figure 9. Expected Earnings and Submission Times of Raven Matrices by Performance Type



Note. The figure shows the expected earnings over time for Raven matrices. We split the sample into three performance types. The grey areas indicate the 95% confidence bounds.