ISSN 2282-6483

Alma Mater Studiorum - Università di Bologna

DEPARTMENT OF ECONOMICS

Managing the Workload: an Experiment on Individual Decision Making and Performance

Veronica Rattini

Quaderni - Working Paper DSE N°1080

Managing the Workload: an Experiment on Individual Decision Making and Performance

Veronica Rattini *

University of Bologna

Abstract

The present research investigates individual decision making regarding jobs scheduling, by means of a laboratory experiment based on the "Admission Test" of the University of Bologna, in which students have to allocate effort among several tasks in a limited timespan. The experiment includes three treatments that differ in the way the test is administered to participants: either with a fixed sequence of questions, or with a fixed time per task, or with no constraints. Results show large and significant heterogeneity in treatment effects. Constraints on the answering sequence or on the time allocation for each task improved the performance of those subjects who failed to efficiently allocate their effort among the tasks, whereas negative effects were found for students who were already good in self-organizing. The study has relevant policy implications for the organization of the workload in the labor force, when different types of workers are employed. Furthermore, important intuitions on the design of the university student-selection mechanisms are also discussed.

Keywords: Individual decision making, Performance, Gittins Index, Mouse-Tracking, Workload.

JEL classification: J20, C91, I20, D80, C80 .

*We would like to thank Maria Bigoni, Stefania Bortolotti, Andrew Caplin, David Cesarini, Vincenzo Denicolò, Guillaume Fréchette, Margherita Fort, Xavier Gabaix, Giovanni Ponti, Ernesto Reuben, Andrew Shotter, who gave useful advise in several discussions.
We are also grateful to the seminar participants at the University of Bologna, New York University, Royal Holloway, to the SONIC experimental group and to the London Behavioral and Experimental Group.
We would like also to thank the students of the Liceo Classico Statale Minghetti for their participation and the University of Bologna for the test's disclosure.

1 Introduction

This paper examines how individuals manage their workload when they have several tasks to solve in a limited timespan. In particular, the aim of the study is to understand if people are effective when choosing how to allocate their effort across tasks through time, or whether better performance could be achieved by restricting the individual's decision space.

Typically, economic models of individual production have assumed that agents choose optimally the cognitive operations for each working domain. However, the evidence described in some top-selling managerial books [Allen (2015), Covey (1991) and Crenshaw (2008)] and in the recent cognitive literature, Crawford and Wiers (2001), recognizes that decisions on workload management require cognitive operations, such as information search and processing, which involve costly and complex valuations, and agents often fail to choose such operations optimally.

The present project is interested in studying the individual decision-making process insofar as it affects productivity in a context of workload management. The fact that individuals may adopt simple heuristics in work organization, is an interesting feature of human behavior *per se*. Nonetheless, understanding the relation between the organizational strategies and the individual performance is essential for the economics of production. The current study experimentally evaluates subjects' behavior in the framework offered by the "Admission Test" to the faculty of Economics of the University of Bologna. In this context, individuals have to solve several tasks in a given timespan. In the baseline condition, students answer the test without any restrictions on their decision space, therefore deciding freely how to allocate the time across tasks and the sequence in which they want to answer. Two treatments are designed to test if restrictions on such space affect their final performance. In such cases, the test will be administered to subjects either with a fixed sequence of questions, or with a fixed time for completing each task.

The application to a field context adds interest in terms of policy evaluation for the case of study. In particular, the replication of the "Admission Test" of the University of Bologna will give the opportunity to test whether the design of such an assessment method has an impact on the selection of the students. In particular, since the current version of the test restricts both the timespan and the sequence of the sections, which are thematically clustered questions, this study will test whether such restrictions affect the students' performances and which type of individuals benefit or are at a disadvantage from such a test format.

Finally, the contribution of the present paper relies also on the adoption of a rich data-gathering method. A sophisticated software design has allowed the collection of comprehensive information on individual behavior using the *mouse-tracking* technique and, therefore, a complete map of the

solving strategy adopted by each student is revealed. Such information is indeed used to provide new insight into subjects' behavior and for refining the empirical analysis.

Results show that, on average, subjects performed better when they could freely choose how to manage their taskload. However, treatment restrictions significantly improve the performance of those individuals who have shown inefficient organizational skills. Such heterogeneity demonstrates that the sample composition would be crucial for interpreting the aggregate results in similar contexts of study.

The paper proceeds as follows: Section 2 revises the literature of interest, while section 3 describes the experimental framework and summarizes the theoretical predictions for such decisional context. Section 4 presents the design and the procedure of the experiment. Section 5 summarizes the results of the experiment and describes the robustness checks performed. Finally, section 6 offers concluding observations.

2 Literature Review

The importance of individual decision making in Economics has been emphasized since the breakthrough work of the Nobel laureate, Herbert Simon, who showed how the decision-making process of individuals might be rationally bounded in several domains [Simon (1955), Simon and Chase (1988)] and clearly pointed out the necessity of characterizing such a process, rather than relying on the Homo Economicus assumptions.

Even though, in recent decades economists started to take into consideration the findings from the psychological field related to the bounded rationality of individuals [for a review DellaVigna (2009)], there is still a lack of extensive studies in workload management and effort allocation. Traditionally, labour economists have defined effort as "the pace or intensity of work", Johnson (1990) and they have usually ascribed such input to individual rational choice, practically governed by economic incentives. However, recent studies on effort allocation and task prioritization have shown that subjects may not be so effective in choosing how to manage their workloads, even when they are incentivized to do so. We will now present and discuss some of those studies that have focused on such decisional domains directly related to the present paper.

It has been shown both empirically and theoretically that, when agents are left free to manage their work independently, real costs in term of performance emerge with respect to the potential output obtained under an imposed schedule of work. The paper most closely related to the present research is Buser and Peter (2012). The authors have examined, in a laboratory experiment, the effects of multitasking on the performance of subjects in two games: a word-search puzzle and Su-

doku. Their results show that, when subjects are left free to to organize themselves independently in the two games, they perform worse than those who face a given sequence of the two. Moreover, the evidence has suggested that forcing individuals to multitask is detrimental with respect to the sequential rationale. However, in this experiment, decisions on the time allocation are restricted and the duration of tasks are always imposed by the researchers. Given that this choice dimension is often encountered in problems of work division, and it could significantly influence the overall individual performance, this paper extends the analysis of Buser and Peter (2012), by considering contexts where time-management is not restricted. In addition, the conclusions provided by the same authors would be valid for the two-tasks case; the present research opens the set-up to the multiple-tasks framework, further enriching the analysis of such topics.

The evidence described in Buser and Peter (2012), that working in parallel on more than one task is detrimental with respect to a sequential working rationale, is further confirmed in the field by the empirical study of Coviello et al. (2015) and by the related theoretical framework in Coviello et al. (2014). In particular, the authors have analysed the work schedule of the Italian judges of the Labour Court of Milan. Judges who work in this Court, exogenously receive a stream of cases that they have to undertake and, as underlined by the authors, they usually work in parallel on more than one assigned project. The evidence has proved that such organization of case-load has a substantial negative impact on the productivity of the Italian judicial system. Moreover, in Coviello et al. (2014) the authors propose a model of dynamic production to prove that the optimal strategy of workload management in such domain would indeed follow a sequential schedule. In their model, the effort allocation is governed by the input rate, i.e the rate at which each task is began, and it is taken as exogenous or determined by co-workers. The present paper complements and extend their analysis by studying workload management in a context where such rate would be endogenously chosen by the individual.

As previously anticipated, limits on agent decisional process not only emerge for choices on the sequence of work, but also for choices regarding the time allocation among several tasks. Tice and Baumeister (1997) show in two longitudinal studies that students identified as "procrastinators", obtained lower grades than those students who do not procrastinate. In principle, there should not be any difference in performance in relation to whether the task is done far ahead of the deadline or just in time. However, the authors find significant differences in performance between the two types of students. As the authors suggest, procrastinators may outweigh the short-term benefit of the postponement and underweight the long-term cost of delaying, allocating, therefore, sub-optimal effort over time; or it may be the case that they simply and naively believe that such a postponement would improve their performance, while it actually constrains their available time.

In addition, it could be also that those who procrastinate lack a clear strategy for prioritizing tasks over time; that is, they have poor time-management skills. Which of these proposed explanations is more likely is still unclear and the present research aims to provide new insight into such question.

The above result is further confirmed by Ariely and Wertenbroch (2002), who proved in a field experiment that when students have to work on several tasks without any strict deadlines per task, they perform significantly worse than those who self-organize their effort through costly deadlines. Moreover, internally imposed deadline are not as effective as external ones in improving task performance. Such experimental findings confirm more robustly the evidence that individuals may not choose effectively the time allocation that would maximize their performance. In addition, the study shows that limiting subjects' decision space by externally imposing a fixed time per task, could be an effective way to raise overall output. Notice, however, that, with the exception of Buser and Peter (2012), the evidence discussed above focuses on tasks to be completed in a long time period. In order to limit future uncertainty regarding available time for task completion, the present study focuses on shorter tasks and on a context of complete information on the total available time.

Important innovations of the present research rely also on the adoption of an advanced data gathering method, the *Mouse-tracking* technique.[1] This method consists of tracking the mouse-click of each subject during the experiment and recording the sequence and the duration of each step in the proceeding working strategy. Due to its completeness and its richness of information acquisition, this method has been widely adopted both by the psychological and economic literature [see for example Arieli et al. (2011), Gabaix et al. (2006) or Johnson et al. (2002)]. By applying such a sophisticated methodology, this study could extend the previous results of the literature, by describing the choices, the sequence and the duration related to each mouse-click and the implicated use of time.

To conclude, the purpose of the present study is to improve our knowledge on the understanding of individual decision making in workload management, by providing evidence from a laboratory experiment, which relies on a field framework where such a choice domain is present. In particular, the current research aims to investigate further the importance of both the time allocation and of the choice of the sequence of work in a context with multiple tasks and with complete information on the total available time, reinforcing and extending the findings raised by the above reviewed literature.

[1] Firstly introduced in Johnson et al. (1989).

3 The Framework

3.1 The TOLC Test

The experiment is based on the format and on the tasks proposed by the electronic version of the Admission Test to the Faculty of Economics at the University of Bologna.

The test is called "TOLC" and is provided by a private agency, i.e CISIA [2]. This test was introduced by the University of Bologna in 2013, changing from a paper-based format to the CISIA electronic version. The importance of such implementation is not only related to its implications for the admission of candidates, but also for its extensive coverage among Italian universities. In 2014, in fact, about 77 universities have decided to adopt such test, namely almost 90% of the national total [3].

The TOLC test is composed by three sections: logic, verbal comprehension and mathematics. Sections are clusters of questions that are focused on the same field of knowledge. The total number of questions is 36, divided into 13 questions of logic, 13 of mathematics and 10 tasks of verbal comprehension. The tasks have a multiple-choices format and there is only one correct answer. The total available time to answer all the questions is one hour and thirty minutes, which is uniformly distributed across sections, i.e 30 minutes per section (not combinable). Moreover, students must follow the order of the sections imposed by the format: logic, verbal comprehension and mathematics. This implies that switching from one section to the others is irreversible, so it is not possible to return to the previous thematic area.

As explicitly stated in the instructions provided by the agency, the logic and the verbal-comprehension tasks do not require previous training or particular skills. The mathematical questions cover the program encountered during the first four years of high school. The questions in the verbal-comprehension section are related to two different essays which students have to read in order to answer, always having the possibility to return to the text.

In the figure below, a screen-shot of the test is shown in order to better clarify the structure of the "Admission Test". From figure 1, it is possible to distinguish: A) the three sections' buttons; B) the text of the first question with the related answer options; C) the buttons corresponding to the other questions in the section; D) the time bar, which scrolls down as the time goes by.

The above description of the test's format aims to underline that the "Admission test" offers a good opportunity to answer the research questions of interest. In particular in this context stu-

[2] For more information: http://www.cisiaonline.it/area-tematica-tolc-cisia/home-tolc-generale/
[3] List from the University of Bologna.

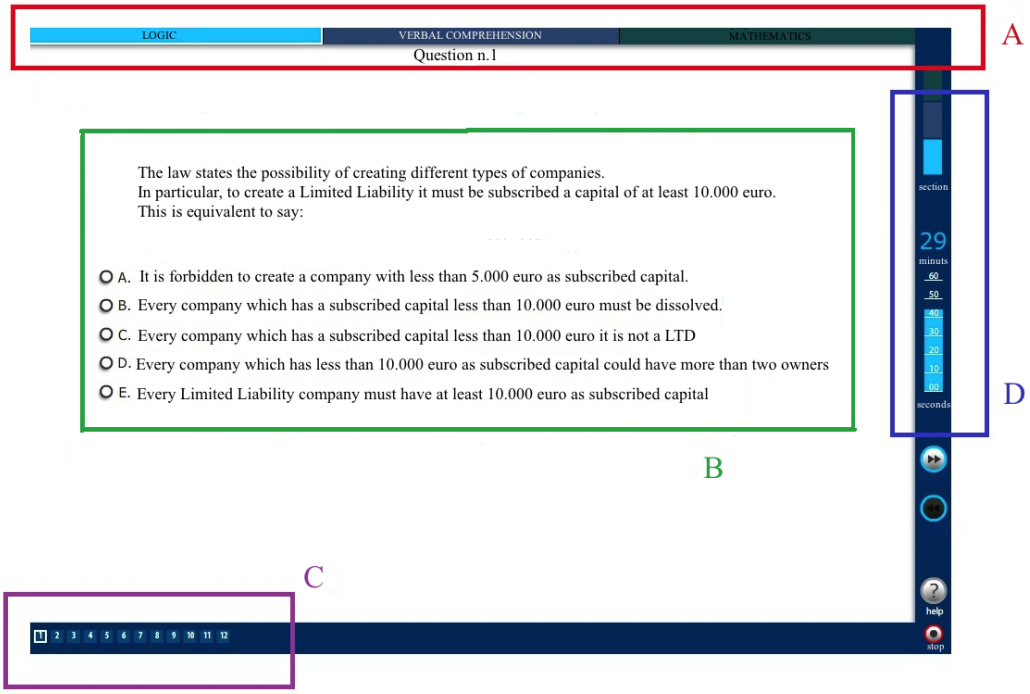


Figure 1: The Admission Test

dents face a problem of workload management, since they have to solve several questions in limited available time, with the possibility to choose how to allocate time across tasks and in which sequence they want to answer. Moreover, differently form previous studies, here there is no uncertainty about the total available time, questions are short and there are more than two tasks. For all the above reasons, the TOLC test would be an ideal set-up for studying decisional problem in workload management and, indeed, to extend and refine previous findings of the literature.

This field framework is also interesting due to the possibility of retrieving evidence useful for the university. Understanding how such new implementation may have impacted the performance of students could have relevant policy implications for the design of the student-selection mechanism.

Finally, the electronic format of the TOLC has also brought the advantage of replicating such environment into the laboratory without inducing any framing effects and overcoming the typical difficulties of recreating the field conditions in laboratory experiments.

3.2 The Experimental Framework

The experimental framework is based on the TOLC test format previously described. In particular, students have to answer a similar number of questions, 34 tasks, which are distributed across the three sections as in the field case. Moreover, no changes are adopted in respect of the total available time, which would be exactly 1 hour and 30 minutes.

The important feature of the experimental version of the test relies on the data collection. In particular, since the objective of the study is to understand how individuals manage their workload, it is fundamental that the researcher observes all the steps that have brought the subjects to take each action. In order to meet the above objectives, we adopted the *mouse-tracking* technique. The mouse-tracking method was firstly used by Payne et al. (1993) and now it has been accredited by a wide set of scientific domains. As previously described, its implementation on this framework has not only the advantage of mapping completely the decision process of individuals, but also of retrieving interesting and unique measures, principally helpful in setting the difference between the present research and previous literature.

The replication of the test and the implementation of the tracking technique were performed using the Ztree software, Fischbacher (2007). In particular, during the experiment, all subjects enter a computer laboratory where they have an isolated workstation with a laptop and a connected mouse. When the test starts, the Ztree software begins to record all the information relating to the click of the mouse used by each experimental subject. In particular, measures on the timing and on the space where the click is located are recorded in the database. This information allows the recording of the time spent by each subject on answering specific questions, the number of different answers to the same question, the adopted sequence of work, or the times they looked at the same question before answering, etc.

The picture below shows a screen-shot of the terminal during the experimental session. It should be noted that the Ztree software allows for perfect replication of the TOLC test framing, since both the structure and the interface were almost equal to the the field case: see figure 1.

The main distinction between the experimental framework and the field case relies on the type of incentives that students face. In particular, while the TOLC test has implications for students' admission to the undergraduate program in Economics at the University of Bologna, the experimental replication provides monetary incentives. Notice, however, that, even if the rewards are different, the payment scheme adopted in the experiment perfectly replicates the TOLC evaluation's scheme. Students, in fact, receive 8 euro as a show-up fee, and the residual part of their earnings are calculated as in the field case: 1 point for each right answer, 0 points for missing and -0,25 for incorrect

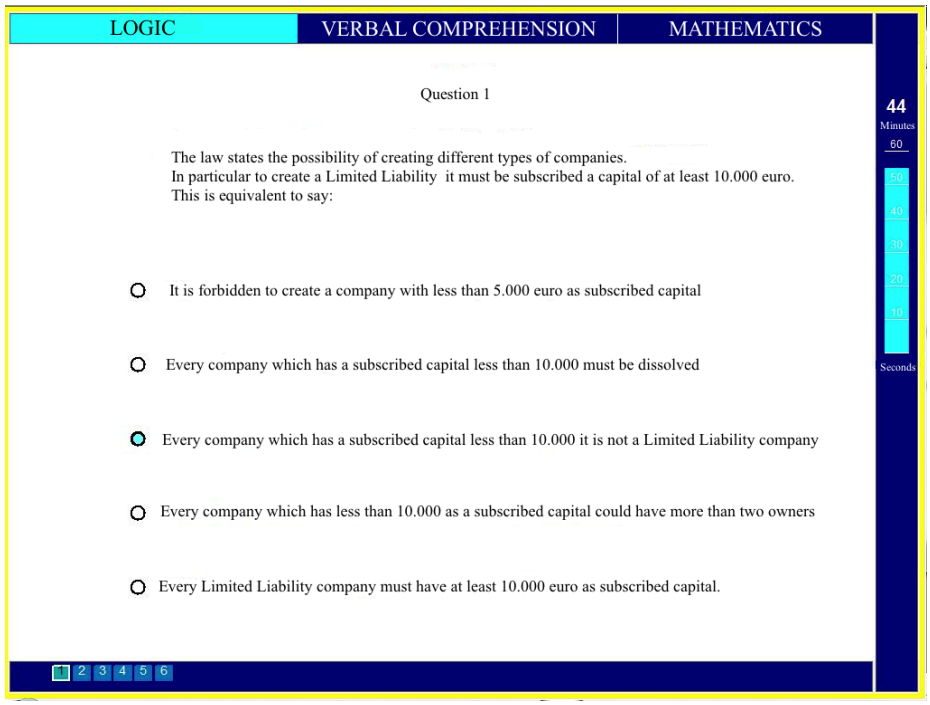


Figure 2: Ztree Replication

answers. The final score will be expressed in points, which are the Experimental Currency Units and they will be converted into euro at the following rate: 2 ECU = 1 euro. This conversion rate implies that the maximum payment would be 25 euro while the minimum is 3.75 [4].

3.3 The Theoretical Predictions

In this framework, students have to decide how to allocate their effort among several tasks, in order to maximize their final payoff, which is given by the sum of the rewards obtained by completing each task. Specifically, at each point in time, each individual has to decide which task should be processed first (*sequence of work*) and how long they should stay on each question (*time allocation*). Typically, this problem is known in the literature as the "*Exploration-Exploitation Dilemma*", where a subjects has to decide how much she wants to search for the best option (*Exploration*) and how much she wants to exploit the current choice (*Exploitation*), given that she has several alternatives to select from (all giving the same reward on completion) and that each option has a random completion time. The so called Gittins Index is the dynamic allocation index that was firstly proposed

[4] These amounts include also the 8*e* show-up fee. The fee has been chosen in order to more than cover the eventual loss from answering all the 34 questions incorrectly, the minimum earnings is, in fact, 3.75*e*.

in Gittins and Jones (1979) to solve the above dilemma. It was initially presented for the "Multi-Armed Bandit" process and then for the "Job Scheduling" problem. In the following paragraphs, the attention is focused on the scheduling problem, given that is the dilemma that most closely resembles the present decisional domain.

In the "Job Scheduling" problem, each agent has a queue of jobs to process and she has to decide how to organize the work in order to minimize the mean number of remaining jobs in the queue. Each job yields a positive reward on completion, which, in turns, happens at random times. Specifically, this is defined by $Ve^{-\gamma t}$ the present value of the reward on completion, V, at time t of a generic job and denote by $F(\cdot)$ the distribution of its service time, with the density given by $f(\cdot)$. The Gittins Index for each job is given by:

$$\nu(x) = \sup_{t>x} \frac{V \int_x^t f(s)e^{-\lambda s}ds}{\int_x^t [1 - F(s)]e^{-\lambda s}ds} \tag{1}$$

where the index represents the probability that the job will be completed within $t - x$ of additional service time, given that it was not completed before, times the reward on completion V.
As explained in Gittins and Jones (1979), the optimal policy is to process, at each point in time, the jobs in order of decreasing indexes $\nu(x)$. In addition, the authors showed that such a policy would be highly simplified for specific families of service time distributions. In particular, they looked at the hazard rate of each job, which is defined as $\rho(s) = f(s)/[1 - F(s)]$, and if jobs belong to the *Decreasing Hazard Rate - DHR* class, it would be optimal to process them in order of least attained service, i.e *Foreground-Background (FB)* policy. In contrast, if the service time distribution belongs to New Better than Used in Expectation - NBUE, as for the case of *Increasing Hazard Rate - IHR*, the *First-Come-First-Served (FCFS)* discipline minimizes the number of jobs in the system (and consequently minimizes the mean delay). Aalto et al. (2009) extends this framework to the class of non-monotonic service time distributions and discusses how the optimal policy would change in such cases. In particular, they consider the case where the hazard rate is piecewise monotonic (firstly increasing and then decreasing), and they showed that the optimal policy would be FCFS+FB(θ^*). Specifically, the optimal schedule would be to serve non-pre-emptively (i.e without switching) the jobs (which initially have received no service) until their attained service reaches θ^*. Jobs with attained service greater than θ^* are served according to FB, and jobs with no attained service have priority over those with attained service of at least θ^*. The proof relies on the assumption of no-crossing between jobs' hazard rates and on the fact that each job receives attention as long as its hazard is increasing, i.e each job receives its own optimal quantum of service θ^*. After this optimal quantum has been reached and each job hazard has started to decrease, jobs are processed in

order of least attained service (FB), starting therefore from the one which has received less service time to the one who has been worked the most.

Such predictions imply that the optimal policy would depend on the distribution of the job hazard rate. In this work, we leave this unrestricted and we will explore empirically what the data tells about the hazard distribution of this kind of jobs, and we will discuss the related implications in terms of optimal behavior in section 5.

4 Experimental Design and Data

4.1 Treatments

The present experimental study is divided into three stages. During the first baseline stage, all the subjects face half of the total questions, i.e 17 tasks, whereof six are from logic, five from verbal comprehension and six from mathematics, with half of the total available time, i.e 45 minutes. Notice that this distribution of questions is in complete accordance with the field case, with the exception that we have excluded two questions to maintain symmetry across stages. In this stage, students proceed to solve the test without having any constraints either on the sequence or on the time per question. For the sake of simplicity, we will define this first stage as "Unconstrained".

After this first part has been completed, subjects will be randomly allocated to three treatments: "Unconstrained", "Fixed Time" and "Fixed Sequence". In the first treatment, subjects answer the remaining 17 questions of the test, again without any restrictions, as in the first stage. In the Fixed Time treatment, they instead answer the remaining questions with a given time per question, which is calculated by dividing the total available time by the number of questions, i.e 40 minutes / 17 tasks = 2 minutes and 22 seconds. Notice that, in this calculation, five minutes were dropped, since this was the maximum time available for reading the text in the verbal comprehension part (this upper bound was computed by considering the registered maximum value from the pilot session). Moreover, in this treatment, when subjects change question, the timer for the switched question would stop and it would re-start as soon as the subject returned to that task.

Finally, a third group of subjects is allocated to the Fixed Sequence treatment. In this case, subjects answer the same remaining 17 questions of the test following a given sequence of the sections, the same as the one provided during the actual admission, i.e first logic, then verbal comprehension and finally mathematics. Moreover, within each section, subjects have to follow a given sequence of questions, following the ascending numeration of the tasks (question 1, then 2, then 3, etc...). Notice that, since the Fixed Sequence treatment has the purpose of denying the subjects the opportunity

to choose the answering sequence, if subjects switch questions, they will not have the possibility to come back to the switched task again (the button related to the changed question would disappear from the screen).

Finally, in the third stage of the experiment, subjects answer a general questionnaire that collects information on gender, birth, residence, risk-preference and impulsivity [measured by the "Cognitive Reflection Test", see Frederick (2005)]. After having completed this final stage, subjects are paid and then leave the experimental session.

The design is summarized in the following table.

Table 1: Design

	Baseline Treatment	Sequence Treatment	Time Treatment
Stage 1	Unconstrained	Unconstrained	Unconstrained
Stage 2	Unconstrained	Fixed Sequence	Fixed Time
Stage 3	Questionnaire	Questionnaire	Questionnaire

Notice that, contrary to the TOLC, the baseline format of the test implies that students do not have a fixed time or a fixed sequence for each section. The reason for this discrepancy relates to the aim of the current research. In particular, the main focus of this study is to understand how individuals *freely* behave in such context, without constraining any aspect of their choice space. For this reason, not imposing any time or order restrictions during the baseline treatment of the experiment, allows us to observe unconditional behavior and specifically to test what are the effects of imposing such bounds on performance. In addition, the findings from this baseline design help in understanding the effects of the current format restrictions of the TOLC test, with respect to the previous unconstrained paper-based test, on applicants' performance and on their admission into universities.

4.2 Procedures

In the present section we will describe the composition of the subject pool and the details related to the experimental sessions.

For the participation in this study, we recruited students currently enrolled at the fifth-year of high-school. At this grade, students might decide to apply for university in few months representing, therefore, a potentially highly interested subjects pool. The recruitment of such pool was conducted in an Italian high-school, located in Bologna, named "Liceo Classico Statale Marco Minghetti".

In accordance with the director of the school, all the students enrolled at the fifth grade by October 2015 were allowed to participate in the present study. The experiment took place at the BLESS Laboratory (Bologna Laboratory for Experiments in Social Science) of the Department of Economics in Bologna from 24th of November to 5th of December 2015.

The recruitment was done in two phases. Firstly, for each enrolled student we collected a form, where contacts (personal e-mail, mobile and home telephone number, address of residence) were provided. In this occasion, students have also to state the preferred mean of communication for receiving details about the date and place of their future participation into the study. After having recorded the above information, students where randomly allocated to treatments. Finally, for each treatment, three dates where offered for participation (each option befall in a different day of the week: Monday, Tuesday, etc.). With such procedure, each student received invitation according to the preferred mean of communication and he/she was asked to choose one of the proposed dates [5].

The experiment started with an introduction explaining the rules of the first stage and with a practising example. Subjects stay in the laboratory for approximately 1 hour and 45 minutes and the average payment was of about 15,30 euro, which is in line with the average salary potentially earned for the same timespan by an high-school student in Italy.

Finally, it is important to notice that the choice of this specific subject pool has brought some important advantages. In particular, the fact that such students will apply to University in few months has probably pushed their participation by means of their intrinsic motivation for experiencing the test. In this way, the standard problem of the saliency of incentives in laboratory is potentially reduced. Moreover, active participation in this study was also enhanced by the fact that typically the training for such kind of test is provided by manuals and books which offer just indicative exercises. While, in this replication, subjects could face the actual difficulty of the TOLC test, since the proposed questions are exactly the ones of the last admission wave.

For the above reasons, even if subjects are incentivized through monetary rewards (following a pay-for-performance scheme), the time at which they are recruited and the type of questions used, suggest that students' effort is further enhanced by means of their intrinsic motives.

[5] The participation rate was of about 65% since 87 subjects participate, out of 134, in one of the 9 sessions.

5 Results

In the previous section, the general theoretical model that is commonly used in the literature to address the Job Scheduling from an optimal perspective was presented.

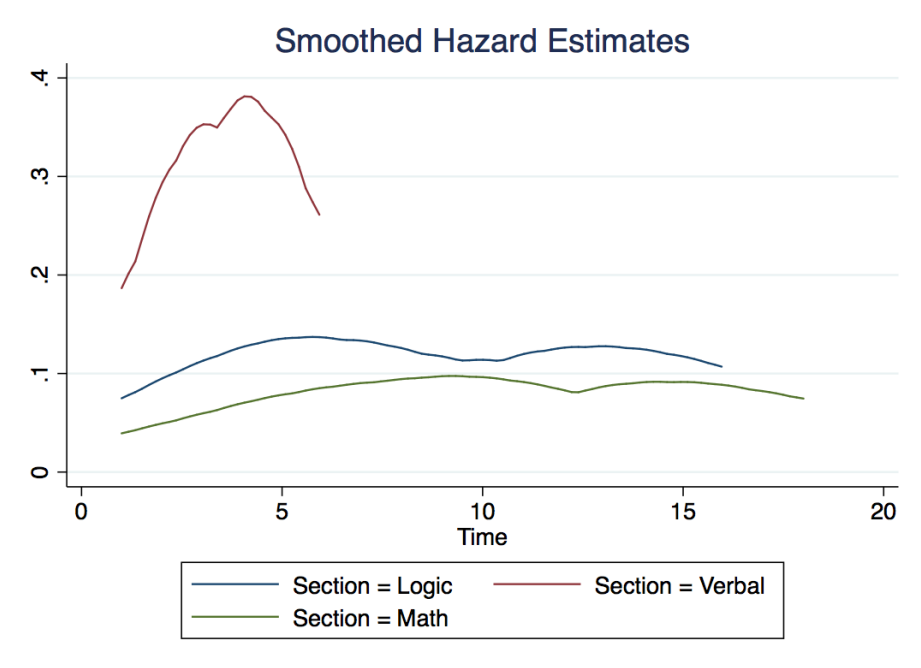
In this section, there will be a discussion of how such reference model of behavior may be applied to the current case and what, then, we should expect from the treatments' effects.

As in the Job Scheduling problem, in this experiment subjects are incentivized to maximize the sum of their final rewards and, at each point in time, they have to decide which task they want to undertake and for how long they want to work on that task. Moreover, additional similarities emerge if we compare the characteristics of the experimental tasks with those of *Jobs*. As for the *Jobs* case, rewards of tasks occur on completion (which, in this case, coincides with the provision of the correct answer), and each task has a random completion time, which is unknown beforehand, so that students do not know the total time needed for completion and they just observe the amount of service received so far. Given that the similarity to the Job Scheduling problem is quite marked, we believe that we can use the framework provided by the *Dynamic allocation index* to understand how students should proceed in the test.

Remember that, as anticipated, the optimal policy would also depend on the characteristics of the distribution of the service time for the tasks under consideration. In particular, we have seen that different rules apply, depending on the monotonicity of the job's hazard rate. Thanks to the data gathered through the *mouse-tracking* technique, we firstly retrieved how the average hazard rate is distributed over time for the case of our experimental tasks and then discussed the prescribed policy and how it is specifically applied in this context.

As may be noticed from picture 3, the average hazard of giving the correct answer initially increases and then, after some critical time, it starts to decrease for all the thematic cluster of questions. Moreover, it seems that some heterogeneity exists across types of questions: while both the logic and the math questions have, on average, flatter and more disperse hazards, the hazard of the verbal questions seems to be steeper and shorter. This evidence suggests that, while, on average, it took students longer to answer the logic and the math sections, without greatly increasing the chances of correctly completing the tasks, students were faster and more accurate in answering the verbal comprehension questions. Notice that, if we think about the context of the "Admission Test", it is intuitive to imagine that the hazard rate for this kind of task has such behavior. In particular, it is plausible that, for each task, initially, there is an increasing chance to give the correct answer but, if, after this initial stage the subject does not provide the answer, it is unlikely that she will know it afterwards.

Figure 3: Average Hazard Rates for Logic, Verbal Comprehension and Math questions



Such characterization of the *hazard rate* in our context will help in understanding what the optimal policy would be for the "Task Scheduling" problem. In particular, the results from Aalto et al. (2009) suggest that, since the hazard rate is piecewise monotonic, students should follow a FCFS+FB(θ^*) policy, where θ^* represents the critical amount of processing time, i.e the optimal quantum Δ^*, that the agent should allocate to each task. Notice that such an optimum would eventually change according to the specific hazard rate of each task and for each individual. In this sense, we do not exclude the possibility that the hazard rate might be heterogeneous, both within subjects and across tasks and across subjects within question. However, since we assume no-crossing between individual specific hazards, the Gittins rule, which assigns specific (optimal) quantum of time to each task, will still define what the optimal (individual) strategy should be.
Below we describe how the FCFS+FB(θ^*) policy would be applied in such context.
At the beginning of the experiment, subjects do not have any information about the exact completion time of each task. What they know is that they have to solve 17 tasks in 40 minutes, that each task is valued the same (1 point) if the correct answer is provided, and that the reward of one task is independent from those of the others. In addition, each question has a multiple-choice format with five available options. With such information set, at time 0 individual i's expectation on the average probability of getting the right answer for task j is equal to 20%: that is the probability of getting one correct answer out of five options.

Under this relatively mild assumptions, we can now describe how the agent should proceed in the "Task Scheduling" problem.

When she starts the test by opening the question of the first task, she will learn how the hazard rate for that task is distributed. In particular, after the first unit of time spent on the task, she receives information on what is the current probability of giving the right answer, given that she has spent one unit of time on that task and in that time she didn't answer (he learns the *hazard rate*). After this first unit of time has been served, the agent will continue to exploit the same task if the probability observed is higher than the average probability of correctly answering the question, which is equal to 20% [6]. If this is the case, then she will stay on that task as long as the hazard increases, and, she will stop as soon as the hazard has started to decrease. Such optimal length of time for continuing working on the task is indeed the one defined by the Gittins quantum policy.

Therefore, if the agent chooses to work on the first task, she will continue either until she completes the task or the hazard for that task has started to decrease. After having processed the first task, the individual switches to the next question and the same reasoning applies. Then, she switches to the third question and so on, until all the questions in the queue have been viewed. At this point, some of the tasks will have been completed (namely, those whose answer was given within the optimal amount of time allocated), whereas some others would be still undone, namely, those that have been postponed or those which were not completed within the time θ^*. The FCFS+FB(θ^*) policy rule prescribes that, from this moment on, jobs with no attained service have priority over those with attained service of at least θ^*. The agent, therefore, will look at the postponed task, in order of decreasing Gittins Index, and she will assign to each one the optimal quantum of service time. The process continues until either the time expires or all the remaining questions have reached the point where their hazards have started to decrease and more time is available. In the former case, the decision process terminates, whereas, in the latter case, the agent should continue to process the tasks in order of increasing service time received, namely from the youngest to the oldest, i.e *Foreground-Background (FB)* policy.

Notice that, even if we can trace all the resolving strategy of each subject, we do not have individual information on the subjective hazard function for each task. Even though we lack such data, interesting conclusions about the individual behavior in this framework may still be drawn by using the FCFS+FB(θ^*) as optimal reference.

The first prediction that arises from the FCFS+FB(θ^*) policy is that the average score obtained

[6] This reasoning relies on the fact that, at time zero, the agent's belief on the probability of getting the correct answer is equal to 20% for all the unseen tasks. So, if after the first unit of time the agent learns that the probability of getting the correct answer for the current task is lower, she will change the task, since more profitable options are available.

by those subjects who have faced the second part of the test under the two treatments' conditions is expected to be lower than the score obtained by those who have faced the test without any restrictions. Such a result arises from the fact that, when students can freely organize their work, they will select and prioritize tasks over time, postponing the difficult questions to the end of the test and allocating to each task the optimal quantum of service time, which by no means has to be equal to the average time per task. For these reasons, we expect that both the constraints imposed by the treatments will prevent subjects from following this strategy and, therefore, will reduce their performance levels.
In the following section, we proceed by firstly presenting summary information on the subject pool and then discussing the treatments' effects.

5.1 Difference-in-Differences

The sample consists of 87 subjects. As explained in section 4, subjects were randomly allocated to three treatments. The summary statistics and the sample balancing is shown in tables A1 and A2 of the Appendix A. Table A2 shows that there are about 30 subjects per treatment group and, thanks to the random allocation, there are no significant differences in all the relevant measured characteristics among the three treatments' groups. In particular, even if different percentages of individuals have taken part in some economic courses across the three groups, the Wilcoxon-Mann-Whitney test fails to reject the null hypothesis that the three samples come from the same population[7]. The same hypothesis is not rejected when we test for difference in risk attitude, in impulsive behavior and in the ability of subjects across groups.

This confirms that the allocation of subjects into the three groups was random in respect of all the above characteristics.

In the following analysis, we will present results of treatments' effects on performance, by looking at the score obtained in solving the tasks of the "Admission Test". In particular, the output of interest is computed by summing up all the points for correct answers and by subtracting the total penalty for the wrong ones.

In table 2, the average score obtained in the two parts of the test is shown both for the Baseline Treatment group and for the Fixed Sequence Treatment group.

From table 2, it is possible to see that, statistically, there are no significant differences among the two groups in the scores for part 1 of the test, suggesting that the two groups do not differ in

[7]Notice that, in order to test whether the random allocation to treatments worked, we have used the Wilcoxon-Mann-Whitney test, since it tests specifically for differences in median values of the three samples, without imposing the normality assumption on the distribution of the outcomes.

Table 2: Average Score for Baseline Treatment and Fixed Sequence Treatment groups

	Part 1			Part 2			
	Baseline	Fixed Sequence		Baseline	Fixed Sequence		
Outcome	Unconstrained	Unconstrained	Δ_1	Unconstrained	Fixed Sequence	Δ_2	DID
Score	10.091	9.741	-0.351	10.667	10.205	-0.462	-0.111
Std. Error	0.684	0.444	0.778	0.684	0.551	0.778	1.100
t			-0.45			-0.59	-0.10
P>t			0.653			0.554	0.920

Notes: *Score* is the outcome variable, which sums the points obtained from the correct answers minus the penalty encountered from the wrong answers. Δ_t represents the average difference in the scores, in each part t, faced by the subjects in the groups "Baseline" and "Fixed Sequence". *DID*: represents the Difference-in-Differences estimate, which is given by $\Delta_t - \Delta_{t-1}$.
Legend: On the left, results related to the fist part of the test, $Part_1$. On the right, results related to the second part of the test, $Part_2$.
Significance levels: $*p < 0.10, **p < 0.05, ***p < 0.01$

terms of "ability". In the second part of the test, students from the Fixed Sequence Treatment have to solve the test through facing a fixed sequence of the questions, as explained in detail in section 4. Even in the second part of the test, the two groups obtained similar results suggesting that, in this context, solving the test through a sequential rationale does not affect individual performance.

In table 3, the results of the Fixed Time Treatment group are presented.

Table 3: Average Score for Baseline Treatment and Fixed Time Treatment groups

	Part 1			Part 2			
	Baseline	Fixed Time		Baseline	Fixed Time		
Outcome	Unconstrained	Unconstrained	Δ_1	Unconstrained	Fixed Time	Δ_2	DID
Score	10.091	9.698	-0.393	10.667	7.147	-3.520	-3.127
Std. Error	0.6854	0.518	0.611	0.684	0.628	0.991	1.164
t			-0.64			-3.550	-2.68
P>t			0.523			0.000***	0.008***

Notes: *Score* is the outcome variable, which sums the points obtained from the correct answers minus the penalty encountered from the wrong answers. Δ_t represents the average difference in the scores, in each part t, faced by the subjects in the groups "Baseline" and "Fixed Time". *DID*: represents the Difference-in-Differences estimate, which is given by $\Delta_t - \Delta_{t-1}$.
Legend: On the left, results related to the fist part of the test, $Part_1$. On the right, results related to the second part of the test, $Part_2$.
Significance levels: $*p < 0.10, **p < 0.05, ***p < 0.01$

As previously explained, the Fixed Time Treatment group has to solve the first part of the test with no restrictions, while, in the second part, students in this group are forced to answer all the questions with a given time per task, i.e 2 minutes and 22 seconds. Notice that, in order to avoid any confounds, decisions related to the sequence of questions are not restricted in this treatment and, therefore, subjects could still choose to switch task. Moreover, if students decide to change question, they will not lose any of the remaining time for the switched task, since the time would be stopped exactly at the moment of the click on the new question.

The aggregate results show, in table 3, that imposing a given time per task, in this context, is significantly detrimental to the overall performance. In particular, it reduces the test score by more

than 3 points (t statistics = -2.68), which corresponds to the 18% of the average score [8].

As expected from the theoretical predictions, we found a negative and significant effect of the Fixed Time treatment for the overall performance, but small and not significant results have instead emerged for the Fixed Sequence treatment.

In order to enrich the analysis and better characterize the results, similar estimations were made by taking into consideration gender' differences. However, no significant results emerged through performing such sub-group analysis.[9]

5.2 Difference-in-Difference-in-Differences

In the previous analysis, we tested whether the predicted average treatment effects were indeed detected in the current experimental context, and we have concluded that imposing a fixed time per task is significantly detrimental, while forcing participants to work sequentially has no effect.

In this part of the analysis, we want to use the information recorded through the mouse-click tracking to better refine the aggregate results. In particular, in the previous discussion, we described what the optimal policy FCFS+FB(θ^*) suggests for the maximization of students' performance. Specifically, we have seen that, if the probability of knowing the task is high enough, the individual should complete the task at the first attempt, otherwise she should postpone the answer and return to the task when all the easier questions have been completed. According to this strategy, subjects should not return to each question too often: if they postpone the task, they will return to it when all the easier questions have been completed, and they will continue to work until its hazard will stop to increase; finally, when the hazard of each task has started to decrease and if more time is still available, subjects should then look, for the third time, at those tasks which have received less service, until their completion, and then move progressively to the ones which have received more and more service. If subjects have followed this strategy, we should observe no more than three attempts at each question. Moreover, we expect that students will leave the difficult questions at the end of test, so that the correlation between the time spent on such questions and the time elapsed since the beginning of the test should be positive.

Summary descriptions of the timing and the sequence of each click performed by each subject are described in table A3 of the Appendix A.

As it is possible to see from table A3 of the Appendix A, on average students performed more than 2 lookups per question in the first part of the test. This measure gives the number of times

[8] Summary statistics are shown in table A1 in the Appendix A

[9] Results are available upon request.

students have returned to the same question during the test. The table also shows information on the average time spent on questions whose answers were missing, wrong and right, respectively. Notice that the former is by far greater than the latter. Finally, on average, students spend about 2 minutes and 20 seconds per question, and the mathematics section seems the one over which subjects generally took most time, while they seemed to find the logic and the verbal comprehension parts easier.

The rest of the analysis continues by using the information detailed above to better characterize the aggregate results.

The first behavioral component that we take into consideration are the lookups. In particular, the predictions from the reference model suggest that students should return to each question a maximum of three times, therefore we have an upper bound on the total number of lookups, which is 51 lookups, i.e 3 times x 17 questions. In order to identify those students who have returned more than three times to one question, we create a dummy variable that takes value 1 if the total number of lookups exceeds this upper bound[10]. As it is possible to see from table A4 of the Appendix A, this categorization allows us to differentiate between those students who have revised each question no more than three times (mean number of lookups smaller than 3) from those who have returned more often to each task on average. As it is possible to notice from table A5, 16% of subjects had seen each task more than 3 times. In table 4, we present the treatments' effects on the probability of giving the correct answer, for those subjects who have looked more than 3 times each question and for those who have not. The table shows that, those students who have looked the questions more frequently (in the first part of the test) have a lower probability of giving a correct answer in the second part ([$Lookups > 3$] x $Part_2$ has a negative and significant coefficient) in respect of those who switched less frequently between questions (the coefficient on $Part_2$ is: + 0.058), when both groups solved the second part of the test without any restrictions. In addition, even if those who had switched less frequently had enhanced their chances of correctly answering the questions in the baseline case, they lose this advantage completely when they are not left free to organize themselves [the coefficients on $FixedSequence$ x $Part_2$ and $FixedTime$ x $Part_2$ are both negative and the latter is also significant and large in magnitude]. On the contrary, if we consider those who have switched more frequently than the predicted level, we see that they reached higher levels of performance both when they are forced to answer sequentially and when they faced a given time per task. In particular, they fill the performance gap that, otherwise, they would have experienced in the

[10]In order to categorize subjects on switching behavior, we have also looked at the number of lookups per task, but we have found very few individuals who have never looked more than three times at all questions, therefore we were restrained from using such different categorization.

unconstrained case with respect to the other group of subjects [both $FixedSequence$ x $[Lookups > 3]$ x $Part_2$ and $FixedTime$ x $[Lookups > 3]$ x $Part_2$ coefficients are positive and significant].

Table 4: Treatments' effects on the probability of giving the right answer - Lookups

	$Pr(Right)$
$Part_2$	0.058*
	(0.032)
$[Lookups > 3]$ x $Part_2$	-0.188**
	(0.078)
$FixedSequence$ x $Part_2$	-0.043
	(0.046)
$FixedTime$ x $Part_2$	-0.207***
	(0.045)
$FixedSequence$ x $[Lookups > 3]$ x $Part_2$	0.219**
	(0.111)
$FixedTime$ x $[Lookups > 3]$ x $Part_2$	0.307***
	(0.111)
$Intercept$	0.644***
	(0.012)
N	87
adj. R-sq	0.211

Notes: *Pr(Right)* is the outcome variable which represents the probability of giving a correct answer during the test. $[Lookups > 3]$ is an indicator variable equal to one if the subject has a mean number of lookups per question greater than three. *Fixed Time* and *Fixed Sequence* are the dummy variable indicating if the subject belong to the respective treatment group. $Part_2$ it is a dummy variable equal to one if the probability in the second part of the test is considered. Fixed effects estimation. Standard errors in parentheses. Significance levels: $*p < 0.10, **p < 0.05, ***p < 0.01$.

Such findings are further confirmed if we look at the final score obtained during both parts of the test. In particular, table 5 shows that the Fixed Sequence and Fixed Time treatments are still significant and relevant, if we look at the final score obtained by those students who switched repeatedly across questions, with respect to the case where they are left free to organize their work.

By looking at the table 5, we notice that both treatments induce the "frequent switchers" to perform as the other group of subjects, even if we know that, without these interventions, they will have performed significantly worse [the coefficient on $[Lookups > 3]$ x $Part_2$ is negative and significant both statistically and economically].

From the evidence detailed above, it is clear that treatments induce heterogeneous effects, depending on which type of student is considered. In fact, the Fixed Sequence treatment is beneficial for that share of students who have switched and looked up each question more than the optimal predicted level, suggesting that preventing them from adopting such behavior, in the second part of the test, is significantly beneficial. Concerning the Fixed Time Treatment, we have found that even this schedule could be beneficial for such group of students, despite its being not specifically designed to address such behavioral dimension. In contrast, such constraint is indeed detrimental to

Table 5: Treatments' effects on the final score - Lookups

	Score
$Part_2$	1.230**
	(0.621)
$[Lookups > 3]$ x $Part_2$	-3.930**
	(1.521)
$FixedSequence$ x $Part_2$	-0.914
	(0.897)
$FixedTime$ x $Part_2$	-4.100***
	(0.878)
$FixedSequence$ x $[Lookups > 3]$ x $Part_2$	4.764**
	(2.159)
$FixedTime$ x $[Lookups > 3]$ x $Part_2$	6.237***
	(2.260)
$Intercept$	9.847***
	(0.235)
N	87
adj. R-sq	0.158

Notes: *Score* is the outcome variable, which sums the points obtained from the correct answers minus the penalty encountered from the wrong answers. $[Lookups > 3]$ is an indicator variable equal to one if the subject has a mean number of lookups per question greater than three. *Fixed Time* and *Fixed Sequence* are the dummy variable indicating if the subject belong to the respective treatment group. $Part_2$ it is a dummy variable equal to one if the score in the second part of the test is considered. Fixed effects estimation. Standard errors in parentheses. Significance levels: $*p < 0.10, **p < 0.05, ***p < 0.01$.

those students who better self-organize their answering strategy by not switching repeatedly between questions. A possible reason for this result may be that the imposition of strict bounds on the time for completing each task may spur students on to be more focused on the task at hand, preventing them from switching repeatedly between questions, which is somehow confirmed, since the mean number of lookups decrease marginally for the Fixed Time group in the second part of the test, see table A6.

In the following analysis, we present the results related to the second decision-making dimension in this context, which is the time allocation. Remember that, during the first part of the test, all the students could freely organize their time to answer all the questions. Moreover, as suggested by the optimal policy $FCFS + FB(\theta^*)$, students who want to maximize their performance should intuitively answer the tasks for which they easily arrive at the solution first. Equivalently said, they should not spend too much time on difficult questions at the beginning of the test, since this could reduce their available time for answering other questions whose answers they might know more easily.

This reasoning implies that we should observe a positive correlation between the time spent on the question and the time elapsed since the beginning of the test (at the moment when the task

is processed for the last time) for the difficult tasks: the higher the time required by the task, the farthest the moment in which she addresses the tasks, from the beginning of the test. In order to understand how students allocate time during the test, we constructed measures of time allocation using the information available from the *mouse tracking*. Firstly, we recorded the time each student used to answer each question, by summing up all the seconds spent on each task, even if not consecutively. Afterwards, we measured the time elapsed from the beginning of the test for those questions whose answers were missing, i.e the difficult tasks. As previously underlined, it is reasonable to expect the correlation between these two measures to be positive. This case, indeed, identifies whether students have properly decided not to get stuck on hard questions at the beginning of the test, and, consequently, to save time for answering other questions. In order to identify this group of students, we created a dummy variable that is equal to zero if the correlation between the time spent on the question (whose answer was missing) and the time passed since the beginning of the test is negative. The distribution of the students in such an indicator variable is collected in table A7 of the Appendix A. As it is possible to notice from the table, and according to the above definition, 25% of subjects have a good time allocation, while the remaining share does not allocate time efficiently across tasks.

In the following table, we present the results of the Diff-in-Diff-in-Diff regression estimates for the groups classified, as explained above, on the time-management dimension. In table 6 we see that imposing a given time per task, as with imposing given sequence of work, is detrimental for those students who have shown good time-management skills (the coefficients of $FixedSequence$ x $Part_2$ and $FixedTime$ x $Part_2$ are both negative and latter also significant). Moreover, fixing a given sequence of work or a fixed time per task seems that could be helpful for enhancing the performance of those who have bad time-management, even if results are not statistically significant. The reason why we have not found statistically significant results for the "Fixed Time" treatment could be the fact that, under such condition, subjects changed the way they answered the test by taking more risk and thereby reducing their probability of correctly answer the task. To confirm such intuition, we have looked at how average probabilities of giving right and wrong answers change across groups and across parts. In particular, from table A8 of the Appendix A, it is possible to confirm that subjects in the"Fixed Time" group have actually increased the probability of wrongly answer the task in the second part (when they face a given deadline per task) with respect to the first part (when they are unconstrained). While the other groups not change their answering strategy, those who faced the pressure of time became more risk lovers, confirming a result found in the psychological literature [see Busemeyer (1985) or Busemeyer and Townsend (1993)].

The findings on the probability of correctly answer the task are further confirmed if we look at

Table 6: Treatments' effects on the probability of giving the right answer - Time-Management

	$Pr(Right)$
$Part_2$	.091*
	(0.053)
$[BadTimeManagement]$ x $Part_2$	-0.091
	(0.064)
$FixedSequence$ x $Part_2$	-0.115
	(0.089)
$FixedTime$ x $Part_2$	-0.194**
	(0.078)
$FixedSequence$ x $[BadTimeManagement]$ x $Part_2$	0.145
	(0.101)
$FixedTime$ x $[BadTimeManagement]$ x $Part_2$	0.051
	(0.092)
$Intercept$	.644***
	(0.011)
N	87
adj. R-sq	0.0159

Notes: *Pr(Right)* is the outcome variable which represents the probability of giving a correct answer during the test. $[BadTimeManagement]$ is an indicator variable equal to one if the subject has shown an inefficient time allocation in the first part of the test. *Fixed Time* and *Fixed Sequence* are the dummy variable indicating if the subject belong to the respective treatment group. $Part_2$ it is a dummy variable equal to one if the probability in the second part of the test is considered. Fixed effects estimation. Standard errors in parentheses. Significance levels: $*p < 0.10, **p < 0.05, ***p < 0.01$.

the final score obtained during both parts of the test. In particular, table 7 shows that students who get stuck on difficult questions at the beginning of the test would have performed much worse in the second part of the test with respect to those who have good time management, if no intervention would have imposed [$BadTimeManagement$x$Part_2$ is negative and significant, while $Part_2$ is of similar magnitude, but positive]. Incremental results of the treatment "Fixed Sequence" may be explained by the fact that, if people know that they can see each question only once, they might become more selective as to which questions they spend their time on. However, further research is necessary to confirm such intuition.

In the following analysis, we look at the intersection between the two classifications previously described, in order to have a complete typology of the students' behavior. In particular, we split the sample into four categories as follows: students who have good time management and who have a mean number of lookups that is less than 3 ("Rational"); students who had allocated time efficiently across tasks but who had switched frequently between questions ("Switchers"); the third group identify those students who have not looked up to questions too frequently but who had poor time management ("Badtime"); finally, the last group is formed of those who have not managed time properly and, at the same time, have looked too frequently at questions ("Switchers & Badtime - S&B").

Table 7: Treatments' effects on the final score - Time Management

	Score
$Part_2$	2.055**
	(1.066)
$[BadTimeManagement]$ x $Part_2$	-2.527*
	(1.274)
$FixedSequence$ x $Part_2$	-2.455
	(1.784)
$FixedTime$ x $Part_2$	-3.930**
	(1.554)
$FixedSequence$ x $[BadTimeManagement]$ x $Part_2$	3.167*
	(1.892)
$FixedTime$ x $[BadTimeManagement]$ x $Part_2$	1.180
	(1.841)
$Intercept$	9.847***
	(0.242)
N	87
adj. R-sq	0.0159

Notes: *Score* is the outcome variable, which sums the points obtained from the correct answers minus the penalty encountered from the wrong answers. [*BadTimeManagement*] is an indicator variable equal to one if the subject has shown an inefficient time allocation in the first part of the test. *Fixed Time* and *Fixed Sequence* are the dummy variable indicating if the subject belong to the respective treatment group. $Part_2$ it is a dummy variable equal to one if the score in the second part of the test is considered. Fixed effects estimation. Standard errors in parentheses. Significance levels: $*p < 0.10, **p < 0.05, ***p < 0.01$.

In table 8, we present the treatment effects for all the groups of subjects described above, for what regards their final score. The intercept represents the score obtained in the first part of the test by the "Rational" subjects, when they answered the test without any restriction. Notice that this group of subjects obtained a higher score by around 2 points ($Part_2$ coefficient), in the second part of the test, when they were still free to answer the test without constraints. Instead, when such subjects are treated they more than lose what otherwise they will have gained (both the coefficients on $FixedTime$x$Part_2$ and $FixedSequence$x$Part_2$ are negative and large, with the former being also statistically significant). If we look at other groups of subjects, instead, we notice that they all achieved lower score than the"Rational", as expected, in the second part of the test, when they could still freely organize their work. In addition, we see that the constraints imposed with the two treatments are indeed incremental for their performance. In particular, the "Switchers & Bad Time" are the ones that are most significantly helped by such restrictions, reaching in the end a performance similar to that of the "Rational" group. From the above evidence, it is clear that taking into account heterogeneity is highly recommended when analysing the effect of such restrictions on workload management. Moreover, the estimates have shown that the average effect will depend on the share of the specific types in the population (see table A9 of the Appendix A to see

the distribution of types in our sample); therefore, knowing their distribution is fundamental for inferring which group is driving the aggregate results and for drawing proper conclusions.

Table 8: Regressions on Treatments' Groups - Score

	Score
Switchers x $Part_2$ x *FixedTime*	8.678*
	(4.730)
BadTime x $Part_2$ x *FixedTime*	1.274
	(1.942)
Switchers&*BadTime* x $Part_2$ x *FixedTime*	6.595**
	(2.898)
Switchers x $Part_2$ x *FixedSequence*	5.500
	(4.841)
BadTime x $Part_2$ x *FixedSequence*	2.919
	(2.190)
Switchers&*BadTime* x $Part_2$ x *FixedSequence*	7.187**
	(2.938)
Switchers x $Part_2$	-6.250*
	(3.331)
BadTime x $Part_2$	-2.235*
	(1.346)
Switchers&*BadTime* x $Part_2$	-5.250***
	(1.923)
FixedSequence x $Part_2$	-3.000
	(1.923)
FixedTime x $Part_2$	-4.928***
	(1.625)
$Part_2$	2.750**
	(1.110)
Intercept	9.847***
	(0.238)
N	87
adj. R-sq	0.210

Notes: *Score* is the outcome variable, which sums the points obtained from the correct answers minus the penalty encountered from the wrong answers. *Intercept* represent the score obtained by the "Rational" subjects of the Baseline group in the first part of the test. *Switchers* is an indicator variable equal to one if the subject has looked more than the predicted optimal level each question in the first part of the test. *BadTime* is a dummy variable equal to one if the subject has shown an inefficient time allocation in the first part of the test. *Switchers*&*BadTime* represents the intersection between the two above explained categories. *Fixed Time* and *Fixed Sequence* are the dummy variable indicating if the subject belong to the respective treatment group. $Part_2$ it is a dummy variable equal to one if the score in the second part of the test is considered. Fixed effects estimation. Standard errors in parentheses. Significance levels: $*p < 0.10, **p < 0.05, ***p < 0.01$.

5.3 Robustness checks

In this section, we will describe some of the robustness checks performed in order to validate the above analysis.

Firstly, we want to underline that the above classification is orthogonal to the number of subjects that enters each category in a given treatment group.

In particular, table A9 of the Appendix A shows that the average shares of types are constant across treatment' groups, suggesting that the results are not driven by ex-ante differences in groups' composition. Given that the number of observations in each group and for each column is not too distant, we can conclude that the categorization is orthogonal to the groups' belonging.

Even if the method adopted to classify subjects is based on theoretical predictions and the checks on the exogeneity of such classification have confirmed the robustness of the results, it could be that such sub-group analysis might still not convince the reader in driving general conclusions. In order to further support the validity of the results, we decided to replicate the above analysis by using a potentially more general classification. In particular, we took advantage of the measures collected in the final questionnaire on specific subjects' characteristics. As previously described, part of the questionnaire was dedicated to the test for impulsive behavior Frederick (2005). We grouped the answers of this test in a unique indicator variable that is equal to 1 if the student was defined as impulsive in all three questions of the test. Summary information on the distribution of students into such categories is shown in table A10 of the Appendix A.

More than 50% of the students are classified as impulsive in all the three questions of the Frederick's test. In the following paragraph, we will show the results of the Difference-in-Difference-in-Difference regression estimates with such groupings. Table A11 of the Appendix A summarizes the evidence. Notice that the table shows the average score obtained in both parts of the test for the two groups of students. The students who did not answer impulsively to the test in the questionnaire obtained a significantly higher score during the test, by more than 2 points, suggesting that this type of behavior is extremely relevant in this context.

Moreover, this group of students seems to lose in terms of performance when it is treated. However, if we look at impulsive students, the opposite results are true, with the "Fixed Sequence" treatment having a large incremental effect.

Such evidence suggests that failing to be impulsive in this context may imply more reflective decisions in the organizational strategy. In particular, as it is possible to notice from table A12, the total number of lookups is higher for the impulsive students, while their scores obtained in the first part is always lower than the one reached by non-impulsive subjects. Such descriptive evidence

might suggest a relation between the impulsivity measure retrieved through the Cognitive Reflection Test and the switching behavior. However, further research should address the validity of this link.

Finally, we have performed a "placebo" analysis by using types of classifications based on variables that should not affect the results, as for example the risk attitude, and indeed we do not find any statistically significant results [11].

To conclude, the results presented in the previous sections are consistent and robust to several checks. Moreover, even adopting a different and, in a way, more general classification, we found significant heterogeneous treatment effects and positive signs for both the imposed schedules, suggesting that they could serve as instruments to correct behavioral mistakes that might emerge in such problems of work-division and attention allocation.

6 Conclusions

The results have shown that, on aggregate, imposing a fixed time to solve each task could be significantly detrimental in terms of overall performance with respect to the case of an unconstrained schedule. Whereas, imposing a sequential type of work has no significant impact with respect to the case where subjects could self-organize their work, when tasks are independent.

Contrary to previous studies, however, the present paper has deepen the analysis, refining the above aggregate results by means of the *mouse-click tracking* technique. Especially, the sub-group analyses have shown that the treatment effects are, in fact, heterogeneous and their net impacts change according to individual' types. In particular, the sequential rationale is significantly beneficial to those students who switch repeatedly between tasks and who have shown bad time management skills. The reason is that, by fixing the sequence of work, they have been prevented from switching repeatedly between tasks and, potentially, they have been indirectly helped in better prioritizing the tasks over time. Given that the real "Admission Test" constrains subjects to answer each section sequentially, it might be that those students who are good at self-organizing their workload lose from this restriction.

As regards the schedule that fixes a given answering time for each question, we still found positive results especially for those students who switched frequently between tasks, and who had shown poor time management skills. Notice that the real assessment test adopted by the University of Bologna fixes a total time for each section. Therefore, since even the Fixed Time treatment is

[11] Estimates available upon request

significantly detrimental to the "Rational" students, which are the ones who generally perform better, the university should be concerned by the fact that, at the passing threshold, "Rational" subjects might be left-out because of this test design.

Given all the above results, the experiment has shown that imposing a fixed working schedule, either a fixed sequence or a deadline per task, in the workplace may enhance the performance of those workers who lack from efficient organizational skills, while it might reduce the performance of those employees who are efficient in prioritizing and organizing the workload. Therefore, in order to maximize the individual and, consequently, the average output, the employer, should firstly asses the types' composition in her group of workers and then propose individual-specific working schedules, if the composition of the group is rather heterogeneous.

Moreover, the experiment has also shown that the design of the assessment method could impact the types of student who succeeds the selection of the "Admission Test". If the university wants to equalize students' organizational abilities, providing general guidelines before the selection starts might be the solution. However, interesting questions are still open as to the reasons why students do not recognize the optimal answering strategy, and which is the best way to "teach" them.

Finally, using a different, but related, classification based on the standard impulsivity test proposed by Frederick Frederick (2005), we found that imposing a given sequence or a given time per task is beneficial for those students who are identified as impulsive, suggesting, therefore, a way to reduce the costs of such behavioral "mistake".

To conclude, such results have put forward new and interesting evidence in term of the heterogeneity of treatments. In particular, future studies that want to extend such research should account for the specific sample composition in order to understand which type is driving the aggregate results and properly to design specific policy for the target group.

References

Aalto, S., U. Ayesta, and R. Righter (2009). On the gittins index in the m/g/1 queue. *Queueing Systems 63*(1-4), 437–458.

Akerlof, G. A. and J. L. Yellen (1990). The fair wage-effort hypothesis and unemployment. *The Quarterly Journal of Economics*, 255–283.

Allen, D. (2015). *Getting things done: The art of stress-free productivity*. Penguin.

Arieli, A., Y. Ben-Ami, and A. Rubinstein (2011). Tracking decision makers under uncertaint. *American Economic Journal: Microeconomics 3*(4), 68–76.

Ariely, D. and K. Wertenbroch (2002). Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological science 13*(3), 219–224.

Baumeister, R. F., T. F. Heatherton, and D. M. Tice (1994). *Losing control: How and why people fail at self-regulation*. Academic Press.

Bellman, R. (1957). Dynamic programming.

Bisin, A. and K. Hyndman (2014). Present-bias, procrastination and deadlines in a field experiment. Technical report, National Bureau of Economic Research.

Bracha, A. and C. Fershtman (2013). Competitive incentives: working harder or working smarter? *Management Science 59*(4), 771–781.

Busemeyer, J. R. (1985). Decision making under uncertainty: a comparison of simple scalability, fixed-sample, and sequential-sampling models. *Journal of Experimental Psychology: Learning, Memory, and Cognition 11*(3), 538.

Busemeyer, J. R. and J. T. Townsend (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review 100*(3), 432.

Buser, T. and N. Peter (2012). Multitasking. *Experimental Economics 15*(4), 641–655.

Covey, S. R. (1991). *The 7 habits of highly effective people*. Simon & Schuster New York, NY.

Coviello, D., A. Ichino, and N. Persico (2014). Time allocation and task juggling. *The American Economic Review 104*(2), 609–623.

Coviello, D., A. Ichino, and N. Persico (2015). The inefficiency of worker time use. *Journal of the European Economic Association 13*(5), 906–947.

Crawford, S. and V. C. Wiers (2001). From anecdotes to theory: A review of existing knowledge on human factors of planning and scheduling. *Human performance in planning and scheduling*, 15–43.

Crenshaw, D. (2008). *The Myth of Multitasking: How "Doing It All" Gets Nothing Done*. John Wiley & Sons.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature 47*(2), 315–372.

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

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, 25–42.

Gabaix, X., D. Laibson, G. Moloche, and S. Weinberg (2006). Costly information acquisition: Experimental analysis of a boundedly rational model. *The American Economic Review*, 1043–1068.

Gittins, J. C. and D. M. Jones (1979). A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika 66*(3), 561–565.

Johnson, E. J., C. Camerer, S. Sen, and T. Rymon (2002). Detecting failures of backward induction: Monitoring information search in sequential bargaining. *Journal of Economic Theory 104*(1), 16–47.

Johnson, E. J., J. W. Payne, J. R. Bettman, and D. A. Schkade (1989). Monitoring information processing and decisions: The mouselab system. Technical report, DTIC Document.

Johnson, G. E. (1990). Work rules, featherbedding, and pareto-optimal union-management bargaining. *Journal of Labor Economics*, S237–S259.

Kocher, M. G. and M. Sutter (2006). Time is money. time pressure, incentives, and the quality of decision-making. *Journal of Economic Behavior & Organization 61*(3), 375–392.

Lipman, B. L. (1991). How to decide how to decide how to...: Modeling limited rationality. *Econometrica: Journal of the Econometric Society*, 1105–1125.

McCall, B. P. and J. J. McCall (1981). Systematic search, belated information, and the gittins' index. *Economics Letters 8*(4), 327–333.

Payne, J. W., J. R. Bettman, and E. J. Johnson (1993). *The adaptive decision maker*. Cambridge University Press.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Radner, R. and M. Rothschild (1975). On the allocation of effort. *Journal of Economic Theory 10*(3), 358–376.

Rothkopf, M. H. (1966). Scheduling with random service times. *Management Science 12*(9), 707–713.

Simon, H. and W. Chase (1988). Skill in chess. In *Computer chess compendium*, pp. 175–188. Springer.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, 99–118.

Spivey, M. J., M. Grosjean, and G. Knoblich (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America 102*(29), 10393–10398.

Strotz, R. H. (1955). Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 165–180.

Tice, D. M. and R. F. Baumeister (1997). Longitudinal study of procrastination, performance, stress, and health: The costs and benefits of dawdling. *Psychological science*, 454–458.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological review 79*(4), 281.

Wald, A. (1950). Statistical decision functions.

Yaroush, R. A. (2003). Stress and cognition: A cognitive psychological perspective.

Zhang, Y., R. S. Goonetilleke, T. Plocher, and S.-F. M. Liang (2005). Time-related behaviour in multitasking situations. *International journal of human-computer studies 62*(4), 425–455.

7 Appendix A

Table A1: Summary Statistics

Variable	N	Mean	St. deviation	Min	Max
Score	87	14.52	6.21	3	29
Gender (*)	87	0.25	0.44	0	1
Right Answers	87	16.19	6.36	5	30
Wrong Answers	87	6.70	3.58	1	18
Eco. Courses	87	0.07	0.25	0	1
Risk	87	6.22	1.51	3	9
Impulsivity 1	87	0.24	0.43	0	1
Impulsivity 2	87	0.28	0.45	0	1
Impulsivity 3	87	0.25	0.44	0	1

Notes: *Score* indicate the sum of all points obtained in the two parts of the test. (*) Gender equal to zero indicate the female sex. Notice that gender imbalance in the subject pool reflects the class composition of the "Liceo Minghetti" and does not depend by sample selection. *Right Answers* indicates the average of points gained from correctly answer the tasks. *Wrong Answers* indicates the average of points lost from not correctly answer the tasks. *Eco.courses* indicates the fraction of subjects who have taken part in any economic course. *Risk* is a subjective measure of risk attitude, which answers the following question: "In general, are you a person ready to take risks, or do you avoid to take risks? Please indicate your answer on a scale of 1 to 10, where 1 means that you do not want to take risk and 10 means that you are ready to take risks". The three *impulsivity* measures are derived from the answers to the three tests in Frederick (2005).

Table A2: Test on Random Allocation

	Baseline Mean	Fixed Sequence Mean	Fixed Time Mean	p-value(*)
Eco. Courses	0.03	0.10		0.2718
			0.06	0.5369
Risk	6.13	6.31		0.2824
			6.24	0.7097
Impulsivity1	0.33	0.23		0.4602
			0.21	0.4156
Impulsivity2	0.33	0.29		0.6978
			0.21	0.2789
Impulsivity3	0.33	0.23		0.4602
			0.25	0.6156
Score (part 1)	10	9.75		0.7077
			9.70	0.9879
Obs.	30	28	29	

Notes: (*) Two-sample Wilcoxon Mann-Whitney test. *Score (part 1)* measure the performance obtained by answering the first part of the "Admission Test".

Table A3: Answering Measures

Variable	Mean	St. deviation	Min	Max
Lookups per question	2.528	1.081	1	9
Total Lookups	50.275	8.716	34	80
Time for missing	186.312	132.355	11.187	653.343
Time for wrong	168.289	63.487	72.984	413.094
Time for right	102.031	21.068	50.324	160.246
Time for missing logic	279.333	136.141	81.213	653.343
Time for missing verbal	77.774	34.340	26.211	127.389
Time for missing math	162.314	124.270	111.879	609.359
Time for wrong logic	167.030	103.737	60.437	600.531
Time for wrong verbal	79.456	36.540	30.812	198.586
Time for wrong math	197.183	84.269	57.795	469.688
Time for right logic	115.430	36.490	51.651	191.231
Time for right verbal	56.378	18.811	23.702	118.421
Time for right math	129.645	49.373	45.906	286.591
Average time	139.448	14.745	84.157	150.555
Average time logic	130.982	34.031	66.396	237.544
Average time verbal	133.106	30.733	53.721	240.067
Average time math	154.257	44.281	64.036	286.591

Notes: All the above information are related to student behavior in the first part of the test. *Lookups per question* is the number of times the student has looked at the same question. *Total Lookups* is the total number of times the student has looked to all the questions. *Switch answer* is the number of times the individual has switched the answer. *Time for missing* indicates the average time spent on questions for which the answer was not provided. *Time for missing logic, verbal, math* measure the same time as the previous variable for each section separately. *Time for wrong* indicates the average time spent on questions for which the final answer was wrong. *Time for wrong logic, verbal, math* measure the same time as the previous variable for each section separately. *Time for right* indicates the average time spent on questions for which the final answer was right. *Time for right logic, verbal, math* measure the same time as the previous variable for each section separately. *Average time* is the mean time spent per question. *Average time logic, verbal, math* measure the same time as the previous variable for each section separately. Notice that all the time measures are expressed in seconds.

Table A4: Mean number of lookups by categorization

Tot_Lookups > 51	Mean Number Lookups
0	2.391
1	3.239

Table A5: Lookups categories

Tot_Lookups > 51	Freq.	Percent
0	73	83.91
1	14	16.09
Total	87	100.00

Table A6: Mean Lookups by groups and parts for the "Switchers"

	Part	
Group	1	2
Baseline	3.235	3.129
Fixed Sequence	3.241	1.794
Fixed Time	3.242	3.000

Table A7: Good and Bad Time-Allocation

Badtime	Freq.	Percent	Cum.
1	65	74.71	74.71
0	22	25.29	100
Total	87	100	

Table A8: Probability of correctly and wrongly answer the tasks across groups and parts

		Part	
Group		First	Second
Unconstrained	Pr(Right)	0.65	0.65
	Pr(Wrong)	0.26	0.27
Fixed Sequence	Pr(Right)	0.64	0.67
	Pr(Wrong)	0.24	0.22
Fixed Time	Pr(Right)	0.63	0.49
	Pr(Wrong)	0.25	0.33

Table A9: Numerosity by Treatments and Types

	"Rational"	"Switchers"	"Badtime"	"Switchers & Badtime"
Fixed Sequence Group	4	1	17	4
Fixed Time Group	7	1	19	4
Baseline Group	8	1	18	3

Table A10: Distribution of students according to their impulsivity

Impulsivity (categorical)	Freq.	Percent	Cum.
3	44	50.57	50.57
2	23	26.44	77.01
1	16	18.39	95.4
0	4	4.6	100
Impulsivity (indicator)			
1	44	50.57	50.57
0	43	49.43	100
Total	87	100	

Notes: The categorical variable shows the share of impulsive students in all the three questions of the Frederick'test. The category equal to three indicates that the student is impulsive in all the questions while the category equal to zero indicates that the subject was never impulsive. Notice that the indicator variable groups together all the students who, at least, were not impulsive in one question.

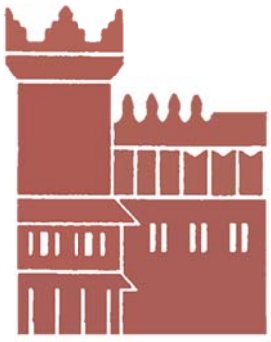
Table A11: Regression estimates based on Impulsivity classification

	Score
$FixedSequence$ x $Part_2$ x $Impulsive$	4.661***
	(1.573)
$FixedTime$ x $Part_2$ x $Impulsive$	4.069**
	(1.615)
$FixedSequence$ x $Part_2$	-1.973*
	(1.013)
$FixedTime$ x $Part_2$	-4.569***
	(1.203)
$Impulsive$ x $Part_2$	-4.708***
	(1.098)
$Part_2$	2.458***
	(0.695)
$Intercept$	9.847***
	(0.223)
N	87
adj. R-sq	0.311

Notes: *Score* is the outcome variable, which sums the points obtained from the correct answers minus the penalty encountered from the wrong answers. *Intercept* represent the score obtained by the non-impulsive subjects in the first part of the test. *Impulsive* is a dummy variable equal to one if the subject was identified as impulsive in all the three questions of the Frederick'test. *Fixed Time* and *Fixed Sequence* are the dummy variable indicating if the subject belong to the respective treatment group. $Part_2$ it is a dummy variable equal to one if the score in the second part of the test is considered. Standard errors in parentheses. Fixed effect estimates. Significance levels: $*p < 0.10, **p < 0.05, ***p < 0.01$.

Table A12: Lookups and Score by type and impulsivity

Types	Impulsivity 1	Impulsivity 0
"Rational"		
Lookups	42.12	39.59
Score part 1	9.458	10.480
"Switchers"		
Lookups	61.25	49.25
Score part 1	4.75	10.75
"Bad Time"		
Lookups	40.71	38.32
Score part 1	9.201	11.065
"Switchers & Bad Time"		
Lookups	48.5	49.2
Score part 1	8.041	9.901



Alma Mater Studiorum - Università di Bologna
DEPARTMENT OF ECONOMICS

Strada Maggiore 45
40125 Bologna - Italy
Tel. +39 051 2092604
Fax +39 051 2092664
http://www.dse.unibo.it