Maastricht University

The role of incentives, preferences and personality in decision making

Citation for published version (APA):

Vogt, B. (2015). *The role of incentives, preferences and personality in decision making*. [Doctoral Thesis, Maastricht University]. Datawyse / Universitaire Pers Maastricht. https://doi.org/10.26481/dis.20150417bv

Document status and date:
Published: 01/01/2015

DOI:
10.26481/dis.20150417bv

Document Version:
Publisher's PDF, also known as Version of record

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 18 Apr. 2024

The Role of Incentives, Preferences and Personality in Decision-Making

Benedikt Vogt

© Benedikt Vogt, Den Haag 2015

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form, or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission in writing from the author.

This book was typeset by the author using LaTeX.

Published by Universitaire Pers Maastricht
ISBN: 978 94 6159 430 3
Printed in The Netherlands by Datawyse Maastricht

The Role of Incentives, Preferences and Personality in Decision Making

DISSERTATION

to obtain the degree of Doctor at
Maastricht University,
on the authority of the Rector Magnificus,
Prof. dr. L.L.G. Soete,
in accordance with the decision of the Board of Deans,
to be defended in public
on Friday 17 April 2015, at 14.00 hours

by

Benedikt Vogt

UNIVERSITAIRE PERS MAASTRICHT

Supervisor:

Prof. dr. B. ter Weel

Co-Supervisor:

Dr. T. Schils

Assessment Committee:

Prof. dr. A. de Grip (Chair)

Prof. dr. T. Dohmen

Prof. dr. R. Dur (Erasmus University Rotterdam)

Prof. dr. D. Webbink (Erasmus University Rotterdam)

Prof. dr. I. de Wolf

This research was financially supported by the Graduate School of Business and Economics (GSBE).

Et jit kij wood ...
To my family.
Thank you.

Contents

Contents **v**

1 Introduction **1**
1.1 Motivation . . . 2
1.2 Aim . . . 3
1.3 Methodology . . . 5
1.4 Relevance . . . 6
1.5 Summary of the Main Findings . . . 7
1.6 Implications . . . 9

2 The Economics of Test Taking: The Effect of Pressure on Decision-Making and Test Performance **11**
2.1 Introduction . . . 12
2.2 Experimental Design . . . 16
2.2.1 Raven Matrices and Numerical Problems . . . 16
2.2.2 Choice Process . . . 17
2.2.3 Varying Time Pressure and Monetary Stakes . . . 20
2.3 Theory . . . 20
2.3.1 Main Idea . . . 22
2.3.2 A Simple Model . . . 23
2.4 Results . . . 25
2.4.1 Baseline Treatment . . . 25
2.4.2 Increasing Stakes and Time Pressure . . . 33
2.4.3 Maximizing Expected Earnings? Adaption to the Test Environment . . . 36
2.5 Little Adjustment of Answering Behavior . . . 38
2.5.1 Probability of Knowing the Correct Answer and Submission Behavior . . . 40
2.5.2 Expected Earnings . . . 43
2.6 Robustness Checks . . . 47

2.6.1 Choice Process . . . 47
2.6.2 Aggregation Bias . . . 51
2.7 Conclusion . . . 51

3 How Do Personality Traits and Preferences Affect Cognitive Test Scores? Evidence from a Laboratory Experiment 55
3.1 Introduction . . . 56
3.2 Conceptual Framework . . . 59
3.3 Experimental Design . . . 61
3.3.1 Measuring personality and preferences . . . 62
3.3.2 Solving Raven matrices . . . 64
3.3.3 Disentangling technology from behavior . . . 65
3.4 Results . . . 66
3.4.1 Descriptive Statistics . . . 67
3.4.2 Determinants of test scores . . . 72
3.4.3 Answering technology . . . 76
3.4.4 Answering Behavior . . . 82
3.5 Conclusion . . . 84

4 Patience and Achievement Test Results 87
4.1 Introduction . . . 88
4.2 Data . . . 91
4.2.1 Measuring Patience . . . 92
4.2.2 Outcome Variables . . . 93
4.2.3 Control Variables . . . 94
4.3 Results . . . 94
4.3.1 Correlation Structure . . . 95
4.3.2 Patience and IQ . . . 95
4.3.3 Patience and High-stake Achievement Test Results . . . 97
4.3.4 Patience and Low-stake Achievement Test Results . . . 100
4.3.5 Differences in Math and Language Test Scores . . . 102
4.4 Potential Mechanisms & Discussion of the Results . . . 107
4.5 Conclusion . . . 110

5 The Effect of Imposed Payment Schemes on Workers' Performance 111
5.1 Introduction . . . 112
5.2 Approach . . . 116
5.2.1 Design . . . 117
5.3 Productivity and preferences . . . 124
5.3.1 Sorting . . . 124
5.3.2 Productivity differences . . . 126
5.3.3 Determinants of sorting . . . 130
5.3.4 Productivity sorting . . . 130
5.4 Exogenously changing payment regimes . . . 134

5.4.1 Productivity changes . . . 134
5.4.2 Productivity differences . . . 136
5.4.3 Heterogeneity across subjects . . . 137
5.4.4 Stress, Effort and Exhaustion . . . 141
5.5 Conclusion . . . 144

Bibliography **147**

Appendices **157**

A Appendix to Chapter 2 **159**
A.1 Subject Pool and Experimental Details . . . 159
A.1.1 Experiment . . . 159
A.1.2 Types of Questions . . . 160
A.1.3 The Red Payment Schemes . . . 162
A.1.4 Screenshots of the Instructions . . . 167
A.2 Additional Results on Numerical Tasks . . . 181

B Appendix to Chapter 3 **185**
B.1 Additional Graphs . . . 185

C Appendix to Chapter 4 **189**
C.1 Intelligence Test . . . 189
C.2 Patience measure . . . 192
C.3 Correlations patterns . . . 193
C.4 Additional Analysis of the Big Five . . . 196

D Appendix to Chapter 5 **199**
D.1 Screenshots . . . 199
D.2 Additional Results . . . 202
D.3 Different Experimental Locations . . . 204

Nederlandse samenvatting **204**

Valorization **213**

Curriculum Vitae **219**

Acknowledgements **221**

Chapter 1

Introduction

1.1 Motivation

One of the core subjects of economics is to understand how and why people make decisions. This does not only involve the traditional framework of utility maximization (e.g. Edgeworth, 1879) but also its ingredients, such as preferences and incentives. Preferences and incentives determine an individual's choices. These choices depend again on the situation in which persons are involved and are revealed in observed behavior.

In recent decades economic theories of decision-making have been tested with real world data. This includes field data, as well as data from laboratory experiments (Harrison and List, 2004; Falk and Heckman, 2009). One core insight of this emerging research area in economics is that decision-making is not very well predicted with the classical ingredients of the utility maximization framework, which involves among others perfectly rational decision makers. The field of behavioral economics has established well-known deviations from this behavior, such as deviations from the expected utility framework (e.g. Kahneman and Tversky, 1979), loss aversion (e.g. Tversky and Kahneman, 1991), social preferences (e.g. Fehr and Schmidt, 1999) and hyperbolic discounting (e.g. Laibson, 1997). These notions have influenced economic models and have been instrumental to a large body of empirical research, both in the field and in the lab (Dohmen, 2014).

A prime example has been the distinction between different types of skills that are important for decision-making. Recently labor economists and economists studying educational choices have been distinguishing non-cognitive and cognitive skills (e.g., Borghans et al. (2008) for an elaborate overview). It turns out that different types of skills influence individual decision-making in different ways. Economists use methods of psychologists to measure both cognitive and non-cognitive skills and incorporate them in their empirical and theoretical analysis (see for instance Dohmen, 2014). Less is known what these measures of cognitive and non-cognitive skills actually measure and how cognitive skills interfere with non-cognitive skills, incentives and effort in individual decision-making.

An example helps to illustrate this development. The psychologist Walter Mischel and his co-authors (Mischel et al., 1972) conducted a simple experiment. 3 to 5 year old children were sitting alone in a room. In front of them was a Marshmallow. They were told that they could either eat this Marshmallow immediately or they could wait for an indefinite period to get a second Marshmallow. In a follow up study Mischel et al. (1989) showed that those children who were waiting for the

second Marshmallow showed better performance in school and could for instance cope better with stressful situations. These experiments triggered a discussion among scientists and policy makers. Both economists and psychologists became interested in the mechanisms behind these findings. Why did some children delay their choice and why did some go for the immediate option? Were the children simply smarter and knew that it was better to wait or did they have a different personality which helped them suppress their need to eat the Marshmallow immediately? And, was this observed behavior typical for the children or did it only apply to this specific setting?

1.2 Aim

These studies help to illustrate the main idea of this dissertation. The aim of this thesis is to shed light on the determinants of decision-making from an empirical point of view. In four studies I investigate the factors that determine individual decision-making and the underlying mechanisms behind these decisions. The studies document and interpret to what extent incentives, preferences and personality traits matter for the decision-making process in the fields of the economics of education and behavioral economics in a broader sense. In all chapters the focus is on individual decision-making. These decisions do not involve social interactions and have no direct consequences for others but only for the individual itself. I make use of well-established methods of assessing preferences, personality traits and cognitive skills.

Figure 1.1 summarizes the main research outline of this dissertation. The idea is that the consequence of a decision is an outcome, for instance an answer to a question on a cognitive test or the choice of an occupation. Before a decision is made a decision-making process takes place. This process is influenced by various factors. Figure 1.1 shows the main ingredients of the decision-making process. Each number in a circle corresponds to a chapter and the position of the chapter in relation to the general problem setting of this thesis. In the following chapters I will provide answers to four questions:

1. To what extent do (financial) incentives change effort and behavior on a test of cognitive abilities? (Chapter 2)

2. How do personality traits and preferences interfere with the decision-making process on a test which measures cognitive skills? (Chapter 3)

3. How is patience related to achievement test scores? (Chapter 4)

4. How do incentives and preferences influence sorting decisions and performance in a real effort task? (Chapter 5)

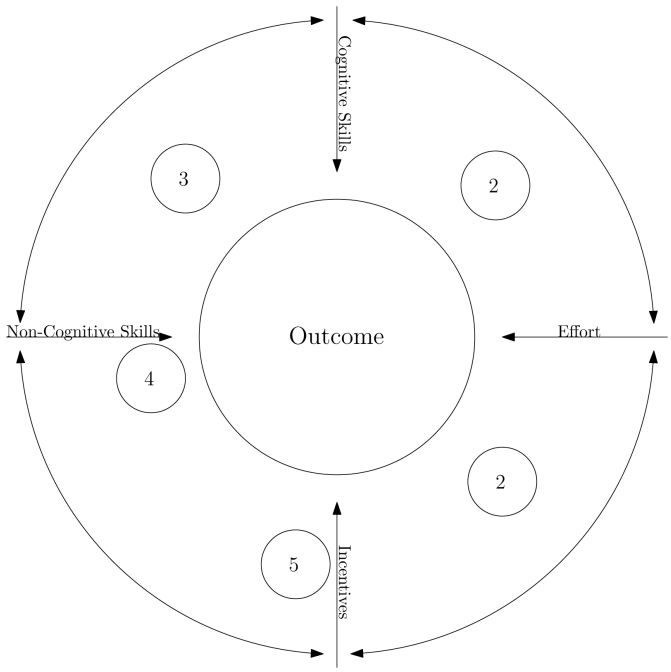


Figure 1.1: Outline of the thesis.

Whereas the neoclassical economic approach is to see an individual's decision-making in form of a utility maximizing process in which the decision maker maximizes her utility for a given set of preferences and incentives over a choice set and her budget constraint, recent developments in economics started to incorporate also personality traits and cognitive ability into this decision-making process. Almlund et al. (2011) formalized this approach in the following way:[1]

$$\sum_{j=1}^{J} R_j \phi_j(\theta, e) - C_j(\theta, e) \tag{1.1}$$

Equation 1.1 states the following. The individual maximizes her reward minus costs of effort for a given amount of J tasks. The value of a task j is obtained

[1] Other approaches are for instance Borghans et al. (2008).

through a utility maximization process which involves the trade-off between the rewards (incentives) $R_j(\cdot)$ and costs $C_j(\cdot)$. The effectiveness of rewards depends on the productivity $\phi_j(\cdot)$. Both, costs and productivity depend on the set of skills θ and effort e. Skills incorporate cognitive skills and non-cognitive skills. Non-cognitive skills include personality traits and economic preferences. Equation 1.1 relates the ingredients of Figure 1.1 to the classical utility maximization framework applied in many economic models. The core consequence is that not only differences in incentives or effort levels, but also differences in cognitive and non-cognitive skills lead to different outcomes. In the course of this thesis I provide evidence that all of the four factors are important for understanding how and why people make certain decisions. In addition, it is shown that this understanding helps to interpret differences in outcomes.

1.3 Methodology

In recent decades laboratory experiments have become a major research tool of economists (Falk and Heckman, 2009). Laboratory experiments offer the opportunity to identify causal effects in a controlled environment. Due to the isolated environment, laboratory experiments provide a perfect opportunity to investigate decision-making. It is important to mention that lab experiments only serve one part of the story. Ideally the findings are backed up with complementary evidence of field experiments (Harrison and List, 2004) or field data as for instance in (Dohmen and Falk, 2011). In this way, laboratory experiments provide evidence on channels of correlations, which are observed in the field but cannot be interpreted as such without making use of the evidence obtained in the lab.

In this dissertation I develop and conduct two laboratory experiments and analyze one cross-sectional data set. Chapters 2 and 3 report on the results of the first experiment, which was conducted in the BEELab with students from Maastricht University. More specifically, in Chapter 2 of this dissertation I investigate experimentally how monetary incentives change the behavior during a test. Chapter 3 supplements this analysis and shows how personality traits and economic preference parameters interfere in the decision-making process during a test. In Chapter 4 I provide complementary evidence to lab findings with field data. I conducted a second set of lab experiments for Chapter 5. The main idea of this project is to investigate the effect of exogenously imposed payment schemes on performance. This experiment took place at two locations with different subject

pools. One part of the experiment was conducted in the BEELab in Maastricht. Another part of the experiment was conducted at several locations and faculties of Hogeschool Zuyd, which is a University of Applied Sciences in the region of Maastricht. Using a more heterogeneous sample has the advantage of increasing the external validity of the results.

1.4 Relevance

In current policy debates about the education system and the labor market many things are examined from an economic point of view. This means that for a restricted number of inputs the output of a production function is optimized. For example, in the field of education, educational attainment of pupils should be maximized for a given set of resources, such as teachers and schools. It seems that in the public debate people take for granted that both inputs, such as teacher quality, and outputs, such as student attainment, are perfectly measurable and the public planner only needs to optimize an educational production function. However, the measurement of educational attainment with an achievement test score is only one way to measure educational attainment. So far not much is known about what achievement tests actually capture. Therefore, investigating the mechanisms of incentives, personality traits and preferences and their influence on a test result seems to be important information in building the education production function. Sound skills measurement is also a crucial first step in policy making.[2]

Policy makers currently place great emphasis on standardized achievement tests to sift and sort people, to evaluate schools, and to assess the performance of entire nations. Academic admissions committees use these tests to screen applicants. Despite its widespread use, the skills that they measure and their importance for success in life are not well-understood. Achievement tests might not adequately capture personality or non-cognitive skills in general (Almlund et al., 2011). These skills have been identified and labeled in a broad and multi-disciplinary literature that mostly exists outside the economic literature and are universally valued across all cultures and societies (Roberts, 2009). Until recently these skills have remained unmeasured and have been largely ignored by researchers and policymakers. However, in recent research economists and psychologists have constructed measures of these skills and provide evidence that they predict meaningful life outcomes (Heckman and Kautz, 2012). Chapters 2, 3 and 4 deal with the relation between

[2]See for instance Green (2013) and van den Berge et al. (2014).

incentives, preferences, personality traits and outcomes on tests scores of cognitive skills.

Besides the role of cognitive and non-cognitive skills in test results, it is also important to look at their role in different settings. One example is the performance of employees and the selection of employees into different occupations. An employer knows neither perfectly how an employee works nor how monetary incentive would motivate people to give their best. However, a crucial condition to optimize outputs for a firm is to understand how employees are motivated and whether monetary incentives work as amplifier or attenuators of their performance (Lazear, 2000; Dohmen and Falk, 2011). Therefore, I conducted a second set of laboratory experiments to study the effect of monetary incentives on self-selection and performance. The experiment should help to understand what happens to the performance of workers if the conditions under which they have to produce output are suddenly changed.

1.5 Summary of the Main Findings

In Chapter 2, which is joined work with Lex Borghans, Huub Meijers and Bas ter Weel, I investigate test taking behavior in a laboratory setting. The idea is that performance on a cognitive test can be considered as economic behavior, because it depends both on ability and on the way people deal with time pressure during a test. Thus, a test score is not only determined by the abilities a test is supposed to measure but also by an individual's test-taking ability. Studies in psychology and economics show that people are responsive to incentives on cognitive tests (Edlund, 1972; Borghans et al., 2008). To document and analyze the economics of test taking, we set up an experiment in which we disentangle the time when an answer to a question is known from the time when the test taker submits the answer to a test question. We change the test environment by varying the time pressure and the monetary stakes for submitting an answer. This way we are able to disentangle the intensity of thinking and the response time to different incentives. We obtain three main findings: First, test takers do not come up faster with an answer if we increase time pressure or monetary stakes. Second, test takers change their submission behavior in the test environment consistent with a simple economic model. Third, changes in the timing of submission are surprisingly small. Especially the small changes in the timing of the answer across different test environments hints at the fact that our test takers are already motivated in

answering these questions and that they do not need to be motivated by higher payments.

Besides incentives, another important set of determinants of tests are preferences and personality traits (see for instance Borghans and Schils (2013)). In a next step I analyze in chapter 3 how personality traits and preferences affect cognitive test scores. Cognitive test scores are used to assess cognitive ability, but are the outcomes of a mental process involving much more than only cognitive ability. In the same experiment as described above we also elicited a battery of personality traits such as the Big Five (Goldberg, 1990) and economic preference parameters such as risk and time preferences. We investigate the decision-making process during a cognitive test in a laboratory experiment in which students have to solve Raven matrices. The design allows us to distinguish between when someone knows the correct answer to a Raven matrix from submitting this answer. We find that openness to experience, neuroticism and an individual's risk preference influence the speed of thinking during a test. Only an individual's discount rate determines the timing of an answer. The results support previous findings in the literature (e.g. Burks et al., 2009; Duckworth et al., 2011) but shed new light on the mechanisms behind the determination process of a test score. They have implications for test scores in the sense that there is potential to improve performance by changing behavior.

Chapter 4 provides complementary evidence from field data to the laboratory study in the lab. In this chapter I investigate the relationship between patience and the score on high and low-stake achievement tests. The main question of interest is to what extent more patient students obtain higher test scores. In a large sample of Dutch secondary school children I use an experimentally validated measure of the internal rate of return of 15 year olds to measure patience. Controlling for cognitive ability and a range of personality traits, I find a strong and significant relationship between patience and the result on a high-stake achievement test. Students who score one standard deviation higher in terms of measured patience obtain more than 16 percent of a standard deviation higher test scores on high-stakes tests. The relationship is smaller in magnitude and insignificant for a low-stake achievement test. We also find strong differences in patience between different levels of education.

Finally, understanding how incentive schemes influence outcomes and behavior is crucial for the understanding of modern economies. In chapter 5, which is joined work with Trudie Schils and Bas ter Weel, we investigate experimentally

how output in a real-effort task varies under a fixed payment and a piece rate. The aim is to investigate how people respond to an imposed change in payment regime. We conduct a laboratory experiment with university students and students from a University of Applied Sciences. The sorting patterns we observe suggest that when people are free to choose they sort according to their productivity and their gender (see e.g. Dohmen and Falk, 2011). The key finding is that, when we impose a payment scheme output does not significantly change. However, imposing a piece rate increases self-reported stress levels, exhaustion and effort. If we increase the pay in the fixed payment scheme and hold the piece rate constant, more productive subjects select the fixed payment scheme. Imposing another payment scheme seems to be ineffective in trying to change performance because workers seem to sort according to productivity. The most productive workers select themselves into the variable payment scheme. Our findings show that the majority of workers seem to exert maximal effort even if their payment is not directly linked to performance. In general this is in line what we often observe in reality. Many contracts are incomplete and agents exert effort even though they know it is not necessarily enforceable by the principal. Previous findings from the lab (e.g., Fehr et al., 1998) and the field (e.g., Kube et al., 2012) interpret this behavior as a gift exchange. The agent reciprocates the principals' kindness with effort levels above the rational level of effort (Akerlof, 1982). Moreover, we find that performance dependent payment cannot genuinely increase output of rather unproductive workers.

1.6 Implications

The findings obtained in Chapters 2, 3 and 4 are important for researchers and policy makers. First, we show that incentives and time pressure slightly change behavior during a cognitive test. In our setting, they do not have consequences for the final test result. Hence, as soon as there is something at stake, the test result does not seem to be influenced by an increase in rewards or time pressure. This has implications for policy makers and researchers when they use and interpret cognitive test scores. Chapter 3 highlights the relationship between non-cognitive skills, which are defined as preferences and personality traits, and test scores. We show that better cognitive test scores go along with favorable economic preferences and personality traits. These results are further supported with evidence from the field in Chapter 4. There are several implications of the results obtained

in these chapters. First, researchers and policy makers should focus on strategies to improve those personality traits which lead to better outcomes. Recent research has shown that personality traits are formed over the life cycle (Cunha et al., 2010) and are malleable (e.g., Roberts (2006); Prevoo (2013)). Carefully conducted interventions, which aim for instance at improving non-cognitive skills, can lead to better outcomes on the individual and the societal level. Second, if education policies only focused on maximizing results of cognitive tests, (at least) equally important determinants of lifetime outcomes such as character skills and personality traits (e.g., Heckman and Kautz (2012)) might stay underdeveloped. Third, the results from the studies in Chapter 2 and 3 can also be interpreted in a more general framework. Since cognitive tests can be seen as problem solving activities these studies also show that the incentives, preferences and personality traits play a role in how people solve problems.

In Chapter 5 we investigate sorting behavior and individuals' responses to imposed incentive schemes. This chapter addresses an important but often overlooked economic concept, namely individual effort provision in response to incentives. This is an extremely relevant question for managers and policy makers. There are two key implications of this particular experiment. First, if payment schemes for certain occupations are changed, different types of workers seem to be attracted to these occupations. It suggests that the allocations of workers to occupations changes in response to different payment schemes (Lazear, 2000; Lazear and Rosen, 1981). This result is of importance for discussions in the education and health sector, and probably of relevance in the public sector as a whole. People sort (in part) in line with the payment conditions. Using incentive payment to improve outcomes could be effective but seems unlikely to improve the performance of the existing employees. Moreover, highly performance-dependent incentive schemes, such as piece rate, do not yield higher outputs compared to a fixed payment, once the most productive individuals are selected into the piece rate. Second, changes in performance schemes could have effects on an employee's health conditions, since they can substantially increase stress levels.

Chapter 2

The Economics of Test Taking: The Effect of Pressure on Decision-Making and Test Performance

2.1 Introduction

Scores on achievement tests and cognitive tests are important inputs in the assessment of students, job applicants and are even used to rank the educational performance of countries. These tests are typically taken under time pressure and the stakes are high for the persons who are assessed. Often a set of questions has to be completed within a limited amount of time and the outcomes are used as major determinants of future educational tracks and career paths and as measures of the educational performance of countries. Performance on a cognitive test depends on time pressure and incentives, which makes answering a question on a cognitive test an economic decision-making problem. Faced with such a test, people are provided with incentives to maximize outcomes and results are interpreted in that way by teachers, potential employers and policy makers. However, surprisingly little is known about how people behave during a cognitive test and under what circumstances test scores are maximized. How do people cope with this pressure? Do the incentives make them think faster, or do they choke thoughts and do people deal with the scarcity of time in an efficient manner? While incentives can push people to show their full potential, scores might also be influenced by the way people deal with them.

We analyze to what extent test takers behave in an economic way when completing a question during a cognitive test. The basic idea is that differences in test performance between different environments result from two reasons. On the one hand, the probability of a correct answer over time can vary in different test environments because the test taker thinks more or less intense about a question. On the other hand, the degree of time pressure and the stakes change the timing of submitting the answer and thus also the test result. If, for instance, low stakes make the test taker invest less time in answering each question, the test result is always lower in a low stake test environment compared to a high stake test environment. However, stakes and time pressure can also influence the intensity of thinking, which in turn could affect the probability of a correct answer. This is important to understand, since behavior on tests can be seen as a decision in which people make a trade-off between the cost of thinking an additional unit of time and the gain of submitting the correct answer.

The aim of this paper is to answer three questions. First, we investigate whether test takers change the speed of coming up with a correct answer under different test conditions. Second, we study if test takers change the timing of

answering in different test environments in an economic way. Third, we document whether this behavior is optimal with respect to payoff maximization. To answer these questions, we develop a novel experimental approach, which is executed in a laboratory setting. Subjects have to answer questions from a cognitive ability test which are randomly assigned to three different incentive schemes for a correct answer. By providing participants, in addition to the test incentives, with an incentive to immediately reveal the answer that comes to mind, we are able to measure the process of deliberation towards a choice and the point in time at which they decide to submit their final choice.

Our main findings can be summarized as follows. The probability of answering a test question is a concave function of the time invested. This probability does not change if we increase time pressure or monetary stakes. Moreover, test takers change their submission behavior in the test environment in the direction predicted by a simple economic model. They invest more time when stakes are high compared to a situation if stakes are low. However, changes in timing are small and the adaption of the test takers' behavior is weaker than we would expect if test takers only maximized their expected payoff. Our results remain robust when we analyze different subsamples and when we take into account possible effects of aggregation bias. To arrive at this main result we set up a cognitive test based on 45 Raven matrices, which are a well-established measure for fluid intelligence (e.g., Carpenter et al. (1990)). 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 question a limited amount of time is available to come up with an answer. We provide subjects with monetary stakes during the choice process and reward the right answer more if it is given earlier. To check the robustness of our method, we replicate the approach using a test in which participants have to solve numerical problems expressed as addition and subtraction operations. [1]

From an economic point of view, solving Raven matrices is a decision-making problem in which subjects decide how much time and effort to invest in finding the solution to the problem. This decision-making problem consists of four components. First, there is the return to investing an additional unit of time for the probability of finding a better solution. This depends on ability and monetary

[1] 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 behaviour during consumer choice processes.

stakes or rewards (e.g., a higher test score or being hired on a vacancy). Second, there is a trade-off between this probability gain and the cost of exerting effort, which depends on the disutility of time and incentives. Third, the decision-making problem consists of the stakes of finding the right solution, which vary with different levels of monetary stakes or rewards. Finally, there is time pressure. There is only a limited amount of time available to come up with the right solution and the sooner the right solution is found, the more matrices can be solved and the higher the final test score will be.

Economic theory predicts that time investment varies with all four components. We provide a model, similar to Borghans et al. (2013), which predicts differences in behavior across different environments. The setup of our experiment provides an environment to test whether or not people adapt their behavior in the way our model predicts. In the conducted laboratory experiment we vary one of the components in each of the treatments and keep the others constant.

Despite the wide use of test scores, investigating the choice process during cognitive tests has been unusual in the economic literature. Generally, test scores are used as outcomes or predictors of educational quality. Subjects are assumed to maximize an objective function, but it remains unclear how choices are generated. Although test scores provide valuable information, our research shows that choices seem to vary across different settings. That is why this paper adds to five different strands of the recent literature in economics and psychology. The first strand investigates the role of incentives on test results. There are numerous studies, both in psychology and economics, which find that incentives have a positive impact on the IQ test results. All studies show that sufficiently high incentives increase the outcome of both IQ tests and achievement tests (e.g., Edlund (1972); Lloyd and Zylla (1988); Borghans et al. (2008, 2013); Duckworth et al. (2011); Segal (2012)).[2] Our paper adds to the understanding of why incentives could help improve test outcomes. The second strand is the vast literature in economics on the relationship between incentives and effort provision. Whereas most studies argue that performance dependent incentives should increase effort provision and hence outcomes (see Prendergast (1999) for an overview and Lazear (2000) for an empirical example), there are a number of studies which show that the relationship is not necessarily strictly positive because incentives can also crowd out the intrinsic motivation to work on a task (e.g., Frey and Oberholzer-Gee (1997), Gneezy and

[2] Almlund et al. (2011) provide an excellent overview of the results of these studies starting on page 57 .

Rustichini (2000) and Bénabou and Tirole (2003)). In our study incentives do change the way people time their answering behavior, but they only have minor effects on the final test outcome. Since most of the latter studies relate to workplace environments, it is also important to mention a third strand of literature that investigates the impact of financial incentives for educational outcomes in the field. Recent work finds positive short term effects of financial incentives on achievement test and graduation rates (e.g.,Rodriguez-Planas (2012)). However, Angrist and Lavy (2009) and Angrist et al. (2009) only find effects for girls and other studies find no effects (Fryer, 2011) or only effects on math test scores (Bettinger, 2011). Our paper can help understanding potential mechanisms, since we control and monitor the decision-making process during a test. A fourth strand of literature is decision-making under time pressure. Kocher and Sutter (2006) 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 experimental findings are in line with their findings. Gabaix et al. (2006), Manzini and Mariotti (2007) and Caplin et al. (2011) use choice-process data to test consumer choice models. They conduct experiments in which subjects are incentivized to not only reveal their choices but also their deliberations to arrive at their final choices. We use similar addition and subtraction operations as a robustness check to the answers to solving Raven matrices as the ones Caplin et al. (2011) have applied. Reutskaja et al. (2011) have a similar goal and apply eye tracking methodologies to understand the computational processes used by consumers to make choices between alternatives. These studies find evidence consistent with our estimates in the sense that subjects do not search optimally. Finally, our paper relates to a recent literature on the relation between non-cognitive skills and test scores: Duckworth and Seligman (2005); Borghans et al. (2008); Duckworth et al. (2011) document that certain personality traits seem to explain a substantial part of differences in performance on a cognitive test. We document substantial heterogeneity in test outcomes, but leave it to future research to investigate to what extent non-cognitive skills influence behavior during the completion of a cognitive test.

The setup of this paper is the following. The most salient features of the experimental design are presented in Section 2.2. Section 2.3 develops a simple theoretical framework about decision-making from an economic perspective. Section 2.4 presents our main results. Section 2.5 analyses why behavior is so inelastic and to what extent the answering behavior is optimal with regard to the mone-

tary incentives provided. In Section 2.6 we provide robustness checks. Section 2.7 concludes.

2.2 Experimental Design

We conducted an experiment at the Behavioral & 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.[3] These are students from various study programs. There were seven sessions in which 130 subjects participated; two participants felt ill during the experiment and had to leave. We conduct the analysis on the remaining set of $n = 128$. The experiment was conducted in November and December 2012. Appendix A in this dissertation 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.

2.2.1 Raven Matrices and Numerical Problems

The goal of this research is to study how people tackle a question during a cognitive test. To do so, we select Raven matrices (Raven, 1962). Raven matrices measure fluid intelligence and have been used to measure cognitive ability in psychology (e.g., Carpenter et al. (1990) and Roberts et al. (2005)).[4] The results of these tests have also been applied in the economic literature as measures of cognitive ability (e.g., Almlund et al. (2011)). Subjects have to select the correct missing figure, which completes the sequence of nine consecutive figures which is designed as a 33 matrix. They have to select the missing figure from a set of eight figures. All figures are connected in a logical manner and there is no prior knowledge needed in order

[3] This is an online recruitment system for experiments (Greiner, 2003).

[4] Fluid intelligence reflects the ability to learn independent of the actual knowledge and the cultural background of the individual. Carpenter et al. (1990) document that Raven matrices measure the ability 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 conscious deliberation and those that are slower and more reflective. Raven matrices refer to the former (e.g., Epstein (1994)).

to be able to solve these matrices. Panel A of Figure 2.1 presents an example of a Raven matrix. We study the decision-making process and performance of subjects on each Raven matrix separately. This is different from how a usual cognitive test is conducted by test takers and evaluated to obtain a cognitive test score. To understand the decision-making processes, we develop a second type of puzzle. Subjects have to solve numerical problems. They face a set of eight different addition and subtraction problems. Each of these problems consists of three terms and each of the terms consists of a number between zero and one hundred. The number is either spelled in Arabic numerals or written out in words on the screen. Subjects had to find the problem which summed up to the highest amount in the set of eight problems. 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. Panel B of Figure 2.1 shows an example of a numerical problem. We use a set of 45 Raven matrices. All subjects had to solve the matrices in three sections of 15 questions. The order in which subjects had to solve the matrices was randomized. Similarly, subjects had to answer three sets of 15 numerical problems, which can be ranked according to complexity. Again subjects had to solve the problems in random order. The most complex numerical problem is the one with the highest number of written out words and answers that are closest to each other. The experiment was divided into two parts. Subjects either had to solve three sets of Raven matrices or numerical problems in the first or second part. The order was randomized across subjects. Before they could start the experiment, they had to go through a trial phase which made them familiar with the problems and the payment schemes. Appendix A presents more information about the Raven matrices and the numerical problems and their rank order.

2.2.2 Choice Process

Figure 2.2 shows a screen shot of the decision screen for a typical Raven matrix and for a numerical task. In every screen the problem set was displayed in the middle of the screen. For each question there is a time limit of 60 seconds, which does not vary during the experiment. The remaining time is indicated with the green bar in the left part of the screen. During standard cognitive tests subjects optimize timing over the whole test. Time investment on each question depends on the expectations about future questions and on the performance on these questions. To avoid this loss of control, we restrict the time to answer a question to 60 seconds. Subjects

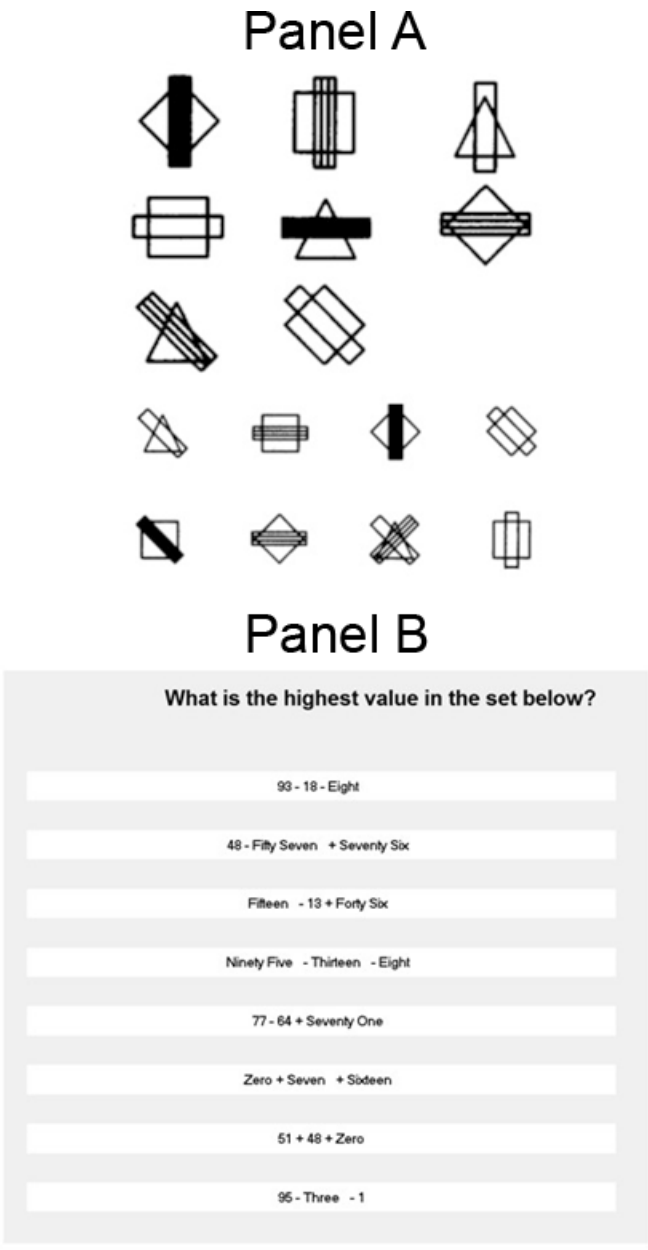


Figure 2.1: Examples of the Tasks

Note. Panel A shows a typical Raven matrix. We took a Raven matrix from Carpenter et al. (1990) (p. 407) and none of our experimental Raven matrices since the psychological tests are meant to be kept discretely. Panel B a typical numerical task.

had to wait until the time elapsed to proceed to the next question. Another advantage of our setup is that we actually obtain data on every question from each subject.

We address three types of decisions during the experiment: the initial choice, the evolution of choices with contemplation time and the final choice. The initial choice subjects make after observing the Raven matrix or the numerical problem is to come up with an immediate first answer. This initial choice is incentivized by what we define as the blue system. The blue system yields a monetary reward of 0.5 cents for every second the correct answer is selected during the 60-second time period. This stimulates an immediate choice after the screen is visible. If a subject for instance chose the correct answer immediately and did not change the selection she would have earned 30 cents only from the blue payment scheme. To induce an immediate choice there was a popup message if subjects did not select an answer within the first five seconds. The information about the blue payment scheme was displayed on the right part of the screen.

The second role of the blue system is that it allows us to track decision-making during the 60-second time period for each question. After the initial choice, subjects could change their answer as often as they liked until the deadline of 60 seconds passed. This allows us to identify how the provisional choices evolved with contemplation time. It also allows us to investigate the investment people were willing to make to find the right answer to the matrix. 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) and Reutskaja et al. (2011)). It differs from the approaches in consumer choice experiments because during a cognitive test the choice is either right or wrong. By contrast, consumers can decide to stop searching for alternatives when they have reached a satisficing level of utility. In our case this would be equivalent to stop changing the preferred solution when subjects are sufficiently sure that the answer is the best they are able to come up with.

At the same time a red system was present. The purpose of the red system was to mimic the time pressure and the stakes people face when conducting a cognitive test. Subjects were incentivized to submit their final solutions to the matrices as soon as possible by pressing the 'submit' button. In the baseline treatment, we linearly decreased the amount of money for a correctly submitted solution from 25 to 5 cents during the 60-second time period available to solve a Raven matrix. Subjects were able to stop the decline by pressing the button

after they had selected a solution. The actual payment of the red payment scheme was displayed on the right side of the screen. The amount of the red payment scheme decreased on the screen according to the respective treatment. Note that after pressing the submit button, choices could be updated in the blue system. So, even after submitting the answer subjects had an incentive to keep on thinking about their submission because the payment of 0.5 cents per second for selecting the correct answer was still running after submission. Appendix A presents more information about the blue and the red system. Each subject had to complete 15 Raven matrices and 15 numerical problems in the baseline treatment.

2.2.3 Varying Time Pressure and Monetary Stakes

The other two sets of 15 Raven matrices and 15 numerical problems were executed under different levels of time pressure and monetary stakes. We did so by changing the incentives in the red system only. First, we changed the level of the monetary stakes, but not the slope, in the red system. Instead of a scheme running from 25 to 5 cents, the schedule now 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 stakes and Low time pressure. The comparison of this incentive scheme to the basic scheme (Low stakes and Low time pressure (LL)) yields information about the effect of the stakes on decision-making. From an economic point of view, we expect the choice behavior to be different across these two incentive schemes. Since the payment for a solution is higher at any point in time, we expect subjects to submit their final choice later in this setting compared to the baseline setting.

Second, we also changed the slope of the monetary stakes to increase both monetary rewards and time pressure. This third scheme ran from 55 to 5 cents. We call this the HH treatment, which stands for High stakes and High time pressure. We expect subjects to submit their final choice earlier than in the HL treatment because time pressure is higher and monetary rewards are lower at any point in time.

2.3 Theory

This section first presents the main theoretical idea of the economics of test taking. Thereafter, we present a simple economic model.

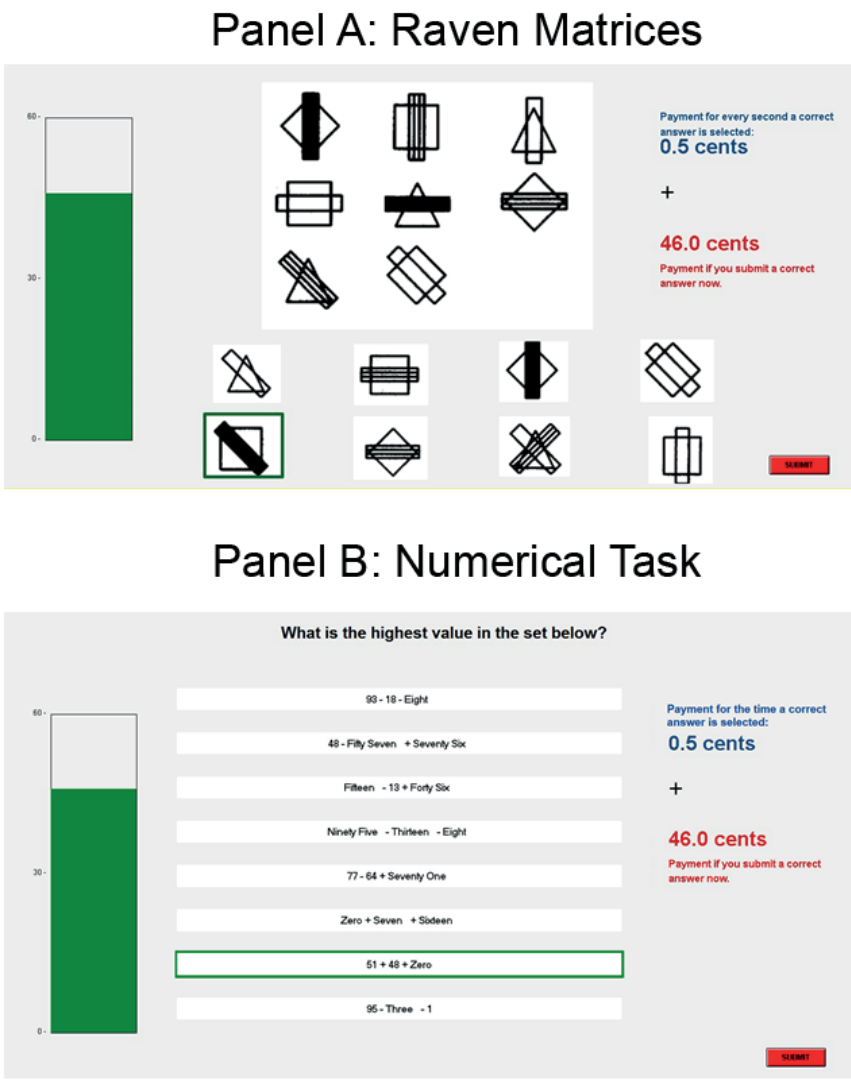


Figure 2.2: Screenshot of the Decision Screen

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

2.3.1 Main Idea

The decision-making process when solving Raven matrices has three phases: (i) the immediate solution to the matrix that comes to mind, (ii) a search phase for a possible better choice, and (iii) a final choice phase. The main input is time. After the immediate choice, subjects begin a choice process during which they search for a possible better solution. The length of this choice process depends on the quality of the first choice, the complexity of the matrix and the disutility of time. The probability that the search for a better solution is continued in the second phase becomes lower with contemplation time.[5] The timing of the final choice depends on the expected increase in performance when continuing searching for better alternatives and the disutility of time.[6]

The quality of the immediate choice that comes to mind depends on ability. More able subjects have a higher probability to find the solution immediately for all levels of complexity. In addition, a more complex type of the problem will reduce the quality of the immediate choice for all subjects. The blue system, which yields a payoff of 0.5 cents per second for the right solution, provides an incentive to reveal the first choice as fast as possible. We assume that a subject's first click is a signal of cognitive ability.

The search phase is costly because of the disutility of time. Subjects face the option of staying with their immediate choice or searching for a possible better solution. Searching for a better solution has two possible outcomes. First, subjects could find out that the immediate choice is the best choice. In this case they will not change their choice, but they do incur search costs. Second, subjects find a better choice and switch to this alternative. The search process can continue until they run out of time or until the disutility of time becomes larger than the probability of finding a better alternative (e.g., Gabaix et al. (2006)). We monitor the search process by counting the number of changes during the search process in the blue system. In the blue system subjects face an incentive to search for the best possible choice because they are rewarded with 0.5 cents per second for each second during the 60-second time period during which they have selected the right

[5] More formally, the probability function of individual i for having the correct answer on question q in test environment τ has the following properties: $0 \leq p_{q,i,}(t|V = v) \leq 1$, $\frac{\partial p_{q,i,}(t|V=v)}{\partial t} > 0$, $\frac{\partial^2 p_{q,i,}(t|V=v)}{\partial^2 t} < 0$ for $\forall t > 0$. The variable $t \in (0, +\infty]$ is the time investment made to solve a matrix. V represents a vector of catchall variables which captures all other factors influencing this probability except for the variable t.

[6] Similar approaches have been developed by Manzini and Mariotti (2007); Caplin and Dean (2011); Reutskaja et al. (2011). However, they do not incorporate the first stage of an immediate solution, which is relevant in the case of solving Raven matrices.

solution to the Raven matrix.

Finally, subjects face an incentive to limit the length of the search process and reveal their final choice. The reason for this is that they face a trade-off between searching for the right answer and the disutility of time. The probability of improving the outcome is decreasing with contemplation time, while the disutility of time is increasing with contemplation time. In addition, the red system provides a monetary incentive to make a final decision as fast as possible. For each unit of time a subject waits the reward for a correct solution to a Raven matrix falls.

2.3.2 A Simple Model

In the following we assume that a test taker is risk neutral when she answers a question. Moreover we assume, that she only cares about the monetary stakes provided in the experiment. The intrinsic motivation in answering a question is crowded out because of the salience of our incentive scheme. The utility function of subject i of answering question q in test environment τ can be written as follows:

$$U_{i,q,\tau}(t) = p_{i,q,\tau}(t) \cdot (F_\tau - \gamma_\tau t) \tag{2.1}$$

In equation 2.1, $p_{q,i,\tau}(t)$ is the probability function of knowing the answer at point t. t measures time in seconds, F is a fixed payment and γ stands for the amount deducted per second.

The test taker optimizes the time when she submits an answer in each respective test environment. The optimal timing of submitting the solution to a Raven matrix is different in different treatments. Maximizing 2.1 with respect to t yields

$$\frac{\partial U_{i,q,\tau}(t)}{\partial t} = \frac{\partial p_{i,q,\tau}(t)}{\partial t} F_\tau - \frac{\partial p_{i,q,\tau}}{\partial t} \gamma_\tau t - p_{i,q,\tau}(t)\gamma_\tau = 0 \tag{2.2}$$

and a positive optimal time investment t^* for sufficiently high $\frac{F_\tau}{\gamma_\tau}$:[7]

$$t^* = \frac{F_\tau}{\gamma_\tau} - \frac{p_{i,q,\tau}(t)}{\frac{\partial p_{i,q,\tau}(t)}{\partial t}} \tag{2.3}$$

[7]We assume that there exists an interior solution. The second-order condition shows that 2.3 is indeed a maximum: $\frac{\partial^2 U_{q,i,\tau}}{\partial^2 t} = \frac{\partial^2 p_{q,i,\tau}(t)}{\partial^2 t}(F_\tau - \gamma_\tau t) - 2\frac{\partial p_{q,i,\tau}(t)}{\partial t}\gamma_\tau < 0$. The first term on the right hand side is smaller than zero since the second derivative of $p(t)$ with respect to t is negative and the second term is smaller than zero as well which makes the whole equation become negative.

Note that the optimal t also depends on the ratio of the value of the probability function and the increase in probability. We explicitly model this probability function dependent on the test environment. The reason for this is that (for instance) low stakes could decrease the intensity of thinking and thus also influence the probability when a subject comes up with a correct answer. Also, very high stakes or time pressure could yield the same result. Some test takers could choke under too high pressure. If the functional form changes under different circumstances is an empirical question.

To predict differences in behavior between treatments, we present comparative statics with respect to the time pressure parameter γ and the incentive parameter F. We write t^* as a function of γ and F and take the first derivative of 2.3 with respect to γ and F. Solving for $\frac{\partial t^*(\gamma)}{\partial \gamma}$ and $\frac{\partial t^*(F)}{\partial F}$ yields two predictions.

First, optimal time investment decreases with increasing disutility of time:

$$\frac{\partial t^*(F,\gamma)}{\partial \gamma} = -\frac{F_\tau}{\gamma_\tau^2} < 0 \tag{2.4}$$

This implies that, in order to behave optimally, subjects should submit their solutions to the Raven matrices earlier when they face higher time pressure. Second, optimal time investment increases with increasing stakes:

$$\frac{\partial t^*(F,\gamma)}{\partial F} = \frac{1}{\gamma_\tau} > 0 \tag{2.5}$$

This implies that, in order to behave optimally, subjects should submit their solutions to the Raven matrices later when monetary stakes are higher. In terms of the three different treatments in the experiment, this simple model predicts that subjects submit their solutions earlier in the treatment with high monetary stakes and high time pressure relative to the treatment with high monetary stakes and low time pressure (first prediction) and that subjects submit their solutions later in the treatment with high monetary stakes and low time pressure relative to the treatment with low monetary stakes and low time pressure (second prediction). The underlying assumption is that the probability function remains the same in all test environments.

2.4 Results

In this section we present our main results. We plot the probability of knowing the correct answer over time for the three different treatments. We also compare submission behavior between different treatments for the Raven matrices and numerical problems and examine whether the behavior is in line with what we expect from economic theory. Finally, we analyze to what extent behavior in each treatment can be understood from an economic point of view.

2.4.1 Baseline Treatment

Probability of Knowing the Correct Answer over Time

Table 2.1 shows the probability of knowing the correct answer over time in the baseline treatment. Panel A reports the average probability of a correct answer over time in terms of the fraction of correct answers and Panel B reports the average cumulative earnings of the blue system. We document the probabilities and earnings for both the Raven matrices and the numerical problems. The Raven matrices are split into three levels of complexity. 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.[8] The unit of observation in Table 2.1 is the matrix or the numerical problem. Standard errors are clustered at the level of questions and subjects. The total number of observations equals $1,920$, which is the result of 128 subjects solving 15 Raven matrices or numerical problems.[9] We document the probability of a correct answer and the cumulative earnings from the blue system at six points during the 60-second time period. The six points 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.

The picture that emerges from Panel A is that the probability of a correct answer and earnings in the blue system rise over time in all columns. The overall performance on the Raven matrices suggests that about half (0.509) of the matrices were solved correctly within 60 seconds. The rate of improvement is highest

[8] 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 end up with 14 easy and moderate questions and 17 difficult questions More information on the items can be found in Appendix A.

[9] We lost seven observations from the solutions to the Raven matrices because of a computer problem in the lab.

Table 2.1: Performance on Raven Matrices and Numerical Tasks

Panel A. Success Rate								
	Raven Matrices							Numerical Tasks
Time elapsed (sec.)	All	Easy vs. Moderate	Easy	Moderate vs. Difficult	Moderate	Easy vs. vs. Difficult	Difficult	All
10	0.156		0.216		0.139		0.098	0.315
	(0.008)	+++	(0.015)	++	(0.014)	+++	(0.012)	(0.011)
0 vs. 10	***		***		***		***	***
20	0.311		0.457		0.285		0.151	0.523
	(0.011)	+++	(0.018)	+++	(0.019)	+++	(0.015)	(0.011)
10 vs. 20	***		***		***		***	***
30	0.409		0.604		0.366		0.204	0.626
	(0.011)	+++	(0.018)	+++	(0.02)	+++	(0.017)	(0.011)
20 vs. 30	***		***		***		**	***
40	0.457		0.664		0.412		0.24	0.687
	(0.011)	+++	(0.017)	+++	(0.02)	+++	(0.018)	(0.011)
30 vs. 40	***		**		ns		ns	***
50	0.499		0.704		0.466		0.271	0.734
	(0.011)	+++	(0.017)	+++	(0.021)	+++	(0.018)	(0.01)
40 vs. 50	***		*		*		ns	***
60	0.509		0.708		0.486		0.28	0.753
	(0.011)	+++	(0.017)	+++	(0.021)	+++	(0.019)	(0.01)
50 vs. 60	ns		ns		ns		ns	***
Observations	1913		740		590		583	1920

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

			Raven Matrices					Numerical Tasks
Time elapsed (sec.)	All	Easy vs. Moderate	Easy	Moderate vs. Difficult	Moderate	Easy vs. Difficult	Difficult	All
10	0.514		0.621		0.505		0.387	0.948
	(0.031)	ns	(0.051)	ns	(0.057)	+++	(0.05)	(0.036)
0 vs. 10	***		***		***		***	***
20	1.758		2.43		1.619		1.047	3.142
	(0.068)	+++	(0.117)	+++	(0.122)	+++	(0.106)	(0.08)
10 vs. 20	***		***		***		***	***
30	3.603		5.151		3.302		1.944	6.058
	(0.109)	+++	(0.182)	+++	(0.191)	+++	(0.164)	(0.122)
20 vs. 30	***		***		***		***	***
40	5.797		8.361		5.253		3.093	9.364
	(0.151)	+++	(0.246)	+++	(0.262)	+++	(0.226)	(0.161)
30 vs. 40	***		***		***		***	***
50	8.217		11.834		7.452		4.4	12.936
	(0.195)	+++	(0.307)	+++	(0.337)	+++	(0.291)	(0.199)
40 vs. 50	***		***		***		***	***
60	10.768		15.407		9.872		5.786	16.67
	(0.239)	+++	(0.369)	+++	(0.417)	+++	(0.359)	(0.236)
50 vs. 60	***		***		***		***	***
Observations	1913		740		590		583	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 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.

in the first 30 seconds (162.2 percent improvement). After 30 seconds the probability equals 0.409, after which it improves to 0.509 after 60 seconds (24.4 percent improvement). There are differences in the performance across the different levels of complexity. For relatively easy matrices the probability equals about 70 percent (0.708), 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. After 30 seconds the rates of improvement are 10.4, 12.0 and 7.6 percent for the easy, medium and difficult Raven matrices, respectively.

Statistical differences between consecutive points in time and within columns are denoted by an asterisk (∗). It turns out that in terms of probabilities there are statistical differences. 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 30^{th} second onwards. Statistical differences between complexity levels are denoted by a plus (+). We report 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. Payoffs are higher for relatively easy Raven matrices and lower for the more difficult ones. The patterns of statistically significant differences confirm the findings of Panel A.

The final column in Panels A and B shows the performance on solving numerical problems. The pattern that emerges from this column is most comparable to the easy Raven matrices. This does not come as a surprise because university students should be able to solve numerical problems of the complexity level we have chosen. In the end the probability of a knowing the correct answer equals 75.2 per cent, with average cumulative earnings of 16.7 cents. Statistically significant differences are obtained between different time intervals.

Table 2.2: Submission Behavior on Raven Matrices and Numerical Tasks

Panel A. Submissions (Percentage)								
			Raven Matrices					Numerical Tasks
Time elapsed (sec.)	All	Easy vs. Moderate	Easy	Moderate vs. Difficult	Moderate	Easy vs. vs. Difficult	Difficult	All
10	0.068		0.099		0.056		0.043	0.064
	(0.006)	+++	(0.011)	ns	(0.009)	+++	(0.008)	(0.006)
0 vs. 10	***		***		***		***	***
20	0.289		0.412		0.246		0.177	0.354
	(0.01)	+++	(0.018)	+++	(0.018)	+++	(0.016)	(0.011)
10 vs. 20	***		***		***		***	***
30	0.543		0.697		0.503		0.388	0.679
	(0.011)	+++	(0.017)	+++	(0.021)	+++	(0.02)	(0.011)
20 vs. 30	***		***		***		***	***
40	0.729		0.853		0.698		0.604	0.863
	(0.01)	+++	(0.013)	+++	(0.019)	+++	(0.02)	(0.008)
30 vs. 40	***		***		***		***	***
50	0.875		0.949		0.863		0.792	0.961
	(0.008)	+++	(0.008)	+++	(0.014)	+++	(0.017)	(0.004)
40 vs. 50	***		***		***		***	***
60	0.963		0.988		0.954		0.942	0.993
	(0.004)	+++	(0.004)	ns	(0.009)	+++	(0.01)	(0.002)
50 vs. 60	***		***		***		***	***
Observations	1913		740		590		583	1920

Panel B. Correct Submissions (Percentage)

	Raven Matrices							Numerical Tasks
Time elapsed (sec.)	All	Easy vs. Moderate	Easy	Moderate vs. Difficult	Moderate	Easy vs. vs. Difficult	Difficult	All
10	0.02		0.039		0.014		0.002	0.036
	(0.003)	+++	(0.007)	++	(0.005)	+++	(0.002)	(0.004)
0 vs. 10	***		***		**		ns	***
20	0.147		0.254		0.114		0.046	0.239
	(0.008)	+++	(0.016)	+++	(0.013)	+++	(0.009)	(0.01)
10 vs. 20	***		***		***		***	***
30	0.286		0.45		0.251		0.113	0.468
	(0.01)	+++	(0.018)	+++	(0.018)	+++	(0.013)	(0.011)
20 vs. 30	***		***		***		***	***
40	0.364		0.539		0.336		0.17	0.595
	(0.011)	+++	(0.018)	+++	(0.019)	+++	(0.016)	(0.011)
30 vs. 40	***		***				***	***
50	0.412		0.581		0.393		0.216	0.653
	(0.011)	+++	(0.018)	+++	(0.02)	+++	(0.017)	(0.011)
40 vs. 50	***		ns		**		**	***
60	0.439		0.597		0.432		0.244	0.671
	(0.011)	+++	(0.018)	+++	(0.02)	+++	(0.018)	(0.011)
50 vs. 60	*		ns		ns		ns	Ns
Observations	1913		740		590		583	1920

Panel C. Earnings (Cents)								
	Raven Matrices							Numerical Task
Time elapsed (sec.)	All	Easy vs. Moderate	Easy	Moderate vs. Difficult	Moderate	Easy vs. vs. Difficult	Difficult	All
10	6.971		9.737		5.473		0.875	12.89
	(0.116)	+++	(0.165)	+++	(0.216)		(0.11)	(0.13)
20	12.095		14.772		10.649		6.508	13.718
	(0.062)	+++	(0.076)	+++	(0.122)	+++	(0.137)	(0.051)
10 vs. 20	***		***		***		***	***
30	8.917		11.102		8.743		5.581	11.601
	(0.051)	+++	(0.074)	+++	(0.091)	+++	(0.094)	(0.042)
20 vs. 30	***		***		***		***	***
40	5.476		7.499		5.329		3.72	9.122
	(0.048)	+++	(0.085)	+++	(0.085)	+++	(0.074)	(0.045)
30 vs. 40	***		***		***		***	***
50	3.289		4.143		3.574		2.575	6.112
	(0.038)	+++	(0.084)	+++	(0.066)	+++	(0.055)	(0.05)
40 vs. 50	***		***		***		***	***
60	1.368	+++	2.124	+++	1.77	+++	0.868	2.955
	(0.024)		(0.069)		(0.045)		(0.027)	(0.054)
50 vs. 60	***		***		***		***	***
Observations	1913		740		590		583	1920

Note. Panel A shows the fraction of people who submit after their answer after the 10^{th}, 20^{th} 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.

Submission Behavior

Table 2.2 presents the performance in the red system in three panels. Panel A reports the cumulative fraction of submitted answers at different points in time during the 60-second time period in which the answer to a question has to be submitted. Panel B reports the cumulative fraction of correctly submitted answers at the same points in time. Finally, Panel C reports the average earnings of answers that were correct and submitted within the different intervals. The features of the numbers in this table are similar to the ones in Table 2.1 above.

Panel A shows that the vast majority of Raven matrices has been answered (96.4 per cent). Subjects failed more often to submit an answer to the most difficult Raven matrices. Answers to almost all numerical problems have been submitted (99.3 per cent). The pattern suggests that on average half of the answers to the Raven matrices have been submitted after 30 seconds. Only for the most difficult matrices this number is lower after 30 seconds (38.8 per cent). In terms of statistically significant differences, we observe that all comparisons are different at the one per cent level.

The numbers in Panel B of Table 2.2 suggest that correct submissions increase with time, but that the pattern is concave. This is generally true for all types of questions. 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 50^{th} 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. For all types the ratio peaks after 30 seconds and declines afterwards. This suggests that those who submit their answers between the 20^{th} and 30^{th} second are more likely to submit the right answer relative to those submitting earlier or later.

Finally, Panel C of Table 2.2 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 2.1, where we document cumulative earnings). The low earnings up to the 10^{th} 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 questions. 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 per cent level.

2.4.2 Increasing Stakes and Time Pressure

In the following we analyze choice and submission behavior in the two other treatments and compare it to our baseline treatment. The main message is that the probability of knowing the correct answer over time does not depend on the reward for submitting an answer. This implies that choice behavior on the blue system does not differ across the three different treatments. However, our test takers wait longer to submit an answer when stakes are higher. This implies that the behavior on the red system varies in the HL and HH treatment compared to our baseline treatment.

Probability of Knowing the Correct Answer over Time

Figure 2.3 shows the probability of knowing the correct answer over the 60 seconds for various scenarios. The Figure is a graphical equivalent to Panel A of Table 2.1 for all treatments. Panels A and B document the probability of knowing the correct answer over time for the Raven matrices and numerical tasks. We construct these functions by taking the mean of the correctly selected answers over all questions and subjects in a respective treatment at each second. The gray areas indicate the 95 per cent confidence intervals. In Panel A we show the curve of the baseline treatment and the curves of the probabilities of the HL treatment and the HH treatment 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 the baseline treatment (LL). These graphs show that the speed of finding a correct answer 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 test environments. All figures show that the fraction of correctly selected answers increases over time. The pattern reveals a concave relationship between the time of selection and the probability of a correct answer.

Submission Behavior

We continue by comparing submission behavior between the three incentive schemes for submitting a correct answer. In Panels C and D of Figure 2.3 we compare submission behavior between the baseline treatment and the treatment of high stakes and low time pressure (HL) and the treatment with high stakes and high time

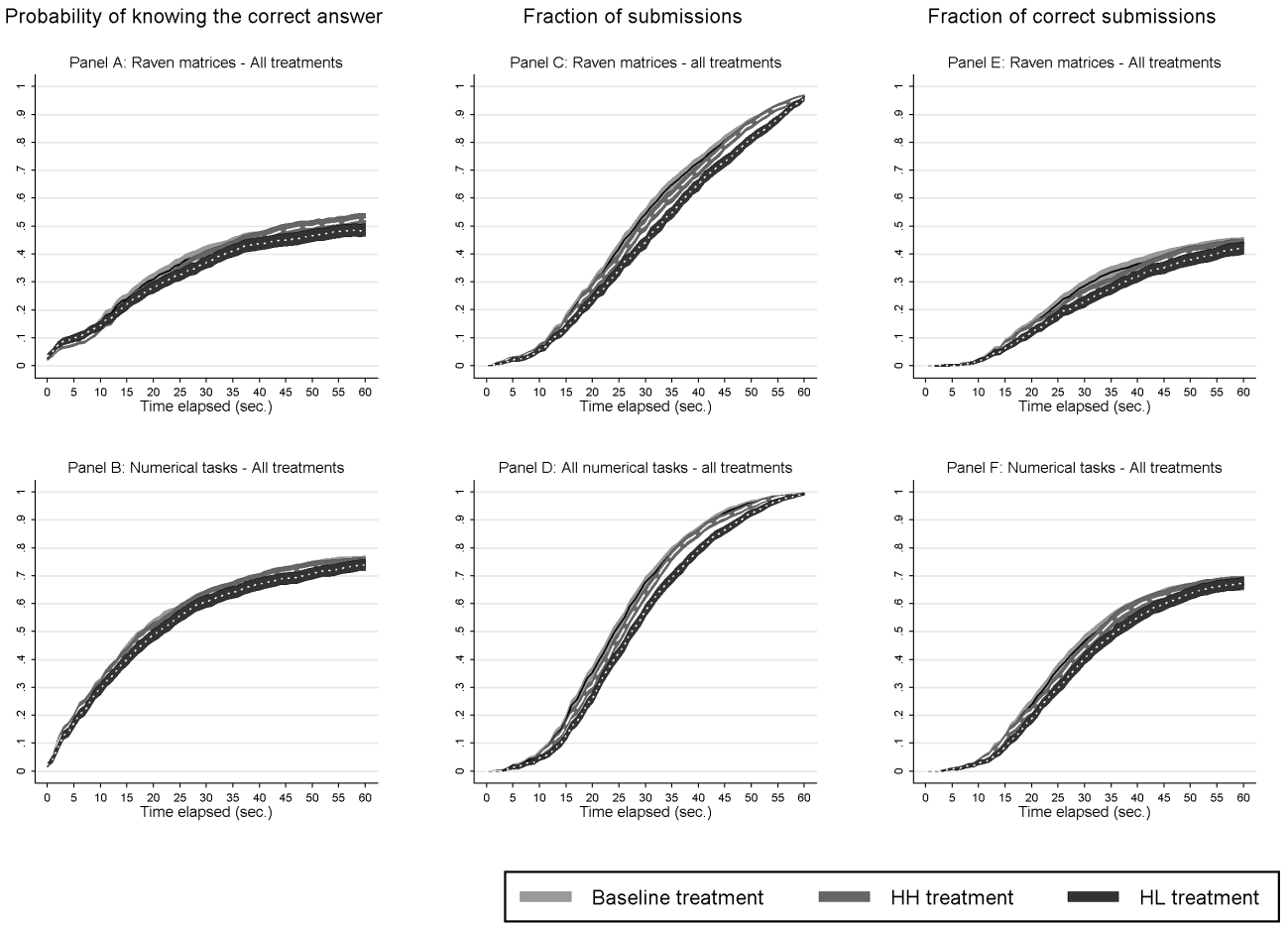


Figure 2.3: Performance on Raven Matrices and Numerical Tasks

Note. Panels A and B report the probability of knowing the correct answer over time for the Raven matrices and the numerical tasks. 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 gray areas indicate 95% confidence bounds.

pressure (HH). Panel C shows the cumulative fraction of submitted answers of the Raven matrices and Panel D shows the fraction of submissions for the numerical tasks. The light gray area with the straight black line shows the cumulative fraction with 95% confidence bounds for the baseline treatment, the gray area with the dashed white line shows the results for the HH treatment and the dark gray area with white dots documents the results for the HL treatment. The pattern in Panel C suggests that subjects change their submission behavior across treatments for the Raven matrices. In the first 20 seconds the curves of all treatments overlap but afterwards answers in the baseline treatment have been significantly faster submitted than in the HL treatment and the HH treatment. Even though subjects submitted fastest in the baseline treatment, we do not obtain statistical significant differences between the baseline treatment and the HH treatment. Panel D shows that behavior is similar for both Raven matrices and numerical tasks. However, answers to the numerical tasks were overall faster submitted than to the Raven matrices.

In a next step we analyze the fraction of correctly submitted answers over time. This provides information about the effect of stakes and time pressure on test scores. Panels E and F of Figure 2.3 show the fraction of correctly submitted answers for the Raven matrices and the numerical tasks in all three treatments. The picture that emerges from these graphs is that the fraction of correctly submitted answers is different across treatments. Panels E and F reveal that in the treatment with higher stakes (HL), compared to the baseline treatment, subjects wait longer until they submit a correct answer. This is in line with what we expect from economic theory. In contrast to our theoretical prediction, we obtain no significant differences in submission behavior between the HL and HH treatment.

In Table 2.3 we look at the submission behavior from a different perspective. We report panel regressions with question and individual fixed effects and the submission time in seconds as the dependent variable. We include treatment dummies for the HL and HH treatment. Columns (1) to (3) report the results for the Raven matrices and columns (4) to (6) report on the numerical tasks. The picture that emerges from Table 2.3 is the following: Test takers change submission behavior significantly in the HL treatment compared to the baseline treatment when they are confronted with the Raven matrices. We do not obtain a significant difference in the timing of submission if we compare the HH treatment and the baseline treatment. The coefficient of the HH treatment and the HL treatment 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 answer is rather small. The maximum change we observe is that subjects wait on average three seconds longer until they submit their answer in the HL treatment compared to the baseline treatment. As Table 2.3 shows this change in timing is equally small for solving Raven matrices and numerical tasks. In the next section we will explore possible reasons for this inelastic adaption behavior.

2.4.3 Maximizing Expected Earnings? Adaption to the Test Environment

In this section we investigate if subjects submitted their answers in an optimal way in different test environments. Theory predicts that, if everything else is kept equal, a higher reward for a correct submission should increase the time until an answer is submitted. Our results suggest that subjects systematically deviate from the optimal submission time. However these deviations do not necessarily yield lower expected earnings because they are relatively small.

Figure 2.7 shows the expected earnings of the Raven matrices in all three treatments. We calculate the expected earnings by multiplying the fraction of correct answers in each treatment with the payoff from the red payment system at each point in time. Panel A shows the results in the baseline treatment. Panels B and C present the results from HH and HL treatment. 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 gray lines indicate the respective 95 per cent confidence bounds. The dotted gray line indicates the time when the expected earnings reach the maximum.

All treatments reveal different times which maximize the expected earnings. In the baseline treatment expected earnings peak at the 26^{th} second for the Raven matrices and the 22nd second for the numerical tasks. In the HH treatment the optimal submission time is the 27^{th} second for the Raven matrices and 25^{th} second for the numerical tasks. The difference between the treatments is the strongest for the HL treatment. Earnings for the Raven matrices peak at the 40^{th} second and for the numerical task at the 39^{th} second. All observed average submission times deviate statistically significantly from the optimal submission times and all panels reveal concave patterns. Subjects submit their answer 4.1 (3.8) seconds later in the baseline treatment when solving Raven matrices (numerical tasks),

Table 2.3: Determinants of Submission Time

	(1)	(2)	(3)	(4)	(5)	(6)
	Raven Matrices			Numerical Tasks		
HH - treatment	0.61	0.939*	0.615	0.876**	1.012**	0.876**
	(0.425)	(0.514)	(0.464)	(0.351)	(0.461)	(0.421)
HL - treatment	2.834***	3.199***	2.854***	3.163***	3.224***	3.163***
	(0.407)	(0.543)	(0.492)	(0.328)	(0.529)	(0.477)
Constant	30.335***	30.103***	22.781***	25.881***	25.816***	23.512***
	(0.241)	(0.29)	(0.986)	(0.182)	(0.274)	(0.743)
Observations	5,741	5,741	5,741	5,760	5,760	5,760
R-squared	0.008	0.01	0.311	0.014	0.019	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 last row reports the p-value of the F-test which tests the equality of both treatment dummies. Robust standard errors are reported in parentheses. We cluster standard errors on the question level in columns (1) and (4). In columns (2),(3),(5) and (6) standard errors are clustered on the individual level. *** p$<$ 0.01, ** p$<$ 0.05, * p$<$ 0.1.

compared to the optimal submission time (two-tailed t-tests, p-value$<$ 0.001). In the HH treatment submissions are 4.0 (1.8) seconds later (two-tailed t-tests, p-value$<$ 0.001). In the HL treatment the submit button is pressed 6.7 (10.0) seconds earlier compared to the point which maximizes expected payoffs time (two-tailed t-tests, p-value$<$ 0.001).

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 average point of submission with the expected earnings at the optimal time of submission. 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. In the baseline treatment this difference is not significantly different from zero both for Raven matrices and numerical 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). In the HL treatment subjects could have earned on average 1.0 cents (0.7 cents) more for each question if they would have thought longer before submitting their answer. The difference is significant at the 1 per cent level (p-value$<$ 0.005 for the numerical tasks). In the HH treatment 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 per cent 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).

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 gain subjects make by adjusting their behavior. Another reason why we do not observe strong changes in submission behavior could be due to heterogeneity in our data.

2.5 Why Do Test Takers Adjust Their Answering Behavior So Little?

We documented that the probability of a correct answer hardly changes across test environments. Moreover, we showed that the submission behavior changes in the way our economic model predicts. However, the adjustment in the timing of submission is rather small and the potential gains in expected earnings are minor.

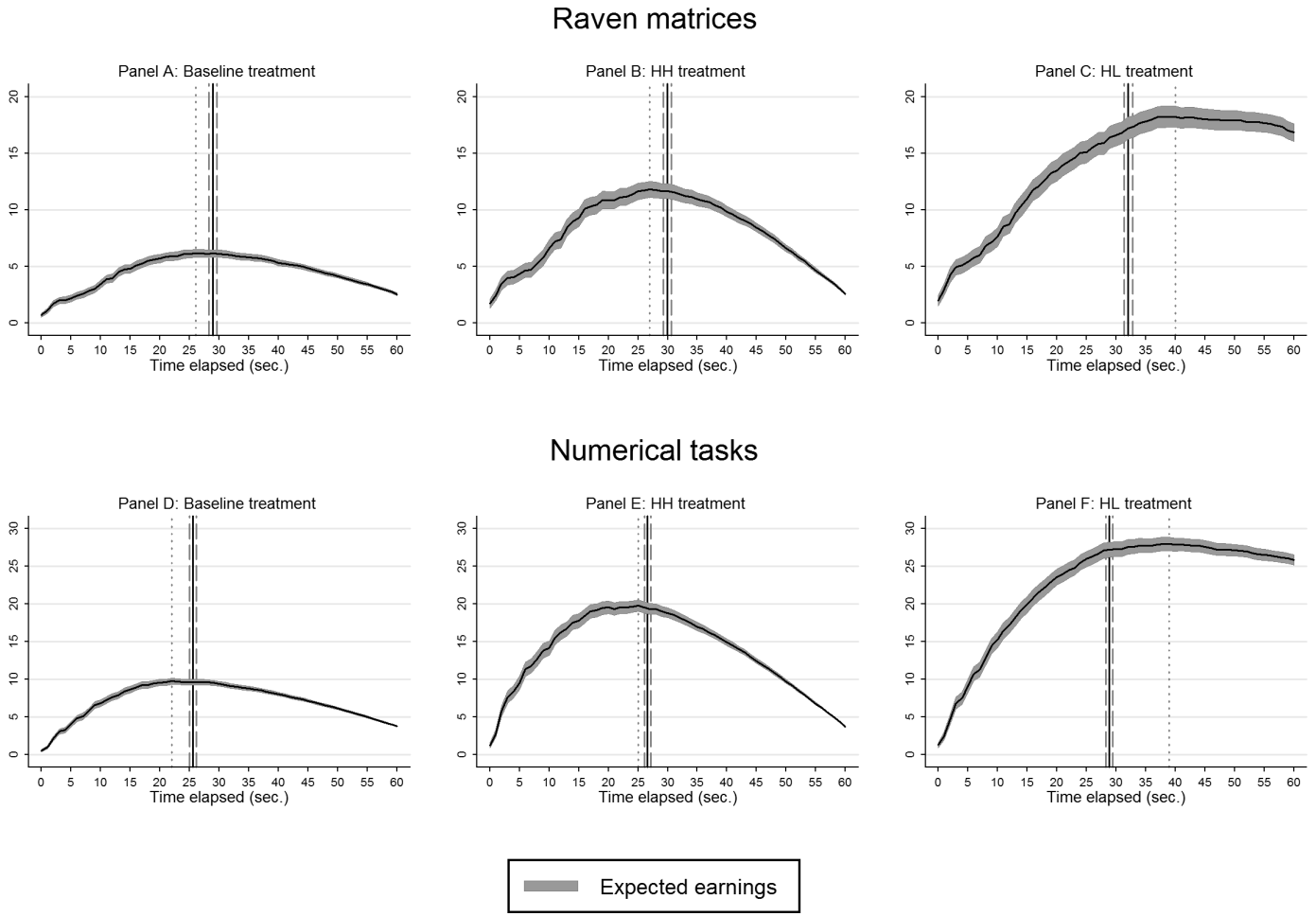


Figure 2.4: Expected Earnings and Submission Times of Raven Matrices and Numerical Tasks

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

We now look at different subsamples of our data to see whether the results change.

2.5.1 Probability of Knowing the Correct Answer and Submission Behavior

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

Heterogeneity in Questions

Panels A, B and C of Figure 2.5 show the probability of knowing the correct answer over time for three different levels of difficulties. Note that 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 questions was randomized across treatments and subjects. The gray areas indicate the 95 per cent confidence intervals. Similar to the aggregate level presented in Figure 2.3, the confidence intervals overlap in all treatments and for all degrees of difficulty. This shows that also for different degrees of difficulty the intensity of thinking does not seem to vary between different incentive schemes for submitting an answer.

Panels D, E and F of Figure 2.5 show submission behavior for different levels of difficulty. These panels show the cumulative fraction of easy, moderate and difficult questions in all three 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. Panel E and F reveal that the submission behavior varies between treatments for moderate and difficult questions. Subjects submit their answer earliest in the baseline treatment and latest in the HL treatment. Similar to our results at the aggregate level, we cannot identify significant differences between the HH and the baseline treatment in submission behavior.

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 answer. Panels G, H and I of Figure 2.5 document the fraction of correctly submitted answers over time for all three treatments.

The pattern that emerges from these panels suggests that the fraction of correctly submitted answers is only different between the baseline treatment and the HL treatment for easy and moderate questions. We do not obtain significant differences in submission behavior of correct answers between the HH treatment and the baseline treatment and the HH treatment and the HL treatment for easy questions. However, moderately difficult questions are significantly faster answered in the baseline treatment than in the HL treatment. Moreover, we do not obtain significant differences between the treatments in the fraction of correctly submitted answers for difficult questions. Overall our analysis of the data shows that subjects submit a correct answer faster if they face lower incentives for easy and moderate questions. Difficult questions do not seem to trigger different behavior in our setup. Most interestingly, the fraction of correctly submitted answers in the 60^{th} second does not differ across treatments.

Heterogeneity in Performance

Panels A, B and C of Figure 2.6 show the probability of knowing the correct answer over time for different levels of performance. We take the fraction of correctly selected answers in the 30^{th} second as a performance measure and split the sample into three groups of equal size.[10] We split the sample in high, moderate and low performance types.

Panels D, E and F in the middle of Figure 2.6 document the cumulative fraction of submitted answers for high, moderate and low performers in all treatments. 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 to moderate and low performance types. Individuals with a high score submit their answer significantly later in the HL and HH treatment compared to the baseline treatment. Subjects with a lower performance also submit their answers in the HL and HH later than in the baseline treatment. However, the difference is smaller than for high performance types.

Next, we analyze submission behavior of a correct answer for different levels of performance. Panels G, H and I of Figure 2.6 show the fraction of correctly submitted answers across all three treatments. The picture that emerges from these panels is that subjects with a higher performance also adapt their submission be-

[10]The results do not change if we take performance at the 20^{th}, 40^{th}, 50^{th} or 60^{th} second as the criterion to split the sample into three performance types.

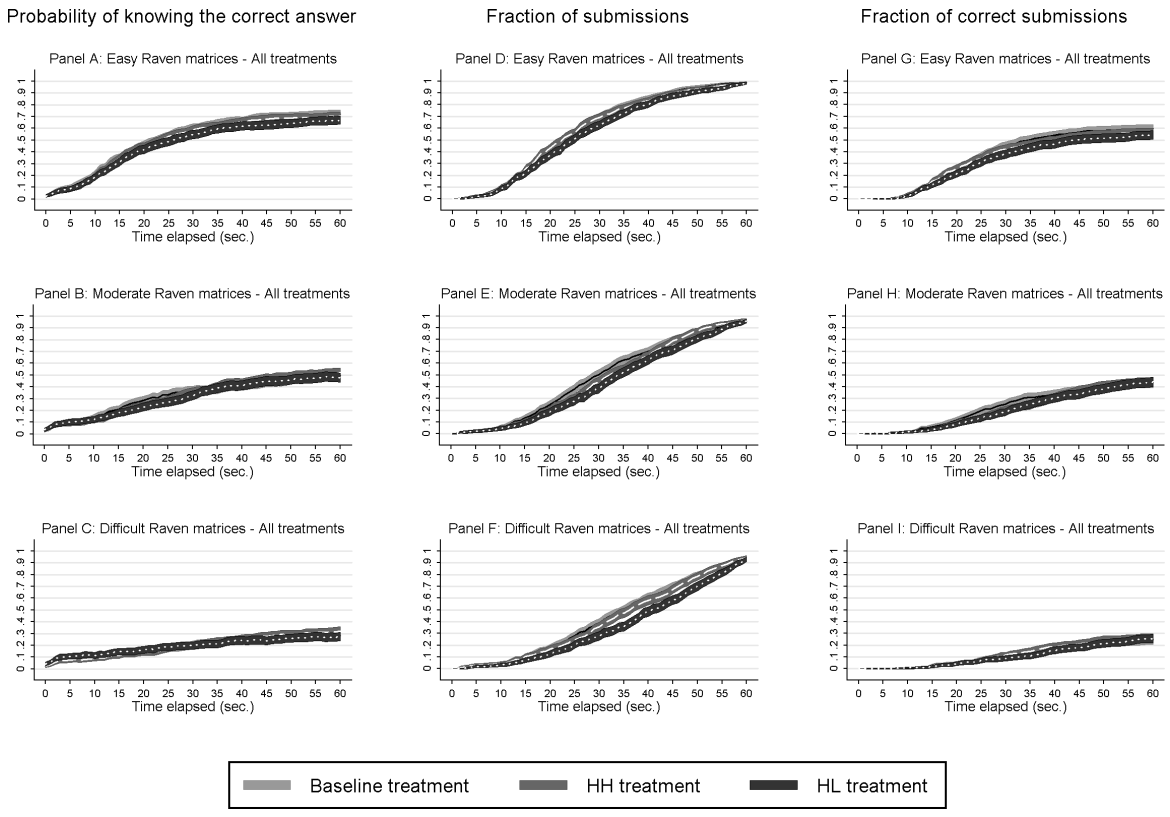


Figure 2.5: Heterogeneity in Questions between Raven Matrices

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 gray areas indicate the 95% confidence bounds.

havior of correct answers stronger to the test environment relative to subjects with a moderate or low performance. High performance subjects wait longer until they submit correct answers in a test environment with higher incentives, compared to subjects with moderate or low performance. These results are in line with what we find for submission behavior in Panels D, E and F.[11]

In conclusion, it seems to be the case that high performers show the strongest adjustment of behavior between different treatments. The next question is whether submission behavior is optimal in terms of payoff maximization.

2.5.2 Expected Earnings

Heterogeneity in Questions

Up to this point we constructed the expected payoff function by aggregating the fraction of correctly submitted answers over all subjects and question types. We now analyze heterogeneity in earnings. Figure 2.7 shows the expected payoffs for the Raven matrices in all three treatments for easy, moderate and difficulty questions. Panels A, B and C show the results for easy questions across all three treatments, Panels D, E and F for moderate and G, H and I for difficult questions. The vertical black line indicates the average submission time and the dotted gray line the optimal submission time for each level and each treatment.

The pattern of submission behavior is heterogeneous for different levels of difficulty. If subjects are faced with an easy question, they submit their answers too early compared to the optimal submission time in the baseline and the HL treatment. In the HH treatment the actual submission time is not significantly different from the optimum. All average submission times for moderate questions deviate from the point where expected earnings peak. In the baseline treatment and the HH treatment answers are always submitted too late. Subjects wait too short to submit their answers in the HL treatment. The picture is different for difficult questions. We do not observe a significant difference between actual and optimal behavior with respect to expected payoff maximization in the baseline treatment and the HL treatment for difficult questions.

The picture that emerges from the deviations in expected payoff is mixed. In the baseline treatment expected earnings for easy questions are marginally higher (0.3 cents, p-value= 0.075, two-tailed t-test) in the optimum compared to the

[11] We also investigate submission behaviour of a correct answer for the numerical tasks. The results are consistent with what we find for the Raven matrices. An equivalent to Figure 2.6 can be found in Appendix A.

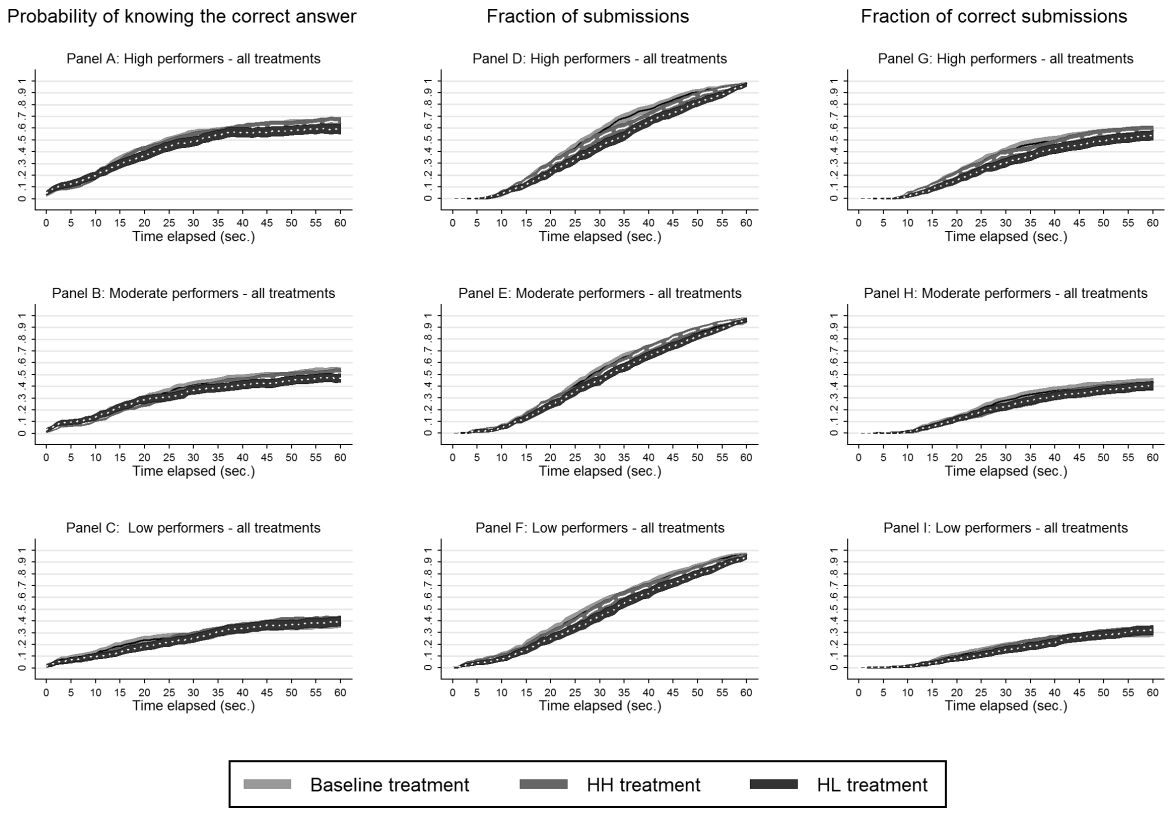


Figure 2.6: Heterogeneity in Performance between Raven Matrices

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 gray areas indicate the 95% confidence bounds.

actual average submission time. In the HH treatment earnings are not significantly different from the optimum, whereas subjects could have earned 2.7 cents more per questions if they would have submitted their answer to easy questions later in the HL treatment (p-value$<$ 0.001, two-tailed t-test). The picture is equivalent for moderate questions. Earnings in the optimum are slightly higher compared to the actual submission time in the baseline treatment. We do not observe significant differences in the HH treatment but expected earnings are 1.6 cents higher at the optimum in the HL treatment (p-value= 0.0145, two-tailed t-test). Difficult questions seem to be answered in a payoff maximizing way. We do not observe differences in expected payoffs for difficult questions in the baseline treatment and the HH treatment, compared to the maximal expected payoff. In the HL treatment subjects could have earned 0.4 cents (p-value= 0.076, two-tailed t-test) more if they would have submitted their answer at the optimum. Since we use a t-test to test for difference between expected payoffs, we already take the correlation of our observations into account. This could potentially yield to significant differences in expected payoffs even if we do not observe significant differences in the submission times.

Heterogeneity in Performance

Figure 2.8 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. Our findings are consistent over different levels of performance. The answers are submitted too late in the baseline treatment and the HH treatment. Moreover, subjects submit their answers too early in the HL treatment.

The picture that emerges from the deviations in expected payoffs is consistent with what we find on the aggregate level. In the baseline treatment only low performance types earn on average 0.5 cents less than in the optimum. In the HH treatment, we cannot identify significant differences in expected payoffs. Only in the HL treatment we identify significantly higher earnings in the optimum, compared to the actual average submission time. Low performance and high performance types would have earned significantly more at the optimal submission time than at their actual average submission time. We do not find significantly higher expected earnings for moderate performance types. An equivalent of Figure

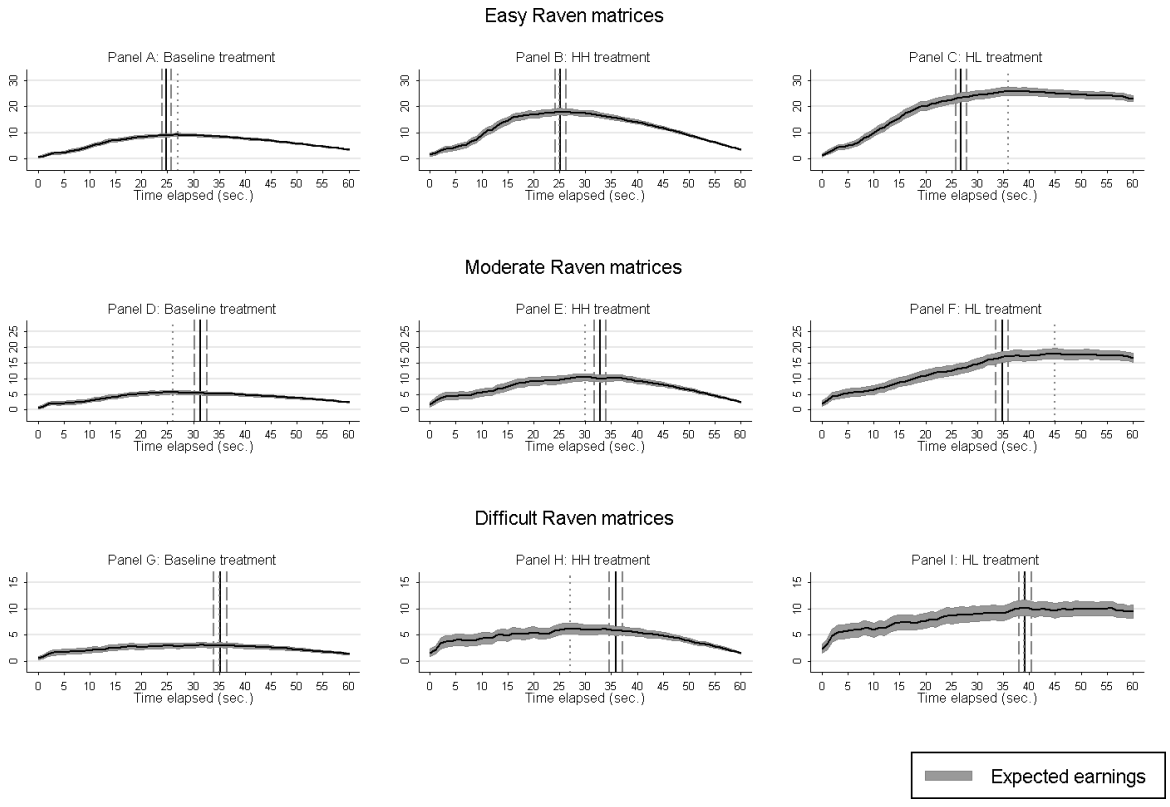


Figure 2.7: Expected Earnings and Submission Times of Raven Matrices by Degree of Difficulty

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

2.8 for the numerical tasks can be found in Figure A.4 in Appendix A.

2.6 Robustness Checks

In this section we provide robustness checks. We first show 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.

2.6.1 Choice Process

To estimate the probability of a correct answer over time, we need to be able to identify an individual's thoughts over time. Since it is not possible to obtain a pure measure of the currently preferred answer, we take behavior in the blue payment scheme as a proxy measure. The blue payment scheme provides an incentive to immediately reveal the answer to a question independent of the actual submitted answer. The data of the selection times show that subjects used the blue payment scheme to indicate their preferred choice in the course of answering a question. 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 question and individual fixed effects and the time of the first choice as the dependent variable. Table 2.4 shows the results. Columns (1) to (3) show the regressions for Raven matrices. The first column only contains question fixed effects. We run a specification only with individual fixed effects in column (2). The specification in column (3) contains question 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.

A second check for independence of the blue payment scheme across treatments is the number of choices subjects make. Table 2.5 shows regression results with the number of choices per question as dependent variable. We again conduct panel regressions with question fixed effects and subject fixed effects. For the Raven matrices none of the treatment dummies is significantly different from zero. The only exception is for numerical tasks in which subjects tend to make fewer choices in the HH treatment compared to the baseline treatment. However, the coefficient is only significant at the 10 per cent level and the point estimate of 0.13 clicks seems to be economically small.

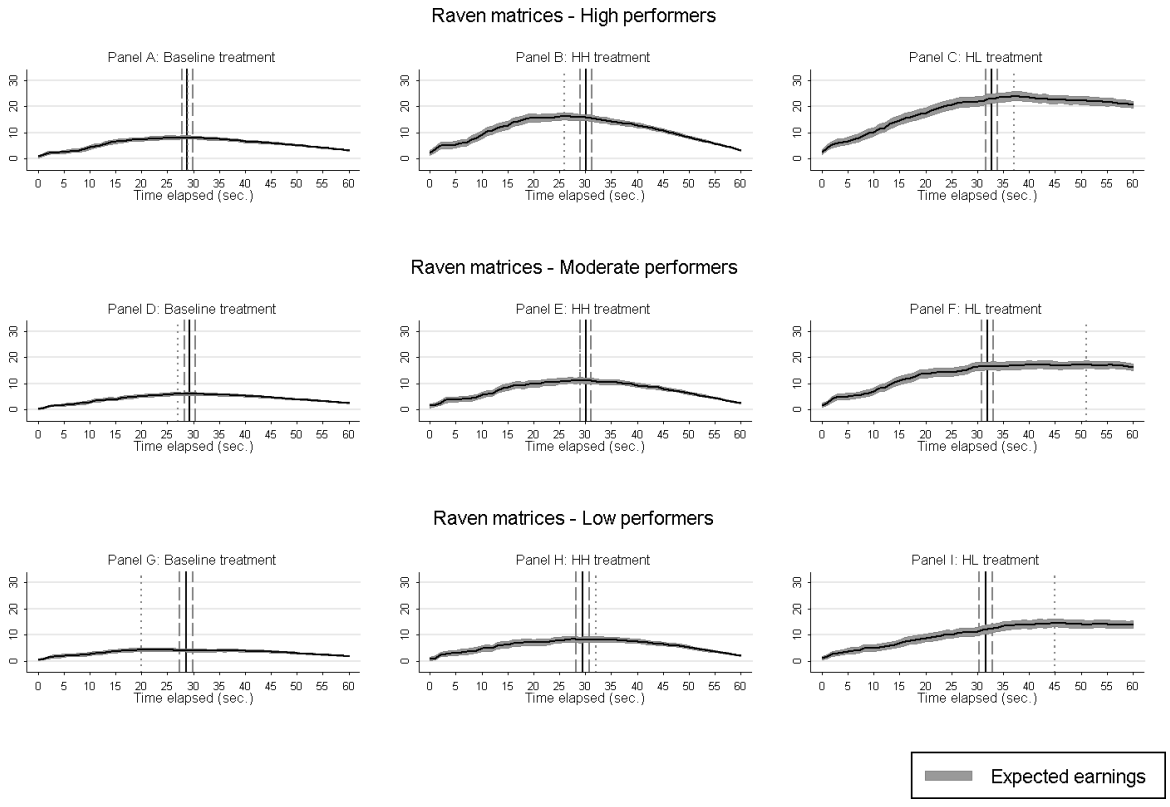


Figure 2.8: Expected Earnings and Submission Times of Raven Matrices by Performance Type

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

Table 2.4: Time of First Choice

	(1)	(2)	(3)	(4)	(5)	(6)
	Raven Matrices			Numerical Tasks		
HH - treatment	0.007	0.059	0.039	0.166	0.173	0.166
	(0.376)	(0.279)	(0.278)	(0.211)	(0.213)	(0.215)
HL - treatment	(0.14	(0.108	(0.149	0.407*	0.394	0.407
	(0.368)	(0.257)	(0.255)	(0.228)	(0.303)	(0.304)
Constant	6.053***	6.025***	5.887***	3.750***	3.752***	3.352***
	(0.23)	(0.157)	(0.602)	(0.122)	(0.164)	(0.227)
Observations	5,741	5,741	5,741	5,760	5,760	5,760
R-squared	< 0.001	< 0.001	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	0.601	0.512	0.462	0.326	0.209	0.171

Note. The table shows panel regression with question fixed and subjects 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 reference group. The last row reports the p-value of the F-test which tests the difference between the coefficients of HH and HL. Robust standard errors are reported in parentheses. We cluster standard errors on the question level in columns (1) and (4). In columns (2),(3),(5) and (6) standard errors are clustered on the individual level. *** p$<$ 0.01, ** p$<$ 0.05, * p$<$ 0.1.

Table 2.5: Number of Choices

	(1)	(2)	(3)	(4)	(5)	(6)
	Raven Matrices			Numerical Tasks		
HH - treatment	0.069	0.075	0.073	-0.141***	-0.128*	-0.141*
	(0.086)	(0.088)	(0.085)	(0.048)	(0.075)	(0.074)
HL - treatment	0.009	0.012	0.01	(0.079)	-0.066	-0.079
	(0.082)	(0.09)	(0.087)	(0.051)	(0.055	(0.054)
Constant	2.809***	2.806***	2.623***	2.718***	2.709***	2.472***
	(0.046)	(0.054)	(0.162)	(0.028)	(0.035	(0.101)
Observations	5,741	5,741	5,741	5,760	5,760	5,760
R-squared	< 0.001	< 0.001	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	0.537	0.379	0.383	0.237	0.431	0.418

Note. The table shows panel regression with question fixed and subjects 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 reference group. The last row reports the p-value of the F-test which tests the difference between the coefficients of HH and HL. Robust standard errors are reported in parentheses. We cluster standard errors on the question level in columns (1) and (4). In columns (2),(3),(5) and (6) standard errors are clustered on the individual level. *** p$<$ 0.01, ** p$<$ 0.05, * p$<$ 0.1.

2.6.2 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.

Panel A of Table 2.6 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 2.6. 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 difficulty of questions. 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.[12] This indicates that our results do not seem to be driven by aggregation bias.

2.7 Conclusion

In this paper we have analyzed test taking from an economic point of view. We propose a new method to shed light on the decision-making process during a cognitive test. By using an experimental set up, we are able to plot the relationship between time investment and the probability of knowing the correct answer to a question on a cognitive ability test. Our findings are consistent with an economic model in which the rate of success depends on time investment. There is, however, substantial heterogeneity in outcomes and more importantly in behavior in our sample of relatively homogenous subjects (all students from one university). This heterogeneity in behavior reflects different choices and different reactions to incentives and time pressure. It could also offer possible scope for improvement

[12]The eleven 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.

Table 2.6: Controlling for Aggregation Bias

Panel A.	Raven Matrices											
Treatment	Sub. time	Opt.	Mean of disaggregated optima										
			1	2	3	4	5	6	7	8	9	10	11
Baseline	30.111	26	27.319	26.702	27.583	27.545	27.2	26.756	28.906	27.495	26.534	25.784	24.3
	(0.044)		(0.007)	(0.006)	(0.006)	(0.009)	(0.009)	(0.015)	(0.038)	(0.036)	(0.037)	(0.036)	(0.036)
HH	31.046	27	29.47	25.052	26.92	26.517	27.694	26.088	26.574	28.27	26.258	25.432	25.103
	(0.044)		(0.005)	(0.014)	(0.002)	(0.012)	(0.009)	(0.014)	(0.037)	(0.033)	(0.035)	(0.036)	(0.036)
HL	33.292	38	39.717	40.659	41.095	46.782	38.642	43.041	34.816	36.35	35.389	36.745	36.66
	(0.046)		(0.008)	(0.015)	(0.006)	(0.026)	(0.014)	(0.031)	(0.042)	(0.038)	(0.039)	(0.041)	(0.043)

Panel B.	Numerical Tasks											
Treatment	Sub. time	Opt.	Mean of disaggregated optima										
			1	2	3	4	5	6	7	8	9	10	11
Baseline	25.816	22	23.234	22.703	24.435	24.946	24.273	24.247	24.507	24.598	23.681	23.549	23.395
	(0.035)		(0.010)	(0.008)	(0.006)	(0.008)	(0.012)	(0.014)	(0.027)	(0.029)	(0.026)	(0.028)	(0.027)
HH	26.828	25	21.531	22.102	23.179	22.886	22.792	22.941	25.099	24.34	25.086	23.147	22.51
	(0.035)		(0.009)	(0.013)	(0.009)	(0.010)	(0.010)	(0.017)	(0.026)	(0.028)	(0.028)	(0.028)	(0.026)
HL	29.04	39	39.492	36.195	39.414	38.259	37.686	39.451	29.982	30.695	31.297	31.938	30.766
	(0.037)		(0.004)	(0.010)	(0.006)	(0.020)	(0.011)	(0.027)	(0.036)	(0.037)	(0.037)	(0.037)	(0.038)

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.

for some subjects. Finally, understanding the economics of test taking provides teachers, employers and policy makers with information as to what value to put on the outcomes of cognitive tests of students, potential employees and to make sense of cross-country patterns.

Our main contribution to the academic literature can be summarized by the following key results. We plot the returns to time on a cognitive test as a concave function of the time invested. In our experiment the test environment does not seem to (statistically) significantly influence the probability that a subject knows the answer. Hence, as of a certain level of incentives, subjects do not come up faster with a correct answer. However, submission behavior differs across test environments as predicted by economic theory. Subjects invest more time when they are faced with higher monetary stakes and decrease time investment when time pressure is increased. Overall, the changes in the timing are relatively small. One potential reason for this could be that in our setup the differences between actual expected earnings and optimal expected earnings are economically small. We also observe heterogeneity in answering behavior if we split the sample into different levels of difficulty of questions and different performance types. Controlling for aggregation bias does not change our results.

An interesting finding of our paper is that we show that subjects adjust their answering behavior to the test 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 stakes do not necessarily maximize test performance. One first conclusion for teachers and employees is that behavior in different test environments is heterogeneous between individuals. Especially low performing subjects adjust the least to new test environments.

Second, our results indicate that as of a certain level of incentives the speed of thinking does not increase. From an economic point of view this suggests that there are certain boundaries in which incentives seem 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 can play a role, while answering a test question.

For policy makers and for applied researchers these results are interesting as well. We show that test results are not only a pure measure of cognitive ability,

but also of the ability how people deal with different environments. Our results indicate that the test environment influences behavior on a test but not necessarily the speed of thinking. An important avenue for future research is to investigate the impact of personality traits and economic preferences on the probability of knowing the correct answer and answering behavior during a cognitive test.

Chapter 3

How Do Personality Traits and Preferences Affect Cognitive Test Scores? Evidence from a Laboratory Experiment

3.1 Introduction

Standardized tests are the most frequently used tool to assess students and job applicants or to compare the educational performance of countries. It is well known that test results are shaped by both cognitive and non-cognitive skills (Borghans et al., 2008). Many studies find that personality traits, such as conscientiousness and neuroticism, and economic preferences, such as time and risk preferences, can be associated with the measured test outcome. However, besides the mere correlation between these attributes and test scores, the mechanisms behind this relation are largely unexplored.

The aim of this paper is to investigate *how* personality traits and preferences (measured by economic preference parameters) influence a test result. The idea is that there are two determinants of a test result, whereas usually these two determinants are collapsed into the term (cognitive) ability. We argue that one has to distinguish between two different determinants which measure an individual's ability. The first determinant is the probability of knowing the correct solution while answering a question, which we define as the answering technology of an individual. The second determinant of a test result is the answering behavior of a test taker. Next to the mere speed of thinking, the decision when to answer a question is a crucial determinant of the test result.

We argue that personality traits and economic preferences influence both determinants. First, they can influence the technology of knowing the correct answer at any point in time. Relatively neurotic people, for instance, could in general be less likely to answer a test question correctly at any point in time. If this is the case, they have an inferior technology relative to people who are relatively less neurotic. Second, personality traits and preferences could influence the behavior during a test. Assume for instance that patient and impatient people are in general all the time equally likely to know the answer when they face a test question. This means that they are equipped with the same technology. However, patient people could invest more time in thinking about a question and hence they are also more likely to know the correct answer to a question. This means that different personality types reveal different behaviors while answering a question, which influence the test score.

For the ideal analysis of the above mentioned mechanism one has to be able to monitor the decision-making process while answering a question in a fully controlled environment. We set up a computerized laboratory experiment in which

each participant has to answer a set of intelligence questions. While answering a test question our participants face two independent payment schemes. These two payment schemes serve as a tool to distinguish between an individual's technology and her behavior while answering a question. This allows us to differentiate between the time when an individual knows the answer to a test question and the time when she submits this answer. The first payment scheme serves as a thought-tracker. We provide an incentive to select the answer which a participant considers to be most likely to be correct. This incentive scheme is active at any moment while she can think about a question. This allows us to plot the technology function of knowing the correct answer over time. The second incentive scheme provides subjects with an incentive to submit an answer to a question by pressing a submit button. If the submitted answer is correct a subject also receives money from this payment scheme. Note that this second decision is the only decision one usually observes in tests. The combination of the two payment schemes allows us to disentangle the technology from behavior during a test.

Besides the extensive monitoring how individuals think and behave when answering a question we elicit their personality and their preferences. We measure personality using the Big Five personality inventory. Moreover, we obtain measures of an individual's risk attitude and time preference using experimentally validated questions.[1] We link the personality and preferences measures to the behavior on the two payment schemes. By doing so we are able to determine whether certain personality traits or preferences interfere with the technology of a test-taker or her behavior.

Our main findings can be summarized as follows. We find that the personality traits openness to experience and neuroticism are related to the speed of thinking of a test-taker. Participants who are more open and emotionally stable seem to have a better answering technology. Moreover, people who can be labelled as risk takers are also equipped with a superior answering technology compared to more risk-averse individuals. However, only an individual's discount rate has a significant impact on answering behavior to a question. Patient people wait longer until they submit their answer to a question. If we look at the aggregate test score only favourable economic preference parameters such as the willingness to take risks and patience have a positive and significant impact on the final test result. Risk preferences influence the test result through the technology. Time preferences influence the test result through the behavior of the test-taker.

[1] We will use the terms time preference, discount rate and (im-)patience interchangeably.

Our paper adds to the literature both in economics and psychology. Numerous papers document a relationship between personality traits and test results. However, the literature remains silent on the reasons why certain personalities perform better on tests than others. This paper builds on the work by (Borghans et al., 2008, 2013). It extends these papers with two new features. First, we introduce a new experimental design which allows us to clearly distinguish between the answering technology and answering behavior. Second, the structure of our data allows us to investigate the full deliberation process when answering a question during a cognitive test.

Our findings are consistent with other recent work. Golsteyn and Schils (2014) find strong differences between personality traits and math and reading achievement using a representative sample of Dutch elementary school children.[2] Their approach is similar to ours, since they decompose these differences into differences in resources and the use of these resources. Their main finding is that boys are better equipped with resources than girls but that girls use their resources more effectively than boys. Duckworth et al. (2011) show that test motivation and non-intellective traits explain a crucial part of the differences in intelligence test scores across subjects.[3] Dohmen et al. (2010) find that risk aversion and impatience are negatively related to measures on a cognitive test. Kirby et al. (2005) find a significant negative relationship between college GPAs and high discount rates. Chamorro-Premuzic et al. (2005) find a positive relationship between openness to experience and performance on Raven matrices.

Our findings are also helpful in explaining differences in task performances when stakes are high, because we monitor behavior during a test. This can be seen as a proxy for differences in problem solving approaches. In laboratory experiments with incentives to exert effort this has been a problem. Typically, the choice of work effort is represented by a monetary function, as we have modeled too. This procedure has been used in tournament experiments (e.g., Bull et al. (1987)) or in efficiency wage experiments (e.g., Fehr and Falk (1999)). In real effort experiments, exerting effort is a task. In Fahr and Irlenbusch (2000), subjects had to crack walnuts, Van Dijk et al. (2001) subjects performed cognitively demanding tasks on the computer (two-variable optimization problems) and in Gneezy et al.

[2] In particular they find a strong negative relationship between the math score and conscientiousness, extraversion and neuroticism. Self-Control, IQ and openness to experience are positively associated with the score on a math and reading achievement tests.

[3] Similar findings are obtained by Borghans and Schils (2013). They show that more agreeable, more motivated and more ambitious students perform longer and better on the PISA test.

(2003) subjects had to solve mazes at the computer. While effort adds realism to the experiment, one should also note that it is realized at the cost of losing control. The experimenter does not know the workers' effort cost. By distinguishing between technology and behavior we are able to derive more precise quantitative predictions.

The remainder of the paper is structured as follows. In Section 3.2 we explain our conceptual framework. Section 3.3 briefly highlights the most important features of the experimental design. Section 3.4 presents the results and Section 3.5 concludes.

3.2 Conceptual Framework

The general idea of this paper is that a test score is determined by an individual's technology and an individual's behavior. For a given technology, we assume that each individual maximizes her utility when solving a test. In the course of the paper we analyze test taking behavior on the question level. Hence we formulate the decision-making problem of a test-taker when answering one question with the following utility function:

$$U(t, \theta, \tau) = p(t|\theta, \tau) \cdot \Pi(t|\theta, \tau) \tag{3.1}$$

The utility function consists of three main ingredients:

1. Time t
2. Skills, preferences and traits which are summarized in θ
3. Test circumstances τ, for instances high and low stakes, under which the test is conducted

We further assume that utility is determined by two components. The technology function $p(\cdot)$ and the reward function $\Pi(\cdot)$. We define the technology function $p(\cdot)$ as the probability of knowing the correct answer at a certain point in time for given circumstances and a given set of skills, preferences and traits. The reward function $\Pi(\cdot)$ reflects the monetary and non-monetary rewards which a test-taker receives for a correct answer on a test question. Rewards decrease with the time invested ($\frac{\partial \Pi(t|\cdot)}{\partial t} < 0$). We further assume that the test environment cannot be influenced by the test-taker and that skills and traits are fixed during the test for

each person. Hence, a rational test-taker only optimizes the time she spends on answering a test question given a certain reward function and her technology:

$$U(t^*, \bar{\theta}, \bar{\tau}) = p(t^*|\bar{\theta}, \bar{\tau}) \cdot \Pi(t^*|\bar{\theta}, \bar{\tau}) \tag{3.2}$$

Equation 3.2 shows that differences in test results between persons can stem from three sources. First, different test-takers can differ in their technology function, which can vary with θ and τ. Second, test-takers can differ in how they behave during a test which also depends on θ and τ. This can also lead to a difference in the optimal timing t^*. The key question of this paper is to assess how preferences and personality traits interfere with the technology and behavior. This means that we are interested in identifying the following expression:

$$\frac{\partial U(t^*, \theta, \tau)}{\partial \theta} = \frac{\partial p(t^*|\theta, \tau)}{\partial \theta} \cdot \Pi(t^*|\theta, \tau) + p(t^*|\theta, \tau) \cdot \frac{\partial \Pi(t^*|\theta, \tau)}{\partial \theta} \tag{3.3}$$

The first part of equation 3.3 captures the effect of skills, preferences and personality on the technology of an individual. The second part of equation 3.3 describes the effect of skills, preferences and personality on the behavior during the deliberation phase. We remain reluctant in making assumptions about potential directions of these effects.

Figure 3.1 plots our conceptual framework in two graphs. We assume that the test environment τ remains constant. Moreover, we assume that both individuals answer the same question. The figure plots the probability of a correct answer on a question for individual i and individual j over time in two panels. This is the technology function $p(t|\theta, \bar{\tau})$. The y-axis indicates the probability of a correct answer and the x-axis time.[4] The individuals differ in their skills, preferences and personality traits θ. The dashed line plots the technology of individual i and the straight line the technology of individual j. In our case individual j has a superior technology to individual i, since the probability of knowing the answer is always higher at any point in time for her.

Panel A plots two different technology functions. The key insight of this figure is that, at any point during the deliberation process, the person with the superior technology has a higher probability to know the correct answer. The straight vertical line indicates the time when the individuals answer a respective question. The equivalent in a test would be the submission of an answer. The point at which

[4]This framework is not exclusive for test-taking behavior but can also be applied to other problem solving activities.

the technology function and the vertical line cross, determines the probability of having a correct answer to a question. In the first scenario both individuals answer at the same time ($t_i^* = t_j^*$). The ultimate effect is that individual j with the superior technology always has a higher probability of knowing the answer to a question and therefore a better test result ($p_j(t^*|\theta_j, \bar{\tau}) > p_i(t^*|\theta_i, \bar{\tau}), \forall t > 0$).

Another scenario one could imagine is two individuals with the same technology but with a different submission time of the answer. This scenario would also yield different probabilities of knowing the answer and ultimately different test results. This refers to differences in the reward functions between personality types: $\Pi(t_j^*|\theta_j, \bar{\tau}) \neq \Pi(t_i^*|\theta_i, \bar{\tau})$. In Panel B of Figure 3.1 both individuals end up with the same probability of knowing the correct answer to the question ($p_i(t_i|\theta_i, \bar{\tau}) = p_j(t_j|\theta_j, \bar{\tau})$). The key insight is that the individual with a superior technology answers the question earlier than the individual with the inferior technology ($t_j^* < t_i^*$) but this is not necessarily reflected in the final test score. The individual with an inferior technology simply waits longer until he makes a decision to answer the respective question. Thus, the probability of answering the test question correctly is equal for both individuals in Panel B regardless of the technology. This result indicates that especially the interaction of answering technology and behavior yields the same test result for different persons. People with inferior technologies can compensate this lack of capacity by thinking longer before they submit. By only observing the number of correct answers at the end of an achievement test one would not be able to disentangle the actual differences in how the test-taker came to this result.

3.3 Experimental Design

The ideal experimental setup to investigate differences in technology and behavior when answering a test question requires measures of personality traits and preferences to determine heterogeneity among people. Next, it requires an experimental design that allows us to disentangle technology from behavior while answering a test question. In the following we document how we measured personality traits and preferences. Finally we explain how we monitored the deliberation process while answering a test question.

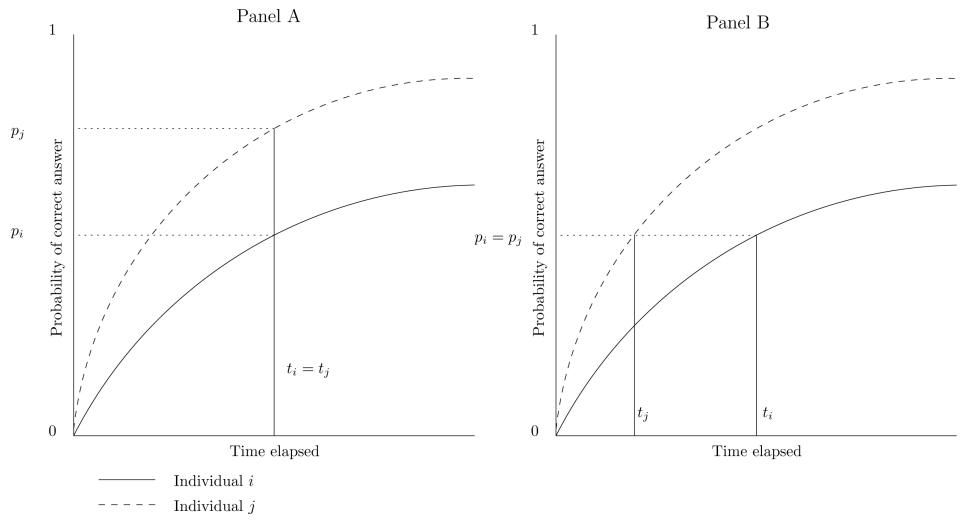


Figure 3.1: Determinants of a test result.

Note. The figure describes the probability of knowing a correct answer to a test question as a function of time. The dashed line and straight line indicate different answering *technologies*. The straight vertical lines indicate answering *behavior*.

3.3.1 Measuring personality and preferences

The first part of the experiment started with a detailed questionnaire on an individual's personality and economic preferences. We assessed personality by using the Big Five personality inventory (Goldberg, 1990). The Big Five personality traits consist of openness to experience, conscientiousness, extraversion, agreeableness and neuroticism. In the literature this has been labeled OCEAN. The questionnaire which we used contains 10 items per trait.[5] Table 3.1 shows the definitions of each trait.

Economic preferences such as risk attitude and time preferences were measured with experimentally validated items designed by Falk et al. (2012). When assessing the risk attitude of an individual each subject had to make five choices between a lottery and a sure payment. Every subject started with the question whether she wants to have a sure payment of 150 Euro or a lottery with a 50 percent chance of winning 300 Euros or a 50 percent chance of receiving nothing. We only varied the sure payment and the lottery remained the same in each situation. If

[5]The questions of the Big Five inventory are available on `http://ipip.ori.org/New_IPIP-50-item-scale.htm`. (Last access on May 20, 2014). The questionnaire is also available upon request.

Table 3.1: The Big Five Personality Traits

Trait	Definition of Trait
Openness to Experience (Intellect)	The tendency to be open to new aesthetic, cultural, or intellectual experiences.
Conscientiousness	The tendency to be organized, responsible and hard working.
Extraversion	An orientation of one's interests and energies toward the outer world of people and things rather than the inner world of subjective experience, characterized by positive affect and sociability.
Agreeableness	The tendency to act in a cooperative, unselfish manner.
Neuroticism (Emotional Stability)	Neuroticism is a chronic level of emotional instability and proneness to psychological distress. Emotional stability is predictability and consistency in emotional reactions, with absence of rapid mood changes.
Source: Almlund et al. (2011).	

the subject chose the lottery, the sure payment increased. On the other hand, if the subject chose the sure payment, the sure payment decreased in the next question. Time preferences were measured in a similar way. Subjects had to make five choices between 100 Euro today and a greater delayed amount in 12 months from today. In the first question each subject made a decision between 100 Euro today or 153.80 Euro in 12 months from now. We only varied the amount in 12 months. If a subject chose the 100 Euro today, the later amount increased to 185 Euro. On the other hand, if the subject chose the 160 Euro in 12 months, this amount decreased to 125.40 Euro in the following question. This method enables us to elicit an individual's internal rate of return. We interpret a low internal rate of return as being more patient and a high rate of internal return as being less patient.

3.3.2 Solving Raven matrices

In the next part subjects had to solve Raven matrices. Before they started they received detailed instructions about the Raven matrices (Raven, 1962) and the payment schemes in the part of the cognitive test. Raven matrices are a standard measure to assess fluid intelligence (Almlund et al., 2011; Neisser et al., 1996; Carpenter et al., 1990). We use them to make our results independent of culture, language and potential differences in acquired knowledge of our subjects.

Before they could start to answer the intelligence questions, subjects had to go through a trial phase which made them familiar with the way we set up the test and the two payment schemes when answering a test question. Note that this is not that uncommon, since many achievement tests (such as the GRE® or TOEFL®) have test procedures in which each test-taker can prepare by using a test preparation manual.[6] Each subject had to solve 45 Raven matrices with different degrees of difficulty. Subjects could not skip a question and had 60 seconds to solve a matrix. The time which was left is indicated with the green bar on the left hand side of the screen shot in Figure 3.2. The order (and thereby the difficulty) of the questions was randomized between subjects. Figure 3.2 shows an example of a typical decision-making screen of the experiment. The green bar on the left indicated how much time was still left to answer the respective question. In the middle of the screen the Raven matrix was displayed. Each matrix consisted of 5 figures which are connected in a logical manner. The idea is that the test taker had to find the missing figure in the bottom right hand corner. The correct solution is among the 8 options below the matrix. The figure which was currently selected was marked with a green frame (in this case this is also the correct solution). On the right hand side of the screen we provided the information about the payment schemes which allowed us to discriminate between answering technology and answering behavior. The text written in blue showed the payoff information of the thought-tracker. The text which is written in red provided information about the payment which a subject would receive for also

[6] In total there were five trial questions. After each trial phase subjects received detailed information how long they selected the correct answer. Moreover we informed them about their submission time, the payoff for a correctly submitted answer at that point and if their submitted answer was correct. From this information they had to calculate their payoff of the respective question. Subjects received detailed feedback and instructions when they answered the questions. They could not continue with the experiment if they did not answer the control questions correctly. Moreover, we sequentially introduced the payment schemes. The first trial phase only consisted of the blue payment scheme. The second trial phase only contained the red payment scheme. In phases three to five subjects had to calculate their payoff from both payment schemes. More details can be found in the Appendix B.1.

pressing the submit button on the bottom right of the screen. In the following we will explain the payment procedure in more detail.

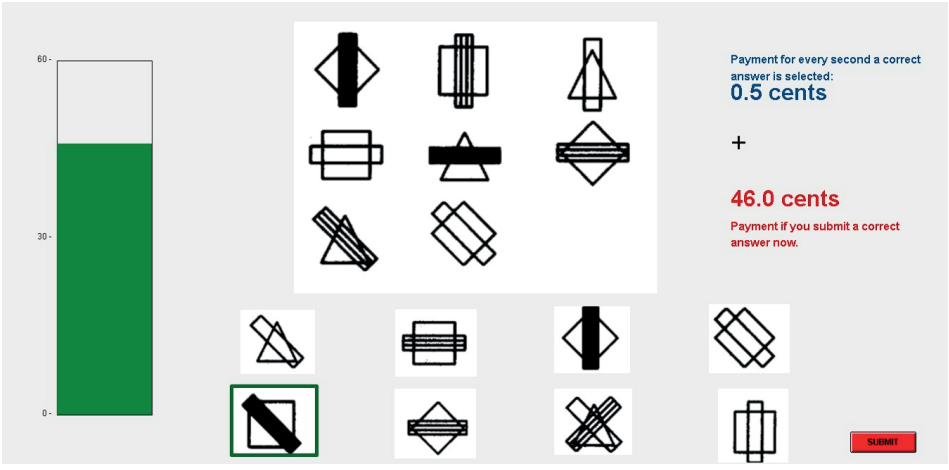


Figure 3.2: Screen shot of the decision-making screen for a typical Raven matrix.

Note. The matrix is taken from Carpenter et al. (1990) to keep the actual Raven matrices confidential.

3.3.3 Disentangling technology from behavior

During answering a question subjects faced two independent payment schemes. The first payment scheme, which we call the blue payment scheme, serves as a measure of monitoring answering technology. It provides an incentive to immediately reveal the answer to a question which the subject thinks is the correct one at any point in time. We paid 0.5 cents for every second a subject selected the correct answer during the 60 seconds when she faced each test question. Subjects could change their answer as often as they liked during that time.[7] As shown in Figure 3.2, the information about the blue payment scheme was presented on the upper right of the screen.

The second incentive scheme, which we called the red payment scheme, was running independently of the first one. We provided subjects with an incentive to submit a correct answer by pressing a submit button. This system serves as a measure for answering behavior. In the red payment scheme we varied the

[7]The payment scheme is similar to the one in Caplin et al. (2011). They use a probabilistic incentive scheme to provide an incentive of an individual's search process. To keep things as simple as possible we decided to use a deterministic payment scheme.

incentives to submit an answer in three treatments, while the blue payment scheme remained constant over the whole experiment. In our Low Stake treatment the amount a subject could receive for submitting a correct answer decreased from 25 cents to 5 cents during the 60 seconds. If a subject for instance selected the correct answer to a test question in the very first second, immediately submitted this answer and did not change her selection her payoff from this question would have been 55 cents (25 cents for submitting the correct answer + $60 * 0.5$ cents for selecting the correct answer over 60 seconds). The information about the current payoff was displayed on the right of the screen (see Figure 3.2). We conducted two other treatments with higher stakes for the submission of a correct answer. In the following we will call them High Stakes treatments. Subjects could receive up to 55 cents only for a correctly submitted answer.[8]

The two incentive schemes allow us to discriminate between technology and behavior. We will use the data from the blue payment scheme to plot the technology function and the data obtained from the red payment scheme to analyze the answering behavior. More information on the procedure of the experiment can be found in Borghans et al. (2014).

3.4 Results

In this section we present the main results of the experiment. It was computerized with zTree (Fischbacher, 2007) and 128 subjects participated.[9] Each individual had to solve 45 Raven matrices. We observe behavior for each individual on each matrix for 60 seconds. Hence, we obtain a panel data set with a total of $345,600$ observations ($128 * 45 * 60$).

[8] In the second treatment which we call the HH treatment (High incentives and High time pressure) the reward for submitting a correct answer decreased from 55 cents to 5 cents during the 60 seconds. In the third treatment which we call the HL treatment (High incentives Low time pressure) the reward for submitting a correct answer decreased from 55 cents to 35 cents. Each treatment contained 15 matrices in a randomized order. We also randomized the treatment order across subjects.

[9] The experiment took place in the BEElab at Maastricht University. We invited subjects with the recruiting software ORSEE (Greiner, 2003). The experiment lasted 2.5 hours and average earnings were 31.85 €(S.D. 4.6).

3.4.1 Descriptive Statistics

Big Five

Figure 3.3 shows the distribution of the Big Five. All measures are standardized with mean zero and standard deviation one. The picture that emerges from this figure that all personality measures seem normally distributed. The gray dotted line in each graph corresponds to the kernel density estimates. The black dashed line shows the normal distribution. We test all distributions for normality using a Shapiro-Wilk test. None of the tests can reject the null hypothesis that the distribution is normally distributed.[10]

A crucial point when measuring the Big Five personality traits is the consistency of the constructs. In practice this means two things. First, the questions which are used to measure the traits should reveal five different constructs which are described in Table 3.2. Second, the items which are supposed to measure one specific trait should be highly correlated. To test the first point we run an exploratory factor analysis on all 50 questions.[11] We obtain 6 factors with an eigenvalue greater than one. However, the last factor which is extracted only has an eigenvalue equal to 1.00295. In addition to that, the cumulative explained variance only increases from .7072 to .7531 from the 5^{th} to the 6^{th} factor. Table 3.2 shows the items which we used for each trait and the respective results of our analysis. The second last row documents the values of Cronbach's alpha (Cronbach (1951)) which measures the consistency of the items for each trait. All values are in the interval between .75 and .86 which is a good indicator for a consistent measurement of each trait. The last row indicates the explained variance by the first factor which is obtained running factor analysis on the 10 items for each trait. Except for the trait openness all values are greater than 80 percent. Moreover, we only obtain one factor for each trait with an eigenvalue greater than 1 if we run the factor analysis only on the ten items per trait. Summing up it can be said that the Big Five questionnaire measures five different traits in a consistent way.[12]

Preferences

Figure 3.4 shows the distribution of the preference measures. The left panel shows the standardized values of the answers to our risk preference elicitation method.

[10] The p-values of the Shapiro-Wilk test for normal data look as follows: Openness $p = .56$; Conscientiousness $p = .66$, Extraversion $p = .27$, Agreeableness $p = .26$, Neuroticism $p = .89$.

[11] We do not rotate the factors.

[12] The Appendix provides additional dendograms for each of the traits separately.

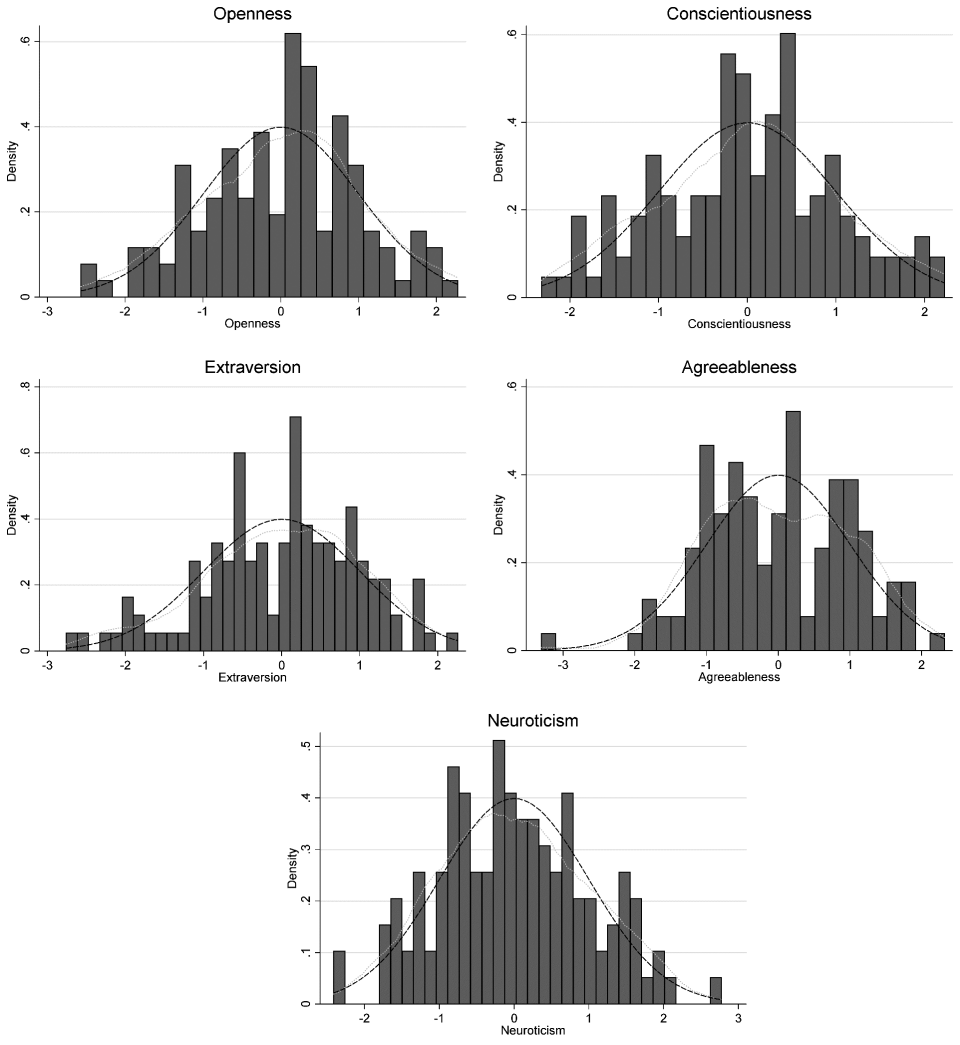


Figure 3.3: Distribution of the Big Five personality traits.

Note. All measures are standardized with mean zero and standard deviation one. The dotted gray line indicates the density estimates of the distribution. The dashed black line shows the density estimates of the normal distribution.

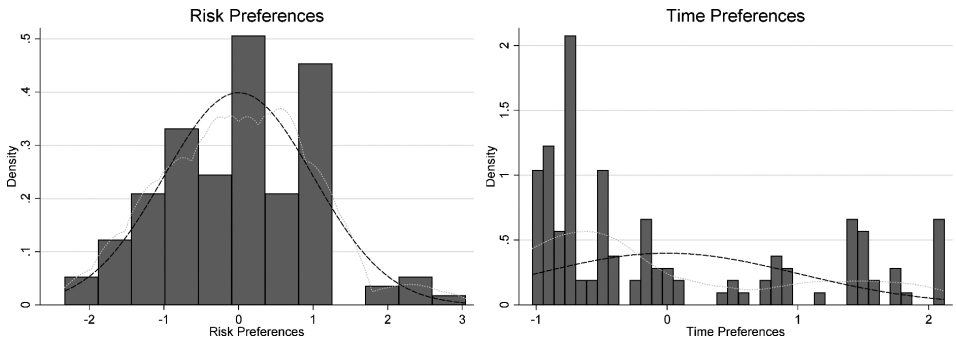


Figure 3.4: Distribution of Risk and Time Preferences.

Note. Both measures are standardized with mean zero and standard deviation one. The dotted gray line indicates the density estimates of the distribution. The dashed black line shows the density estimates of the normal distribution.

The gray dotted line shows the kernel density estimates and the black dashed line shows the plot of the normal distribution. A Shapiro-Wilk test for normality cannot reject the null that the data is normally distributed ($p = .68$). The right panel shows the standardized values of the answers to our time preference elicitation method. The gray dotted line shows the kernel density estimates. The dark dashed a corresponding normal distribution. The picture that emerges from this graph is that the distribution is right-skewed. A Shapiro-Wilk test for normality rejects the null hypothesis that the data is normally distributed ($p < 0.0001$).

IQ score

Standard IQ scores or achievement test scores are usually normally distributed. Figure 3.5 shows the standardized distribution of correctly submitted answers across all treatments. Note that this distribution is very similar to the one from standard intelligence tests. The black dashed line plots a normal distribution and the gray dotted lines the kernel density plot. We test for the normality of the distribution of correctly submitted answers. A Shapiro-Wilk test cannot reject the null that the data is normally distributed ($p = 0.71$). We do not obtain significant differences in the distributions between the treatments. The respective figures can be found in the Appendix to this chapter.

Table 3.2: Reliability of the Big Five

	Openness	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Items	I have a rich vocabulary	I do chores right away	I am the life of the party	I feel little concern for others (R)	I get stressed out easily
	I have difficulties understanding abstract ideas (R)	I'll leave my things lying around (R)	I do not talk a lot (R)	I am interested in people	I am relaxed most of the time (R)
	I have a vivid imagination	I live my life according to schedules	I feel comfortable around people	I insult people (R)	I worry about things
	I am not interested in abstract ideas (R)	I neglect my obligations (R)	I keep in the background (R)	I sympathize with others' feelings	I seldom feel blue (R)
	I have excellent ideas	I have an eye for details	I start conversations	I am not interested in other people's problems (R)	I am easily disturbed
	I do not have a good imagination (R)	I am accurate in my work	I have little to say (R)	I have a soft heart	I get upset easily
	I am quick to understand things	I forget to put things back where they belong (R)	I talk to a lot of different people at parties	I am not really interested in others (R)	I change my mood a lot
	I use difficult words	I am always well prepared	I do not like to draw attention to myself (R)	I take time out for others	I have frequent mood swings
	I spend time reflecting on things	I often make a mess of things (R)	I do not mind being the centre of attention	I feel others' emotions	I get irritated easily
	I am full of ideas	I like order	I am quiet around strangers (R)	I make people feel at ease	I often feel blue
Cronbach's alpha	0.7541	0.7832	0.869	0.7513	0.8304
Explained Variance	69.18%	89.07%	95.35%	85.32%	82.47%

Note. The table shows the items of the IPIP personality questionnaire. An (R) indicates that this item is a reversed measure of the trait. We obtained one factor per personality trait using Kaisers criterion to drop all components with an eigenvalue lower than 1. The "Explained Variance" is the proportion of the variance explained by this one factor.

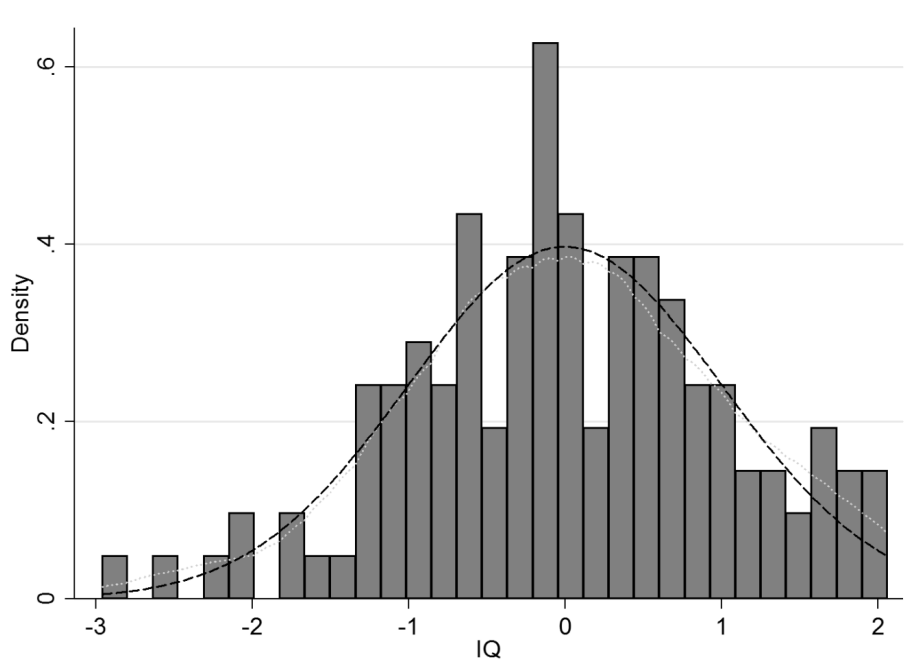


Figure 3.5: Distribution of IQ in the experimental sample

Note. The figure shows the distribution of the standardized number correctly submitted answers in all treatments. The black dashed line plot the normal distribution. The gray dotted lines plots the kernel density plot of this distribution. A Shapiro-Wilk test cannot reject the null that the data is normally distributed ($p = 0.71$).

Correlation Structure

Recent evidence suggests that there is not only a correlation among the different personality traits but that personality traits are also associated with economic preference parameters. Moreover, Becker et al. (2012) show that there is a complementary relationship between personality traits and economic preference parameters. This suggests that it is important to investigate whether our data reveals a similar pattern and to see whether we might suffer from potential collinearities in our analysis.

Table 3.3 shows the relationship between the measures of the personality traits and the economic preference parameters. Looking within the measures of personality traits, it turns out that only openness to experience seems to be positively correlated with extraversion and conscientiousness. Interestingly, neuroticism is significantly negatively correlated with the willingness to take risks. There seems to be a positive relationship between an individual's risk attitude and openness to experience. Most strikingly, an individual's discount rate does not significantly correlate with any of the other personality and preference measures. Overall, our results are in line with those from the experimental data presented in Becker et al. (2012), which suggest complementarity between economic preference parameters and personality traits.

The correlation sign between risk attitude and the discount rate is negative as one would expect (see for instance Dohmen et al. (2010)). Moreover, our data reveals a relationship between preferences and the intelligence test score. Risk attitude and the intelligence test score are positively correlated and the discount rate and the intelligence test score are negatively correlated. Overall, Table 3.3 reveals only a weak cross-correlation pattern.

3.4.2 Determinants of test scores

If personality traits, or non-cognitive skills in general, are related to cognitive test scores we know that these tests measure something else than pure cognitive ability. Table 3.4 shows the relationship between preferences, personality traits and the final test result. The dependent variable is the number of correctly submitted answers in all treatments.[13] All independent variables are standardized with mean

[13]The results remain the same if the dependent variable takes the value 1 if the answer to a question was submitted correctly and if we control for possible effects of the three treatments. The regression table is available upon request.

Table 3.3: Correlation Structure of Personality Traits, Preferences and IQ

	Discount Rate	Risk Attitude	O	C	E	A	N
Risk Attitude	-0.114						
Openness	-0.118	0.187*					
Conscientiousness	0.0645	0.00346	0.186*				
Extraversion	0.0818	0.0876	0.341***	0.0874			
Agreeableness	0.0876	-0.168	0.169	-0.049	0.114		
Neuroticism	0.0145	-0.245**	-0.118	0.0822	-0.167	0.113	
IQ	-0.198*	0.189*	0.0688	0.0404	-0.0039	0.0658	-0.0996

Note. All measures are standardized with mean 0 and variance 1. * $p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Table 3.4: Determinants of Intelligence Test Score

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All Treatments		Low Stake Treatment		High Stake Treatments		(5)-(3)	(6)-(4)
Discount Rate	-1.225**	-1.195*	-0.489**	-0.436*	-0.736*	-0.759*	-0.25	-0.32
	(.5840)	(.6260)	(.2420)	(.2550)	(.4170)	(.4400)		
Risk Attitude	1.091**	0.977*	0.459**	0.417*	0.631*	0.56	0.172	0.143
	(.4830)	(.5420)	(.2040)	(.2350)	(.3520)	(.3800)		
Openness	0.43	-0.113	0.195	0.0714	0.235	-0.184	0.04	-0.251
	(.5610)	(.5600)	(.2450)	(.2510)	(.3890)	(.3810)		
Conscientiousness	0.249	0.434	-0.0554	-0.00299	0.304	0.437	0.354	0.437
	(.5920)	(.5630)	(.2480)	(.2350)	(.4130)	(.4030)		
Extraversion	-0.0239	-0.177	-0.122	-0.195	0.0986	0.0183	0.218	0.208
	(.5500)	(.5390)	(.2150)	(.2250)	(.3910)	(.3890)		
Agreeableness	0.413	0.802	0.138	0.272	0.275	0.53	0.137	0.258
	(.5410)	(.5390)	(.2180)	(.2350)	(.3970)	(.3980)		
Neuroticism	-0.615	-0.544	-0.193	-0.147	-0.422	-0.397	-0.23	-0.25
	(.5270)	(.5820)	(.2250)	(.2490)	(.4150)	(.4510)		
Constant		18.78***		6.284***		12.49***		
		(.5380)		(.2310)		(.3810)		
Observations	128	128	128	128	128	128		
R-squared		0.085		0.073		0.07		

Note. Dependent variable is the number of correctly submitted answers. All values of the independent variables are standardized with mean 0 and s.d. 1. Columns (1) and (3) report coefficients from separate regressions for each independent variable. Columns (7) and (8) report the differences between the Low Stake and the High Stake Treatments. Robust standard errors in parentheses. ** $p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

zero and standard deviation one. Each of the columns shows the results of a linear regression.

In column (1) we include each personality trait and preference measure separately as independent variables. Hence, each point estimate shows the result from a separate regression. An individual's risk attitude and the discount rate seem to play an important role in determining an individual's test score ($p < 0.05$). A one standard deviation increase in an individual's discount rate is associated with at least one correct answer less in all treatments. The effect is very similar in size and significance level for risk attitude. Personality traits do not yield significant results. However, the effects go in the expected direction. Neuroticism for instance is negatively associated with the number of correctly submitted answers. This is in line with findings from previous studies (Borghans and Schils, 2013). The results remain the same if we add all independent variables simultaneously (see column (2) in Table 3.4). The significance of the risk attitude and time preference coefficients slightly drops ($p < 0.1$).

Since the reaction of personality types and preferences could be heterogeneous to the test environment, we look at Low Stake and High Stake test environments separately. First, we look at the association between preferences and the number of correctly submitted answers in a Low Stake test environment. We present the results of the Low Stake environment in columns (3) and (4) of Table 3.4 and the results of the High Stake environment in columns (5) and (6). The size of the estimates drops for all independent variables. However, the signs and level of significance remain consistent in all cases except for conscientiousness. Most importantly, the relative size of the coefficients compared to the total number of correctly submitted answer does not drop. In all treatments a one standard deviation increase in the discount rate is associated with a 6% increase in the test score. The relative effect size of the discount rate is the same in the Low Stake and High Stake test environment.[14] Only in column (6) the coefficient on risk attitude becomes insignificant. This could be due to the fact that risk attitude and neuroticism are significantly negatively correlated and the standard error of the estimator of risk attitude increases because of multicollinearity.[15]

In the last two columns of Table 3.4 we compare the estimates of the Low Stake treatment with the estimates of the High Stake treatments. In Column (7) we report the differences of the regression coefficients from the separate regressions in

[14] We obtain this result by dividing the coefficient by the constant: $\frac{-1.195}{18.78} - 0.0637, \frac{-0.436}{6.284} \approx -0.0694, \frac{-0.759}{12.49} \approx -0.0608$.

[15] For further details on this see Greene (2012) (p. 89 ff.).

columns (5) and (3). Column (8) shows the differences of the regression coefficients of the simultaneous regressions in columns (6) and (4). None of the differences is different from zero. However, the picture that emerges from both columns is that the effect of personality traits and preferences seems to be slightly higher in the High Stake treatments. This could also be due to the fact that the High Stake treatments contain more questions than the Low Stake treatment.

3.4.3 Answering technology

The effects of technology and answering behavior can go into opposite directions and still yield the same test result. The submission of a correct answer is determined by the speed with which an individual comes up with a correct answer and the point when an individual decides to submit an answer. Hence the actual test result is a combination of technology and behavior. Individuals have different technologies but also different answering behaviors, which means that the ultimate test result might be the same.

Figure 3.6 shows the relationship between the fraction of correctly selected answers at any point in time (in the blue system) and the time elapsed in seconds. We construct the technology function by taking the average of correctly selected answers in each second. The relationship in Figure 3.6 can also be seen as an empirical representation of $p(t|\theta, \tau)$.[16] The light gray area around the straight black line indicates the 95% confidence bound. We draw two main conclusions from Figure 3.6. First, the probability of a correct answer increases over time. Second, the relationship between the probability of a correct answer and the time elapsed seems to be concave.

In Figure 3.7, we look what happens to $p(t|\theta, \tau)$ for different levels of θ. We first examine the relationship between personality traits and the fraction of correctly selected answers. We split the sample into below and above the median of the distribution of each respective trait and preference parameter. In Panel A we present the technology function for different levels of the trait openness to experience. We show that subjects who are more open to experience seem to have a higher chance of knowing the answer to a test question. The difference increases as time elapses. However, statistical significance is only reached in the 45^{th} second. Panels B- D present the relationship for conscientiousness, extraversion and agreeableness. We do not find an impact on the technology for these personality

[16] Note that at this moment we still assume that θ and τ do not change.

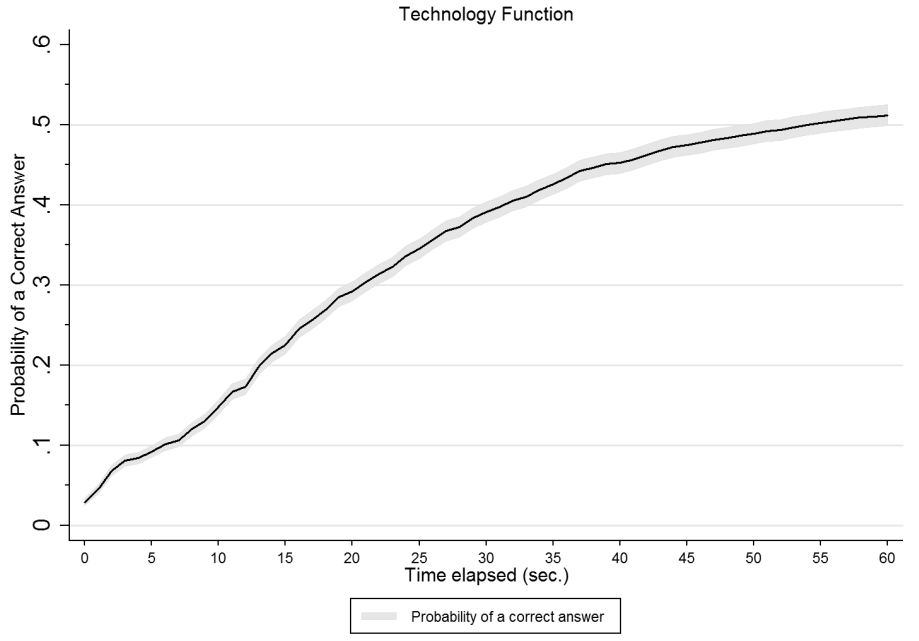


Figure 3.6: The relationship between time and the probability of a correct answer on a question.

Note. The gray area indicate the 95 per cent confidence bounds.

traits. Panel E shows the relationship between the time a test-taker thinks about a question and the probability of a correct answer for different levels of neuroticism. A lower level of neuroticism is associated with a superior answering technology. If we take the average of the difference over the whole 60 seconds, individuals with an above median neuroticism scale have a 2.8% lower probability of knowing the correct answer over time. In our test this would mean that they have at least one correct answer less because their technology is worse than the technology of emotionally stable individuals.[17] However, since the confidence intervals overlap in some points in time, we do not always obtain statistical significant differences.

Second, we examine the same relationship for economic preferences. Panel A of Figure 3.8 shows the relationship between the probability of knowing the correct answer over time and an individual's risk attitude. The picture that emerges from Panel A is that, subjects who are risk averse are on average 3.5 percentage points

[17] We obtain this number by multiplying $2.8\% * 45 = 1.26$.

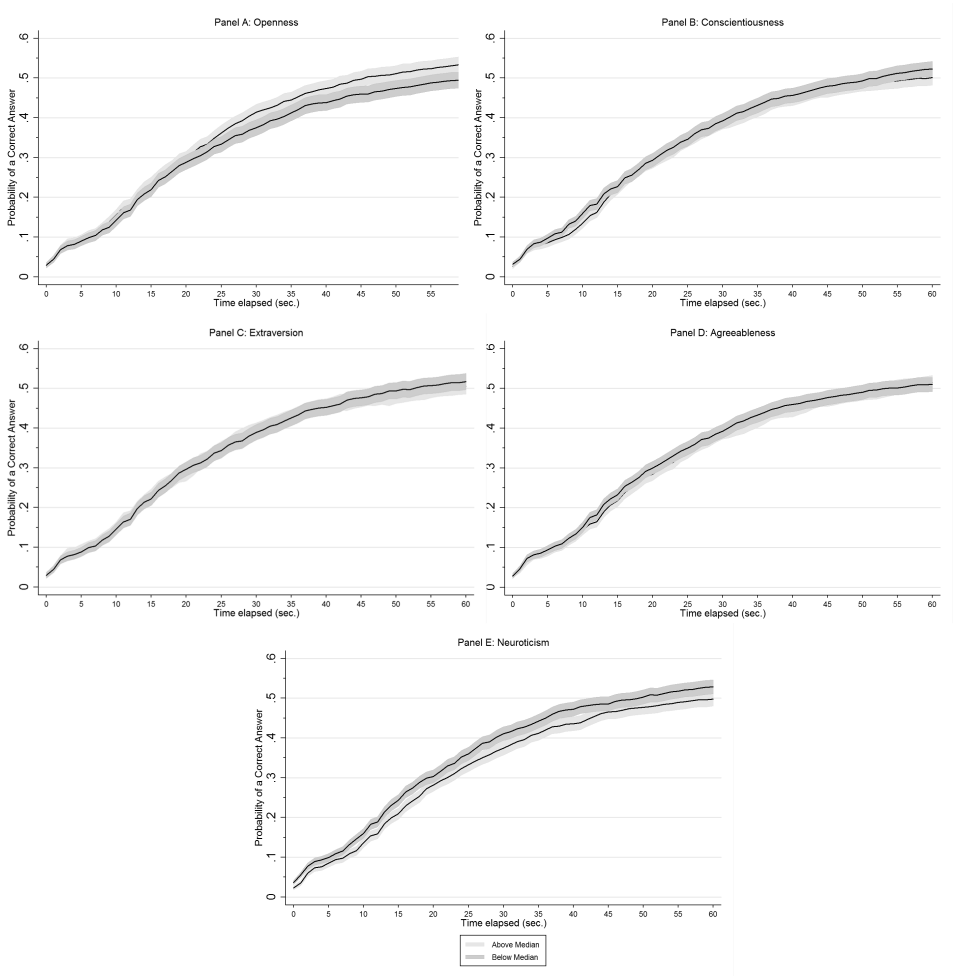


Figure 3.7: The relationship between the Big Five and the probability of knowing the correct answer over time.

Note. The figure plots the fraction of correctly selected answers in each second. We split the sample into above and below the median of the respective trait. The gray areas indicate the 95% confidence interval.

less likely to know the correct answer to a test question.[18] The difference is highly significant after the 37^{th} second.[19] Panel B of Figure 3.8 shows the relationship between an individual's discount rate and the probability of knowing the correct answer to question over time. More patient subjects have a higher probability of knowing the answer in the first five seconds. The overall probability of knowing the answer remains lower for impatient subjects. However, the difference is not statistically different. Overall, we find a highly significant impact of an individual's risk attitude on the probability of knowing the correct answer to a test question over time.

Finally, Table 3.5 shows the results of a linear regression with the fraction of correctly selected answers every 10^{th} second as dependent variable. We regress the fraction of correctly selected answers on personality traits and preferences. All regressions contain question fixed effects. All independent variables are dummy variables which take the value 1 if the actual value is above the median of the distribution. The picture that emerges from Table 3.5 confirms our results. First, a high willingness to take risk is associated with a superior answering technology. As of the 50th second test-takers who are more willing to take risks are significantly more likely to know the correct answer ($p < 0.1$). Second, there seems to be a weak relationship between the discount rate and technology, not revealed in the Figure 3.8. Patient test-takers seem to have a higher probability of knowing the answer, once we control for their risk attitude and their personality. Third, there seems to be a systematic positive relationship between openness to experience and the probability function of a correct answer. However, conscientiousness, extraversion and agreeableness, are not systematically related to the technology function. The estimate of neuroticism is significantly negative in the 10^{th}, 30^{th} and 40^{th} second ($p< 0.1$) and remains stable in size until the 60^{th} second. Thus, there seems to be a negative relationship between the degree of neuroticism and answering technology.

[18] We obtain this number by taking the average of the differences in probability between below and above the median of the risk attitude distribution.

[19] Since the willingness to take risks and neuroticism are significantly negatively correlated they might capture a similar personality facet. We analyse which items of the neuroticism are significantly associated with our risk measure and create a new neuroticism scale with these three items and compare whether there are differences in the technology in the same manner as in Figures 3.7 and 3.8. Our results do not reveal a significant difference in technology. Thus, risk attitude seems to capture another facet of personality than neuroticism.

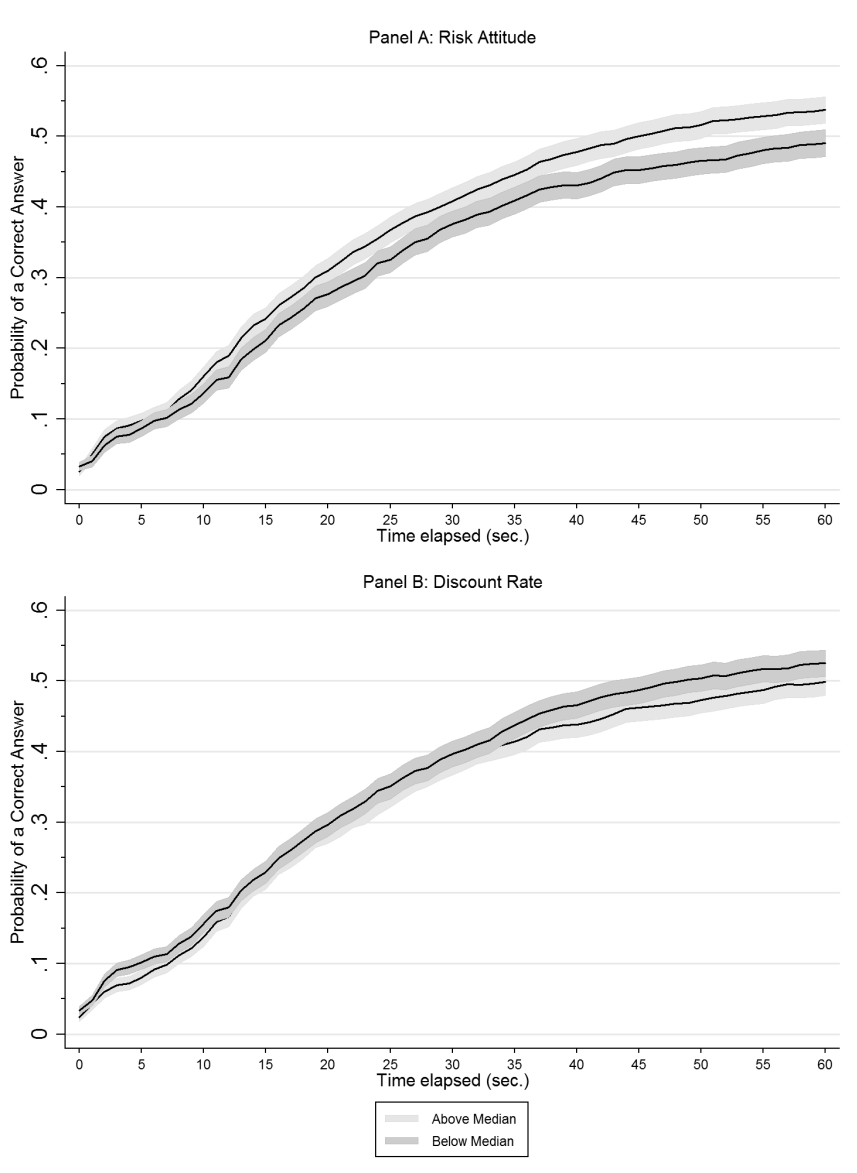


Figure 3.8: The relationship between economic preferences and the probability of knowing the correct answer over time.

Note. The figure plots the fraction of correctly selected answers in each second. We split the sample into above and below the median of the respective preference. The gray areas indicate the 95% confidence interval.

Table 3.5: The Relationship between Personality, Preferences and Answering Technology

	10 th sec.		20 th sec.		30 th sec.		40 th sec.		50 th sec.		60 th sec.	
Discount rate	-0.019	-0.017	-0.009	-0.004	-0.011	-0.005	-0.027	-0.021	-0.03	-0.019	-0.026	-0.015
	(.012)	(.012)	(.018)	(.018)	(.020)	(.020)	(.021)	(.020)	(.021)	(.021)	(.022)	(.022)
Risk attitude	0.012	0.006	0.023	0.019	0.024	0.015	0.027	0.016	0.044**	0.034*	0.041*	0.032
	(.013)	(.012)	(.019)	(.019)	(.021)	(.020)	(.021)	(.021)	(.021)	(.020)	(.022)	(.022)
Openness	0.013	0.009	0.01	0.011	0.039**	0.036*	0.035*	0.031	0.037*	0.036*	0.034	0.033
	(.012)	(.012)	(.018)	(.019)	(.019)	(.019)	(.020)	(.020)	(.020)	(.020)	(.021)	(.021)
Conscientious-	-0.024**	-0.024**	-0.003	-0.002	-0.003	-0.005	-0.008	-0.008	-0.007	-0.008	-0.022	-0.023
ness	(.012)	(.012)	(.018)	(.017)	(.020)	(.020)	(.021)	(.020)	(.021)	(.020)	(.022)	(.021)
Extraversion	0.003	-0.002	-0.012	-0.022	0.004	-0.015	0.002	-0.016	-0.012	-0.029	-0.012	-0.028
	(.012)	(.013)	(.018)	(.019)	(.020)	(.021)	(.020)	(.022)	(.021)	(.022)	(.022)	(.022)
Agreeableness	-0.007	-0.001	-0.016	-0.014	-0.003	0.002	-0.014	-0.008	-0.003	0.001	0.003	0.009
	(.012)	(.013)	(.018)	(.019)	(.020)	(.021)	(.021)	(.022)	(.021)	(.021)	(.022)	(.022)
Neuroticism	-0.023*	-0.024*	-0.021	-0.021	-0.035*	-0.033	-0.036*	-0.034	-0.025	-0.025	-0.03	-0.032
	(.012)	(.013)	(.018)	(.019)	(.020)	(.022)	(.021)	(.023)	(.021)	(.023)	(.022)	(.023)
Constant		0.177***		0.550***		0.718***		0.774***		0.800***		0.854***
		(.036)		(.050)		(.047)		(.043)		(.042)		(.040)
Observations	5,760	5,760	5,760	5,760	5,760	5,760	5,760	5,760	5,760	5,760	5,760	5,760
R-squared		0.084		0.197		0.235		0.239		0.238		0.249

Note. The table shows results from a linear regression. The dependent variable takes the value 1 if a correct answer was selected. All regressions contain question fixed effects. The column headlines indicate the point in time (10^{th}, 20^{th}, ..., 60^{th} second). All independent variables are dummies which take the value 1 if the actual value of the variable is above the median. Standard errors are clustered at the individual level ($N = 128$). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

3.4.4 Answering Behavior

Behavior during a test also determines the final result. We investigate the relationship between personality traits, preferences and answering behavior by analyzing information obtained from the behavior on the red payment scheme. In this payment scheme we paid subjects for the submission of a correct answer. Table 3.6 shows the determinants of submitting an answer to a question. The dependent variable is the time elapsed in seconds when a subject submitted the answer to a question. All regressions contain question fixed effects and controls for the treatments. We cluster standard errors at the individual level to take into account correlations between questions and individuals. In columns (1), (3) and (5) we regress the submission time on each trait of the personality traits and each preference separately. In columns (2), (4) and (6) we add all traits and preferences at the same time. We aggregate the submission times from three different treatments in columns (1) and (2).

We find that the discount rate is significantly negatively associated with submission time, which suggests that patient test-takers submit significantly later. Since we cannot identify differences in technology between patient and impatient test-takers, but do find a significant impact on the test-score, the crucial point is that patient people wait longer until they answer a question but do not know the answer faster. The opposite picture emerges from risk preferences. We find no significant association between submission time and risk attitude. However, we find strong differences in technology, which suggests that the impact of risk attitude on test result seems to be driven by technology and not by behavior. In our baseline specification in columns (1) and (2) none of the personality traits is significantly associated with submission time. It seems that openness is associated with waiting longer to submit an answer. Conscientious people wait shorter to submit their answer.

The effects of personality traits and economic preference parameters might interact with differences in treatments. We split our sample into Low Stake treatment and High Stake treatments in columns (3) to (6). It turns out that agreeable test-takers wait longer until they submit their answer and conscientious test-takers submit their answer earlier ($p < .1$) in our Low Stake treatment. In columns (7) and (8) we calculate the differences between the treatments and test whether they are significantly different from zero. We find weak evidence that more neurotic and more agreeable people react differently in different environments. However, none of the differences is statistically different from zero. Our results are consis-

Table 3.6: Determinants of Submission Time

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All Treatments		Low Stake Treatment		High Stake Treatments		(5)-(3)	(6)-(4)
Discount Rate	-1.425*	-1.469*	-1.282	-1.31	-1.483*	-1.543*	-0.2	-0.23
	(.827)	(.827)	(.889)	(.866)	(.824)	(.836)		
Risk Attitude	-0.471	-0.622	-0.714	-0.836	-0.343	-0.502	0.37	0.33
	(.680)	(.708)	(.773)	(.787)	(.681)	(.719)		
Openness	0.397	0.345	0.779	0.776	0.186	0.0815	-0.593	-0.695
	(.627)	(.618)	(.676)	(.676)	(.647)	(.649)		
Conscientiousness	-0.776	-0.723	-1.351*	-1.429*	-0.501	-0.377	0.85	1.05
	(.660)	(.607)	(.763)	(.727)	(.642)	(.586)		
Extraversion	0.0584	0.111	0.513	0.5	-0.114	-0.0222	-0.623	-0.52
	(.597)	(.598)	(.687)	(.679)	(.589)	(.605)		
Agreeableness	0.623	0.544	1.116*	0.798	0.395	0.437	-0.721	-0.361
	(.556)	(.590)	(.609)	(.662)	(.560)	(.591)		
Neuroticism	0.0753	0.0141	0.257	0.293	-0.0473	-0.15	-0.297	-0.443
	(.661)	(.699)	(.722)	(.729)	(.666)	(.715)		
Constant		22.81***		23.41***		22.95***		
		(1.079)		(1.592)		(1.413)		
Observations	5,760	5,760	1,920	1,920	3,840	3,840		
R-squared		0.244		0.228		0.257		
Question FE	YES	YES	YES	YES	YES	YES		

Note. The table shows results of a linear panel regression. The dependent variable is the time of submission in seconds. All values of the independent variables are standardized with mean 0 and s.d. 1. Columns (1) , (3)and (5) report coefficients from separate regressions for each independent variable. Columns (7) and (8) report the differences between the Low Stake and the High Stake Treatments. We add treatment dummies in columns (1),(2), (5) and (6). All columns contain question fixed effects. Standard errors are clustered at the level of the individual (128 clusters). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

tent with respect to the preference measures. The picture that emerges from Table 3.6 is that especially the discount rate plays an important role in determining the answering behavior. Subjects who are impatient submit their answer significantly earlier, which is detrimental for their test scores.

3.5 Conclusion

In this chapter we investigate the behavioral mechanisms behind a test score. Compared to previous studies (e.g., Borghans et al. (2008)) we use a new experimental method to shed light on the deliberation process while answering a question. While we were mainly interested in the impact of incentives and time pressure on technology and behavior in a recent paper (Borghans et al., 2014), this paper looks at the role of personality traits and economic preferences in the determination process of a test score. Since we are able to discriminate between the answering technology and answering behavior during a test we can assign different roles to personality traits and preferences in the determination process. If we look at the aggregate test score we find weak evidence that personality traits measured with the Big Five are associated with the aggregated test score. However, one of the main determinants for a good test result seems to be an individual's discount rate and her risk preference. These results are in line with previous findings (e.g., Dohmen et al. (2010); Burks et al. (2009)). The new feature of this study is that we can explain why and how preferences and certain personality traits influence test results. Our experiment reveals that patient individuals do not have a superior technology during a test but they simply wait longer until they submit their answer. In contrast, risk lovers seem to have a higher probability of knowing the answer at any point in time while answering a question. More open and emotionally stable people seem to have a superior answering technology.

Our study has three major caveats. First we use a university student sample. Data from a more representative population is needed to further support our findings. Second, our data focuses only on intelligence questions. It is important to investigate whether similar results are obtained when subjects answer different types of questions. Third, it is not entirely clear what our preferences measures capture. Does the correlation between an individual's patience and a better test outcome imply that we capture actually patience as a separate preference or personality trait or do more able individuals better understand the trade-off between

smaller earlier rewards and greater delayed awards? We think that these are important questions for further research projects.

Further research should focus on a more representative sample to increase the external validity of our results. Moreover, studies from brain scanners (e.g., FMRI scanners) or eye tracking devices can complement our methodology. Evidence from the field linking preference and personality measures with test outcomes should be used to further validate our findings.

Chapter 4

Patience and Achievement Test Results

4.1 Introduction

Imagine two equally smart students in primary school who are both taking an important test, which is a major determinant of their secondary school track. One of the students is impatient and rushes through the test. The other student, however, takes her time to think about each question. In the end, she obtains a higher test score and is allowed to go to a higher level secondary school, whereas the other student ends up in a lower school track. Achievement tests are an important tool to measure a student's ability, to sift and sort children in different educational tracks and to measure the quality of the education system.[1] Evidence on determinants of test scores, besides cognitive abilities of the test takers, is scarce.

The aim of this paper is to empirically investigate the relationship between patience and the result of an achievement test. We analyze whether this relationship holds for high and low stake achievement tests. This is important because differences in high and low stake tests are not well established yet. The idea is that patience is a personality trait which is related to self-control and ambition but not necessarily related to pure cognitive abilities (Duckworth et al., 2012; Moffitt et al., 2011). We link an economic measure of patience to achievement test scores. Our study has three key features. First, we use an experimentally validated measure of an individuals' patience, namely the internal rate of return for money. Second, our data set contains scores of two different achievement tests and a measure for IQ. One of the tests is a high-stake achievement test, which is important for the future career of the students in the Netherlands (Cito test).[2] The second one is a low-stake achievement test, which contains items of the PISA and the COOL study.[3] The high-stake test was taken at the age of 12. The low-stake test was taken in the course of the survey at the age of 15. Third, our data set contains a variety of controls such as measures of personality traits of the students, obtained from the Big-Five questionnaire, and the socioeconomic status of the parents. It is representative for the student population in secondary education in the south of the Netherlands. Since the survey was mandatory for the students, the data do

[1] The US College Board reports an official number of 1.6 million students taking the SAT test in 2013 (Board, 2013).

[2] In the Netherlands 97% of the students in primary school take this achievement test which is a crucial determinant for their secondary school track. In 2014 this summed up to a total of 160,000 students (Cito, Cito).

[3] PISA is an international achievement test which is used to compare the educational attainment between countries. COOL is an achievement test, which is used for research purposes to compare the performance of schools in the Netherlands.

not suffer from selection into the survey.

Our empirical strategy is as follows. We first investigate a potential relationship between patience and cognitive skills. Since we measure patience as an economic preference parameter, we investigate the relationship between cognitive skills and time preference (e.g., Benjamin et al. (2013); Dohmen et al. (2010); Burks et al. (2009)). Next, using simple OLS regression models with a big set of controls we examine the relationship between patience and achievement test scores. In a last step we investigate differences in the distribution of our patience measure between different school tracks. The analyzes in this research are mainly descriptive and complement the evidence from the lab in the previous chapters.

Our main results are as follows. First, our patience measure is correlated with IQ. Students who score one standard deviation higher on the IQ test tend to score 10 percent of a standard deviation better on our patience measure. The sign of the estimate is in line with previous findings in the literature (Benjamin et al. (2013); Dohmen et al. (2010); Burks et al. (2009); Borghans et al. (2008)). Second, we find an economically and statistically significant link between patience and high-stake test results. Students who score one standard deviation higher in our patience measure, score 16.7 percent of a standard deviation higher on the high-stake achievement test. This point estimate is more than half of the size of the point estimate for IQ. The relationship is much weaker for the low-stake achievement test. In fact the relationship between our patience measure and low-stake test scores only holds for the mathematics score but not for language. These results can also be driven because of a bigger measurement error in the low stake test. In the last section, we provide evidence for potential channels of patience. We also find differences in the degree of patience between different school tracks. Students from higher school tracks report substantially higher degrees of patience compared to students from lower tracks. This is in line with the findings of Golsteyn et al. (2014) who find differences in discount rates for different levels of educational attainment.

This paper contributes to the literature in economics and psychology. The first strand is the empirical literature in economics and psychology about patience and various lifetime outcomes, such as tests scores. Achievement tests are important outcomes in modern societies since they determine for instance school tracks, university study entries or hiring decisions. We add to this literature by linking an economic measure of patience with low and high-stake test results and control for various observables such as non-cognitive abilities, gender and age.

Non and Tempelaar (2014) find weak evidence that patient university students obtain better grades and have a higher chance of passing the first sit exams. In a recent paper Golsteyn et al. (2014) find in a representative Swedish panel that lower discount rates at the age of 13 predict higher educational attainment, higher wages, wealth and better health. In a follow up paper Akerlund et al. (2014) find a statistically and economically significant relationship between high discount rates and the engagement in criminal behavior. Sutter et al. (2013) find in a sample of young Austrian adolescents that favorable time preference parameters (higher degrees of patience) are associated with higher saving rates, better math grades and lower spending on smoking and drinking. Burks et al. (2012) find a positive relationship between impatience and trainee-program drop-out and unemployment for US-truck drivers. Duckworth et al. (2012) find that controlling for IQ, higher self-control is associated with better grades, but not with higher achievement test results in a US sample. Self-control in childhood has been found to predict higher wealth and better health in a representative longitudinal data set from New Zealand (Moffitt et al. (2011)). Similarly, Castillo et al. (2011) find that children with lower discount rates are significantly less likely to receive disciplinary referrals. Bettinger and Slonim (2007) investigate which individual attributes correlate with incentivized measures of time preferences. They find that girls and older students seem to be significantly more patient. Test scores seem to have no relationship with the degree of patience. However their total sample is small (N=191) and the number of observations at the age of 15 is only 77.[4] Kirby et al. (2005) find a weak negative association between higher discount rates and the average grade in US colleges.

The second strand is the literature in economics and psychology on the measurement of an individual's patience. Borghans et al. (2008) state that [t]he science of measuring preferences is almost a century behind the science of measuring IQ. It is of crucial importance to know which preference measures predict outcomes best. Economists think of patience as a preference parameter whereas psychologists tend to interpret it as a personality trait. In general there are various distinct methods to measure patience. The first method is used by economists and psychologists. Individuals are confronted with the choice between a smaller earlier reward and a delayed greater reward. The reward can be money or a consumption

[4]They also investigate whether students are able to make rational choices when they are faced with the typical questions economist use to elicit discount rates. They find that more than 25% of their participants make inconsistent choices. We enforce rational choices through our stair case elicitation method.

good such as cookies. When using this method economists and psychologists try to measure an individual's discount rate. They define the discount rate as the parameter of a function with which an individual devaluates a payoff in the future.[5] Despite the fact that there is a debate how the discount function looks like, there is also a debate how to accurately measure an individual's discount rate. In the seminal study by Mischel et al. (1989) children at the age of 4 had the choice between one Marshmallow now or two Marshmallows in 15 minutes. Those children who were able to wait, developed better competences and had higher scholastic performances at the age of 10. Many papers replicated these findings using similar measures (see for instance Duckworth et al. (2013) and the citations therein). However, psychologists were among the first to also use real monetary incentives to elicit an individuals' discount rate (see Frederick (2005)).

Other studies use observer reported scales of teachers or parents to assess self-control among children or young adolescents. Moffitt et al. (2011) for instance use parents', teachers' and self-assessments of an individual on various ratings such as impulsivity, fleeting attention and lacking persistence to create a scale of self-control. Duckworth and Kern (2011) summarize the measures of self-control and patience in four main categories: executive function tasks, delay of gratification tasks, self-report questionnaires and informant-report questionnaires. Our method as most of the economic measures of patience and self-control belongs to the category of delay of gratification tasks.

The remainder of the paper is structured as follows. Section 4.2 describes our data set and the measures we use. Section 4.3 presents the results, section 4.4 provides a discussion and Section 4.5 concludes.

4.2 Data

We use rich survey data combined with administrative data of secondary schools in the south of the Netherlands. The survey and the test were fully computerized and students had to fill in the questionnaire during a school lesson. The major advantage of this collection method which results from a unique cooperation with

[5]The theoretical concept was first formalized in a model by Samuelson (1937). He introduces one parameter which devaluates the payoff in future periods. We refer to the discount factor as the parameter δ with which an individual devaluates the payoff in each period: $U_t(x) = \delta^t v(x)$. The discount rate is the actual value with which payoff in time t is devaluated with. Later work in psychology and Economics introduced the concepts of hyperbolic (Ainslie (1975); Strotz (1955)) and quasi hyperbolic discounting (Laibson (1997)) which means that periods which are closer to today are devaluated stronger than periods which are further away from today.

almost all schools in this region is that our data does not suffer from selection into the survey. The approximate duration of the questionnaire was 50 minutes. A total of 9,092 students participate in the survey and a randomized subsample of 490 students received our time preference questions.[6] The average age of our participants is 15.3 years (standard deviation 0.56, $min = 13.8$, $max = 17$) and 52.5 percent of them are female.[7]

4.2.1 Measuring Patience

In this paper we use an economic measure for patience. We define patience as the ability to delay immediate gratification. Economists typically use questions in which subjects have to make a decision between a smaller earlier amount and a delayed greater amount. Time horizons and the level of incentives differ between studies. Ideally the researcher uses incentivized measures of time preferences (see for instance Dohmen et al. (2012)). In many large scale studies it is not suitable to confront participants with sweets or to pay them according to their decisions. The reasons for that are mostly time and budget constraints. The key feature of our study is that we use an experimentally validated measure of time preferences.

Some studies also use unincentivized measures of discount rates where individuals had to make decisions between earlier smaller amounts and delayed greater amounts (e.g., Golsteyn et al. (2014); Non and Tempelaar (2014)). However, for these measures it is not clear whether individuals would give similar answers if the amounts were actually paid to them. Moffitt et al. (2011) use a battery of 9 items to assess an individual's degree of self-control. They use a rating of teachers, parents and self-reported answers of items such as fleeting attention, lacking persistence, impulsivity to assess self-control of children between the age of $3-11$. This method of assessment has the advantage of reducing measurement error of the respective trait of interested since one has multiple observers for the same item. On the other hand it is often not feasible in huge data sets to get this detailed information.[8] This is why we use a rather short but very efficient version

[6] The samples do not seem to be significantly different in observables. We compare age, gender, education level of the parents, IQ, Big-Five personality traits, high and low stake achievement test scores. We only obtain a slight significant difference in the high achievement test between the whole sample and the sample which received the time preference question. The randomly selected sample seems to perform about 1% of a standard deviation better (t-test, $p < 0.1$) on the achievement test. The results are available upon request.

[7] Further information on the data set can be found in Feron et al. (2014) and Borghans and Schils (2013).

[8] Another potential advantage of eliciting discount rates is the possible calibration of parameters such as δ and β (Laibson (1997); Samuelson (1937)) in utility functions.

of assessing an individual's patience. Falk et al. (2012) found that the answers to these unincentivized questions explain most of the variance from incentivized experiments to elicit an individual's discount rate compared to other conventional elicitation methods.

Before our participants answered the questions we explained to them that they have to make five hypothetical choices between a smaller amount they would receive now and a greater later amount in a year from now.[9] We vary only the later amount depending on what a person answers to each of the five questions. The first question each individual faces is the following: Please choose: Would you like to get 100 Euro today or 154 Euro in 12 months? If a person chooses the 100 Euro today the amount for the delayed option increases to 185 Euro in the second question. If a person chooses the 154 Euro the amount of the delayed option decreases to 125 Euro in the second question. This method enables us to pin down the upper and lower bound of an individual's internal rate of return.

4.2.2 Outcome Variables

Our data set also contains results of two tests which were conducted at different points in time and under different conditions. Table 4.1 shows an overview of the tests and their characteristics. The first test is a high-stake achievement test called Cito test which takes place at three days with each day lasting approximately three hours. This sums up to a total of nine hours. We obtain this test result from administrative data. 97% of all children at the end of primary school take this test and it is one of the major determinants for the secondary school track in the Netherlands.[10] The Cito test has a focus on measuring math and language skills. However it also contains a battery of other skill measures such as geography, history and study-skills. Children take the test when they are on average 12 years old, hence the Cito test result is a past measure of achievement. The second test is a low-stake achievement test which combines items of the PISA test and the COOL test.[11] The achievement test is used by the OECD to compare the student's scholastic performance in math and language at the age of 15. COOL is

[9]The exact wording of the introduction for the questions was as follows (translated from Dutch): Please answer the following questions about the patience you have to postpone things. You are always asked if you want a certain amount today or a greater amount in one year. As soon as you made your choice we will present the next situation. In total you have to make five of these decisions with different amounts..

[10]Further information on the four main parts, covering language, study abilities, mathematics and general knowledge can be found in Cito (Cito).

[11]PISA stands for Program of International Student Assessment.

an achievement test which is similar to PISA. It is used by Dutch researchers to investigate the development of pupils from 5 to 18.[12]

Table 4.1: Overview of the Tests

Name	Content	Mean age when taken	Duration in minutes
CITO	High-stake achievement test, math and language skills,	12	540
COOL and PISA	Low-stake achievement test, math and language skills	15	30

4.2.3 Control Variables

Our data set contains a rich set of control variables. It includes a measure of cognitive skills. The IQ test which was part of the questionnaire contained amongst others similar items as the Raven matrices (Raven (1962)). Examples of the items are provided in the Appendix. Moreover, we obtain measures of the Big Five personality traits (Goldberg (1992)) and the school track of the students. We will make use of this information to investigate potential mediators of patience.

4.3 Results

In this section we present the results of our study. We start off by showing the raw correlation structure between our variables of interest. Next, we use a similar empirical strategy as Burks et al. (2012). We first document the relationship between our patience measure and various covariates, such as the score on an intelligence test, personality traits and scores on the achievement tests. Afterwards we investigate the relationship between the score on a high-stake and a low-stake achievement test and patience, cognitive skills and personality traits.

[12] More information can be found on `www.cool5-18.nl`.

4.3.1 Correlation Structure

The determination process of an achievement test is complex. Thus, it is helpful to investigate the raw correlation structure of these variables. Since economic preference parameters could also capture facets of an individual's personality, we investigate the cross correlation between our personality measures and the patience measure. Table 4.2 shows the raw correlations between our variables of interest. All variables except for gender and age are standardized with mean zero and standard deviation one. Openness to experience, conscientiousness and agreeableness are significantly negatively correlated with our discount rate measure. The correlations between the Big Five personality inventory and the discount rate are weak. The correlation signs are similar to the one in Becker et al. (2012). However, the size of the correlation is slightly higher.[13] Overall, the picture in Table 4.2 suggests a rather complementary relationship between an individual's discount rate and personality traits.[14]

The high-stake achievement test score is significantly negatively correlated with our patience measure, which suggests that patient individuals seem to have a higher score on these tests. However, the low-stake test score is not significantly correlated with the discount rate. The correlations between the high and low-stake test are positive and highly significant. Summing up, we find a rather low correlation between our patience measure and personality traits. Next, the high and low-stake achievement test seem to measure a similar set of skills since the correlation between the test scores is positive and significant. However, the correlation between the high-stake test and the discount rate is significantly negative whereas the correlation between the low-stake test and the discount rate is not statistically significant.

4.3.2 Patience and IQ

Many studies argue that individuals with higher cognitive abilities are also more able to value future outcomes and delay their immediate sense for gratification (see for instance Benjamin et al. (2013); Dohmen et al. (2010); Borghans and Golsteyn (2007)). One of the crucial arguments of this paper is that patience is not necessarily driven by pure cognitive ability. This is similar to the argument brought up by Frederick (2005). However, we need to investigate the relationship

[13] We use Table 3 in Becker et al. (2012) to compare our findings.

[14] We also provide scatter plots of the most important relationships to document possible non-linear relationships. The graphs are provided in the appendix.

Table 4.2: Cross correlations between discount rate, personality traits and test outcomes

	Discount Rate	O	C	E	A	N	Age	1 if female	IQ	High-stake
Openness	-0.209***									
Conscientiousness	-0.074	0.115								
Extraversion	-0.056	-0.023	0.021							
Agreeableness	-0.153**	0.279***	0.223***	0.281***						
Neuroticism	0.1	-0.066	-0.118	-0.087	0.180**					
Age	0.084	-0.057	-0.12	0.097	0.016	0.082				
1 if female	-0.005	-0.042	-0.035	0.125	0.302***	0.334***	0.157**			
IQ	-0.122	0.026	-0.019	0.073	-0.006	0.029	-0.045	0.015		
High-stake Test	-0.183**	0.037	-0.141*	0.0463	-0.006	-0.094	-0.151**	-0.079	0.450***	
Low-stake Test	-0.104	0.063	-0.011	-0.114	0.003	-0.016	-0.252***	0.031	0.306***	0.416***

All measures except for the gender dummy and age are standardized. *$p < 0.1$, **$p < 0.05$,***$p < 0.01$. The correlations were obtained using 173 observations.

between patience and cognitive ability. Since our study participants had to do an intelligence test we are able to investigate the relationship between patience and cognitive abilities. Table 4.3 shows the results. The dependent variable are the standardized values of our discount rate measure. In column (1) we only include the standardized IQ score as independent variable and in the following column we control for personality traits (column 2). We add age and gender in column (3). The picture that emerges from Table 4.3 is that a higher score on the IQ test is significantly positively associated with a lower discount rate. Thus, students with a higher measured cognitive ability exhibit higher degrees of patience. This relationship holds if we control for personality traits and gender. Individuals who score higher on the trait openness to experience also seem to have a significantly lower discount rate. The effect size seems to be equally strong compared to our measure if IQ. Moreover, higher levels of neuroticism are associated with lower degrees of patience.

4.3.3 Patience and High-stake Achievement Test Results

The key question of this paper is to investigate the relationship between patience and achievement test scores. Table 4.4 shows results from a linear regression with the standardized Cito test score as the dependent variable. The test score from the Cito test is one of the main determinants of placement into a secondary school track. Because of the early tracking system in the Netherlands, a higher test score allocates the child to a higher educational track. This makes the test a high-stake test. We first analyze the aggregate score of mathematic and reading skills. Then, we will analyze the score in mathematics and language separately.

In the first column we only include the standardized values of the patience measure as an independent variable. The estimates suggest that an increase of one standard deviation in the discount rate is associated with a decrease of 0.23 of a standard deviation in the overall test score ($p < 0.01$). Note that these estimates could be biased due to misspecification of our regression model. We separately regress the test-score on our IQ measure in column (2). We add IQ and patience in column (3). Compared to the previous specifications the coefficients remain robust. We find that IQ is significantly positively associated with the overall test score. The effect size of the discount rate estimate does not change significantly compared to the first specification but slightly decreases in magnitude. If we add the Big Five personality traits in column (4) the estimate of patience further drops. From the Big Five personality traits openness to experience is positively associ-

Table 4.3: Covariates of the Discount Rate

	(1)	(2)	(3)
IQ	-0.1352**	-0.1188**	-0.1094*
	(.0580)	(.0569)	(.0566)
Openness		-0.1130**	-0.1214***
		(.0428)	(.0448)
Conscientiousness		-0.0851	-0.0869
		(.0641)	(.0641)
Extraversion		0.0994*	0.0969*
		(.0500)	(.0510)
Agreeableness		-0.0942**	-0.0802*
		(.0452)	(.0466)
Neuroticism		0.1165***	0.1215***
		(.0408)	(.0411)
Age			0.0939
			(.0806)
1 if female			-0.0911
			(.1036)
Constant	-0.0237	-0.0234	-1.4111
	(.0569)	(.0564)	(1.2376)
Observations	454	454	454
R-squared	0.0184	0.0695	0.074

Note. All measures except for gender and age are standardized. Standard errors are clustered at the school level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4.4: Determinants of high stake achievement test score

	(1)	(2)	(3)	(4)	(5)
Discount Rate	-0.2343***		-0.1922***	-0.1677***	-0.1671***
	(.0578)		(.0546)	(.0554)	(.0563)
IQ		0.3403***	0.3148***	0.2986***	0.2883***
		(.0566)	(.0549)	(.0541)	(.0515)
Openness				0.2035***	0.2003***
				(.0572)	(.0588)
Conscientiousness				-0.0807*	-0.0921**
				(.0440)	(.0457)
Extraversion				-0.0021	0.0039
				(.0461)	(.0468)
Agreeableness				-0.0109	0.0105
				(.0500)	(.0494)
Neuroticism				-0.0919**	-0.0703*
				(.0348)	(.0351)
Age					-0.1902*
					(.0992)
1 if female					-0.1482
					(.0969)
Constant	0.0699	0.0627	0.0589	0.0554	3.0371**
	(.0959)	(.0885)	(.0851)	(.0810)	(1.4999)
Observations	397	397	397	397	397
R-squared	0.0582	0.124	0.1625	0.2111	0.2266

Note. All measures except for gender and age are standardized. The dependent variable is the standardized sum of the score of the sum on the mathematics and reading part of the cito test. Standard errors are clustered at the school level. *** $p < 0.01$, ** $p < 0.05$, $*p < 0.1$

ated with the overall score on the high-stake achievement tests. Neuroticism and extraversion are negatively associated with the test score. In the next specification in column (5) we add gender and age in years as additional controls. It turns out that the age of the test-taker is negatively associated with the test score. One potential reason for this finding is that older students could be repeaters of a class. Our results from Table 4.4 indicate that patience is statistically and economically significantly correlated with the score of the high-stake achievement test. Controlling for IQ, personality traits, gender and age a one standard deviation increase in the discount rate is associated with a 0.17 standard deviation decrease in the high-stake achievement test score. The magnitude of our patience estimate is equal to 58% of the relationship between IQ and the achievement test score. Thus, more patient subjects score higher on a high-stake achievement test.

4.3.4 Patience and Low-stake Achievement Test Results

An important question is whether the relationship between patience and test scores also holds for tests which are conducted under different conditions. In the previous section we found support for a significant relationship between being patient and a better test outcome in a high-stake achievement test. Our data set also contains measures of a low-stake achievement test, which is similar to the PISA and COOL test.

In the previous analysis in Table 4.2 we document a high and significant correlation between the high and low-stake test results ($\rho = 0.416$). Analyzing the relationship more in depth, Figure 4.1 shows the correlation between the high-stake and low-stake achievement tests. First, it shows a scatter plot with the standardized high-stake achievement test scores on the y-axis and low-stake achievement test scores on the x axis in the middle of the graph. The straight line indicates a linear regression and the gray area around that line shows the 95 percent confidence interval. If there was no difference in the test scores the slope of the regression line should not be statistically different from 1. However, the slope of the regression line is 0.47 and highly significantly different from one ($p < 0.001$).[15] On the top of the graph we document the corresponding distribution of the standardized values of the low-stake achievement test. On the right hand side we plot the distribution of the standardized values of the high-stake achievement test. The distribution of

[15] Note that the scatter plot in Figure 4.1 contains 200 observations and the reported correlation in Table 4.2 only 173 since we restrict the number of observations in Table 4.2 to those where all variables are available. The results obtained in Figure 4.1 do not change significantly if we restrict the analysis to the 173 observations.

the low-stake test seems to follow a normal distribution. However, the distribution of the high-stake test seems to be right-skewed. Note that this is due to the vertical position of the distribution. If the distribution was plotted on the horizontal axis it would be left-skewed. However, this truncation of the distribution is typical for this kind of test and is not due to an artefact of our sample (e.g., Borghans et al. (2014)).

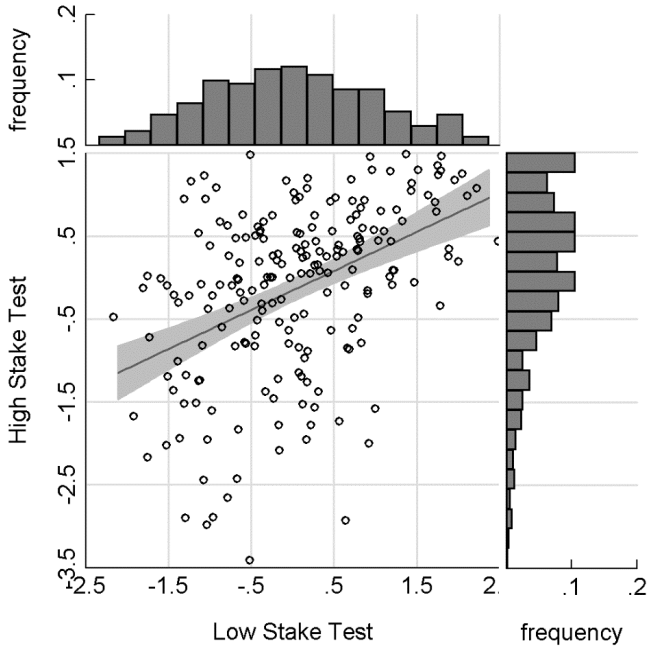


Figure 4.1: The relationship between high-stake and low-stake test scores

In the following we will analyze the relationship between the low-stake test and patience. Table 4.5 shows results of a linear regression with the standardized test score of the low-stake achievement test as the dependent variable. Our empirical strategy is the same as in the previous section. We start off by using only our patience measure as independent variable in the first specification in column (1). In this specification the discount rate seems to be negatively associated with the test score. However, the estimate is not significant and is only half of the size compared to the high-stake test score. In column (2) we only regress the low-stake test score on our IQ measure. As for the high-stake test, IQ is significantly positively associated with the low-stake test score. In column (3) we add IQ and patience to

our model. Patience remains insignificant and IQ remains significant. Both point estimates slightly drop but do not change a lot. We add controls such as Big-Five, gender and age in columns (4) and (5). In all specifications, IQ is positively and highly significantly ($p < 0.01$) associated with the score on the low-stake achievement test. The size of the point estimate of the discount rate drops and remains insignificant. However, extraversion seems to be negatively associated with the test score. Moreover the age of the test taker is also negatively associated with the test score. Compared to the results of the high-stake achievement test patience seems to matter less for the aggregate low-stake achievement test score. However, the sign of the estimate is still negative which can potentially indicate a similar relationship between patience and test performance on a low-stake achievement test.

4.3.5 Differences in Math and Language Test Scores

High-stake Test Score

The relationship between patience and a test score could be different between different test subjects. In this section we analyze two sub scores of the high-stake achievement test separately. Table 4.6 shows the results. Columns (1) to (4) show estimation results with various controls for the mathematics test score and columns (5) to (8) for the language score.

In the columns (1) and (6) we only include the patience measure as an independent variable. The estimates suggest that an increase of one standard deviation in the discount rate is associated with a 20 percent of a standard deviation decrease in the math test score ($p < 0.05$) and a 26 percent of a standard deviation decrease in the language score ($p < 0.01$). This is followed by a specification which only includes the intelligence measure in columns (2) and (7). Similar to the previous regression results IQ is positively and significantly associated with both the language and the math score on the high-stake achievement test. Note that these estimates could be biased due to omitted variables in our regression model. Thus, we first run regressions with IQ and the patience measures simultaneously. Both point estimates decrease slightly but remain significant for both test scores. We successively add Big Five personality traits (columns (4) and (9)) as controls. We find that IQ is significantly positively associated with the mathematics and language test score. The effect size of our discount rate estimate does not change significantly compared to the first two specifications but the level of significance

Table 4.5: Determinants of low stake achievement test

	(1)	(2)	(3)	(4)	(5)
Discount Rate	-0.0977		-0.0636	-0.0566	-0.0413
	(.0794)		(.0743)	(.0797)	(.0795)
IQ		0.3128***	0.3043***	0.3171***	0.3052***
		(.0718)	(.0717)	(.0721)	(.0724)
Openness				0.0271	0.0293
				(.0811)	(.0839)
Conscientiousness				-0.0222	-0.0421
				(.0774)	(.0779)
Extraversion				-0.1623**	-0.1425**
				(.0620)	(.0613)
Agreeableness				0.0468	0.0255
				(.0863)	(.0865)
Neuroticism				-0.0408	-0.0498
				(.0669)	(.0620)
Age					-0.4041***
					(.1179)
1 if female					0.1737
					(.1581)
Constant	0.1073	0.1283	0.1248	0.1301	6.2327***
	(.0877)	(.0809)	(.0790)	(.0801)	(1.7738)
Observations	173	173	173	173	173
R-squared	0.0109	0.0938	0.0983	0.1218	0.1769

Note. All measures except for gender and age are standardized. The dependent variable is the standardized sum of the scores on the mathematics and reading part of the low-stake test. Standard errors are clustered at the school level. *** $p < 0.01$, ** $p < 0.05$, *$p < 0.1$

decreases. In the next specification in columns (5) and (10) we include gender and age as controls. The age of the test taker seems to be negatively associated with the test score. This could be due to repeaters. Interestingly the relationship between patience and the mathematics and language score is quite similar. People who are more open to experience score better in the mathematics and the language part. However, neuroticism is only significantly negatively associated with the score on the mathematics test. Interestingly women seem to perform significantly worse on the mathematics test score. Overall these findings are consistent with what we find on the aggregate test score.

Low-stake Test Scores

This section presents the analysis of the two separate scores for language and mathematics on the low-stake achievement test. Table 4.7 shows results of a linear regression with the standardized test results of the low-stake achievement test as the dependent variable. Columns (1) to (5) document the estimation results with various controls for the mathematics test score. Columns (6) to (10) show the results for the language test score. The empirical strategy is the same as in the previous section. We only include our patience measure as independent variable in the first specifications in column (1) for math and column (5) for the language test score. The discount rate seems to be negatively associated with the language and the math score. This estimate is similar in magnitude compared to the high-stake test. However, the relationship is only significant for the math score (p¡0.01) and not for the language score. We also run a separate regression only with IQ as independent variable. As for the high-stake test score IQ is highly significantly positively associated with the low-stake test scores in the math part and the language part. Interestingly, the size of the point estimate for the language score is only half of the size of the estimate of the score on mathematics. We add controls such as Big-Five, age in years and gender in columns (3)-(5) and (7)-(10). After controlling for personality and IQ we only find a significant negative relationship between an individual's discount rate and the score in mathematics but not for the language score. However, extraversion seems to be negatively associated with the score on the mathematics part and the language part of the test. We also find weak evidence that openness to experience is positively associated with the score in mathematics. Similar to the high-stake test, neuroticism is only negatively associated with the score on the mathematics test.

Table 4.6: Determinants of high stake achievement test score

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
			Mathematics					Language		
Discount Rate	-0.2052**		-0.1742**	-0.1694	-0.1915*	-0.1881**		-0.1548*	-0.1534*	-0.1368
	(.0877)		(.0830)	(.1025)	(.0974)	(.0800)		(.0796)	(.0825)	(.0850)
IQ		0.4034***	0.3908***	0.3818***	0.3974***		0.4130***	0.3984***	0.3814***	0.3797***
		(.0805)	(.0792)	(.0803)	(.0721)		(.0950)	(.0945)	(.0816)	(.0806)
Openness				0.1436	0.1076				0.132	0.1464
				(.0937)	(.0889)				(.0967)	(.0975)
Conscientious-				-0.0736	-0.1177				-0.2006**	-0.2066**
ness				(.0842)	(.0862)				(.0747)	(.0790)
Extraversion				0.0187	0.0595				0.0284	0.0402
				(.0662)	(.0695)				(.0701)	(.0746)
Agreeableness				-0.0282	0.0683				0.1371	0.1118
				(.0666)	(.0707)				(.0892)	(.1001)
Neuroticism				-0.1714***	-0.1101*				-0.0558	-0.0656
				(.0596)	(.0539)				(.0627)	(.0716)
Age					-0.1676					-0.1980*
					(.1522)					(.1160)
1 if female					-0.5472***					0.1559
					(.1776)					(.1616)
Constant	0.067	0.0504	0.0296	0.0197	2.8799	-0.0665	-0.042	-0.0431	-0.0738	2.8731
	(.1306)	(.1113)	(.1100)	(.1079)	(2.3178)	(.1308)	(.1087)	(.1051)	(.0992)	(1.7295)
Observations	159	159	159	159	159	157	157	157	157	157
R-squared	0.0422	0.1791	0.2093	0.2525	0.3236	0.0395	0.1682	0.1948	0.2621	0.2799

All measures except for gender and age are standardized. The dependent variable is the standardized score of the mathematics and reading part of the high-stake test. Standard errors are clustered at the school level. *** $p < 0.01$, ** $p < 0.05$, * p< 0.1

Table 4.7: Determinants of low stake achievement test score

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
			Mathematics					Language		
Discount Rate	-0.1577***		-0.1216**	-0.0803	-0.082	-0.0678		-0.045	-0.0205	-0.0132
	(.056)		(.054)	(.055)	(.058)	(.064)		(.061)	(.063)	(.063)
IQ		0.3028***	0.2862***	0.2697***	0.2598***		0.1818**	0.1757**	0.1831**	0.1654**
		(.061)	(.062)	(.059)	(.063)		(.070)	(.071)	(.069)	(.069)
Openness				0.1305*	0.1277*				0.0646	0.0816
				(.065)	(.067)				(.052)	(.051)
Conscientiousness				0.0262	0.013				0.0159	0.0083
				(.048)	(.047)				(.062)	(.064)
Extraversion				-0.1563**	-0.1407**				-0.1074**	-0.1047**
				(.058)	(.057)				(.046)	(.050)
Agreeableness				0.0363	0.0493				0.1095	0.0755
				(.075)	(.075)				(.067)	(.065)
Neuroticism				-0.1842***	-0.1468***				0.0165	0.0061
				(.053)	(.051)				(.062)	(.062)
Age					-0.2533**					-0.2260**
					(.096)					(.099)
1 if female					-0.1401					0.2204*
					(.135)					(.128)
Constant	0.0719	0.078	0.0775	0.0824	4.0233***	0.0891	0.0828	0.0798	0.0741	3.4113**
	(.092)	(.087)	(.085)	(.082)	(1.449)	(.070)	(.065)	(.065)	(.065)	(1.492)
Observations	278	278	278	278	278	285	285	285	285	285
R-squared	0.025	0.085	0.0996	0.1606	0.1829	0.0048	0.0326	0.0347	0.0605	0.0866

All measures except for gender and age are standardized. The dependent variable is the standardized score of the mathematics and reading part of the high-stake test. Standard errors are clustered at the school level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

4.4 Potential Mechanisms & Discussion of the Results

In our analysis we established a negative relationship between high-stake achievement test scores and an individual's impatience. We also documented that this relationship is slightly weaker for a low-stake achievement test score. Potential reasons for this difference can be the difference in duration of the tests. Whereas the high-stake test takes about nine hours over three days the low-stake achievement test only takes only about 30 minutes. This could have two potential effects. The first one is measurement error in the dependent variable. Our high-stake and low-stake tests differ in the stakes and the length. Thus the measurement error of math and language ability could be higher in the low- stake test than in the high-stake test. This would lead to an inefficient estimation that could explain why our estimates of patience remain insignificant (Greene (2012)). The second effect could be that patience can play a bigger role for the outcome if the time investment to succeed in a task is high.

We are reluctant in interpreting our results in a causal way. Nonetheless, for both researchers and policy makers it would be important to gain insights in potential mechanisms behind the relationship between patience and achievement test-scores. In this section we provide two potential explanations why patience could matter for a better test result.

First, there could be a mechanism between patience and the score during the actual answering process of the achievement test. Borghans et al. (2014) provide evidence from a laboratory experiment that patient individuals do not seem to be equipped with a better technology when they answer the test but they just think longer about a question and thus have a higher test score. Hence, differences in test results would stem from a different behavioral pattern during the test for patient and impatient individuals.

Second, there could be various ways how the preference parameter or personality trait patience is formed. A conservative assumption would be that the patience parameter is determined by birth. Another assumption would be that patience is malleable over the life cycle. Recent evidence from the literature in economics and psychology suggests that personality traits are malleable and also develop over the life cycle. People tend to become more conscientious and more agreeable when they become older (Borghans et al. (2008)). Evidence from the Perry preschool study suggests that early childhood interventions can change the non-cognitive

skills of individuals (Cunha and Heckman (2009)). A study by Perez-Arce (2011) for instance showed that being randomly assigned to better schools increased patience of Mexican students. One potential mechanism behind our results could be that different forms of schooling also shape an individual's patience and in this way they influence the achievement test score. In the Netherlands students are tracked into different levels of education at the age of 12. The tracks differ in their difficulty and the curricula. The lowest tracks focus on vocational education whereas the highest tracks prepare for studies at the university. We are interested in the relationship between patience and the school track. Thus, we split our data into four different tracks.[16] Figure 4.2 shows the distribution of our patience measure split up by the four different tracks. The picture that emerges from Figure 4.2 is that there are major differences in the distribution of patience between school tracks. Note that higher values indicate lower degrees of patience. There is a common trend that patience increases with the track level. Whereas students in the two lower vocational education tracks (vmbo lwoo/bl/kl and vmbo gl/tl) seem to be rather impatient, students in the two highest tracks (havo and vwo) exhibit substantially higher degrees of patience. The difference in patience between the two lowest and two highest tracks is significant at the 1 percent level.[17]

Since we analyze a cross-sectional data set we face the problem of reversed causality. Indeed, many studies (Benjamin et al. (2013); Dohmen et al. (2010)) use patience as the dependent variable to explain outcomes such as IQ or educational attainment. However, these studies are rather interested in the determination of economic preference parameters. Other studies also use outcomes such as crime, educational attainment and health as dependent variable and patience as independent variable (e.g. , Akerlund et al. (2014); Golsteyn et al. (2014)). We are interested in the determinants of an achievement test besides cognitive ability. It is unlikely that a good achievement test outcome actually made our participants more patient. Moreover, we argue that patience is a non-cognitive skill. Non-Cognitive skills are formed through a complex skill formation process (Cunha et al. (2010)) and one achievement test is quite unlikely to influence this formation as strongly as we find in our analysis. Another fact which is in favor

[16]The two lowest tracks are named vmbo (lwoo/bl/kl) and vmbo (gl/tl). Vmbo stands for voorbereidend middelbar beroepsonderwijs which means preparatory secondary vocational education. The two highest tracks are named havo (hogere algemeen voortgezet onderwijs) which means higher general continued education and vwo (vorbereidend wetenschappelijk onderwijs) which means preparatory scientific education.

[17]We run a linear regression with the discount rate as dependent variable and school track dummies as independent variables. The results are available upon request.

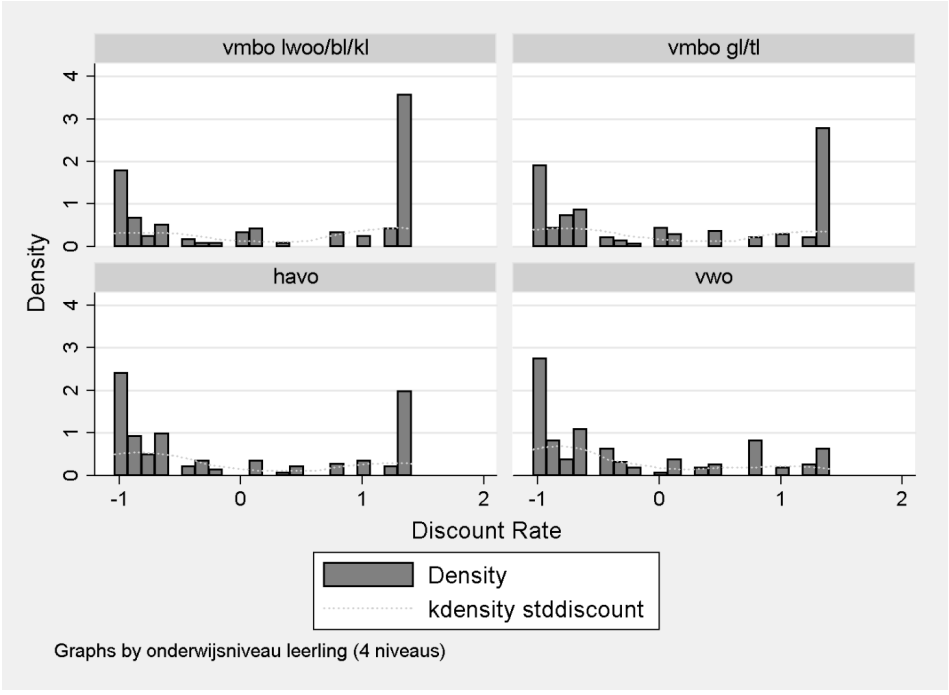


Figure 4.2: Patience by school track.

Note. We distinguish between four school tracks (lowest to highest): vmbo lwoo/bl/kl, vmbo gl/tl, havo, vwo.

of our argumentation is that the correlation between our patience measure and the achievement test result is very similar for the tests which were taken at two different points in time.

Another potential concern about any measure is its reliability. Psychologists usually test the reliability of their measures with a test re-test correlation. For economic measures this evidence is scarce. However, in a recent paper Wölbert and Riedl (2013) find a high test-retest correlation (Spearman's Rho 0.61) between discount rate measures taken at two different points in time. Similarly Kirby (2009) finds test-retest correlations of .7 with a one year time distance between the measurements. The stability of these measures is promising evidence for their validity.

A particularly interesting finding is the fact that our patience measure strongly varies between school tracks. The reasons for that are still unexplored. However, our findings suggest that education could potentially influence the non-cognitive

skill formation process. Experimental evidence is needed to provide clear causal relationships. Lastly, we remain silent about the relationship between socioeconomic background and the relationship between patience, cognitive skills and achievement test results. Recent evidence by Deckers et al. (2014) suggests that children from families with a higher socio-economic status are significantly more patient and perform better on an intelligence test.

4.5 Conclusion

In this paper we investigate the relationship between patience and achievement test scores. We add to the current state of research by showing that more patient students seem to score better on a high-stake achievement test. The relationship between patience and the test result is similar but weaker pronounced for a low-stake achievement test. Our results are relevant for policy makers. We find a strong relationship between patience and the success on achievement tests. Since achievement tests serve as major predictors for later outcomes such as wages, physical and mental health (Borghans et al. (2014)) it is important to understand what determines higher scores. Our results suggest that patience seems to be one of the crucial factors. Potential policies from our findings could be to train patience and the delay of gratification. Mischel et al. (1972) propose several methods such as distracting children from the respective reward by thinking fun things or hiding the reward facilitated the delay. Another promising method of enhancing patience and decreasing discounting is mental contrasting (Oettingen et al. (2010).

The current study provides a broad agenda for future research. To establish a clear causal link there is a strong need for randomized intervention studies. On the one hand these studies should focus on methods to train patience. Hence further research should focus on intervention studies which try to increase the evaluation of future events through, self-control strategies and increasing the awareness of future events such as mental contrasting. On the other hand more evidence from randomized studies in the field such as school lotteries and patience measures such as in Perez-Arce (2011) is needed to learn more about the mechanisms behind our findings. It would be important to know until what age patience continues to develop and what factors are important for this development. Another interesting step is to link our measure of patience to actual grades in school and graduation rates to test whether a similar relationship holds as we find for achievement test results.

Chapter 5

The Effect of Imposed Payment Schemes on Workers' Performance

5.1 Introduction

Performance dependent payment is in different degrees prevalent in many occupations and seems to be gaining momentum.[1] Direct incentive payment, such as piece rates, is especially prevalent in occupations in which output is relatively easily measured. For example, in the mailing sector employees are paid according to the number of letters they deliver and journalists, who write blogs, are often paid according to the number of words they write. On the other hand, discussions about the effects of pecuniary incentives for workers in the financial sector have been questioning the desirability of these incentives. As a result, banks have been forced to change their payment schemes towards a higher fraction of fixed payment. Recently it has become more and more popular to introduce performance dependent payment in the public sector as well. There is for example a lively policy debate whether or not it is beneficial to change the payment schemes of teachers from a fixed payment scheme to performance dependent payment (e.g., Lazear (2000); Woessmann (2011); Fryer (2013)). The effect of changing the payment mode in an occupation has not been studied extensively, since most studies ignore the effects of sorting. Employees could have sorted into a specific occupation because they prefer a fixed payment mode, which yields an effect on outcomes of both incentives and sorting when introducing performance dependent payment in that occupation.

This research studies performance differences in which people are forced to work under a certain payment scheme and compares performance relative to their preferred payment mode. The main aim to study these performance differences is to measure and understand which types of workers sort into which payment schemes and to identify the effects of imposed payment schemes on relative output. Addressing this question is crucial for understanding the effectiveness of changing the payment scheme in occupations for the existing workforce in those occupations. The theoretical and empirical literature suggests that workers sort according to productivity and preferences (Dohmen and Falk (2010, 2011); Larkin and Leider (2012)). Changing payment schedules leads to different sorting patterns in the labor market, which potentially overestimate the theoretically expected incentive effect of moving from a fixed payment scheme to a performance dependent payment mode. In the short run and for less mobile workers (e.g., those with firm-specific

[1] Gittleman and Pierce (2013) document that in the United States more than 40 percent of the payment of employees is performance related. The incidence of direct incentive payment fluctuated between 5 and 8 percent from 1994 to 2013.

human capital or older workers), it is difficult to change occupations overnight. This means that many workers will be faced with new working conditions, which could influence their performance.

We make use of a laboratory experiment to study the effects of imposed payment schemes relative to a worker's preferred payment scheme. To obtain a clear distinction between payment modes we selected two schemes: a fixed payment in which payment is completely independent of performance and a piece rate in which pay for performance is maximized. The setup of the experiment follows the design developed by Dohmen and Falk (2011) but differs from it by introducing an imposed and preferred payment mode.[2] It consists of four phases. In the first phase individual productivity levels are measured. The second phase randomly assigns subjects to a piece rate or a fixed payment scheme. We also observe which of the two payment schemes the subjects prefer and how much output is produced during a period of 10 minutes. In the third phase of the experiment subjects are allowed to make a choice between the two payment schemes. Again we observe output during a period of 10 minutes. To make sure our design does not suffer from order effects, half of the population is allowed to make a choice for a payment scheme in the second phase and faces a randomly selected payment scheme in the third phase. After each phase we obtain self-reported measures of work effort, stress, and exhaustion. The final phase consists of the measurement of risk preferences, personality traits, reciprocity, and a set of Raven matrices to measure cognitive ability.

The work task consists of multiplying one-digit numbers by two-digit numbers. The main advantage of using this work task is that it is a real effort task, which is easy to measure and to explain to subjects. It does not require previous knowledge and there are no learning effects involved during the short period of time the experiment takes. Finally, measured output is characterized by a relatively large degree of heterogeneity in productivity (the average productivity across all treatments equals 23.32 correctly solved problems, with a standard deviation of 11.99).

We study only two treatment conditions, a fixed payment independent of performance and a piece rate which is highly dependent on performance, which cover the extremes of the spectrum of observed payment schemes in reality. This allows us to first study the sorting patterns when the choice is between fixed payment

[2]Note that we take their z-Tree code and adapt it to our experimental setup. Moreover, for the data analysis we partly use their code to reproduce their results with our data.

and a piece rate scheme. Second, it allows us to compare performance between a preferred and imposed payment scheme. Since the treatments in the second and third phase are identical, this setup can be used to study the effects of imposed and preferred sorting patterns as a response to different payment schemes in a uniform and comprehensive framework.

Our results reveal that individuals sort according to productivity, with the most (least) productive ones sorting into the piece rate (fixed payment) scheme. In addition, output levels are significantly higher under the piece rate regime when individuals are free to choose the payment scheme they prefer. We do not observe order effects. This implies that it does not matter if individuals have to perform under the imposed treatment in the second or third phase. Our most important result is that cumulative output levels do not significantly differ across the two treatments. Whether or not individuals are able to choose their mode of payment does on average not yield different levels of output. Average output levels of those who prefer a piece rate are similar under both regimes and are statistically significantly higher than for those who prefer a fixed payment scheme. Imposing a piece rate on individuals who prefer a fixed payment scheme does not yield an increase in productivity. This suggests that changing the mode of payment does not seem to significantly change productivity, at least not in the short run. The vast majority has chosen the payment mode that maximizes their earnings.

This latter result is corroborated when we increase or decrease the amount that can be earned in the fixed payment scheme. If we lower the amount, which means that the piece rate becomes more attractive, only low-productive subjects select into the fixed payment scheme. If we double the amount, only the very productive types select into the piece rate. The sorting patterns show that subjects are aware of their productivity levels and sort accordingly. We do not find a significant relationship between risk preferences and the sorting behavior. Women seem to be less likely to select themselves in a piece rate payment scheme. If we compare people who are at the margin of selecting themselves in a piece rate or a fixed payment productivity and gender are the major determinants. Women at the productivity margin are less likely to sort into the piece rate. Our results remain robust if we control for the type of university of our subjects.

Finally, we find that stress and effort levels are significantly higher when individuals perform under a piece rate regime. This holds to a lesser extent for levels of exhaustion. More productive workers seem to experience lower levels of stress. These estimated coefficients do not seem to be influenced by differences in

preferences and treatments (or combinations of the two).

This research is related to the literature on performance under different payment modes in a laboratory setting. The literature is silent about the effects of exogenous changes in payment modes on performance in real effort tasks. Bull et al. (1987) investigate (in an experiment with hypothetical effort levels) the effect of an exogenous selection in either a piece rate scheme or a rank-order tournament payment. They find that the variance in effort in the tournament is higher than in the piece rate scheme. Eriksson et al. (2009) provide complementary evidence to this paper in that they find that the between subject variance of effort provision decreases if individuals have the choice between selecting into a piece rate scheme or into a tournament. They argue that the sorting decision enhances efficiency since it decreases the heterogeneity of the contestants in the tournament. Eriksson and Villeval (2008) analyze the effect of performance pay on incentives and sorting in a laboratory experiment with hypothetical effort levels. They find a positive effect of performance pay on hypothetical provided effort levels. In these three studies subjects did not work on a real effort task but chose effort levels according to a hypothetical cost function. Dohmen and Falk (2011) investigate sorting behavior. Their evidence from a real effort experiment suggests that productivity is the driving factor of selection into a variable payment scheme. However, especially for those individuals who are at the margin with regard to their productivity between sorting into a variable or a fixed payment scheme they find that more risk averse people tend to select the fixed payment. The study which comes closest to our paper is the paper by Cadsby et al. (2007). They conduct a real effort experiment and exogenously sort subjects into a variable payment scheme. In contrast to our study they find a positive effect on performance of an exogenously selected variable payment scheme. Moreover, they do not elicit stress, effort and exhaustion.

Our estimates reveal that an experimental setup is useful to analyze the effects of sorting in a controlled way. The design is in line with Lazear et al. (2012). They impose dictator games or let people sort into dictator games. The main finding is that giving subjects the opportunity to sort out of sharing environments significantly reduces sharing. We add to this literature by providing new evidence on imposing payment schemes on subjects and comparing it to the situation in which they have the opportunity to sort. In a recent paper Banuri and Keefer (2013) investigate differences between public sector workers and private sector workers in India. They find that workers tend to select into the public sector because they are more pro-socially motivated but not less productive. This evidence is supported

by Dur and Zoutenbier (2012) who find that public sector employees who work in caring industries are more altruistic than private sector employees. We add to these findings by investigating what preferences matter for sorting decisions in the lab. Evidence from the field shows that performance dependent incentive schemes can enhance output. Lazear (2000) for instance finds that a change from fixed payment to a piece rate in an auto glass company increases output per worker by 44 percent. However, half of the effect is due to the incentives and half of the effect is driven by sorting effects. Lavy (2002) finds that performance dependent pay of teachers increases students' attainment in Israel. Woessmann (2011) finds higher student attainments in OECD countries where teacher compensation is partly performance dependent. However, it is not clear which potential hidden costs come with such incentive contracts. Besides a different composition of the work force due to sorting effects psychological costs due to the performance dependent payment can arise (see for instance Larkin et al. (2012)). Another important dimension of salient incentive pay is the potential crowding out effect of intrinsic motivation (Bénabou and Tirole (2003); Frey and Oberholzer-Gee (1997)). With our study we are able to analyze the incentive effects of performance dependent payment since we exogenously change the incentive environment. Moreover, we are able to make statements about the potential psychological costs of changes in payment scheme since we measure stress, effort and exhaustion levels.

This paper proceeds as follows. Section 5.2 presents the experimental design. Section 5.4 shows the results and Section 5.5 concludes.

5.2 Approach

The study of measuring changes in performance or productivity as a result of exogenously changing the payment mode is tough when using labor-market data. There are a number of reasons for this challenge. First, policy changes with regard to changing payment schemes need time to affect the composition of the workforce in an occupation or sector of industry. This makes it hard to draw direct causal conclusions about whether or not the change in payment mode or other factors have changed outcomes. In the short run, labor markets are to some extent rigid, which makes it hard for workers to become mobile. In addition, mobility can be unattractive for workers who have invested in firm-specific human capital, who have a short horizon to capture the returns from investing in new skills or in a new job and for workers who are well-protected by legal institutions, such as

employment protection legislation. This could be one reason why studies that investigate changes in payment modes hardly find any effect on outcomes relatively shortly after its introduction (e.g., Fryer (2013)). The second challenge is that incentives in the workplace are often hard to measure. The reason is that workers are exposed to many different kinds of incentives, which all affect their output in both positive and negative ways. Third, individual output is hard to measure and is often judged subjectively by the firm's management. Finally, changes in payment schemes need to be implemented exogenously to study their effects on output and other behaviors. This is debatable because firms do not seem to be able to change payment schemes overnight. Yet, from a societal or firm's policy point of view it is important to understand how people respond to changes in payment schemes. We believe that our laboratory experiment offers a tool to complement observational data. First, in the lab we are able to study the immediate impact on output of changing payment schemes when workers are immobile and forced to perform under different regimes. This approach is similar to the one used in Lazear et al. (2012) to study if subjects who are reluctant to share avoid situations, or are even prepared to pay for avoiding these circumstances, when they are able to. Second, we are able to define the incentives upon which subjects can base their sorting decision. These decisions are observed immediately and can be confronted with the outcomes of a situation in which subjects are forced to perform under an imposed payment regime. Third, we measure output from a simple work task, which yields objective and well-defined measures of productivity. Finally, the laboratory allows us to change the specific circumstances with respect to the payment mode only. This makes it easier to study the effects of exogenously implemented changes in incentives on the performance of the same population.

5.2.1 Design

Figure 5.1 shows an overview of the experimental procedure. We discuss the most salient features of the experiment in this section. The instructions, a number of screenshots and additional information are presented in the Appendix.

Work task

The real effort task consists of multiplying a two-digit with a one digit number. Following Dohmen and Falk (2011), there are five different degrees of difficulty.[3]

[3]Note that we created new problem sets based on their definition of the degrees of difficulty.

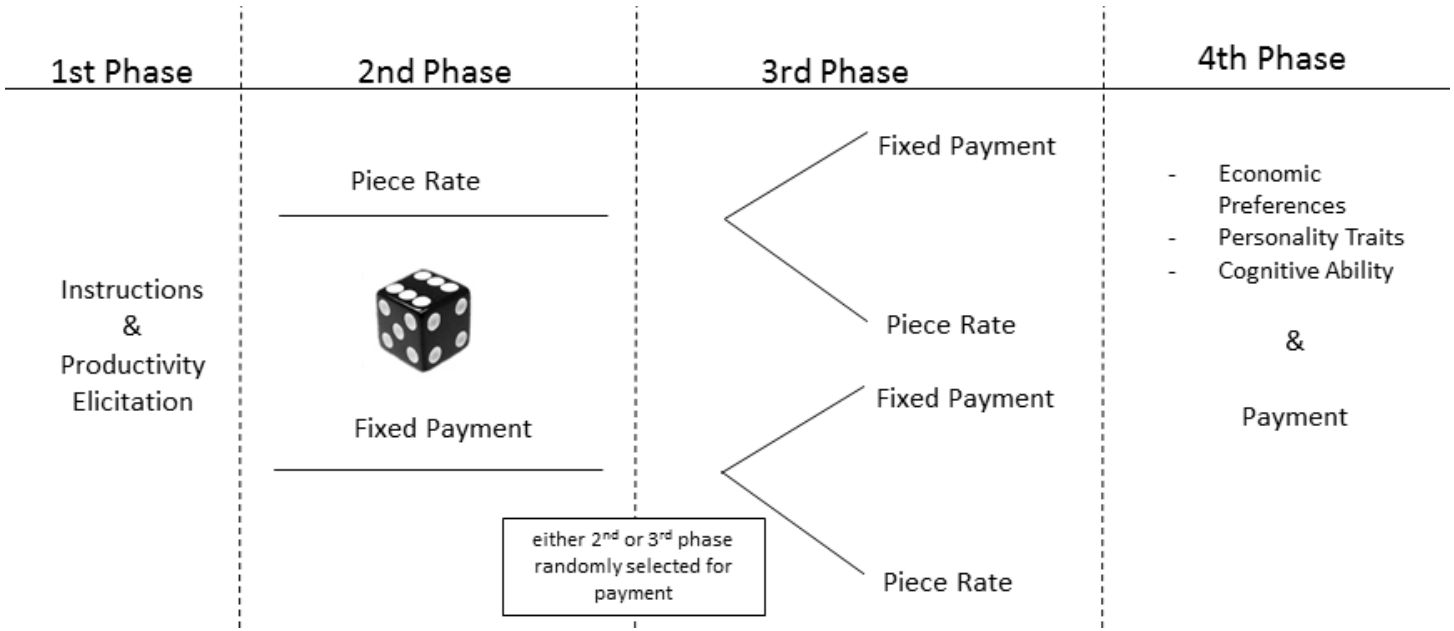


Figure 5.1: Experimental Design

Note. The order of the 2nd and 3rd phase was randomized within each session.

The advantage of this task is that the task is easy to explain to subjects, that output is straightforward to measure and that learning effects during the experiment are absent or only very small. The nature of the task requires effort, which implies that subjects have to work and that they are uncertain about the reward in the piece rate regime. This feature of the experiment makes it more likely to subjects sort into the payment regime they prefer and that they relatively dislike performing under the other regime. The experiment of Dohmen and Falk (2011) uses the same real effort task and other real effort experiments also apply relatively easy to measure tasks that require effort. One disadvantage of the task is that some subjects, or groups of subjects with particular background characteristics, have distaste for this particular work task. This would induce them to sort into fixed payment regime and exert as little effort as possible. We only observe differences with respect to gender. Women are statistically significantly less likely to sort into the piece rate compared to men. This could be because they have distaste for the task, but also because they shy away from competition (Niederle and Vesterlund (2007); Buser et al. (2014)) or because they are less productive in this task (which could be a result of distaste, see (Gneezy and Rustichini (2004)). We do not find a relationship between an individuals' risk attitude and the sorting behavior. This is consistent Dohmen and Falk (2011). The multiplication problems appeared on a computer screen and subjects had to fill in the answer to the problem. After they entered the solution they had to press the OK-button. When they solved the problem correctly the next problem appeared. When they failed to find the correct solution, they had to try again until they found the correct answer. During the different phases of the experiment subjects were aware of their cumulative score. In Section D of the Appendix a screenshot is presented.

Setup

The experiment consisted of four phases, which by and large follows the experiment conducted by Dohmen and Falk (2011). We differ in the sorting pattern of the experiment by distinguishing an imposed payment scheme and a voluntarily chosen payment regime. Hence, Phase 1 and 3 are identical to Dohmen and Falk (2011). In Phase 2 and 4 we adjust their design to answer our research question. We gave detailed instructions about the real effort task and elicited the productivity of each individual on the multiplication task. Subjects were informed about the fact that the computer randomly selects a payment scheme in the second phase. Subjects were randomly selected into either a piece rate or a fixed payment regime. We

call this treatment the EXO treatment. In the next phase they could choose the payment scheme themselves. We call this treatment the ENDO treatment. These two phases lasted for 10 minutes. Individuals were informed about the fact that the computer randomly selects one treatment which will be paid out. We flipped the order of the two treatments within a session to control for order effects in output. In the last phase subjects answered a detailed questionnaire and afterwards they were paid.

Phase 1

We introduced the real effort task. Subjects had to multiply a one-digit with a two-digit number. There were five different levels of difficulty (see Figure 5.4). We explained in detail what the work task is about and how a subject could submit the answer to each problem set. Next, subjects had the possibility to learn their productivity in three steps. First, they were asked to multiply numbers as fast as possible without receiving a payment. Second, they were asked to multiply numbers as fast as possible with a monetary incentive to do so. Third, they were asked to multiply as many numbers as possible for a period of five minutes. Afterwards we elicited stress, effort and exhaustion with three simple questions. At this point individuals know their productivity level.

Phase 2

Subjects were randomly assigned to a piece rate or a fixed payment scheme. We call this part the EXO treatment. They were informed that the computer would either select a fixed payment scheme or a piece rate for the performance remuneration of the following 10 minutes. All subjects were informed that the fixed payment equals 400 points and that in the piece rate mode each correctly solved problems yields 10 points. The break-even point of 400 points at which risk-neutral subjects should be indifferent between payment schemes is reached when they correctly solve 40 multiplication problems in 10 minutes. The exchange rate from points to Euros is the following: 10 points equals 0.02 Euro.[4]

If subjects were selected in the fixed payment scheme they received a message on the computer screen which said The computer chose the fixed payment for you. You will receive 400 points independent of the number of problems that you solve correctly (so regardless of whether you solve 0, 17 or 152 problems). If subjects were selected in the piece rate they received the following message on the computer screen: The computer chose the variable payment for you. You will

[4]We adapted the exchange rate of Dohmen and Falk (2011) taking inflation and purchasing power in the Netherlands into account.

receive 10 points for each problem that you solve correctly during the 10 minutes of time.

Before subjects started with the working phase we asked them if they would have chosen the selected payment scheme themselves if they had been allowed to make their own choice. Afterwards we confronted them with hypothetical choice lists as in Dohmen and Falk (2011) conditional on what they had chosen. If they were selected in the piece rate and indicated yes we confronted them with a choice list which contained an increasing fixed payment with increments of 50 starting from 450 and going up to 800 points. If they were selected in the piece rate and answered the question if this was also their preferred payment with no we confronted them with a choice list with a decreasing fixed payment starting at 350 points and going down to 0 in increments of 50. In case a subject was selected in the fixed payment scheme the respective choice list was displayed depending on what a subject answered. Afterwards the working phase started. After the working phase we elicited stress, effort and exhaustion in the same way as before.

Phase 3

Subjects now had the freedom to sort into a fixed payment scheme or a piece rate regime. We call this phase the ENDO treatment. The payment modes are the same as in the EXO treatment. Before the working phase started we confronted subjects with the same choice lists as in the EXO treatment depending on their selected payment mode. When the 10 minutes were over, we again elicited levels of stress, effort and exhaustion using three simple items.

Phase 4

We elicited a battery of preferences, as well as cognitive ability. Figure 5.2 shows the distribution of these measures. First, subjects had to solve 15 Raven matrices (Raven (1962)) to measure their cognitive ability. We paid 15 points for each correctly solved matrix. The distribution of correctly selected answers is shown in Panel A of Figure 5.2.[5] Next, we elicited a battery of preferences in a non-incentivized way. Our questionnaire includes experimentally validated measures for economic preference parameters (Becker et al. (2012)). All three questions are taken from their study. Panels B to D show the distributions of the respective answers to these questions. We elicited an individual's general risk attitude by asking the question "In general, are you a person who is willing to take risks or do you try to avoid risks?" (see also Dohmen et al. (2011)). Subjects had

[5]In one of the sessions we experienced network problems during the IQ test. We excluded these observations from this analysis.

to give an answer on a scale from 0 to 10. A zero indicated "completely unwilling to take risks" and a ten indicated "very willing to take risks". Trust was measured with one item question asking "How would you assess your willingness to trust strangers?". Higher values indicated a higher willingness to trust. Reciprocity was measured using the following question. "Please think about what you would do in the following situation. You are in an area you are not familiar with, and you realize that you lost your way. You ask a stranger for directions. The stranger offers to take you to your destination. Helping you costs the stranger about 20 Euro in total. However, the stranger says he or she does not want any money from you. You have 6 presents with you. The cheapest present costs 5 Euro, the most expensive one costs 30 Euro. Do you give one of the presents to the stranger as a "thank you"-gift? If so, which present do you give to the stranger?" Higher values on these question indicate a more expensive gift. At the end subjects received their payoff and left the lab. In a last step we measured an individual's personality using a 50 item Big 5 questionnaire (Goldberg (1992)).[6]

Subjects

The number of participants equals $n = 165$. The subject pool consists of students from Maastricht University ($n = 93$) and Hogeschool Zuyd ($n = 72$), a higher vocational education school which educates young people from the region of Maastricht, the Netherlands. All analyses include a dummy variable to take care of possible differences across these two populations. A total number of 13 sessions took place at both locations. We invited subjects from Maastricht University by making use of the recruitment software ORSEE (Greiner (2003)). At the Hogeschool Zuyd we send emails through the official university scheduling and communication department. Afterwards subjects could register online. To run the experiment we made use of zTree (Fischbacher (2007)). The experiment lasted approximately 90 minutes. The average payoff was €27.28 (S.D. €8.37) for the Maastricht University students and €25.72 (S.D. €7) for the Hogeschool Zuyd students.[7]

[6] We used the New IPIP 50 item scale which can be found at: `http://ipip.ori.org/New_IPIP-50-item-scale.htm` (last access on September 1^{st} 2014).

[7] The payoffs are not statistically significant ($p = .2063$, t-test).

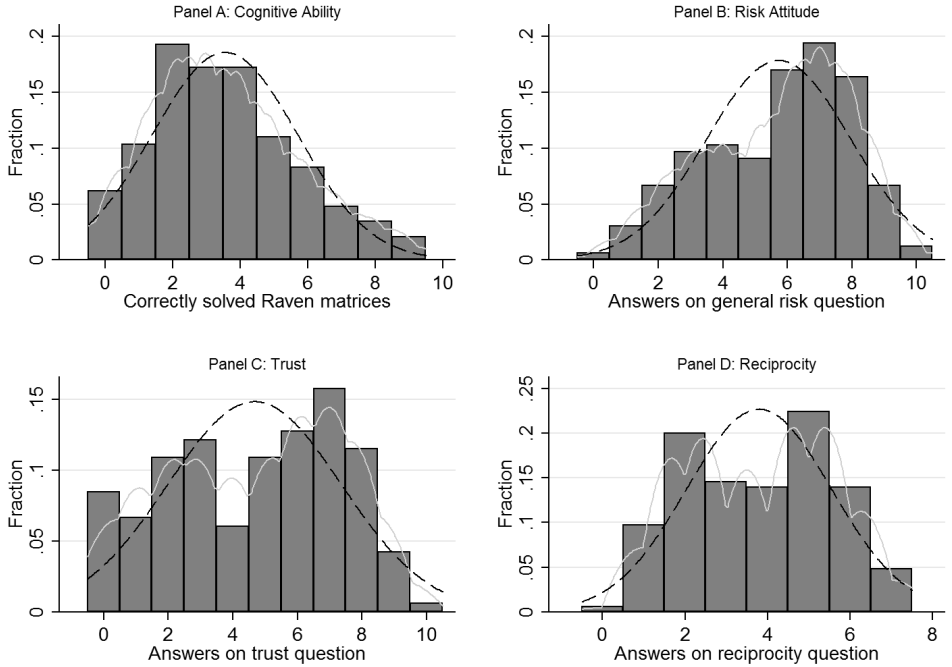


Figure 5.2: Distributions of the preference measures.

Note. Panel A shows the distribution of correctly solved Raven matrices. Panel B shows the answers on a scale from 0 to 10 (higher values indicate again stronger preferences for risk) to the question In general, are you a person who is willing to take risks or do you try to avoid risks? Panel C shows the distribution of answers on scale from 0 to 10 to the question How would you assess your willingness to trust strangers? Higher values indicate a higher willingness to trust. Panel D shows the answers to the question Please think about what you would do in the following situation. You are in an area you are not familiar with, and you realize that you lost your way. You ask a stranger for directions. The stranger offers to take you to your destination. Helping you costs the stranger about 20 Euro in total. However, the stranger says he or she does not want any money from you. You have 6 presents with you. The cheapest present costs 5 Euro, the most expensive one costs 30 Euro. Do you give one of the presents to the stranger as a "thank you"-gift? If so, which present do you give to the stranger? The numbers correspond to the value of the present. The higher the number the more expensive the present. The straight gray line show kernel density estimates of the corresponding distribution. The black dashed line shows the normal distribution.

Descriptive Statistics

Table 5.1 provides raw correlations of the variables we use in the analysis. The picture that emerges from this table is that productivity in the work task is positively correlated with cognitive ability. Women and students from the University of Applied Sciences seem to have a lower score on the test for cognitive ability. The willingness to trust is positively correlated with cognitive ability. Our measure of reciprocity is significantly positively correlated with the willingness to take risks and the willingness to trust. Students from the University of Applied Sciences seem to be less willing to take risk. Moreover, the University of Applied Sciences seems to have a higher proportion of women. We also add the standardized measures of stress, effort and exhaustion after the productivity elicitation phase. First these measures are highly and significantly correlated. Second, more productive subjects seem to report lower stress, effort and exhaustion levels. Third students from the University of Applied Sciences seem to report lower stress, effort and exhaustion levels.

5.3 Productivity and preferences

In this section we document results about differences in preferences and productivity between subjects. We do so by using the results from the ENDO treatment in which subjects are able to make their own sorting choices. We expect a positive output effect of the piece rate regime because more productive types are more likely to select into a piece rate regime.

5.3.1 Sorting

We define two groups: those who prefer a fixed payment ($n = 70$) and those subjects who prefer a piece rate ($n = 95$). Panel A of Figure 5.3 documents the cumulative distribution function of all subjects in phase 1. This phase lasted for 5 minutes, in which 10 points could be earned for a correct answer to a multiplication problem. The break-even productivity at which risk-neutral subjects should be indifferent between choosing a fixed payment or a piece rate in the third phase equals 20 correctly solved problems, which is indicated by a vertical line in Panel A of Figure 5.3. In Panel B of Figure 5.3 the gray squares (black triangles) indicate the productivity of subjects who chose the fixed payment (piece rate) later on in the ENDO treatment. The pattern that becomes apparent from the figure suggests

Table 5.1: Correlation between variables of interest

	Productivity	Cognitive Ability	Risk Attitude	Trust	Reciprocity	Female	Applied sciences	Stress	Effort
Cognitive Ability	0.1993**								
Risk Attitude	-0.0159	-0.0465							
Trust	-0.1165	0.1812**	0.1272						
Reciprocity	-0.0191	0.001	0.2605***	0.1778**					
Female	-0.1073	-0.1950**	-0.0241	-0.131	-0.0783				
Applied sciences	-0.0212	-0.2367***	-0.3244***	-0.0599	-0.1026	0.1434*			
Stress	-0.3778***	0.0141	0.0305	0.0392	0.0264	0.059	-0.2474***		
Effort	-0.1735**	0.08	0.216***	0.1307	0.0697	-0.0411	-0.4119***	0.3992***	
Exhaustion	-0.1458*	0.0094	0.0695	0.0008	-0.0357	-0.0504	-0.224***	0.491***	0.3777***

Note. Correlations. Risk attitude, trust, reciprocity, stress effort and exhaustion are standardized with mean zero and s.d. one. Stress, effort and exhaustion levels are taken from the productivity elicitation phase. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

that subjects who select themselves into the piece rate mode later on are more productive in the elicitation phase than those who select themselves into the fixed payment scheme. The difference is statistically significant (Wilcoxon rank-sum test, p-value$<$ 0.0001).

Panel C of Figure 5.3 shows the cumulative distribution function of the piece rate and the fixed payment in the actual ENDO treatment. Since this phase lasted for 10 minutes, break-even productivity equals 40 (for risk-neutral individuals). The picture that emerges from this figure is similar to the one from the productivity elicitation phase in Panel B. Overall output is higher in the piece rate regime compared to the fixed payment scheme. A Wilcoxon rank-sum test reveals, that the difference is highly statistically significant (p$<$ .0001).

5.3.2 Productivity differences

The fact that more productive types sort into the piece rate regime does not come as a surprise. It is interesting to analyze the productivity differences further by distinguishing between the period of time subjects need to solve problems of different degrees of difficulty. This sheds light on the fact whether or not productivity differences could be related to differences in cognitive ability.

Figure 5.4 shows the average calculation times split up by the five degrees of difficulty and by payment scheme in the ENDO treatment. First, independent of the payment scheme, subjects need more time to calculate the correct solution for more difficult questions. This is visible from the bars in each panel with the darker bars being longer than the lighter bars. Second, there are strong differences in the calculation times between subjects in the piece rate and the fixed payment scheme. Subjects who sort into the piece rate regime solve the multiplication problems always faster than subjects who work under the fixed payment. All differences are statistically significant (t-test, all p-values$<$ 0.001). Finally, subjects who sort into the fixed payment scheme solve fewer multiplication problems (on average almost 15) in the productivity elicitation phase compared to those who sort into the piece rate.

Table 5.2 documents means and standard deviations of differences in observable characteristics. The top (middle) panel of Table 5.2 shows summary statistics for those who prefer a fixed payment (piece rate). The bottom panel documents the differences between the two groups and the p-value of a Wilcoxon rank-sum test. Next to the differences in average productivity that are borne out in Figure 5.4, subjects who sort into the piece rate regime obtain higher scores on the cognitive

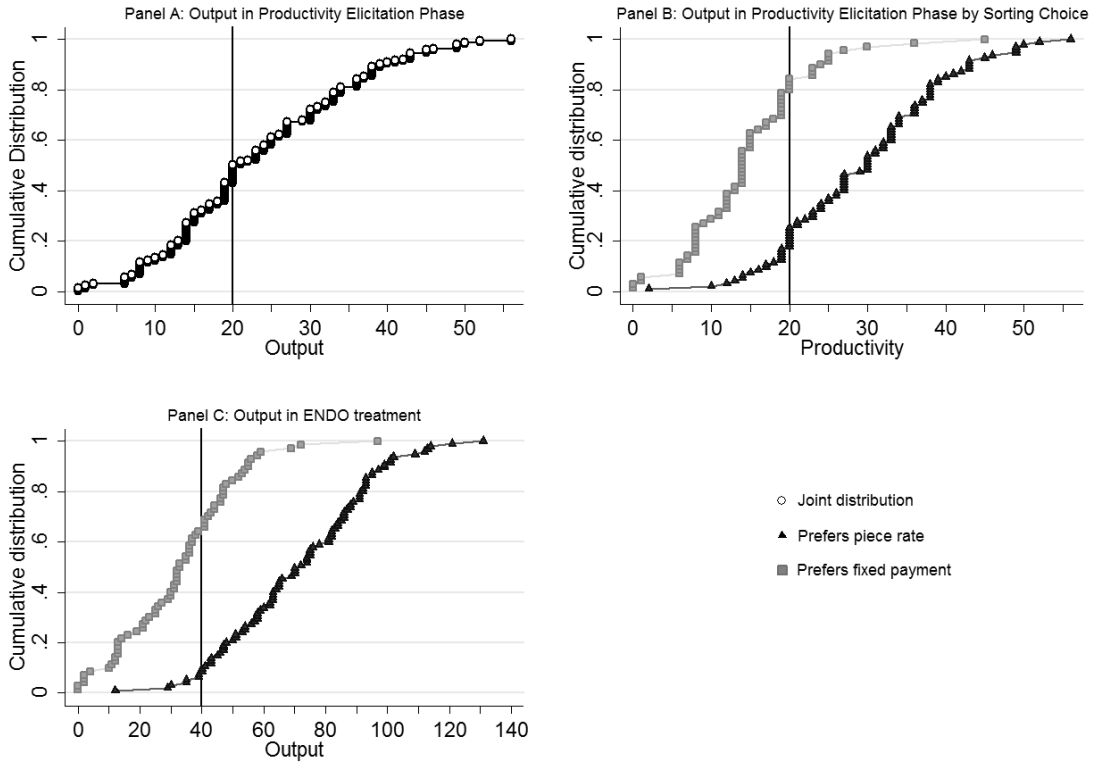


Figure 5.3: Cumulative distribution of the output in the productivity elicitation phase and the ENDO treatment.

Note. Panel A shows the cumulative distribution function of the output in the productivity elicitation phase. In Panel B we split the output of the productivity elicitation phase into two groups. In the first group are those subjects who chose a piece rate in the ENDO treatment and in the second groups are those subjects who chose a fixed payment in the ENDO treatment. Panel C shows the equivalent figure of the cumulative output in the ENDO treatment split by payment scheme choice.

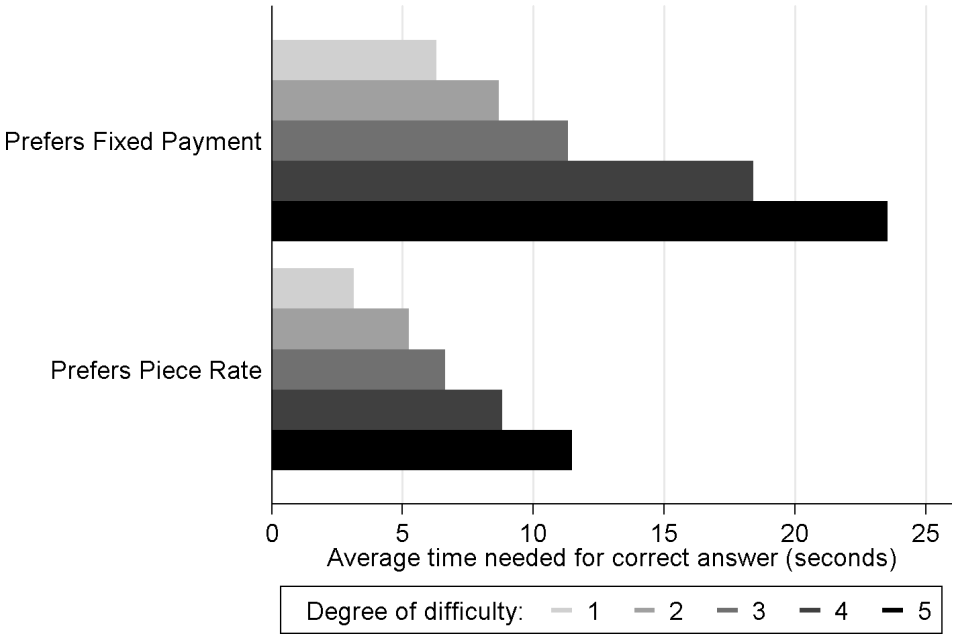


Figure 5.4: Calculation times by degrees of difficulty and payment scheme in the ENDO treatment.

Note. The figure shows the average calculation time needed for each of the five degrees of difficulty for the subjects who select the fixed payment and for those subjects who select the piece rate separately in the ENDO treatment. This figure is based on Figure 2 in Dohmen and Falk (2011).

test they have to take. Subjects who sort into a piece rate solve more Raven matrices (p=0.040). Also, significantly more women sort into the fixed payment scheme ($p = 0.093$). Finally, stress and effort levels are higher in the piece rate regime. Performing in a piece rate regime requires more effort because subjects are paid for performance. Apparently the subjects also face higher stress levels. Since these sorting patterns are based on voluntary choices, these stress levels seem to be taken into account by the people who sorted into this payment mode. In terms of unconditional differences between the two groups, we do not observe any statistically significant differences across the other variables we have measured.

Table 5.2: Differences in observables between preferences for fixed payment and for variable payment

Subject prefers		Product.	Cognitive Ability	Risk Attitude	Trust	Reciprocity	Female	Applied Sciences	Stress	Effort	Exhaustion
	Mean	14.757	3.085	0.058	4.786	3.657	0.657	0.471	-0.257	-0.175	-0.122
Fixed	S.D.	7.941	1.869	1.026	2.53	1.825	0.478	0.503	1.035	1.018	0.984
payment	N	70	59	70	70	70	70	70	70	70	70
	Mean	29.632	3.86	-0.043	4.579	3.884	0.526	0.411	0.19	0.129	0.09
Piece	S.D.	10.476	2.286	0.988	2.823	1.719	0.502	0.495	0.935	0.972	1.007
rate	N	95	86	95	95	95	95	95	95	95	95
	Difference	-14.874	-0.775	0.101	0.207	-0.227	0.131	0.061	-0.447	-0.305	-0.212
	p-value	0	0.04	0.466	0.699	0.374	0.093	0.437	0.006	0.054	0.178

Note. The Table shows the differences in means between the group of subjects which prefers the fixed payment and the group which prefers the piece rate. The values of stress, effort and exhaustions are standardized values from the first period. P-values reported are obtained from a Wilcoxon ranksum test.

5.3.3 Determinants of sorting

Table 5.3 shows the determinants of the sorting decision by means of a regression analysis. The dependent variable takes the value one if the piece rate was chosen. We report marginal effects of a probit model evaluated at the mean of the independent variables. All regressions contain session fixed effects and we cluster the standard errors on the session level. Moreover, each regression model contains a dummy which controls for the treatment order and controls for the study track.

We first investigate the correlation between productivity levels from the elicitation phase and sorting into the piece rate (column (1)). Because gender seems to influence the sorting into the piece rate (see Table 5.2), we add this covariates to the model in column (2). Column (3) shows the effects for adding risk preferences, reciprocity and trust, as controls to the equation. In column (4) we add an interaction term of gender and cognitive ability. In a last step we control for the Big 5 personality traits (column (5)). The estimation results do not change, which seems to point at the fact that personality traits do not influence the sorting decision independent of the covariates we already included in the regression models. The overall picture that emerges from the estimation results in Table 5.3 is that more productive subjects are more likely to sort into the piece rate regime. We find weak evidence that women tend to be less likely to select a piece remuneration.

Table 5.4 shows the results for subjects who are at the margin of selecting a piece rate or a fixed payment. We follow the definition of marginal types by Dohmen and Falk (2011). The dependent variable takes the value one if the piece rate was chosen in the ENDO treatment. We restrict the sample to marginal types with regard to productivity (column 1) and marginal and non-marginal types with regard to response time (column 2 and 3). It turns out that for subjects who are at the margin reciprocity and gender seem to significantly influence their decision to select the piece rate. If women are at the margin of selecting the piece rate they are less likely to choose it. Whereas more reciprocal subjects tend to be more likely to choose the piece rate if they are at the margin.

5.3.4 Productivity sorting

The sorting decision made by subjects in the ENDO treatment is based on a threshold of solving 40 multiplication problems, which equals a score of 400 points. When we change the attractiveness of the fixed payment regime, the average productivity of those who sort into the piece should change. We therefore further investigate

Table 5.3: Determinants of choosing a piece rate remuneration

	(1)	(2)	(3)	(4)	(5)
Productivity	0.0375***	0.0371***	0.0383***	0.0386***	0.0388***
	(.0072)	(.0076)	(.0078)	(.0086)	(.0093)
1 if Female		-0.1388*	-0.1257	-0.0572	-0.0791
		(.0765)	(.0844)	(.0768)	(.0597)
Risk Attitude			-0.0454	-0.0467	-0.0584
			(.0340)	(.0492)	(.0442)
Trust			0.0102	0.0079	0.0107
			(.0125)	(.0141)	(.0135)
Reciprocity			0.0264	0.0536	0.054
			(.0314)	(.0333)	(.0375)
Cognitive Ability				0.0128	0.0166
				(.0191)	(.0161)
Observations	165	165	165	145	145
Controls	YES	YES	YES	YES	YES
Big Five	NO	NO	NO	NO	YES

Note: The table shows marginal effects of a probit estimation evaluated at the mean. The dependent variable takes the value one if a piece rate was chosen. All regressions contain session dummies as controls, controls for the treatment order and controls for the study track. Robust standard errors clustered on the session level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 5.4: Sorting decision for subjects at the margin

	(1)	(2)	(3)
	Marginal type (productivity)	Marginal type (response time)	Non-marginal type (response-time)
Productivity	0.046***	0.039***	0.035***
	(0.007)	(0.008)	(0.006)
Risk	-0.023	-0.004	-0.019
	(0.022)	(0.023)	(0.023)
Trust	-0.009	-0.014	0.024
	(0.017)	(0.022)	(0.033)
Reciprocity	0.053*	0.027	0.029
	(0.031)	(0.043)	(0.033)
1 if Female	-0.049	-0.253***	0.104
	(0.080)	(0.079)	(0.089)
Observations	99	81	84
Pseudo R-squared	0.186	0.296	0.445

Note. Probit regression with marginal effects evaluated at the mean of the independent variable. The dependent variable takes the value 1 if the piece rate was chosen. Column (1) defines the marginal type according to his or her productivity. Subjects who solved between 10 and 30 problems in the productivity elicitation phase are defined as marginal types. Columns (2) and (3) define a marginal type according to the response time when to decide for a payment scheme. Subjects with an above median response time were defined as marginal types. The table is based on Dohmen and Falk (2011).

the role of productivity to sort into one of the two payment modes.

We elicited individuals' preferences to sort into the piece rate for different levels of the fixed payment. A risk neutral utility maximizer would be indifferent between choosing a piece rate and a fixed payment as soon as the payoff in the piece rate mode equals the payoff in the fixed payment regime. Using the information from the productivity elicitation phase of 5 minutes, subjects are able to calculate how many problems they would be able to solve in 10 minutes. A rational risk neutral worker should be indifferent between a piece rate and an offered fixed payment as soon as productivity exceeds the fixed payment divided by 20.[8] This is what we observe in the figures above with many points around 20 and 40 (Figure 5.3). This suggests that for low levels of the fixed payment only relatively unproductive workers prefer the fixed payment, whereas relatively productive workers

[8]Note that the piece rate is 10 points per correctly solved answer. Since the productivity elicitation phase lasts 5 minutes and the actual working phase lasts 10 minutes the worker simply has to compare her productivity multiplied by 20 and check if she would reach the offered fixed payment.

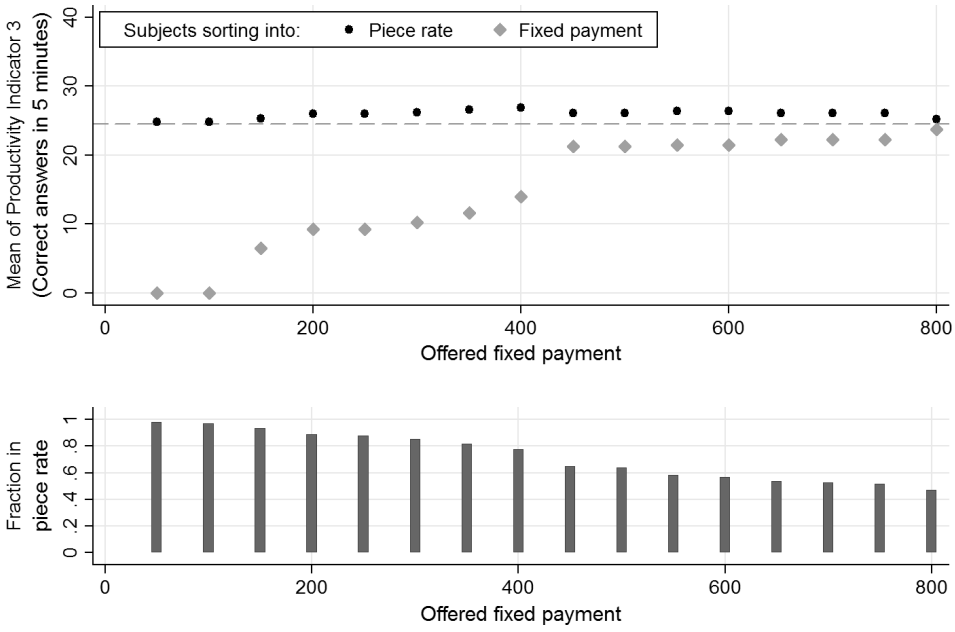


Figure 5.5: Sorting choices depending on productivity and offered fixed payment.

Note. The top panel of this figure shows the productivity of the subjects and their sorting decision for given levels of a hypothetically offered fixed payment. The gray diamonds indicate the average productivity in the productivity elicitation phase of those subjects who select the fixed payment for a given offered fixed payment. The black dots indicate the average productivity of those subjects choosing the piece rate for a given offered fixed payment. The dashed line indicates the overall average productivity of our sample. The bottom panel of the figure shows the corresponding overall fraction of subject choosing the piece rate for a given offered fixed payment. We created this figure based on figure 4 in Dohmen and Falk (2011).

prefer the piece rate. A higher fixed payment would induce subjects with higher levels of productivity to sort endogenously into the fixed payment regime. Figure 5.5 shows the results from the hypothetical choice lists in the ENDO treatment. The dashed gray line plots the average productivity of all workers from the productivity elicitation phase. The top panel plots the average productivity of those who actually either prefer the fixed payment (gray diamonds) or the piece rate (black dots) for a given level of the fixed payment. The panel at the bottom shows the fraction of subjects who prefer to be remunerated according to a piece rate for a given level of the fixed payment. Up to a fixed payment of 200 points more than 80 percent of the subjects would have liked to sort into the piece rate regime. If the fixed payment is 800 points 51.1 percent of the subjects would sort into a fixed payment regime. This pattern is in line with the actual distribution of the output

in the 10 minutes working phase. 41.1 percent of our subjects solve more than 80 multiplication problems. Hence, it is reasonable to sort into the piece rate regime if the fixed payment would be equal to 800 points.[9]

5.4 Exogenously changing payment regimes

Our core result concerns performance differences between exogenous and endogenous sorting in payment schemes. We have obtained a set of estimates suggesting that more productive types are more likely to sort into the piece rate regime. Now, we explore cases in which we force subjects to work in regimes that are not necessarily in line with their preferences.

5.4.1 Productivity changes

To start off, we document cumulative distribution functions which are similar to the ones in Figure 5.3 but now for different imposed treatments. Figure 5.6 shows cumulative distribution functions of the output in the two different payment schemes in the EXO treatment. The left panel shows the performance of all subjects in the fixed payment scheme. The black triangles indicate the cumulative distribution under the fixed payment scheme in the EXO treatment for subjects who sort into the piece rate in the ENDO treatment.[10] The gray squares indicate the cumulative distribution function of subjects who actually sort into the fixed payment in the ENDO treatment and are selected into the fixed payment scheme in the EXO treatment. The graph shows that subjects who prefer to work in the piece rate mode are more productive than those who prefer a fixed payment even if the remuneration is independent of performance. A Wilcoxon ranksum test reveals that this difference is statistically significant (p-value < 0.01). The right panel of Figure 5.6 shows the corresponding situation for the piece rate payment scheme in the EXO treatment. The black triangles indicate the performance of subjects who sort into a piece rate compensation in the ENDO treatment and the gray squares show the performance of subjects who sort into a fixed payment in the ENDO treatment. The picture that emerges from this panel indicates that there are strong differences in performance in the piece rate regime too. Subjects who prefer to work in a fixed payment regime perform worse than those who prefer to

[9] We elicited these preferences as well in the EXO treatment. The choices look exactly the as in Figure 5.5. The equivalent figure is available in the Appendix.

[10] All results do not change if we take the hypothetical choice in the EXO treatment.

work in a piece rate regime. The difference between the distributions is statistically significant (Wilcoxon ranksum test, p-value$<$ 0.001). The two graphs in Figure 5.6 display a vertical line at a level of output equal to 40. This is the number of multiplication problems subjects had to solve in the piece rate regime to break even with the earnings from the fixed payment mode. In the piece rate payment scheme all, except for one individual, manage to pass the threshold above which it pays to sort into the piece rate regime. Also in the fixed payment scheme most individuals who prefer to work in a piece rate scheme, solve more problems than is necessary to meet this threshold.

Figure 5.7 documents similar results from a different perspective. Instead of splitting the sample into performance in different payment schemes we split the sample into subjects who prefer to work in a piece rate regime and subjects who prefer a fixed payment scheme.[11] This way we compare the cumulative distribution functions of workers with the same preferences in different incentive schemes. The left (right) panel of Figure 5.7 compares the performance of subjects who prefer a piece rate (fixed payment) in both regimes. Interestingly, overall output is only slightly higher (four units) in the piece rate than in the fixed payment in the left panel of Figure 5.7. A Wilcoxon ranksum test reveals no statistically significant difference ($p = 0.556$). The right panel of Figure 5.7 shows the same picture for those who prefer a fixed payment. The picture that emerges from this panel is that subjects who prefer a fixed payment but are selected into a piece rate regime perform only slightly better compared to the fixed payment (1.5 units on average). However, the difference is not statistically significant (Wilcoxon ranksum test, $p = 0.874$) and seems to hold for low-productive types only. Most of those individuals who sort into the fixed payment scheme in the ENDO treatment do not solve more than 40 problems correctly if they are forced to work in a piece rate regime.

These results are interesting for various reasons. First Figure 5.6 shows that imposing a payment scheme does on average neither increase nor decrease the performance of our subjects. Second, as Figure 5.7 reveals, subjects who prefer to work in a piece rate scheme do on average not decrease their performance in a fixed payment scheme.

[11] Again we took the choice of the payment scheme in the ENDO treatment. Our results do not change if we take the hypothetical choices in the EXO treatment.

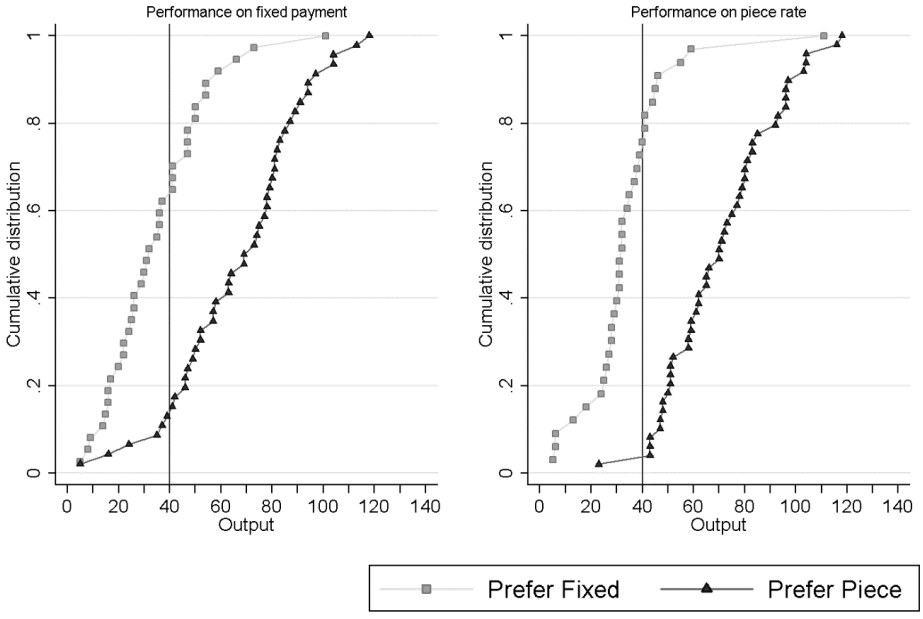


Figure 5.6: Performance in EXO treatment by payment schemes.

5.4.2 Productivity differences

Next to the finding that more productive types are more likely to sort into piece rate schemes, we have obtained a finding that productivity does not seem to change when subjects are sorted into a payment scheme which does not correspond with their preferences. To investigate to what extent this is related to performance on different types of questions we show analyzes in which we distinguish different levels of difficulty.[12]

Figure 5.8 shows the results for the EXO treatment. We first split the sample

[12]We also perform a randomization check of the 2^{nd} phase. The corresponding Figure D.7 can be found in the Appendix. We first distinguish between the EXO and ENDO treatment. Second, we split the sample by payment scheme. If the randomized allocation to fixed payment and piece rate in the EXO treatment worked we would have an equal balance of all productivity types in the piece rate payment scheme and the fixed payment. Thus we should not observe differences in calculation times between the fixed payment and the piece rate in the EXO treatment. This is what we find. There are no significant differences in the calculation times between fixed payment and piece rate in the EXO treatment. However, calculation times are lower in the fixed payment of the EXO treatment than in the fixed payment of the ENDO treatment. This stems from the fact that more productive workers now also work under a fixed payment. All differences are significant at the 1% level (t-test). Moreover, overall calculation times in the piece rate payment scheme of the EXO treatment are higher than in the piece rate payment scheme of the ENDO treatment. All differences are significant with a p-value$<$ 0.001. This is due to the fact that also less productive workers are exogenously selected in the piece rate payment scheme.

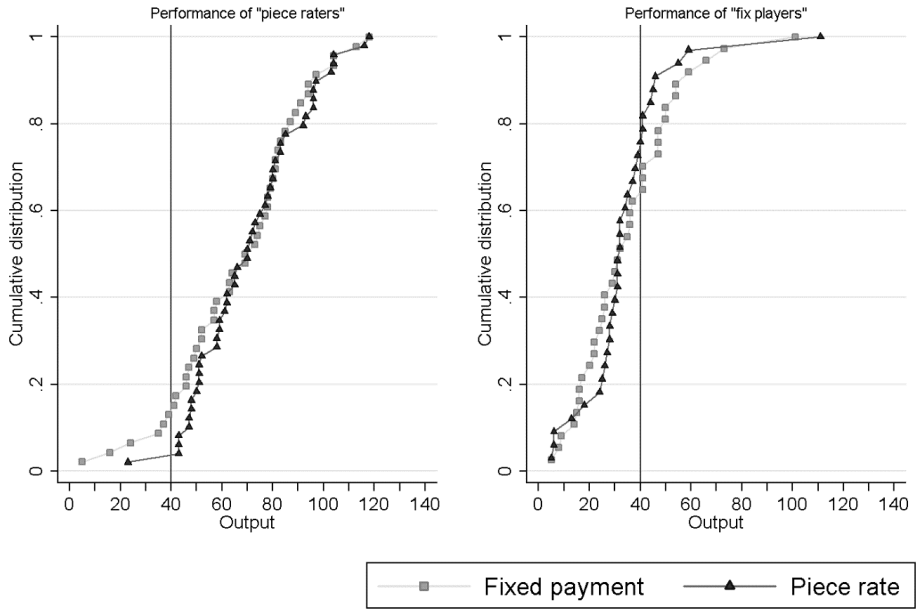


Figure 5.7: Performance in EXO treatment by preferences.

according to the preferred payment scheme in the ENDO treatment. Then we split the sample into the two different payment schemes. The picture that emerges from Figure 5.8 shows that subjects who prefer a piece rate become only slightly less productive if they are forced to work in a fixed payment regime (the difference is not statistically significant for all different levels of difficulty). However, the difference in productivity for subjects who prefer a fixed payment is remarkable. Average calculation times seem to increase if these subjects are forced to work in a piece rate scheme. Subjects who prefer a fixed compensation and who are pushed into a piece rate scheme need now more time to calculate the correct solution to the multiplication problems. Potential reasons for this result could be that subjects choke under the pressure of the performance dependent payment scheme. Except for calculation problems with difficulty 5 the differences are all significant at the 1 percent level.

5.4.3 Heterogeneity across subjects

The average differences in performance as depicted by the cumulative distribution function could mask heterogeneity across subjects. In this subsection we compare

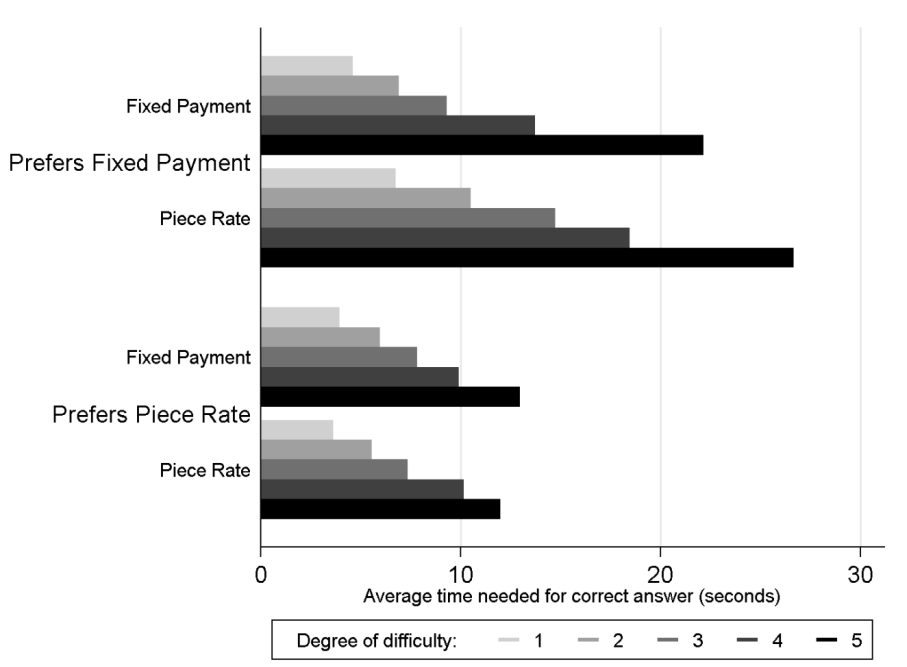


Figure 5.8: Average calculation time in the EXO treatment by preference for payment scheme.

individual performance in the EXO and the ENDO treatments. Since all subjects participated in both treatments, we are able to conduct a within subject comparison. Panels A and B in Figure 5.9 and show the analyses for different payment schemes in the EXO treatment; Panel A for the piece rate and Panel B for the fixed payment as exogenously imposed payment schemes. The overall conclusion is that subjects do not substantially change behavior when they are forced to work in a different payment scheme compared to when they can sort themselves into a payment scheme. Forcing workers in a payment scheme seems to slightly improve the performance of relatively low-performers in both regimes.

Panel A shows the relationship between the piece rate scheme in the EXO treatment and sorting into either a piece rate or fixed payment scheme in the ENDO treatment. We distinguish between subjects who sort into a piece rate (dark gray dots) and a fixed payment (light gray diamonds) scheme. We have drawn a horizontal and vertical line at 40, which mark the performance thresholds above which a piece rate scheme becomes beneficial to the subjects. It is immediately clear that only one subject who sorts into the piece rate does not meet the threshold in both the EXO and ENDO treatment. Investigation of the gray diamonds reveals that most subjects do not meet the threshold, which makes it rational (given risk-neutrality) that they sort into the fixed payment scheme.

The diagonal black line shows the 45 degree line. If all points were on this line, subjects would not change their behavior between the EXO and the ENDO treatment. Panel A shows that some subjects who sort into a piece rate scheme perform significantly different if they are imposed on a piece rate. The dashed dark gray line represents the regression line. The slope of the line is 0.75 which is statistically significantly different from 1 ($p < 0.01$). The intercept is equal to 17.35 ($p < 0.01$). However, subjects who sort into a fixed payment do not change performance if they are forced to work in a piece rate regime. The slope of the regression line is equal to 0.81 and not statistically significantly different from 1 ($p > .35$). The intercept is equal to 7.47 which is not statistically significantly different from zero ($p > 0.25$). The slope of the line of subjects who sort into the piece rate is smaller than 1. This suggests that the low-performers in this group of piece raters perform better when they are forced work in a piece rate regime compared to when they sort themselves into this payment scheme. For high-performers the story seems to be the opposite. Although not statistically significant, the majority of the gray diamonds representing subjects who sort into the fixed payment scheme are above the diagonal for low-performing subjects.

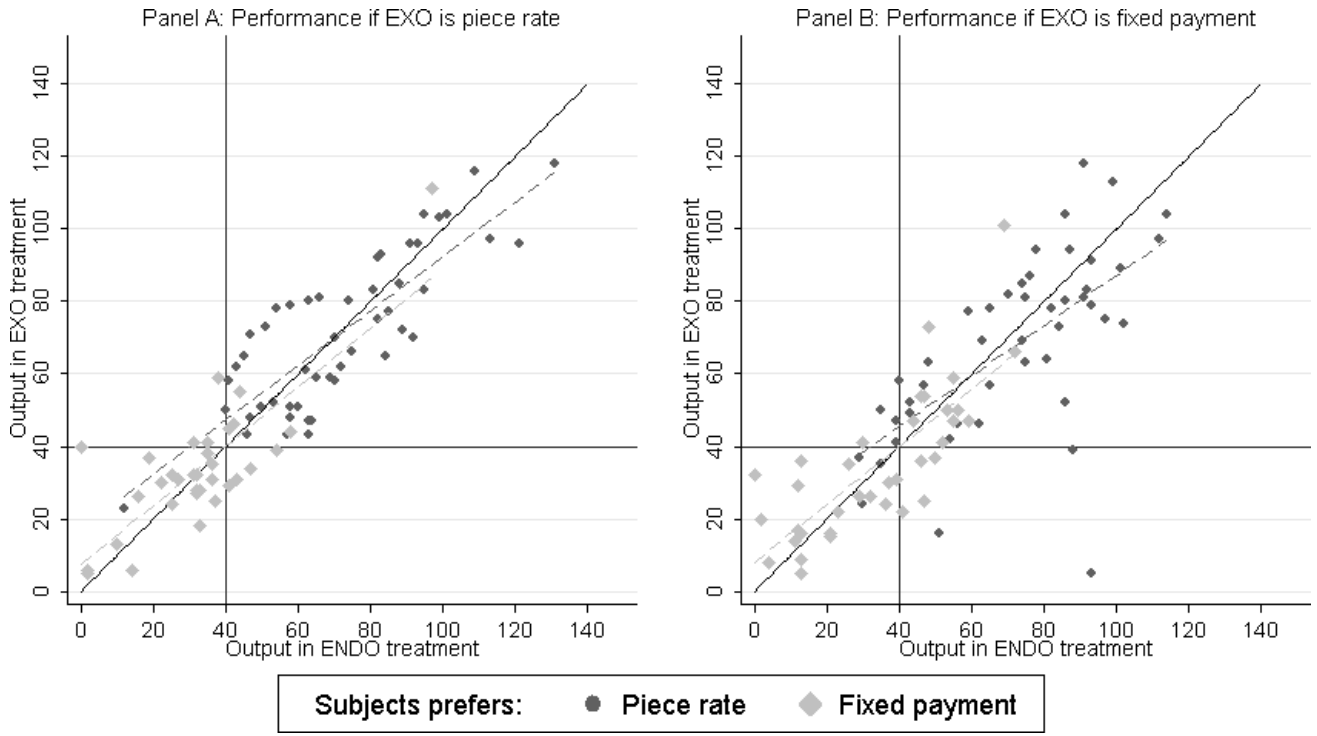


Figure 5.9: Relationship between performance in the EXO and ENDO treatment.

Note. Panel A shows the relationship between the performance in the EXO and the ENDO treatment if the EXO treatment was the piece rate. Panel B shows the equivalent figure if the EXO treatment was the piece rate. A dark gray dot indicates if a subject chose the piece rate in the ENDO treatment. A light gray diamond indicates if a subject chose the fixed payment in the ENDO treatment.

Panel B shows similar results as Panel A but now for the case in which the fixed payment is imposed in the EXO treatment. The horizontal axis shows performance in the ENDO treatment and vertical axis the performance in the EXO treatment. Again we have drawn a horizontal and vertical line at the threshold values of 40. What is clear is that now more subjects fail to meet the threshold of 40. When we look to the left of the vertical line, it turns out that six subjects sort into the piece rate scheme but do not meet the threshold. If we look at the dots below the horizontal line, we observe that three of them do not meet the threshold in both regimes, whereas three additional subjects who sort into the piece rate are relatively unproductive in the fixed payment scheme compared to the piece rate regime. Subjects who sort into the fixed payment seem to be more productive in this figure compared to Panel A.

The diagonal black line indicates the 45 degree line. The picture that emerges from Panel B is that subjects who sort into a piece rate scheme do slightly change their output if they are imposed to the exogenous fixed payment scheme. The intercept of the dark gray dashed line is 17.71 ($p < 0.05$). The slope is equal to .70 and statistically significantly different from 1 ($p < 0.01$). This leads to the same conclusion as in Panel A. However, subjects who sort into a fixed payment do not change their behavior if they are exogenously imposed to work in a fixed payment scheme. The slope of the regression line is .80 which is not statistically significantly different from one ($p > .14$). The intercept is 8.126 ($p < .1$), which is on the margin of significance. This latter results implies that low-performing subjects who sort into a fixed payment perform slightly better if they are forced to work in a fixed payment scheme.

5.4.4 Stress, Effort and Exhaustion

Changing payment modes exogenously does not seem to influence average performance relative to endogenous sorting into a payment scheme. However, changing payment schemes could have an effect on other outcomes, such as stress levels, effort and exhaustion. We measure stress, effort and exhaustion after each 10 minute working phase. Table 5.5 shows the impact of the payment schemes and treatments on these measures. All dependent variables are standardized with zero mean and a standard deviation of one. All regressions contain controls for the treatment order, a dummy for the phase, cognitive ability, study track, personality and session fixed effects. Standard errors are clustered on the session level.

The first three columns of Table 5.5 show the results of the self-reported stress

levels after the 10 minutes working phases. The first column reveals that higher levels of productivity are associated with lower stress levels. Also, being in a piece rate payment scheme leads to substantially higher levels of self-reported stress. Being in a piece rate is associated with an 80.2 per cent of a standard deviation increase of self-reported stress levels. The dummy variable for whether the treatment is exogenously imposed yields a small positive but statistically insignificant coefficient. This suggests that working in a piece rate scheme leads to higher stress levels, but that forcing subjects to work in a particular payment scheme does not induce higher stress levels. The specification shown in column (2) is the result of a model in which we have added a dummy which takes the value one if a worker prefers a fixed payment scheme. We also add an interaction with the piece rate. Finally, we add a dummy for gender. We do not find statistically significantly different effects compared to the estimated model displayed in column (1). The only difference is that women report higher levels of stress than men.

We find a similar picture for the self-reported measures of effort (columns (4) to (6)). Working in a piece rate scheme increases effort levels by at least 61.8 percent of a standard deviation in our baseline specification in column (4). Productivity also seems to be slightly negatively related with the self-reported effort levels. The test for the joint significance of the interaction terms yields a positive and significant result. This suggests that the exogenous assignment to a piece rate induces higher self-reported levels of effort. Women reveal higher self-reported levels of effort than men.

Another important dimension is the level of workers' exhaustion and the effect of our treatment on the self-reported level of exhaustion. Columns (7) to (9) show the determinants of the self-reported level of exhaustion after the 10 minutes working phases. First, we find a statistically insignificant but negative effect on the level of exhaustion of the EXO treatment. Also, productivity does not seem to be related with the self-reported exhaustion levels. However, there seems to be a weak positive relationship between being in a piece rate and the levels of exhaustion. If we test for the joint significance of the interaction terms with the preference for a fixed payment but being assigned to a piece rate, we obtain an statistically insignificant positive association between self-reported levels of exhaustion and being imposed on a piece rate scheme. Again women report higher levels of exhaustion.

Table 5.5: Determinants of stress, effort and exhaustion

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Stress			Effort			Exhaustion	
Productivity	-0.0164*	-0.0176*	-0.0167*	-0.0149**	-0.0146*	-0.0139*	0.00227	0.00263	0.00337
	(.008)	(.009)	(.009)	(.007)	(.007)	(.007)	(.008)	(.010)	(.010)
EXO Treatment	0.016	-0.0246	-0.0277	-0.0755	-0.00948	-0.0121	-0.112	-0.0788	-0.0815
	(.097)	(.116)	(.116)	(.065)	(.095)	(.094)	(.077)	(.117)	(.116)
Piece rate	0.802***	0.706***	0.682***	0.618***	0.747***	0.727***	0.381**	0.450*	0.430*
	(.097)	(.197)	(.182)	(.152)	(.218)	(.206)	(.141)	(.238)	(.234)
Preference for fix payment		-0.136	-0.209		0.136	0.0743		0.0804	0.0181
		(.276)	(.283)		(.251)	(.261)		(.375)	(.377)
Prefers fixed x piece rate		0.165	0.171		-0.278	-0.273		-0.14	-0.135
		(.264)	(.245)		(.259)	(.251)		(.345)	(.341)
1 if Female			0.385***			0.326**			0.328**
			(.112)			(.139)			(.119)
Observations	330	330	330	330	330	330	330	330	330
R-squared	0.272	0.273	0.301	0.294	0.297	0.316	0.106	0.106	0.126
Joint coefficient interactions	-	.735***	.644**	-	.605**	.528*	-	0.391	0.313
Controls	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note. The table shows linear regressions with the standardized behavioral measures of stress, effort and exhaustion as dependent variables. The last row indicates the coefficient of the test following linear combination: $E(stress|EXO = 1,$ Piece Rate $= 1,$ Preference for fixed payment $= 1) - E(stress|Exo = 0,$ Piece Rate $= 0,$ Preference for fixed payment $= 1) = EXO +$ Piece Rate $+$ Prefers fixed payment $+$ EXO x Prefers fixed payment $= 0$. $E(stress|EXO = 1,$ Piece Rate $= 1,$ Preference for fixed payment $= 1) - E(stress|Exo = 0,$ Piece Rate $= 0,$ Preference for fixed payment $= 1) = E\hat{X}O +$ Piece$\hat{}$ Rate $+$ Prefers fixe$\hat{}$d payment $+$ EXO x Prefers$\hat{}$ fixed payment $= 0$ The equivalent holds for effort and exhaustion. Our measures were the following questions. Effort. "How much effort did you exert solving the question during the previous 10[5] minutes?". Stress: "How stress did you feel?". Exhaustion: "How exhausted did you get?" Robust standard errors clustered on the session level in parentheses. Our controls include the treatment order, a dummy for the phase, personality, cognitive ability, type of the university and session fixed effects. ***$p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.5 Conclusion

In this paper we investigate sorting behavior and the effects of imposing either a piece rate or a fixed payment on workers. To do so we conduct a controlled laboratory experiment to investigate the role of an imposed payment scheme on the output, stress, effort and exhaustion levels of workers. Our experiment yields four main findings. First, imposing a variable payment scheme does not increase overall performance. More specifically the most productive workers seem to decrease their performance if they either imposed on a piece rate and a fixed payment scheme. Those workers who prefer a fixed payment but are imposed on a piece rate only slightly change their output. They do not change their output if they are imposed on a fixed payment. Third, the driving factor behind a sorting decision is an individual's productivity. Conditional on productivity preferences do not seem to matter for the sorting decision. Fourth, the act of imposing a piece rate does not increase stress levels, effort levels and exhaustion. However, working in a piece rate always increases stress levels compared to a fixed payment.

Our findings show that the majority of workers seem to exert maximal effort even if their payment is not directly linked to performance. Moreover, performance dependent payment cannot genuinely increase output of rather unproductive workers. In general this is in line what we often observe in reality. Many contracts are incomplete and agents exert effort even though they know it is not necessarily enforceable by the principal. Previous findings from the lab (e.g. Fehr et al. (1998)) and the field (e.g. Kube et al. (2012)) explain this behavior as a gift exchange. The agent reciprocates the principals' kindness with effort levels above the rational level of effort (Akerlof (1982)). Since self- reported levels of stress and effort always increase if a piece rate is at play our findings also have important implications for managers and policy makers. Since the incentive effect of our piece rate is rather negligible but the increase in stress and effort levels is substantial, changes in performance schemes could have effects on employee's health conditions.

Interesting future avenues for research would be to investigate the effects of imposing different payment schemes such as tournaments and revenue-sharing payment schemes. Since these payment schemes become more and more popular it is important to investigate their effects on output and stress levels. Other potential continuations of this research would be to extend the time of the working phase, to change the conditions of the working phase and to vary the nature of the real effort task. A longer working phase might trigger different effects of the payment

schemes. Moreover, giving subjects the opportunity of selecting an outside option besides the working task might at even more realism to the experimental conditions. Although the nature of the task comes close to many white color like working tasks it is important to investigate if our findings hold if the task is changed.

Bibliography

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin 82*(4), 463–496.

Akerlof, G. A. (1982). Labor contracts as partial gift exchange. *The Quarterly Journal of Economics 97*(4), 543–569.

Akerlund, D., B. Golsteyn, H. Gronqvist, and L. Lindahl (2014). Time preferences and criminal behavior. *IZA Discussion Paper No. 8168.*.

Almlund, M., A. L. Duckworth, J. J. Heckman, and T. D. Kautz (2011). Personality psychology and economics. In *Handbooks of the Economics of Education*, pp. 1–158. Amsterdam, North Holland: Elsevier.

Angrist, J., D. Lang, and P. Oreopoulos (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics 1*(1), 136–63.

Angrist, J. and V. Lavy (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *The American Economic Review 99*(4), 1384–1414.

Banuri, S. and P. Keefer (2013). Intrinsic motivation, effort and the call to public service.

Becker, A., T. Deckers, T. Dohmen, A. Falk, and F. Kosse (2012). The relationship between economic preferences and psychological personality measures. *Annual Review of Economics 4*(1), 453–478.

Bénabou, R. and J. Tirole (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies 70*(3), 489–520.

Benjamin, D. J., S. A. Brown, and J. M. Shapiro (2013, December). Who is behavioral? cognitive ability and anomalous preferences. *Journal of the European Economic Association 11*(6), 1231–1255.

Bettinger, E. and R. Slonim (2007, February). Patience among children. *Journal of Public Economics 91*(12), 343–363.

Bettinger, E. P. (2011, July). Paying to learn: The effect of financial incentives on elementary school test scores. *Review of Economics and Statistics 94*(3), 686–698.

Board, C. (2013). *Total Group Profile Report*.

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. t. Weel (2008). The economics and psychology of personality traits. *Journal of Human Resources 43*(4), 972–1059.

Borghans, L. and B. H. Golsteyn (2007). Time discounting and imagination. *Maastricht University, Working Paper*.

Borghans, L., B. H. Golsteyn, J. J. Heckman, and J. E. Humphries (2014). What do grades and achievement tests measure? *Maastricht University, Working Paper*.

Borghans, L., B. H. Golsteyn, and U. Zölitz (2014). School quality and the development of cognitive skills between age four and six. *Maastricht University, Working Paper*.

Borghans, L., H. Meijers, and B. Ter Weel (2008). The role of noncognitive skills in explaining cognitive test scores. *Economic Inquiry 46*(1), 2–12.

Borghans, L., H. Meijers, and B. ter Weel (2013). The importance of intrinsic and extrinsic motivation for measuring IQ. *Economics of Education Review 34*, 17–28.

Borghans, L., H. Meijers, B. Vogt, and B. Ter Weel (2014). The economics of test taking: The effect of pressure on decision making and test performance. *Maastricht University, Working Paper*.

Borghans, L. and T. Schils (2013). The leaning tower of pisa decomposing achievement test scores into cognitive and non-cognitive components. *Maastricht University, Working Paper*.

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

Burks, S., J. Carpenter, L. Gtte, and A. Rustichini (2012). Which measures of time preference best predict outcomes: Evidence from a large-scale field experiment. *Journal of Economic Behavior & Organization 84*(1), 308–320.

Burks, S. V., J. P. Carpenter, L. Goette, and A. Rustichini (2009). Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proceedings of the National Academy of Sciences 106*(19), 7745–7750.

Buser, T., M. Niederle, and H. Oosterbeek (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, qju009.

Cadsby, C. B., F. Song, and F. Tapon (2007). Sorting and incentive effects of pay for performance: An experimental investigation. *Academy of Management Journal 50*(2), 387–405.

Caplin, A. and M. Dean (2011). Search, choice, and revealed preferences. *Theoretical Economics 1*(6), 19–48.

Caplin, A., M. Dean, and D. Martin (2011). Search and satisficing. *The American Economic Review 101*(7), 2899–2922.

Carpenter, P. A., M. A. Just, and P. Shell (1990). What one intelligence test measures: A theoretical account of the processing in the raven progressive matrices test. *Psychological review 97*, 404–431.

Castillo, M., P. J. Ferraro, J. L. Jordan, and R. Petrie (2011, December). The today and tomorrow of kids: Time preferences and educational outcomes of children. *Journal of Public Economics 95*(1112), 1377–1385.

Chamorro-Premuzic, T., J. Moutafi, and A. Furnham (2005). The relationship between personality traits, subjectively-assessed and fluid intelligence. *Personality and Individual Differences 38*(7), 1517–1528.

Cito. *Terugblik en resultaten 2014 - Eindtoets Basisonderwijs Groep 8* (2014 ed.). Arnhem.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika 16*(3), 297–334.

Cunha, F. and J. J. Heckman (2009). The economics and psychology of inequality and human development. *Journal of the European Economic Association 7*(2-3), 320–364.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

Deckers, T., A. Falk, F. Kosse, and H. Schildberg-Hörisch (2014). How does socioeconomic status shape a child's personality? *Mimeo*.

Delfgaauw, J. and R. Dur (2008). Incentives and workers motivation in the public sector*. *The Economic Journal 118*(525), 171–191.

Dohmen, T. (2014). Behavioural labour economics: Advances and future directions. *Labour Economics*.

Dohmen, T. and A. Falk (2010). You get what you pay for: Incentives and selection in the education system*. *The Economic Journal 120*(546), F256–F271.

Dohmen, T. and A. Falk (2011). Performance Pay and Multidimensional Sorting-Productivity, Preferences and Gender. *American Economic Review 101*(2), 556–590.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2010). Are risk aversion and impatience related to cognitive ability? *The American Economic Review 100*(3), 1238–1260.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2012). Interpreting time horizon effects in inter-temporal choice. *IZA Discussion Paper No. 6385*.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association 9*, 522–550.

Duckworth, A. L. and M. L. Kern (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality 45*(3), 259–268.

Duckworth, A. L., P. D. Quinn, D. R. Lynam, R. Loeber, and M. Stouthamer-Loeber (2011, April). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*.

Duckworth, A. L., P. D. Quinn, and E. Tsukayama (2012). What no child left behind leaves behind: The roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *Journal of Educational Psychology 104*(2), 439–451.

Duckworth, A. L. and M. E. Seligman (2005). Self-discipline outdoes iq in predicting academic performance of adolescents. *Psychological science 16*(12), 939–944.

Duckworth, A. L., E. Tsukayama, and T. A. Kirby (2013, July). Is it really self-control? examining the predictive power of the delay of gratification task. *Personality and Social Psychology Bulletin 39*(7), 843–855.

Dur, R. and R. Zoutenbier (2012). Intrinsic motivation of public sector employees: Evidence for germany. *Rotterdam University Working Paper*.

Dur, R. and R. Zoutenbier (2014). Working for a good cause. *Public Administration Review 74*(2), 144–155.

Edgeworth, F. Y. (1879, July). The hedonical calculus. *Mind 4*(15), 394–408.

Edlund, C. 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, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist 49*(8), 709–724.

Eriksson, T., S. Teyssier, and M.-C. Villeval (2009, July). Self-selection and the efficiency of tournaments. *Economic Inquiry 47*(3), 530–548.

Eriksson, T. and M. C. Villeval (2008). Performance-pay, sorting and social motivation. *Journal of Economic Behavior & Organization 68*(2), 412–421.

Fahr, R. and B. Irlenbusch (2000). Fairness as a constraint on trust in reciprocity: earned property rights in a reciprocal exchange experiment. *Economics Letters 66*(3), 275–282.

Falk, A., A. Becker, T. Dohmen, D. Huffman, and U. Sunde (2012). An experimentally validated preference module. Technical report, Working Paper.

Falk, A. and J. J. Heckman (2009). Lab experiments are a major source of knowledge in the social sciences. *Science 326*(5952), 535–538.

Fehr, E. and A. Falk (1999). Wage rigidity in a competitive incomplete contract market. *Journal of Political Economy 107*(1), 106–134.

Fehr, E. and A. Falk (2002). Psychological foundations of incentives. *European Economic Review 46*(45), 687–724.

Fehr, E., G. Kirchsteiger, and A. Riedl (1998). Gift exchange and reciprocity in competitive experimental markets. *European Economic Review 42*(1), 1–34.

Fehr, E. and K. M. Schmidt (1999, August). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics 114*(3), 817–868.

Feron, E., T. Schils, and B. Ter Weel (2014). Matching students and education levels: the role of standardized test scores and teacher evaluation. *Maastricht University, Mimeo*.

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. *The Journal of Economic Perspectives 19*(4), 25–42.

Frey, B. S. and F. Oberholzer-Gee (1997). The cost of price incentives: An empirical analysis of motivation crowding- out. *The American Economic Review 87*(4), 746–755.

Fryer, R. G. (2011). Financial incentives and student achievement: Evidence from randomized trials. *Quarterly Journal of Economics 126*(4), 1755–1798.

Fryer, R. G. (2013, April). Teacher incentives and student achievement: Evidence from new york city public schools. *Journal of Labor Economics 31*(2), 373–407.

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.

Gerhards, L. and N. Siemer (2014). Private versus public feedback - the incentive effects of symbolic awards. *Economics Working Papers - Aarhus University*.

Gittleman, M. and B. Pierce (2013). How prevalent is performance-related pay in the united states? current incidence and recent trends. *National Institute Economic Review 226*(1), R4–R16.

Gneezy, U., M. Niederle, and A. Rustichini (2003, August). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics 118*(3), 1049–1074.

Gneezy, U. and A. Rustichini (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics 115*(3), 791–810.

Gneezy, U. and A. Rustichini (2004). Gender and competition at a young age. *The American Economic Review 94*(2), 377–381.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology 59*(6), 1216–1229.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment 4*(1), 26–42.

Golsteyn, B. H., H. Grnqvist, and L. Lindahl (2014). Adolescent time preferences predict lifetime outcomes. *The Economic Journal*, n/a–n/a.

Golsteyn, B. H. and T. Schils (2014). Gender gaps in primary school achievement. a decomposition into endowments and returns to IQ and non-cognitive factors. *Economics of Education Review*.

Green, F. (2013). *Skills and Skilled Work: An Economic and Social Analysis*. Oxford University Press.

Greene, W. H. (2012). *Econometric Analysis* (7 ed.). New York.

Greiner, B. (2003). *An Online Recruitment System for Economic Experiments*, Volume GWDG Bericht 63 of *Forschung und wissenschaftliches Rechnen 2003*. Kurt Kremer, Volker Macho.

Greiner, B. (2004). The online recruitment system orsee 2.0-a guide for the organization of experiments in economics. *University of Cologne, Working paper series in economics 10*(23), 63–104.

Harrison, G. W. and J. A. List (2004). Field experiments. *Journal of Economic Literature 42*(4), 1009–1055.

Heckman, J. J. and T. Kautz (2012). Hard evidence on soft skills. *Labour Economics 19*(4), 451–464.

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47*(2), 263–291. ArticleType: research-article / Full publication date: Mar., 1979 / Copyright 1979 The Econometric Society.

Kautz, T., J. J. Heckman, R. Diris, B. ter Weel, and L. Borghans (2014). Fostering and measuring skills.

Kirby, K. N. (2009). One-year temporal stability of delay-discount rates. *Psychonomic Bulletin & Review 16*(3), 457–462.

Kirby, K. N., G. C. Winston, and M. Santiesteban (2005). Impatience and grades: Delay-discount rates correlate negatively with college GPA. *Learning and Individual Differences 15*(3), 213–222.

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

Koszegi, B. (2014). Behavioral contract theory. *Journal of Economic Literature 52*(4), 1075–1118.

Kube, S., M. A. Marchal, and C. Puppe (2012). The currency of reciprocity: Gift exchange in the workplace. *The American Economic Review 102*(4), 1644–1662.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics 112*(2), 443–477.

Larkin, I. and S. Leider (2012). Incentive schemes , sorting, and behavioral biases of employees: Experimental evidence. *American Economic Journal: Microeconomics 4*(2), 184–214.

Larkin, I., L. Pierce, and F. Gino (2012). The psychological costs of pay-for-performance: Implications for the strategic compensation of employees. *Strategic Management Journal 33*(10), 1194–1214.

Lavy, V. (2002). Evaluating the effect of teachers group performance incentives on pupil achievement. *Journal of Political Economy 110*(6), 1286–1317.

Lazear, E. (2000). The power of incentives. *American Economic Review 90*(2), 410–414.

Lazear, E. and S. Rosen (1981). Rank-order tournaments as optimum labor contracts. *The Journal of Political Economy 89*(5), 841–864.

Lazear, E. P., U. Malmedier, and R. A. Weber (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics 4*(1), 136–63.

Lloyd, M. E. and T. M. Zylla (1988, October). 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, P. and M. Mariotti (2007). Sequentially rationalizable choice. *The American Economic Review 97*(5), 1824–1839.

Mischel, W., E. B. Ebbesen, and A. Raskoff Zeiss (1972). Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology 21*(2), 204–218.

Mischel, W., Y. Shoda, and M. I. Rodriguez (1989). Delay of gratification in children. *Science 244*(4907), 933–938.

Moffitt, T. E., L. Arseneault, D. Belsky, N. Dickson, R. J. Hancox, H. Harrington, R. Houts, R. Poulton, B. W. Roberts, S. Ross, M. R. Sears, W. M. Thomson, and A. Caspi (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences 108*(7), 2693–2698.

Neisser, U., G. Boodoo, T. J. Bouchard Jr, A. W. Boykin, N. Brody, S. J. Ceci, D. F. Halpern, J. C. Loehlin, R. Perloff, and R. J. Sternberg (1996). Intelligence: Knowns and unknowns. *American psychologist 51*(2), 77.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics 122*(3), 1067–1101.

Non, A. and D. Tempelaar (2014). Time preferences, study effort and academic performance. *ROA Working Paper. Maastricht University.*.

Oettingen, G., D. Mayer, and J. Thorpe (2010). Self-regulation of commitment to reduce cigarette consumption: Mental contrasting of future with reality. *Psychology & Health 25*(8), 961–977.

Perez-Arce, F. (2011). The effect of education on time preferences. *RAND working paper*.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of economic literature 37*(1), 7–63.

Prevoo, T. (2013). *The relevance, variability, and malleability of personality traits*. Maastricht University Dissertation. Maastricht: Maastricht University Press.

Raven, J. (1962). *Advanced Progressive Matrices*. H. K. Lewis & Co. Ltd, London.

Reutskaja, E., R. C. Nagel, C. F. Camerer, and A. Rangel (2011). Search dynamics in consumer choice under time pressure: An eye-tracking study. *American Economic Review 101*(2), 900–926.

Roberts, B. W. (2006). Personality development and organizational behavior. *Research in Organizational Behavior 27*, 1–40.

Roberts, B. W. (2009, April). Back to the future: Personality and assessment and personality development. *Journal of Research in Personality 43*(2), 137–145.

Roberts, R. D., P. M. Markham, G. Matthews, and Z. Moshe (2005). Assessing intelligence: Past, present, and future. In *Handbook of Understanding and Measuring Intelligence*, pp. 333–360. Thousand Oaks, CA: Sage Publications Inc.

Rodriguez-Planas, N. (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.

Samuelson, P. A. (1937). A note on measurement of utility. *The Review of Economic Studies 4*(2), 155–161.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science 58*(8), 1438–1457.

Strotz, R. H. (1955). Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies 23*(3), 165–180.

Sutter, M., M. G. Kocher, D. Gltzle-Rtzler, and S. T. Trautmann (2013). Impatience and uncertainty: Experimental decisions predict adolescents' field behavior. *The American Economic Review 103*(1), 510–531.

Tversky, A. and D. Kahneman (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics 106*(4), 1039–1061.

van den Berge, W., T. Ooms, B. ter Weel, R. Daas, and A. B. Dijkstra (2014). *Investeren in skills en competenties: Een voorstudie voor programmering van onderzoek en beleid*. Amsterdam: CPB, Universiteit Amsterdam.

Van Dijk, F., J. Sonnemans, and F. Van Winden (2001). Incentive systems in a real effort experiment. *European Economic Review 45*(2), 187–214.

Woessmann, L. (2011). Cross-country evidence on teacher performance pay. *Economics of Education Review 30*(3), 404–418.

Wölbert, E. and A. Riedl (2013). Measuring time and risk preferences: Realiability, stability, domain specificity. *CESifo Working Paper*.

Zoutenbier, R. (2015). *Work Motivation and Incentives in the Public Sector*. Erasmus University of Rotterdam Dissertation. Rotterdam: University of Rotterdam.

Appendices

Appendix A

Appendix to Chapter 2

This section contains additional information about the experimental procedure and additional results of the experiment. In Section A.1 we provide detailed information on the procedure of the experiment. This includes additional data on our subject pool and the schedule of the experiment in Section A.1.1. Section A.1.2 contains information about the question types we used in the experiment. We also report success rates per item both for numerical tasks and Raven matrices. In Section A.1.3 we provide information about the payment schemes and in Section A.1.4 screenshots of the instructions. In Section A.2 we provide additional results on the numerical tasks.

A.1 Subject Pool and Experimental Details

A.1.1 Experiment

Our subjects were students from Maastricht University. We recruited them using the software ORSEE (Greiner, 2004). The experiments were conducted at the 23^{rd} of November, 4^{th} and 5^{th} of December 2012 in BEELab at Maastricht University. Table A.1 shows an overview of the characteristics of our subject pool. Due to the international orientation of the faculty only 10.9 percent of the participants were Dutch. The vast majority of the participants are German. Most of our subjects (67 percent) were students of the School of Business and Economics. The mean age was 22 and 38 percent of our subjects were female. The average monthly budget is 650 euros. Figure A.1 shows the schedule of the experiment. The computer determined with which question type (Raven matrices or numerical tasks) a subject started. Afterwards they read detailed instructions about the tasks, the payment and the treatments. Before they could start with the actual experiment, they had to play a trial phase which consisted either of five Raven matrices or five numerical tasks. They were asked to calculate their payoff for every puzzle conditional on their performance. This phase was meant to make them aware of the difference

Table A.1: Descriptive Statistics of the Subject Pool

	Mean	St. Dev.
Women	0.383	(0.488)
Age	22.102	(1.988)
Monthly Income (Euro)	648.977	(311.73)
Earnings (Euro)	31.85	(4.602)
Chinese	0.039	(0.195)
Dutch	0.109	(0.313)
German	0.602	(0.492)
Italian	0.008	(0.088)
Polish	0.047	(0.212)
Other	0.195	(0.398)
Business	0.5	(0.502)
Economics	0.219	(0.415)
Medicine	0.008	(0.088)
Psychology	0.016	(0.125)
Law	0.086	(0.281)
European Studies	0.078	(0.269)
Other Studies	0.094	(0.293)
Observations	128	

between the blue and the red payment scheme and to help them understand how their payoff was determined. They could not continue the experiment if they did not answer the questions regarding their payoff correctly. After every block of 15 questions subjects had to wait for one minute before they could continue with the next set of questions. Before the first question of every block started subjects were reminded of the payoff from the blue and the red system. They had to answer control questions regarding their payoffs before the block started. After the first set of 45 questions they had a short brake of another 3 minutes. Afterwards the block of the next set of 45 questions was started. When they were done with the last question, they were asked to answer some control questions. Afterwards they received their payoff information on the red and the blue system and the number of correctly submitted answers for each question type on the screen. They were paid and left the lab.

A.1.2 Types of Questions

We took Raven matrices from two different sources. First, we took questions from set I and set II from the original versions (Raven, 1962). Second, we took Raven matrices from an online version. These Raven matrices were very similar in the degree of difficulty to the ones from the original set. Panel A of Table A.1 shows the success rate of each item in the experiment. We define the success rate as correctly submitted answers. Columns (2) to (4) document the success rate,

Time	Stage	Description
	Instructions	Numerical task/Ravens
	Trial Phase	Numerical task/Ravens
	Numerical task/Ravens	3 Treatments in randomized order. 15 puzzles per treatment. LL, HH, HL
	Break	
	Instructions	Ravens/Numerical task
	Trial Phase	Ravens/Numerical task
	Ravens/Numerical task	3 Treatments in randomized order. 15 puzzles per treatment. LL, HH, HL
	Final Questionnaire	Control Questions
	Payment	

Figure A.1: Schedule of the Experiment

standard errors and the number of observations from our experiment. Columns (5) to (7) show the results for the corresponding matrices from the test manual of the Raven matrices. At the bottom of Panel A we document the success rates for all the Raven matrices from our experimental sample and the overall success rate from the sample in the test manual. We test whether the difference is statistically significant. The cell on the bottom right of Panel A document the t-value of this test. Since it is smaller than 1.96, we reject the hypothesis that the fraction of correctly submitted answers is significantly larger in our sample compared to the sample in the Raven test manual. Unfortunately standard errors on the question level are not reported in the test manual. Thus, a statistical comparison at the question level is not possible.

We created numerical problems based on an idea developed by Caplin et al. (2011). We adapted the degree of complexity, since the working time was restricted to 60 seconds per question and we only paid for the correct solution. The most complex numerical task corresponds to a level of complexity 3 in Caplin et al. (2011). We report the success rates in Panel B of Table A.1. To our surprise there is one item which none of our participants answered correctly. However, we checked the choices during answering this item and it turns out that the correct answer is selected by three persons at the beginning. However, when it comes to submission none of our participants submitted the correct answer. One possible explanation for this behavior is that they select a random answer at the beginning. Afterwards, they start searching for the correct answer by searching from the top to the bottom (see Caplin et al. (2011) for a comparison). It might be that the time was over before they came to the correct solution. One subject did indeed choose the correct answer in the last second but submitted a different answer earlier in time.

A.1.3 The Red Payment Schemes

Figure A.2 plots the functional form of the red payment schemes in three treatments. The dotted line plots the payoff at each second for the baseline treatment. The dashed line plots the payoff function of the high stakes and high time pressure treatment and the straight line the plots the payoff function for the high stakes and low time pressure treatment.

Table A.2: Success Rate of Raven Matrices and Numerical Tasks by Item

Panel A. Raven Matrices

	Experiment			Data from Raven Manual			Difference	t-value diff.
Item	Success Rate	Standard Error	N	Success Rate	Standard Error	N		
1	0.758	0.038	128	0.672	n/a	1015	0.086	n/a
2	0.685	0.041	127	0.624	n/a	1015	0.061	n/a
3	0.74	0.039	127	0.67	n/a	1015	0.07	n/a
4	0.606	0.044	127	0.615	n/a	1015	-0.009	n/a
5	0.547	0.044	128	0.558	n/a	1015	-0.011	n/a
6	0.563	0.044	128	0.597	n/a	1015	-0.035	n/a
7	0.512	0.045	127	0.581	n/a	1015	-0.069	n/a
8	0.362	0.043	127	0.449	n/a	1015	-0.087	n/a
9	0.535	0.044	127	0.493	n/a	1015	0.042	n/a
10	0.492	0.044	128	0.442	n/a	1015	0.05	n/a
11	0.378	0.043	127	0.302	n/a	1015	0.076	n/a
12	0.32	0.041	128	0.365	n/a	1015	-0.045	n/a
13	0.445	0.044	128	0.399	n/a	1015	0.046	n/a
14	0.409	0.044	127	0.255	n/a	1015	0.154	n/a
15	0.227	0.037	128	0.264	n/a	1015	-0.037	n/a
16	0.227	0.037	128	0.199	n/a	1015	0.028	n/a
17	0.25	0.038	128	0.272	n/a	1015	-0.022	n/a
18	0.305	0.041	128	0.245	n/a	1015	0.06	n/a
19	0.18	0.034	128	0.168	n/a	1015	0.012	n/a
20	0.252	0.039	127	0.269	n/a	1015	-0.017	n/a
21	0.244	0.038	127	0.168	n/a	1015	0.076	n/a
22	0.305	0.041	128	0.175	n/a	1015	0.13	n/a
23	0.078	0.024	128	0.036	n/a	1015	0.042	n/a
24	0.219	0.037	128	n/a	n/a	n/a	-	n/a

(Panel A ctd.)								
25	0.461	0.044	128	n/a	n/a	n/a	-	n/a
26	0.614	0.043	127	n/a	n/a	n/a	-	n/a
27	0.719	0.04	128	n/a	n/a	n/a	-	n/a
28	0.805	0.035	128	n/a	n/a	n/a	-	n/a
29	0.633	0.043	128	n/a	n/a	n/a	-	n/a
30	0.339	0.042	127	n/a	n/a	n/a	-	n/a
31	0.453	0.044	128	n/a	n/a	n/a	-	n/a
32	0.797	0.036	128	n/a	n/a	n/a	-	n/a
33	0.422	0.044	128	n/a	n/a	n/a	-	n/a
34	0.409	0.044	127	n/a	n/a	n/a	-	n/a
35	0.402	0.044	127	n/a	n/a	n/a	-	n/a
36	0.625	0.043	128	n/a	n/a	n/a	-	n/a
37	0.484	0.044	128	n/a	n/a	n/a	-	n/a
38	0.118	0.029	127	n/a	n/a	n/a	-	n/a
39	0.25	0.038	128	n/a	n/a	n/a	-	n/a
40	0.103	0.027	126	n/a	n/a	n/a	-	n/a
41	0.094	0.026	128	n/a	n/a	n/a	-	n/a
42	0.125	0.029	128	n/a	n/a	n/a	-	n/a
43	0.906	0.026	128	0.844	n/a	1015	0.062	n/a
44	0.606	0.044	127	0.84	n/a	1015	-0.234	n/a
45	0.827	0.034	127	0.818	n/a	1015	0.009	n/a
All	0.441	0.033						
All Raven from manual	0.452	0.043		0.435	0.046		0.017	0.248
Other Raven	0.425	0.053						

		Panel B. Numerical Tasks	
Item	Success Rate	Standard Error	N
1	0.766	0.038	128
2	0.852	0.032	128
3	0.656	0.042	128
4	0.297	0.041	128
5	0.781	0.037	128
6	0.914	0.025	128
7	0.656	0.042	128
8	0.844	0.032	128
9	0.891	0.028	128
10	0.859	0.031	128
11	0.273	0.04	128
12	0.008	0.001	128
13	0.703	0.041	128
14	0.922	0.024	128
15	0.813	0.035	128
16	0.899	0.027	128
17	0.648	0.042	128
18	0.633	0.043	128
19	0.695	0.041	128
20	0.484	0.044	128
21	0.938	0.021	128
22	0.914	0.025	128
23	0.828	0.033	128
24	0.609	0.043	128
25	0.867	0.03	128
26	0.883	0.029	128
27	0	0	128

		Panel B. Numerical Tasks	
Item (Panel B ctd.)	Success Rate	Standard Error	N
28	0.578	0.044	128
29	0.789	0.036	128
30	0.539	0.044	128
31	0.805	0.035	128
32	0.531	0.044	128
33	0.961	0.017	128
34	0.781	0.037	128
35	0.891	0.028	128
36	0.563	0.044	128
37	0.477	0.044	128
38	0.719	0.04	128
39	0.711	0.04	128
40	0.031	0.015	128
41	0.516	0.044	128
42	0.938	0.021	128
43	0.477	0.044	128
44	0.656	0.042	128
45	0.734	0.039	128
All	0.674	0.006	5760

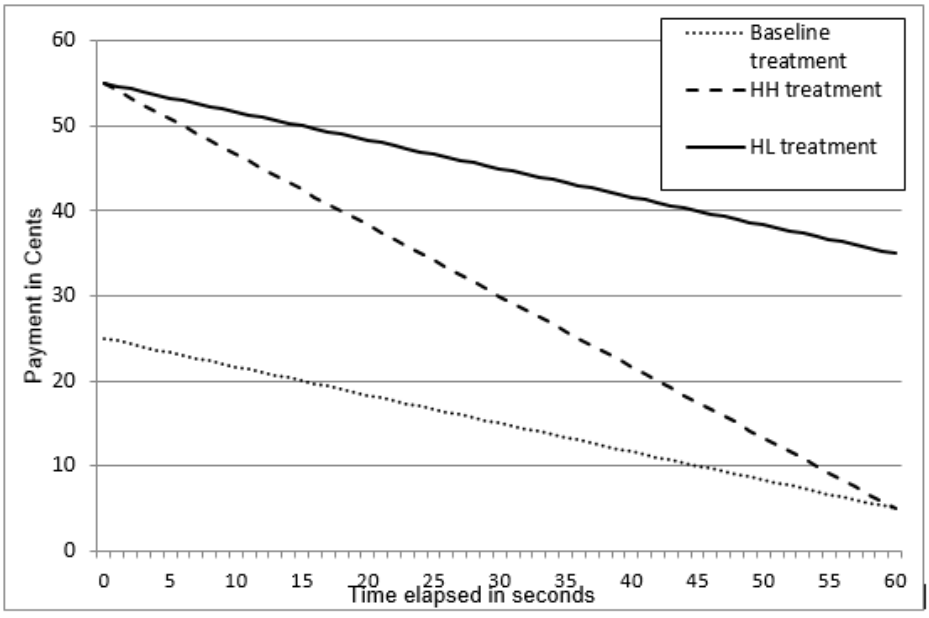
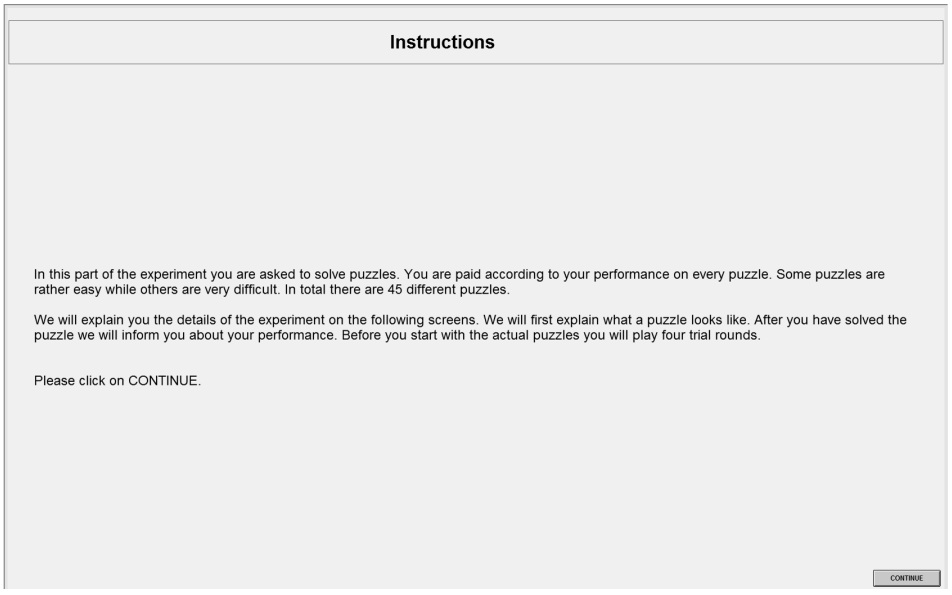
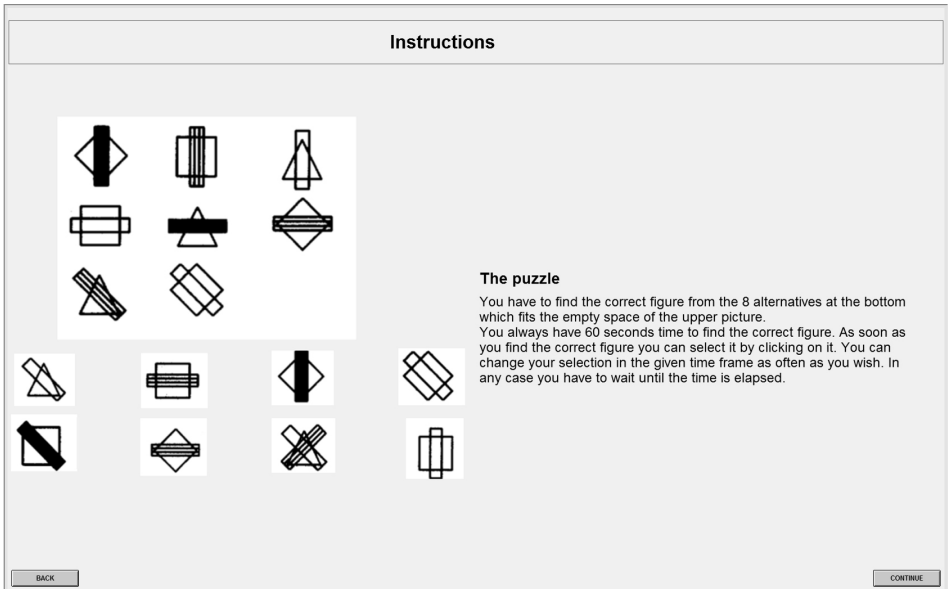


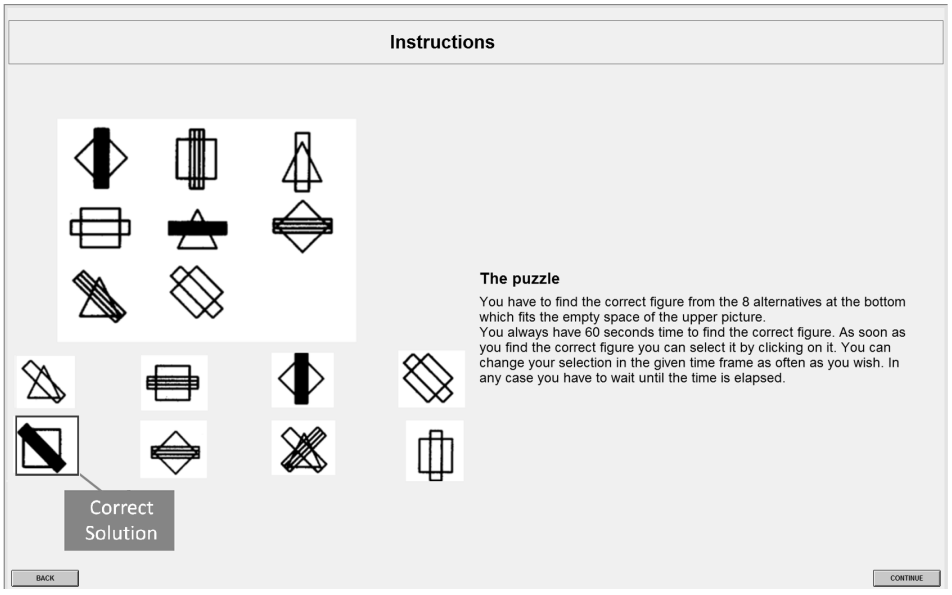
Figure A.2: Different Payment Regimes of the Red Payment Scheme

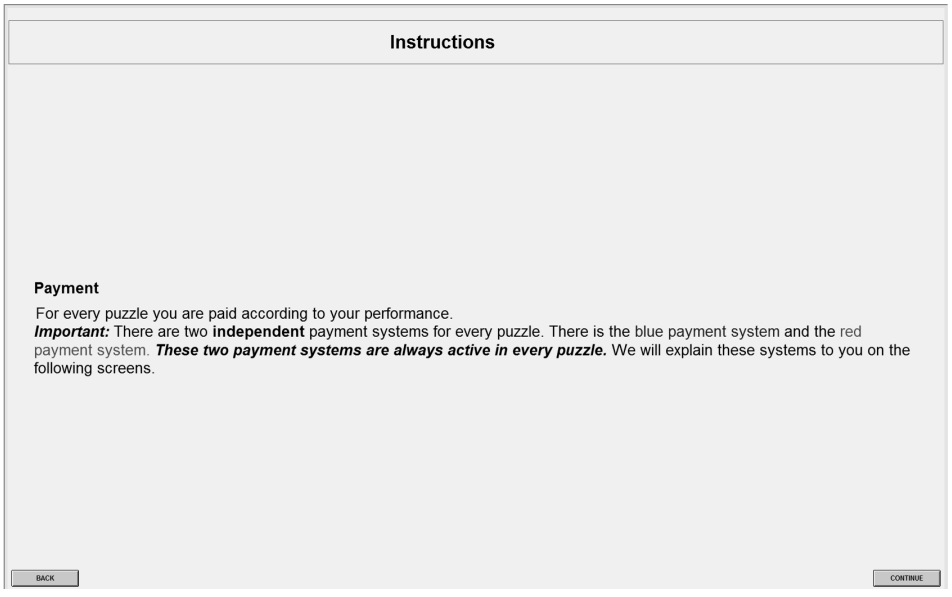
A.1.4 Screenshots of the Instructions

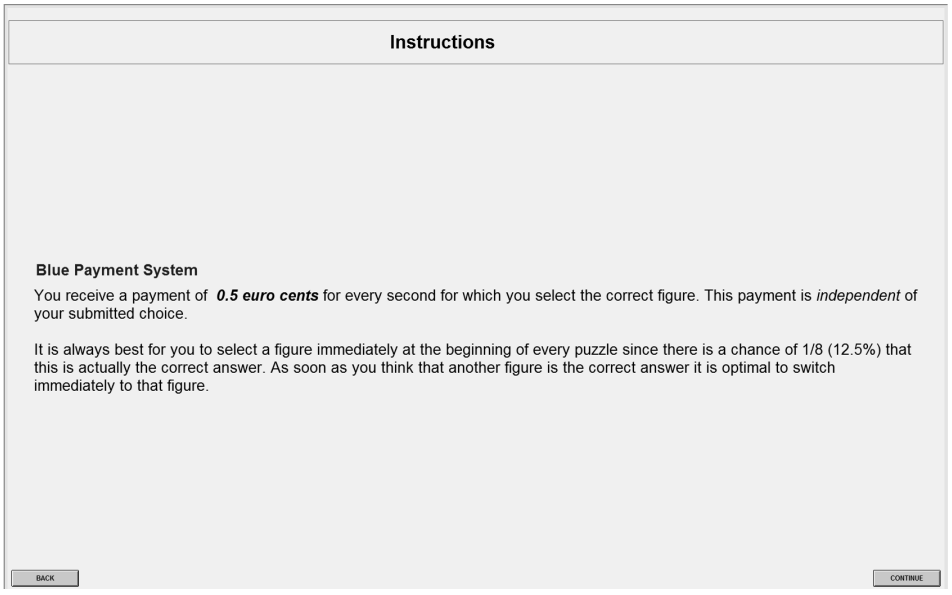
In the following we show screenshots of the experiment. It is important to mention that we always take the same Raven matrix in these screenshots. In the actual experiment we used of course different matrices. However, the original ones shall be kept confidential and the matrix we use as an example is already published in Carpenter et al. (1990).

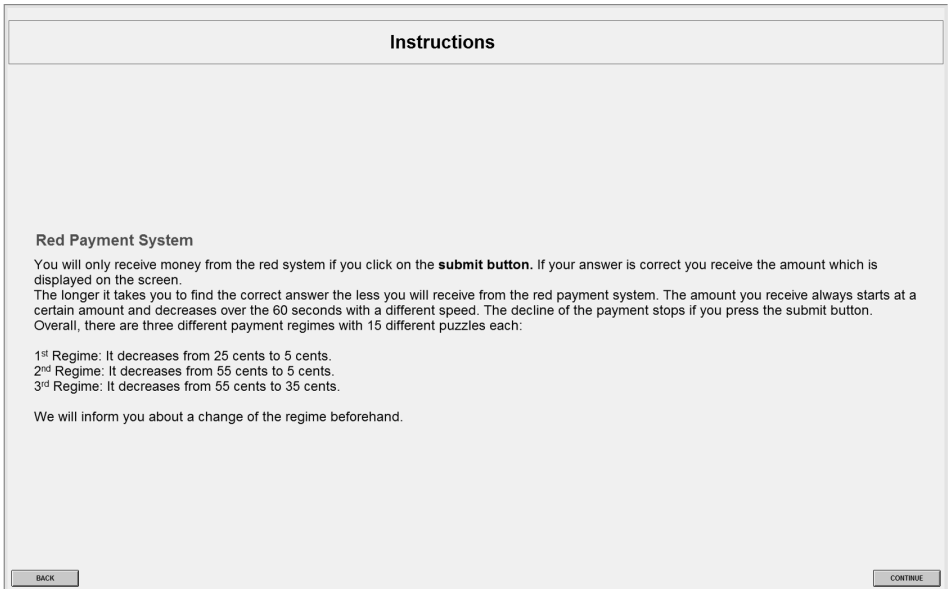

Instructions
In this part of the experiment you are asked to solve puzzles. You are paid according to your performance on every puzzle. Some puzzles are rather easy while others are very difficult. In total there are 45 different puzzles.
We will explain you the details of the experiment on the following screens. We will first explain what a puzzle looks like. After you have solved the puzzle we will inform you about your performance. Before you start with the actual puzzles you will play four trial rounds.
Please click on CONTINUE.
CONTINUE

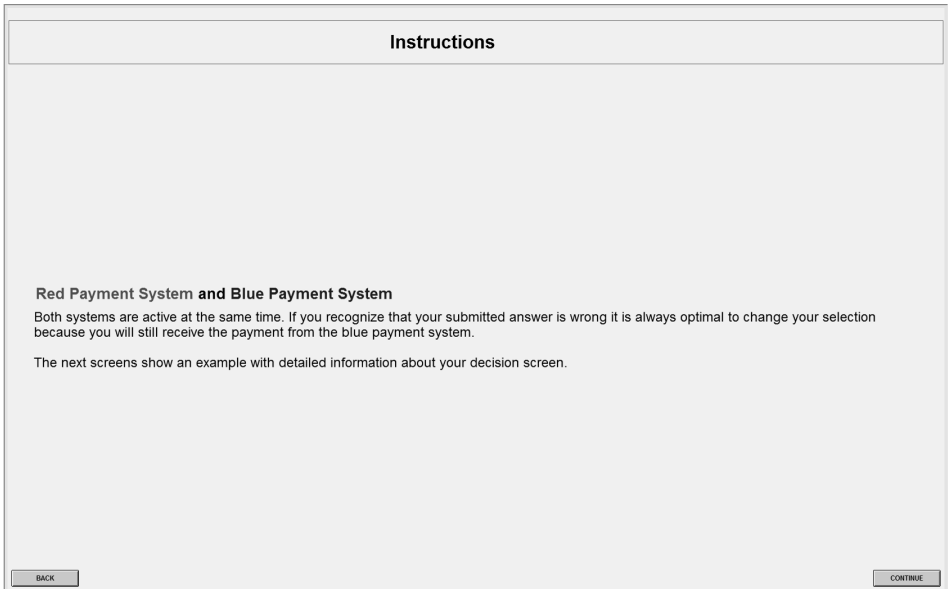

Instructions
The puzzle
You have to find the correct figure from the 8 alternatives at the bottom which fits the empty space of the upper picture.
You always have 60 seconds time to find the correct figure. As soon as you find the correct figure you can select it by clicking on it. You can change your selection in the given time frame as often as you wish. In any case you have to wait until the time is elapsed.
BACK
CONTINUE

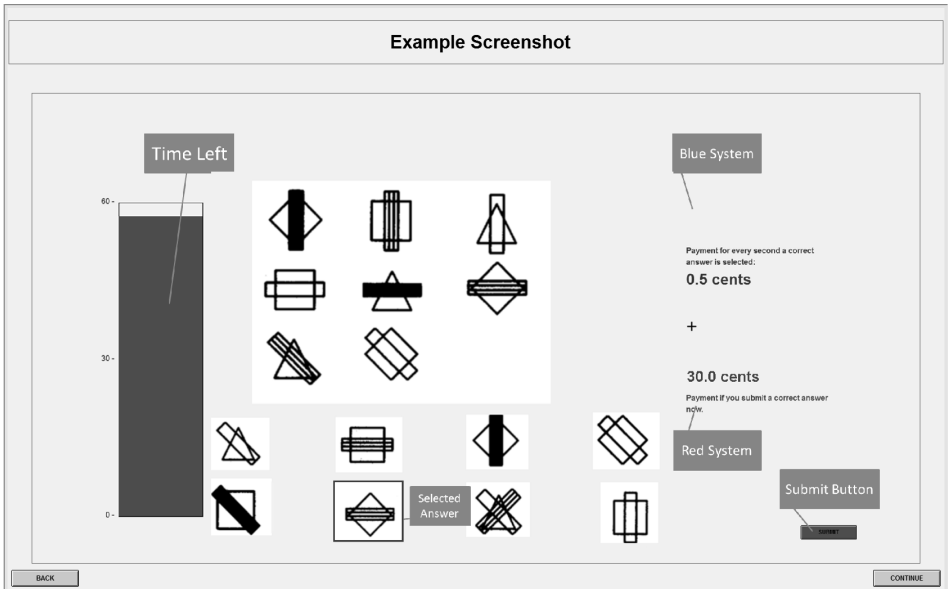

Instructions
The puzzle
You have to find the correct figure from the 8 alternatives at the bottom which fits the empty space of the upper picture.
You always have 60 seconds time to find the correct figure. As soon as you find the correct figure you can select it by clicking on it. You can change your selection in the given time frame as often as you wish. In any case you have to wait until the time is elapsed.
Correct Solution
BACK
CONTINUE

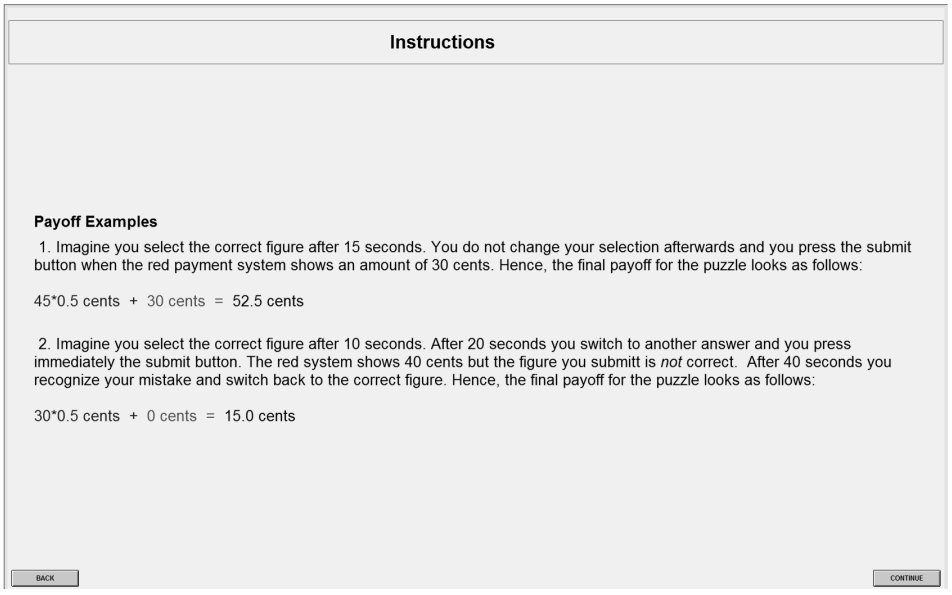

Instructions
Payment
For every puzzle you are paid according to your performance.
Important: There are two independent payment systems for every puzzle. There is the blue payment system and the red payment system. These two payment systems are always active in every puzzle. We will explain these systems to you on the following screens.
BACK
CONTINUE

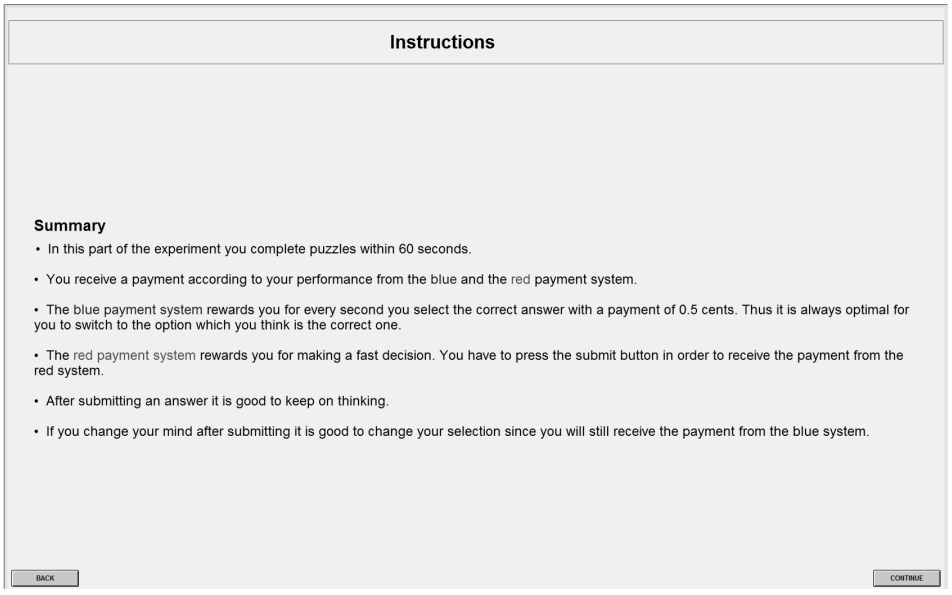

Instructions
Blue Payment System
You receive a payment of 0.5 euro cents for every second for which you select the correct figure. This payment is independent of your submitted choice.
It is always best for you to select a figure immediately at the beginning of every puzzle since there is a chance of 1/8 (12.5%) that this is actually the correct answer. As soon as you think that another figure is the correct answer it is optimal to switch immediately to that figure.
BACK
CONTINUE

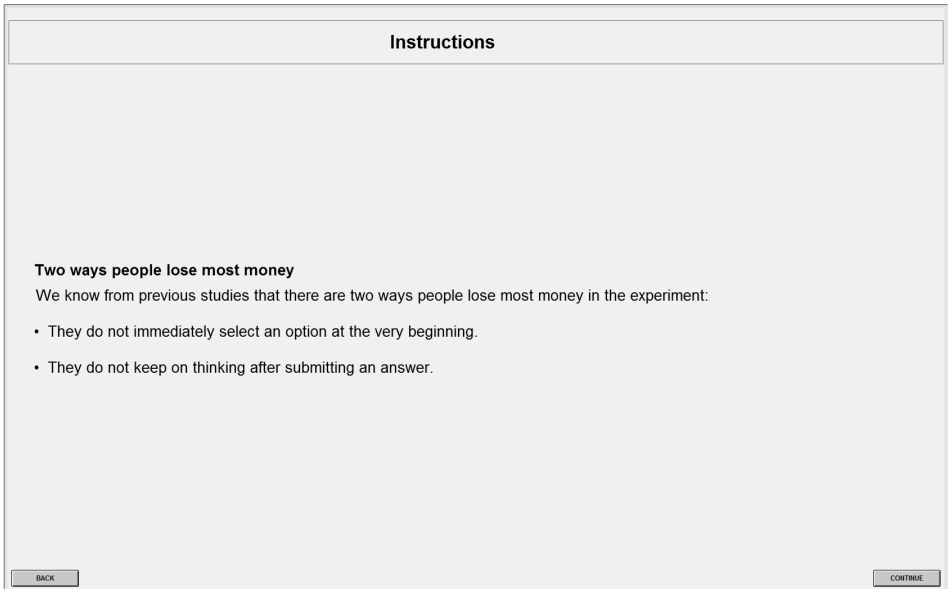

Instructions
Red Payment System
You will only receive money from the red system if you click on the submit button. If your answer is correct you receive the amount which is displayed on the screen.
The longer it takes you to find the correct answer the less you will receive from the red payment system. The amount you receive always starts at a certain amount and decreases over the 60 seconds with a different speed. The decline of the payment stops if you press the submit button.
Overall, there are three different payment regimes with 15 different puzzles each:
1st Regime: It decreases from 25 cents to 5 cents.
2nd Regime: It decreases from 55 cents to 5 cents.
3rd Regime: It decreases from 55 cents to 35 cents.
We will inform you about a change of the regime beforehand.
BACK
CONTINUE

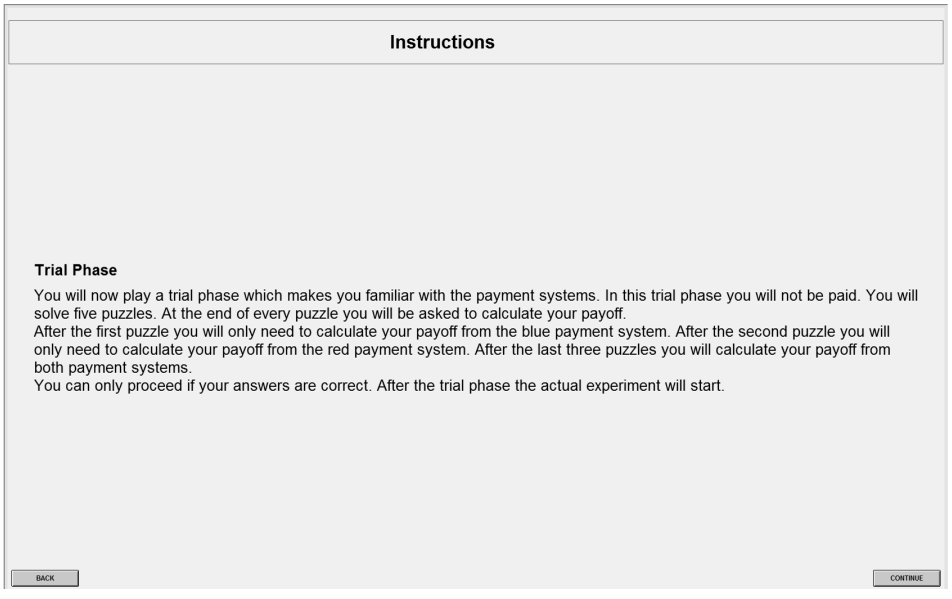

Instructions
Red Payment System and Blue Payment System
Both systems are active at the same time. If you recognize that your submitted answer is wrong it is always optimal to change your selection because you will still receive the payment from the blue payment system.
The next screens show an example with detailed information about your decision screen.
BACK
CONTINUE

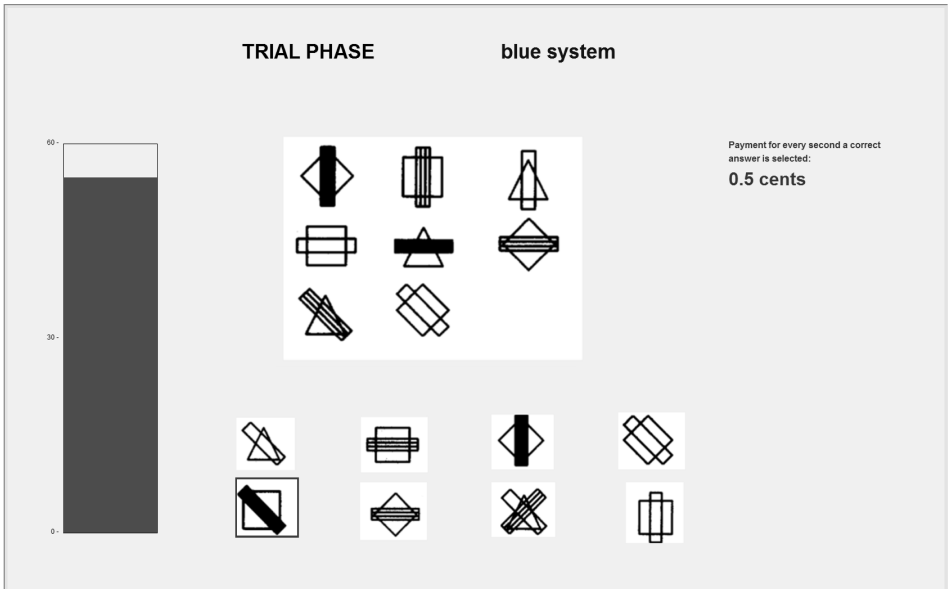

Example Screenshot
Time Left
60 -
30 -
0 -
Blue System
Payment for every second a correct answer is selected:
0.5 cents
+
30.0 cents
Payment if you submit a correct answer now.
Red System
Submit Button
Selected Answer
BACK
CONTINUE

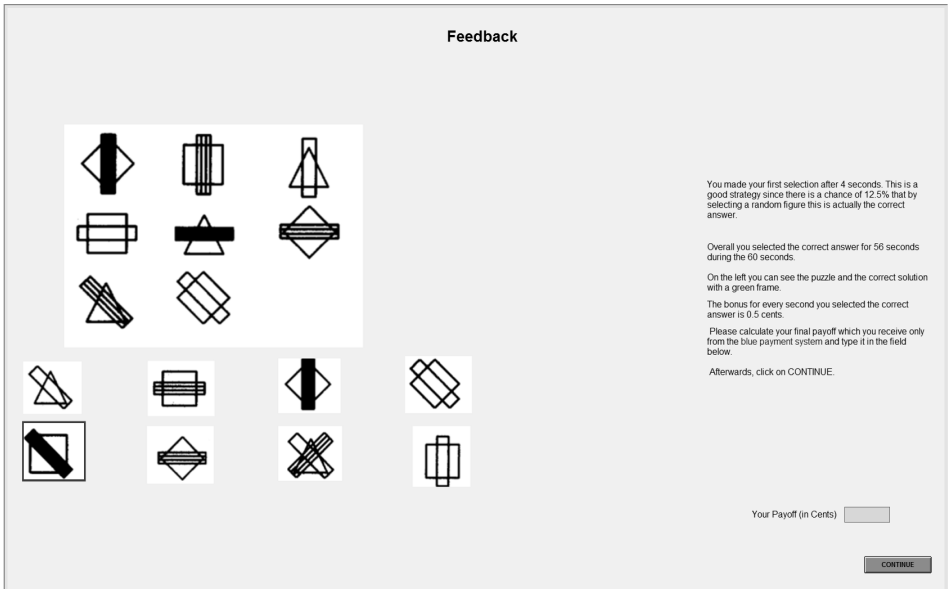

Instructions
Payoff Examples
1. Imagine you select the correct figure after 15 seconds. You do not change your selection afterwards and you press the submit button when the red payment system shows an amount of 30 cents. Hence, the final payoff for the puzzle looks as follows:
45*0.5 cents + 30 cents = 52.5 cents
2. Imagine you select the correct figure after 10 seconds. After 20 seconds you switch to another answer and you press immediately the submit button. The red system shows 40 cents but the figure you submitt is not correct. After 40 seconds you recognize your mistake and switch back to the correct figure. Hence, the final payoff for the puzzle looks as follows:
30*0.5 cents + 0 cents = 15.0 cents
BACK
CONTINUE

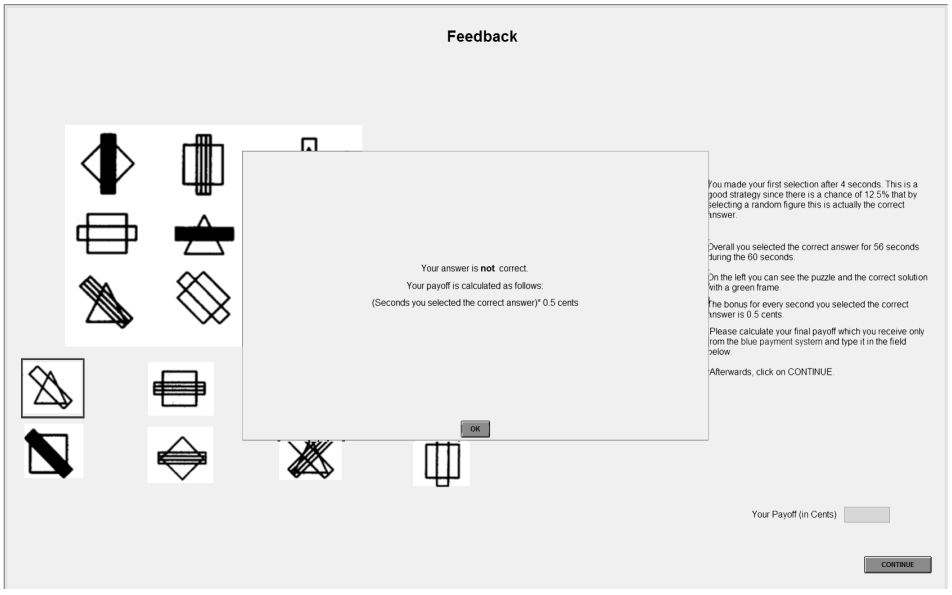

Instructions
Summary
• In this part of the experiment you complete puzzles within 60 seconds.
• You receive a payment according to your performance from the blue and the red payment system.
• The blue payment system rewards you for every second you select the correct answer with a payment of 0.5 cents. Thus it is always optimal for you to switch to the option which you think is the correct one.
• The red payment system rewards you for making a fast decision. You have to press the submit button in order to receive the payment from the red system.
• After submitting an answer it is good to keep on thinking.
• If you change your mind after submitting it is good to change your selection since you will still receive the payment from the blue system.
BACK
CONTINUE

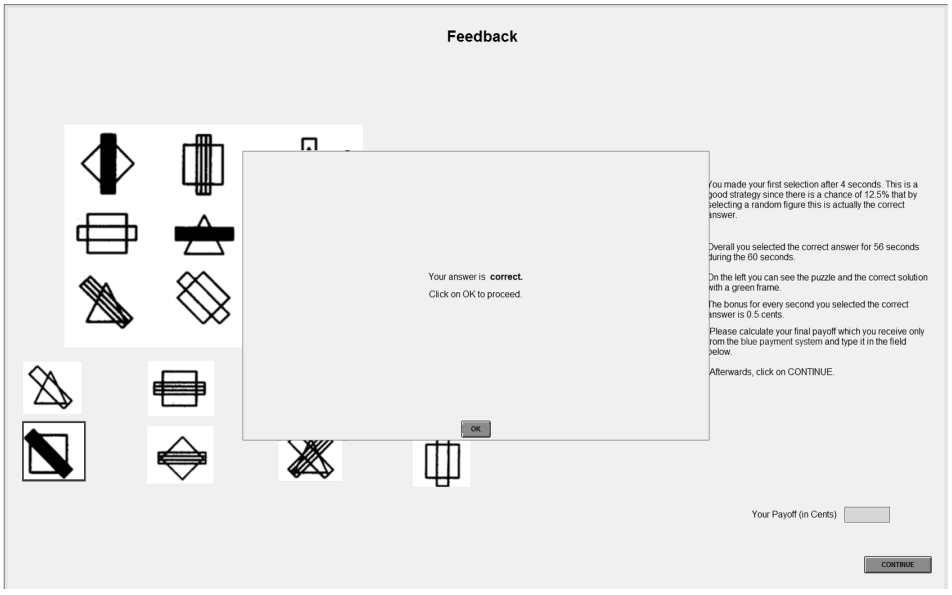

Instructions
Two ways people lose most money
We know from previous studies that there are two ways people lose most money in the experiment:
• They do not immediately select an option at the very beginning.
• They do not keep on thinking after submitting an answer.
BACK
CONTINUE

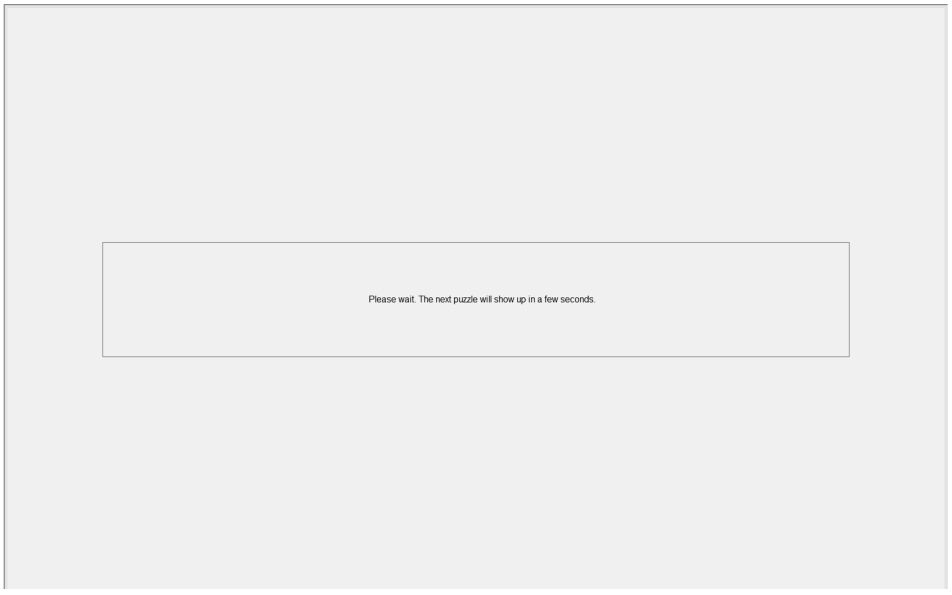

Instructions
Trial Phase
You will now play a trial phase which makes you familiar with the payment systems. In this trial phase you will not be paid. You will solve five puzzles. At the end of every puzzle you will be asked to calculate your payoff.
After the first puzzle you will only need to calculate your payoff from the blue payment system. After the second puzzle you will only need to calculate your payoff from the red payment system. After the last three puzzles you will calculate your payoff from both payment systems.
You can only proceed if your answers are correct. After the trial phase the actual experiment will start.
BACK
CONTINUE

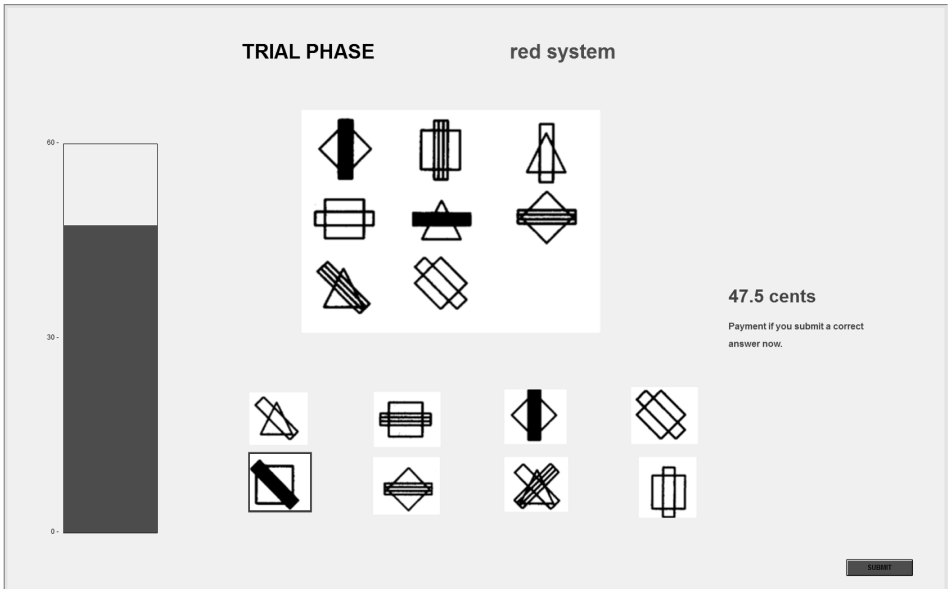

TRIAL PHASE
blue system
60 -
30 -
0 -
Payment for every second a correct answer is selected:
0.5 cents

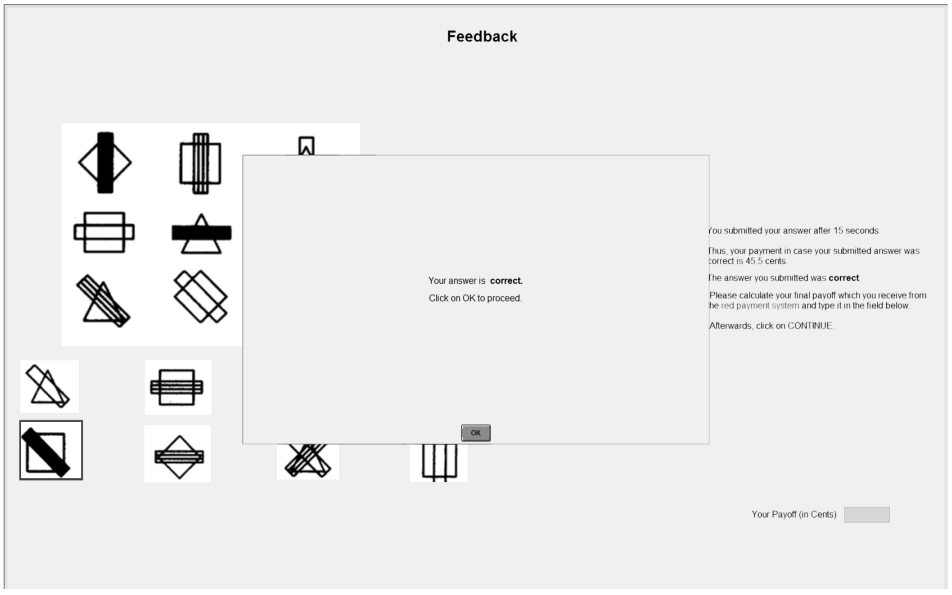

Feedback
You made your first selection after 4 seconds. This is a good strategy since there is a chance of 12.5% that by selecting a random figure this is actually the correct answer.
Overall you selected the correct answer for 56 seconds during the 60 seconds.
On the left you can see the puzzle and the correct solution with a green frame.
The bonus for every second you selected the correct answer is 0.5 cents.
Please calculate your final payoff which you receive only from the blue payment system and type it in the field below.
Afterwards, click on CONTINUE.
Your Payoff (in Cents)
CONTINUE

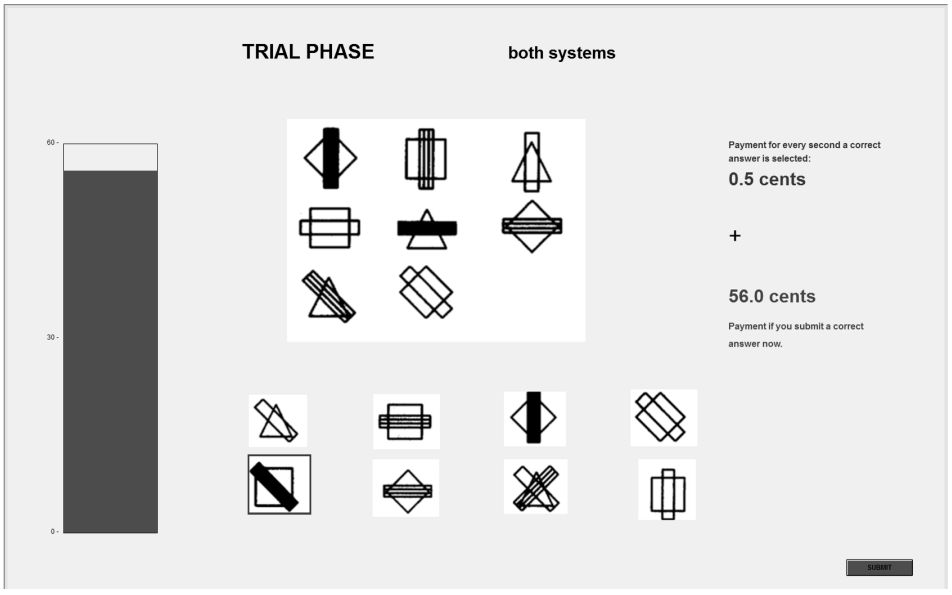

Feedback
Your answer is not correct.
Your payoff is calculated as follows:
(Seconds you selected the correct answer)* 0.5 cents
OK
You made your first selection after 4 seconds. This is a good strategy since there is a chance of 12.5% that by selecting a random figure this is actually the correct answer.
Overall you selected the correct answer for 56 seconds during the 60 seconds.
On the left you can see the puzzle and the correct solution with a green frame.
The bonus for every second you selected the correct answer is 0.5 cents.
Please calculate your final payoff which you receive only from the blue payment system and type it in the field below.
Afterwards, click on CONTINUE.
Your Payoff (in Cents)
CONTINUE

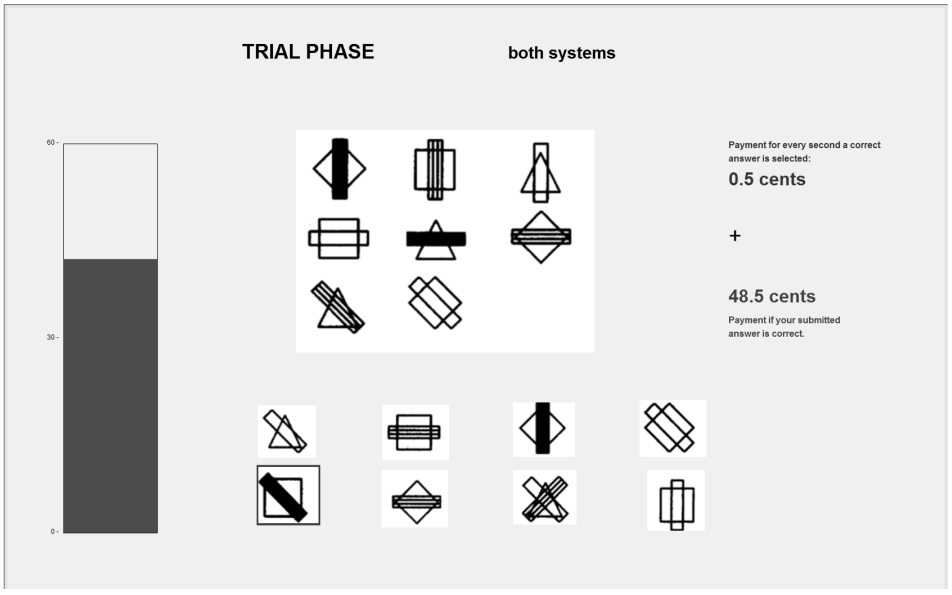

Feedback
Your answer is correct.
Click on OK to proceed.
OK
You made your first selection after 4 seconds. This is a good strategy since there is a chance of 12.5% that by selecting a random figure this is actually the correct answer.
Overall you selected the correct answer for 56 seconds during the 60 seconds.
On the left you can see the puzzle and the correct solution with a green frame.
The bonus for every second you selected the correct answer is 0.5 cents.
Please calculate your final payoff which you receive only from the blue payment system and type it in the field below.
Afterwards, click on CONTINUE.
Your Payoff (in Cents)
CONTINUE

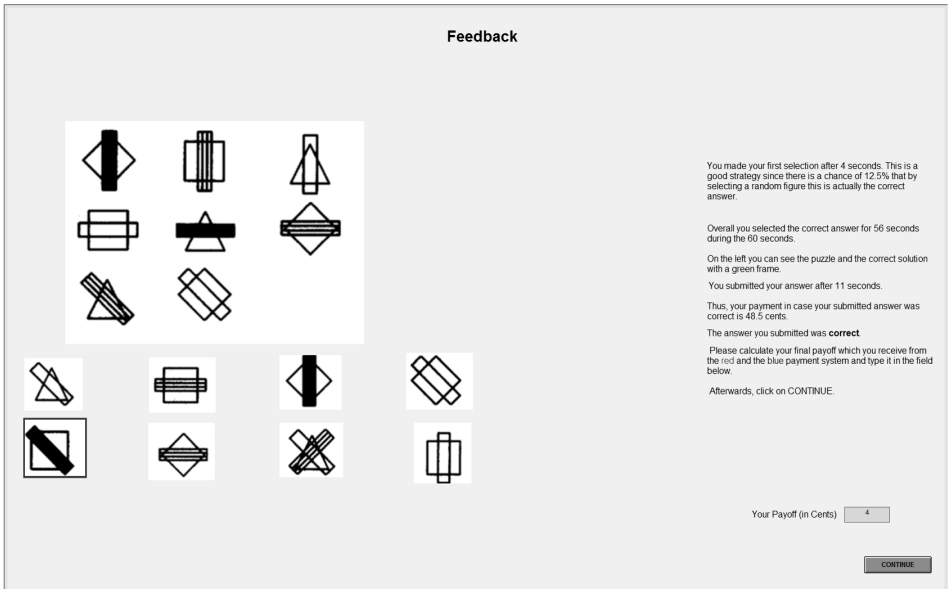

Please wait. The next puzzle will show up in a few seconds.

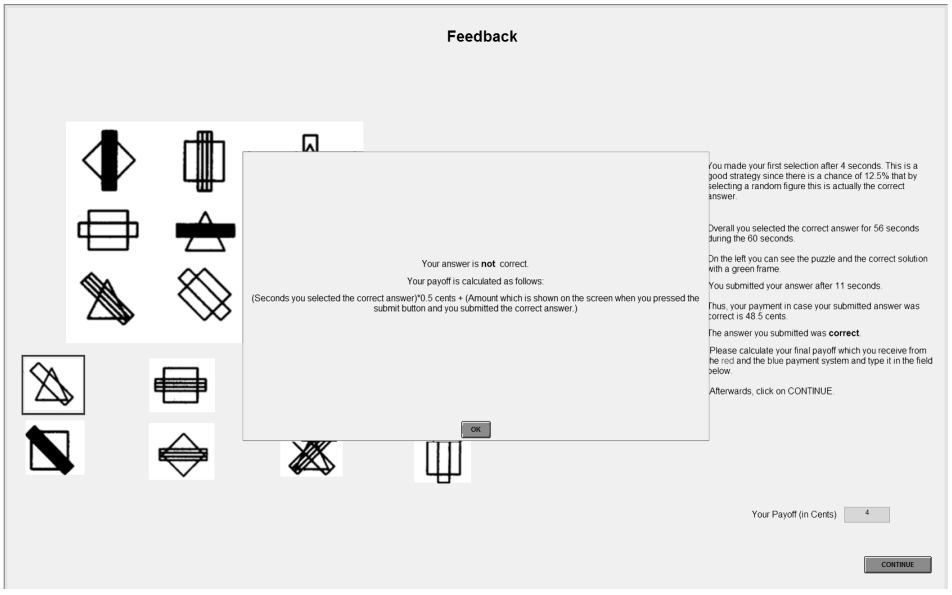

TRIAL PHASE
red system
60 -
30 -
0 -
47.5 cents
Payment if you submit a correct answer now.
SUBMIT

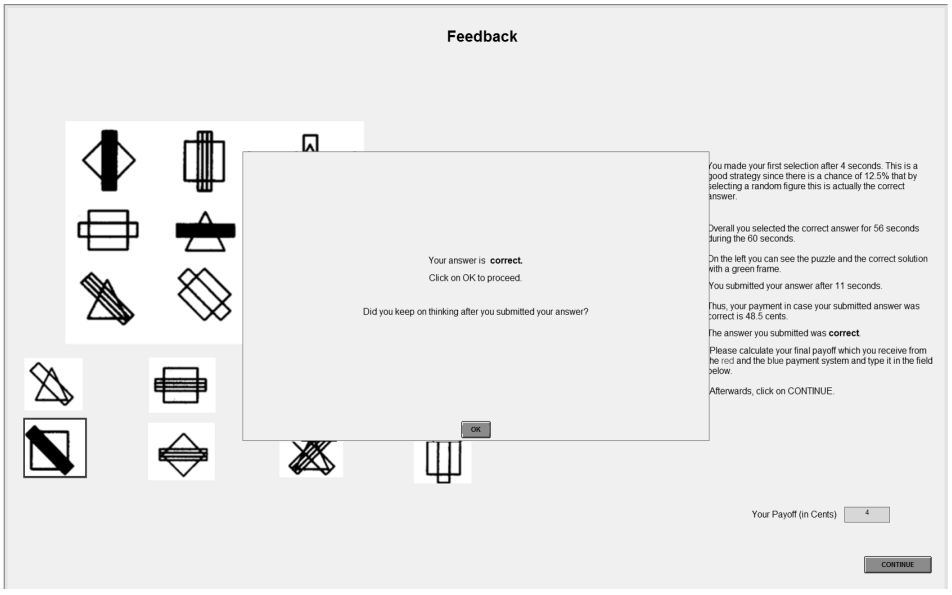

Feedback
Your answer is correct.
Click on OK to proceed.
OK
You submitted your answer after 15 seconds.
Thus, your payment in case your submitted answer was correct is 45.5 cents.
The answer you submitted was correct.
Please calculate your final payoff which you receive from the red payment system and type it in the field below.
Afterwards, click on CONTINUE.
Your Payoff (in Cents)

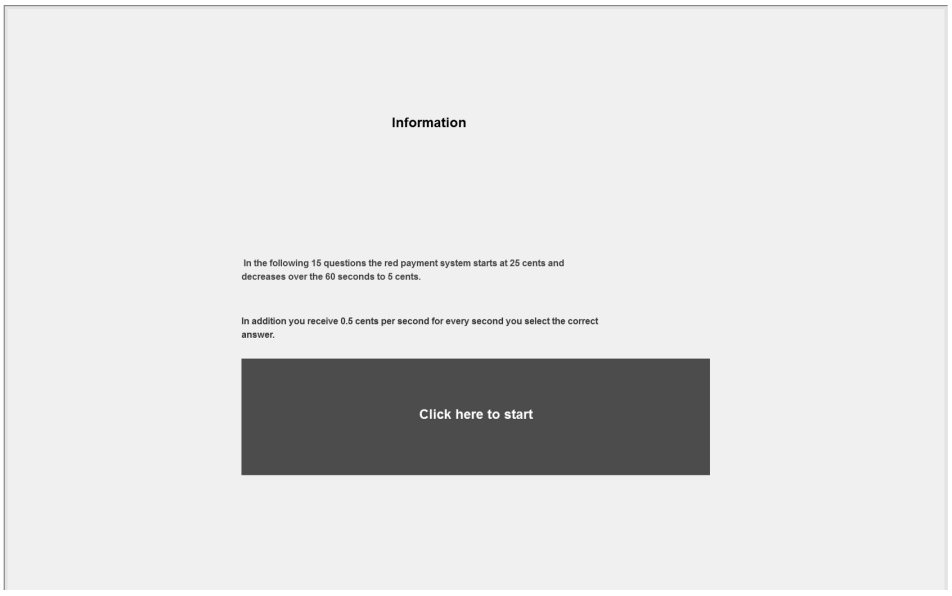

TRIAL PHASE
both systems
60 -
30 -
0 -
Payment for every second a correct answer is selected:
0.5 cents
+
56.0 cents
Payment if you submit a correct answer now.
SUBMIT


TRIAL PHASE
both systems
60 -
30 -
0 -
Payment for every second a correct answer is selected:
0.5 cents
+
48.5 cents
Payment if your submitted answer is correct.


Feedback
You made your first selection after 4 seconds. This is a good strategy since there is a chance of 12.5% that by selecting a random figure this is actually the correct answer.
Overall you selected the correct answer for 56 seconds during the 60 seconds.
On the left you can see the puzzle and the correct solution with a green frame.
You submitted your answer after 11 seconds.
Thus, your payment in case your submitted answer was correct is 48.5 cents.
The answer you submitted was correct.
Please calculate your final payoff which you receive from the red and the blue payment system and type it in the field below.
Afterwards, click on CONTINUE.
Your Payoff (in Cents)
4
CONTINUE


Feedback
Your answer is **not** correct.
Your payoff is calculated as follows:
(Seconds you selected the correct answer)*0.5 cents + (Amount which is shown on the screen when you pressed the submit button and you submitted the correct answer.)
OK
You made your first selection after 4 seconds. This is a good strategy since there is a chance of 12.5% that by selecting a random figure this is actually the correct answer.
Overall you selected the correct answer for 56 seconds during the 60 seconds.
On the left you can see the puzzle and the correct solution with a green frame.
You submitted your answer after 11 seconds.
Thus, your payment in case your submitted answer was correct is 48.5 cents.
The answer you submitted was **correct**.
Please calculate your final payoff which you receive from the red and the blue payment system and type it in the field below.
Afterwards, click on CONTINUE.
Your Payoff (in Cents)
4
CONTINUE


Feedback
Your answer is **correct.**
Click on OK to proceed.
Did you keep on thinking after you submitted your answer?
OK
You made your first selection after 4 seconds. This is a good strategy since there is a chance of 12.5% that by selecting a random figure this is actually the correct answer.
Overall you selected the correct answer for 56 seconds during the 60 seconds.
On the left you can see the puzzle and the correct solution with a green frame.
You submitted your answer after 11 seconds.
Thus, your payment in case your submitted answer was correct is 48.5 cents.
The answer you submitted was **correct**.
Please calculate your final payoff which you receive from the red and the blue payment system and type it in the field below.
Afterwards, click on CONTINUE.
Your Payoff (in Cents)
4
CONTINUE


Information
In the following 15 questions the red payment system starts at 25 cents and decreases over the 60 seconds to 5 cents.
In addition you receive 0.5 cents per second for every second you select the correct answer.
Click here to start

A.2 Additional Results on Numerical Tasks

Figures A.3 and A.4 present additional results for the numerical tasks. The figures are similar to the ones presented for Raven matrices in Section 2.5. The results obtained are similar to the results displayed in Figures 2.6 and 2.7.

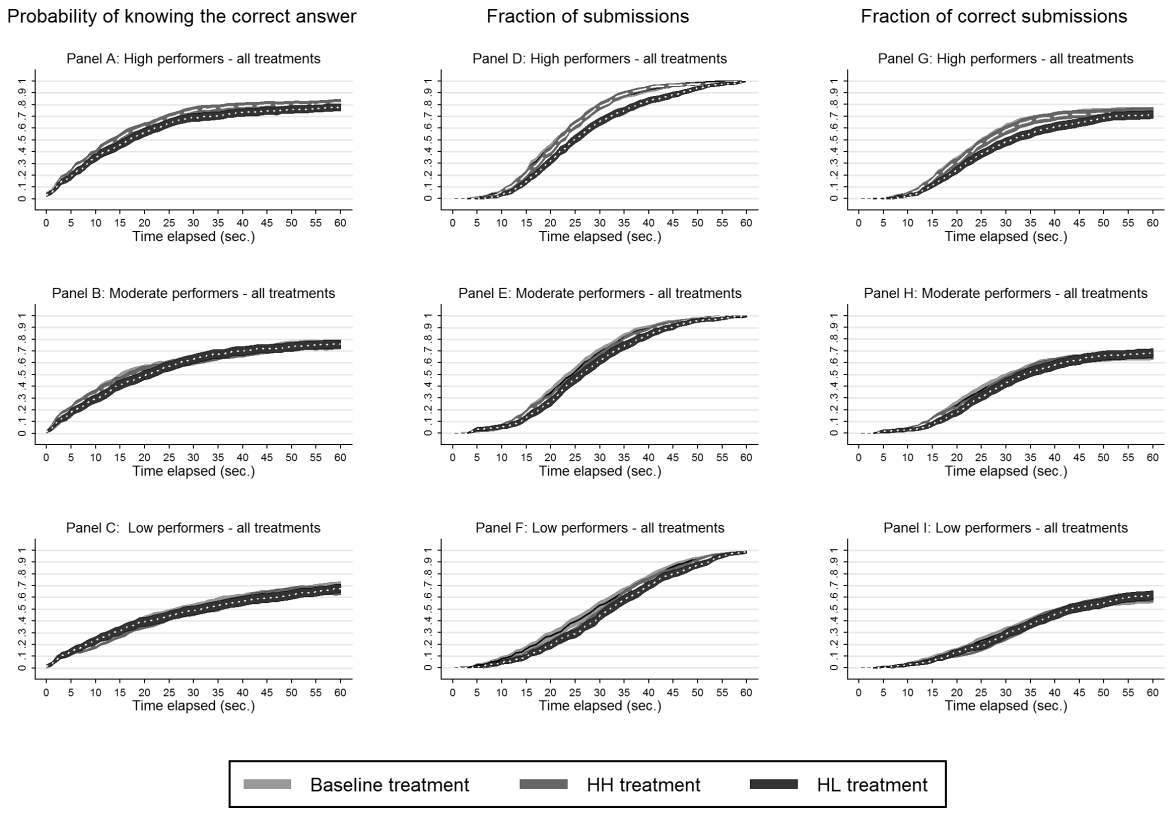


Figure A.3: Heterogeneity in Performance between Numerical Tasks

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 gray areas indicate the 95% confidence bounds.

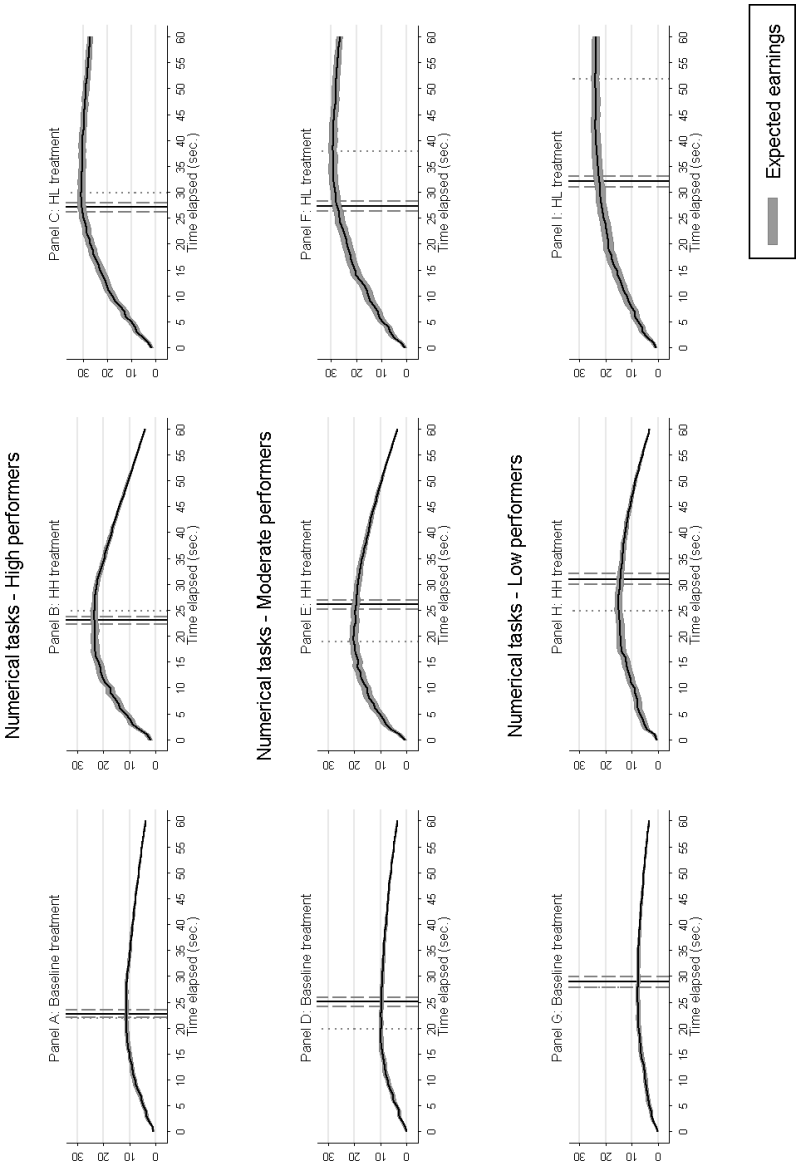


Figure A.4: Expected Earnings and Submission Times of Numerical Tasks by Performance Type

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

Appendix B

Appendix to Chapter 3

B.1 Additional Graphs

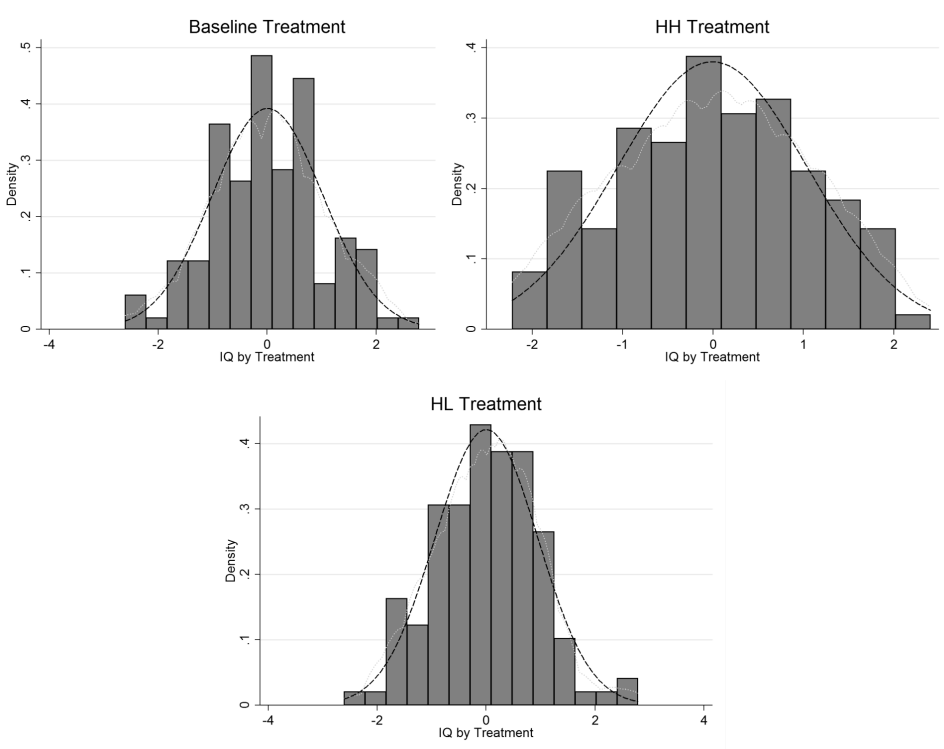


Figure B.1: Distribution of IQ by treatment

Note. The figure shows the distribution of the standardized number correctly submitted answers in all treatments. The black dashed line plot the normal distribution. The gray dotted lines plots the kernel density plot of this distribution. A Shapiro-Wilk test cannot reject the null that the data is normally distributed (Low time pressure Low stakes treatment p= 0.98; High time pressure High stakes treatment, p= 0.14, High incentives Low stakes treatment, p= 0.98).

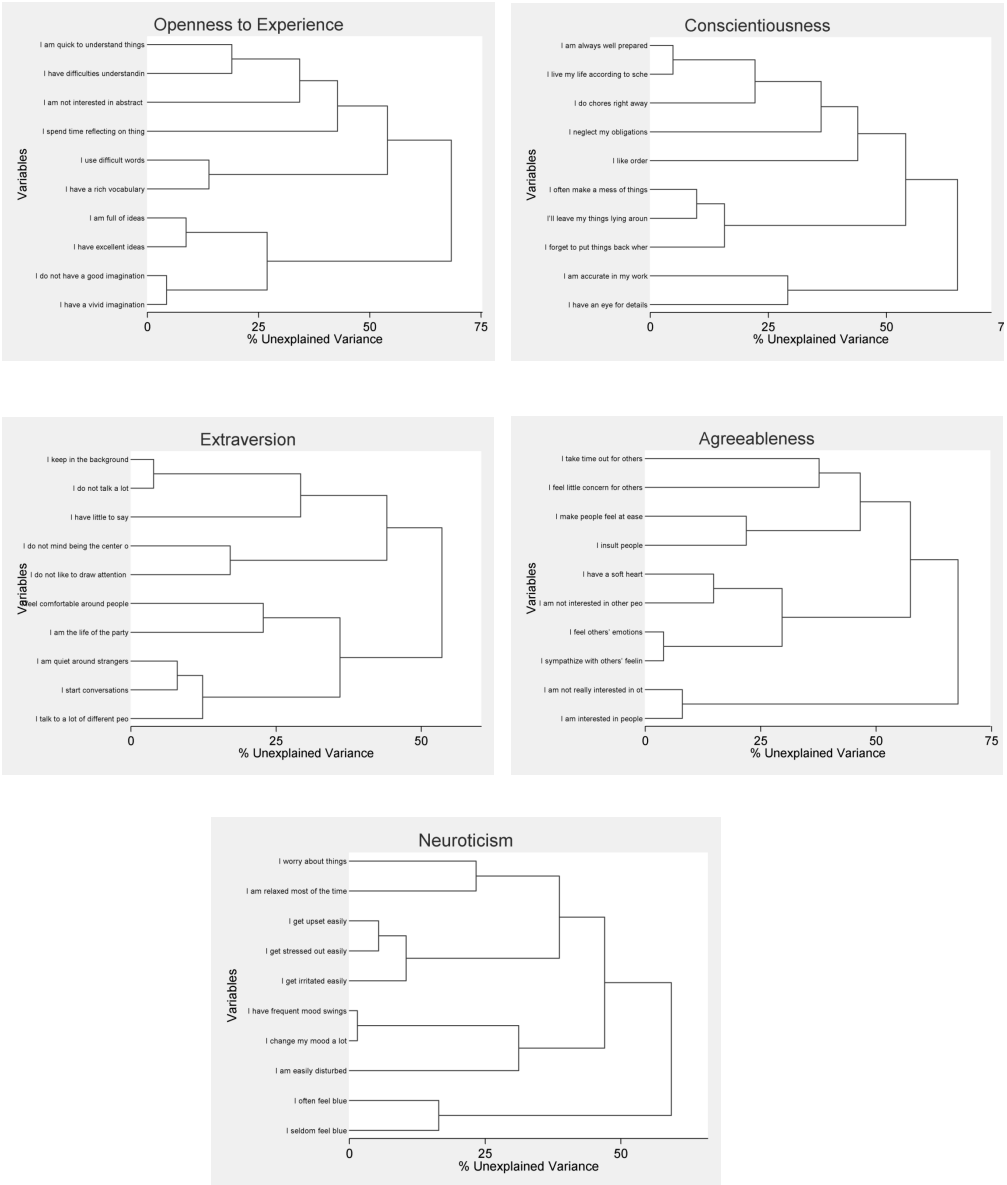


Figure B.2: Dendograms of the Big Five

Appendix C

Appendix to Chapter 4

C.1 Intelligence Test

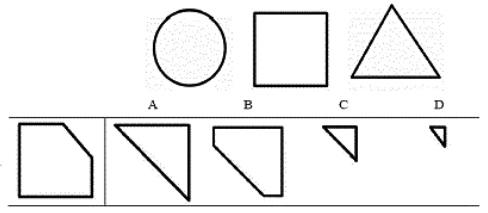
Eerste onderdeel

In dit onderdeel komen vijf soorten vragen voorbij. Hieronder zie je van elke soort vraag een voorbeeld. Lees dit dus even goed door, dan weet je dadelijk precies wat je moet doen.

Taak 1: Figuren maken

Bij deze test zie je voor de streep telkens een figuur die niet af is. Deze figuur moet een cirkel, vierkant of driehoek worden met de figuren die achter de streep staan. Let op: deze figuren kunnen linksom of rechtsom gedraaid zijn.

Voorbeeld:

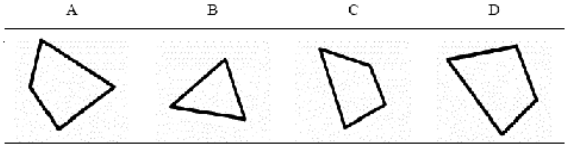


Het goede antwoord is hier C, want daarmee wordt het een vierkant.

Taak 2: Welk figuur hoort er niet bij?

Hieronder zie je vier figuren. Boven de figuren staan de letters A, B, C en D. Eén figuur is een anders dan de anderen. Zoek de letter van de figuur die anders is dan de rest.

Voorbeeld:



Je ziet in het voorbeeld dat alle figuren uit vier lijnen bestaan behalve B, want die heeft er maar drie. B is hier dus het goede antwoord.

Figure C.1: Examples of the IQ test

Taak 3: Getallen invullen

In deze taak zie je steeds een reeks getallen waarin telkens één getal is weggelaten. Je moet dan dat getal kiezen uit de vier mogelijke antwoorden dat past in het lege hokje.

Voorbeeld:

4	7	6	6	9	8	8	--

Het goede antwoord moet hier zijn **11**.

Taak 4: Welk woord hoort erbij?

Voor de streep staan drie woorden die bij elkaar horen. Achter de streep staan vier woorden, waarvan er één past bij de drie woorden voor de streep. Bij de woorden achter de streep horen letters; die staan er boven: A, B, C en D.

Voorbeeld:

			A	B	C	D
lopen	wandelen	slenteren	liggen	zitten	hurken	stappen

Welk woord achter de streep hoort bij de drie woorden die voor de streep staan? Dat is het woord stappen, want stappen is ook lopen. Welke letter hoort daar bij? Dat is D.

Taak 5: Welk woord hoort op de stipjes?

Links voor de verticale streepjes staan twee woorden die bij elkaar horen. Achter de verticale streepjes staat één woord waarbij jij het woord moet vinden dat op de stippeltjes moet komen. Dit woord staat bij de woorden onder A, B, C, of D.

Voorbeeld:

				A	B	C	D
bal	-	appel	dobbelsteen-	ster	peer	kubus	beker

Welk woord hoort net zo bij dobbelsteen als appel bij bal? Dat is een kubus, want een kubus is net als een dobbelsteen een soort vierkant doosje, terwijl een bal en een appel allebei rond zijn. C is dus het goede antwoord.

Figure C.2: Examples of the IQ test

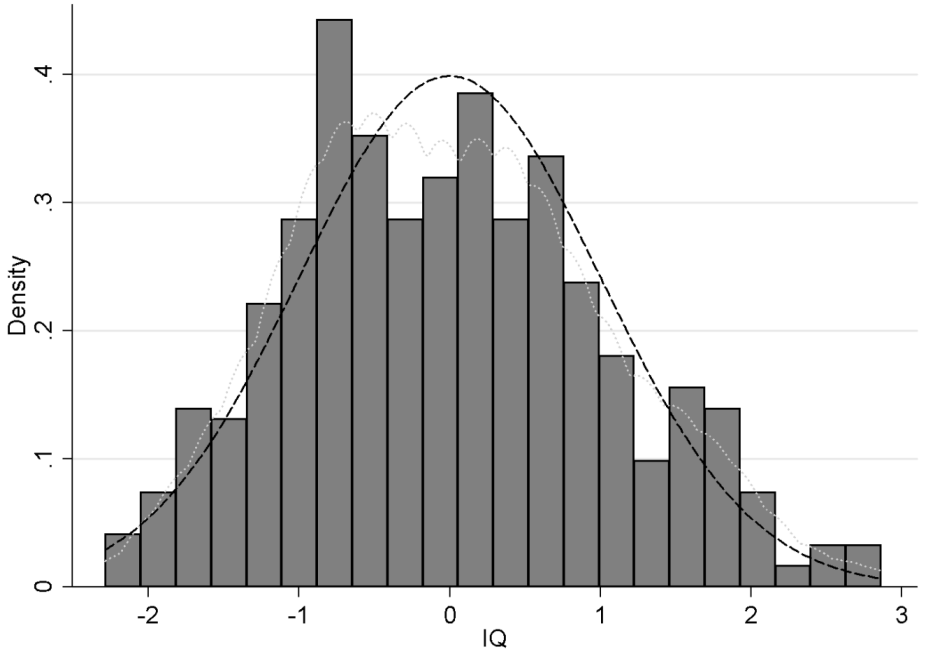


Figure C.3: Distribution of IQ in the sample

Note. The measure is standardized with mean 0 and standard deviation 1.

C.2 Patience measure

Figure C.4 shows the distribution of our patience measure. Higher values on that scale indicate lower degrees of patience. We observe that 25 percent of the people reveal an internal rate of return which is smaller or equal to 5% per annum. On the other hand 25% of our subjects reveal to have an internal rate of return which is greater than 115% percent. This means that they still chose 100 today if we offered them 215 in a year from today. Note that the reported numbers at the extremes of the distribution serve as upper and lower bounds of an individuals' internal rate of return. The gray dotted line indicates the kernel density estimate and the dashed back line the kernel density of the normal distribution. As it can be seen in Figure C.4 the distribution is not normal, but tends to be bimodal. A Shapiro-Wilk test rejects the null hypothesis that the patience measure is normally distributed ($p < 0.01$).

The general picture that emerges from Figure C.4 is that there is substantial variation in the discount rate. We interpret the reported internal rates of return as an individuals degree of patience. The amounts we presented to our subjects are still in a range that 15 year olds are able to understand the differences. Hence

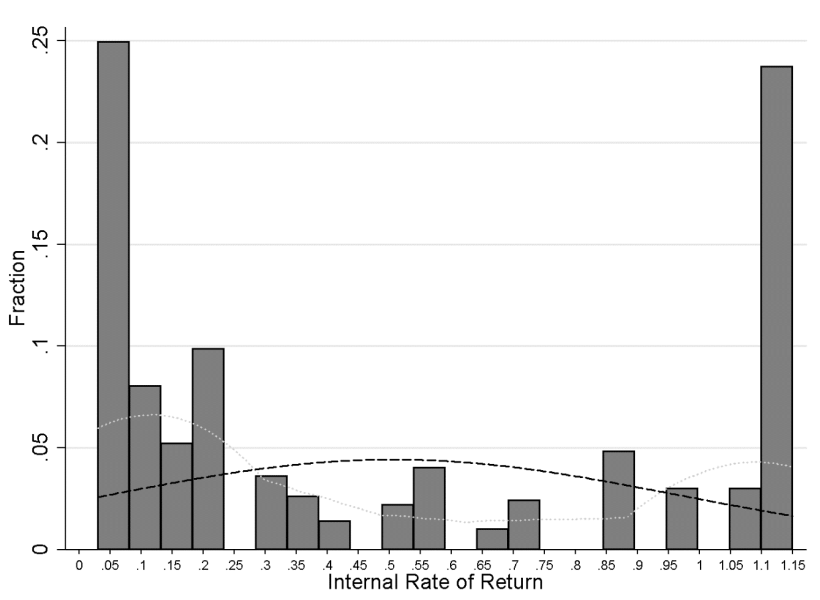


Figure C.4: Distribution of the patience measure.

Note. The gray dotted line indicates the kernel density estimates. The black dashed line shows the corresponding normal distribution.

it is reasonable to assume that a person who chooses €100 today instead of €215 in a year from today is classified as impatient compared to a person who chooses €103 in a year from today instead of choosing €100 today.

C.3 Correlations patterns between IQ, Patience, High- and Low Stake Achievement test

This section reports various scatter plots to further explore the correlation pattern between our IQ measure, the patience measures and the high and low achievement test scores. All measures are standardized. The idea is to explore whether there are potential non-linear relationships between the variables. However, as the scatterplots suggests a linear relationship as assumed in the regressions represents the relationship between the variables of interest quite well.

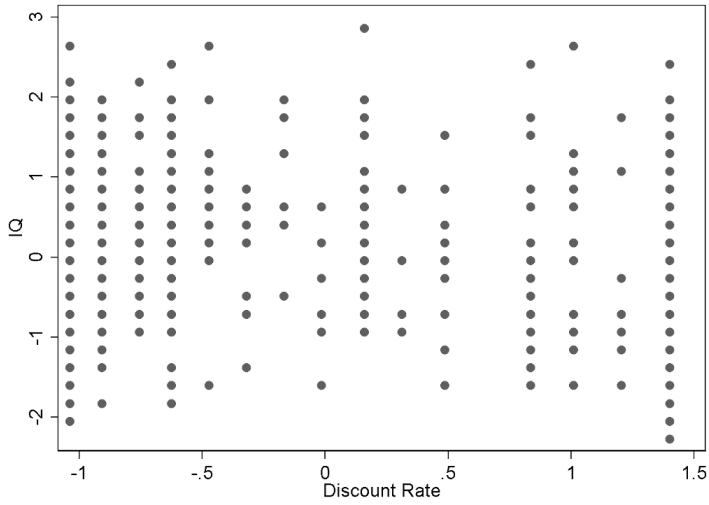


Figure C.5: Correlation between IQ and Patience.

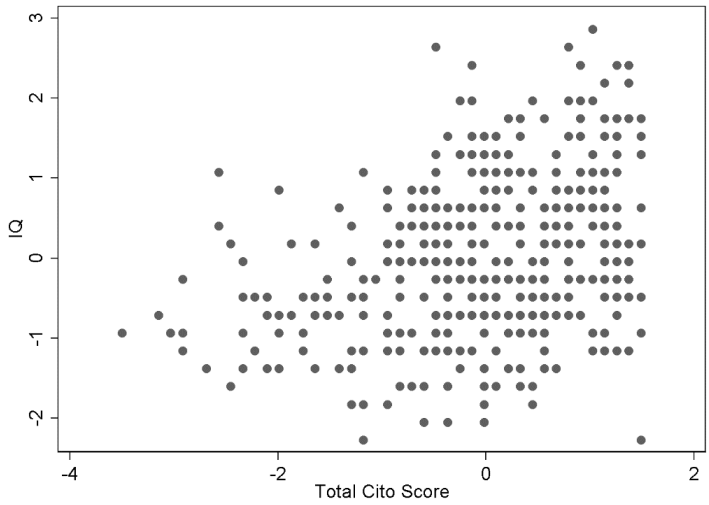


Figure C.6: Correlation between IQ and High-Stake Test Score.

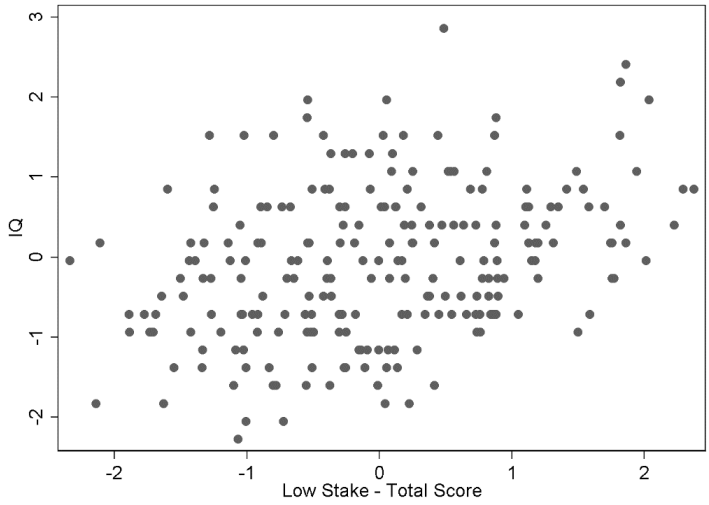


Figure C.7: Correlation between IQ and Low-Stake Test Score.

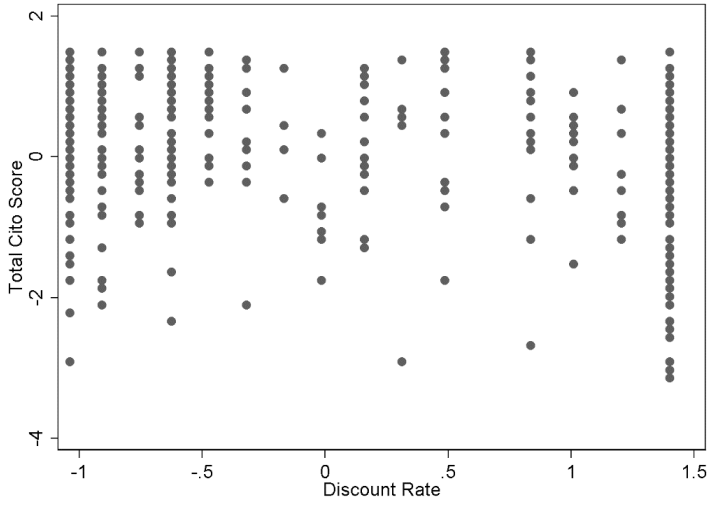


Figure C.8: Correlation between Patience and High-Stake Test Score.

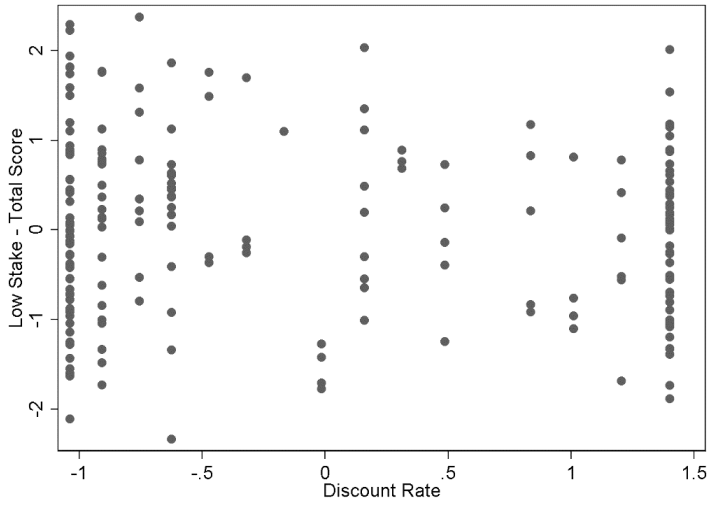


Figure C.9: Correlation between IQ and Low-Stake Test Score.

C.4 Additional Analysis of the Big Five

This section shows are more in depth analysis of the Big Five personality traits. The Big Five personality traits capture the facets Openness to Experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism. In short Openness is a trait which captures intellect and curiosity. Conscientiousness captures the tendency to be hard working and organized. Extraversion reflects the tendency to be outgoing and approaching other people. Agreeableness describes the tendency to cooperative and unselfish. Neuroticism reflects emotional stability. For a detailed description see for instance (Almlund et al. (2011); Goldberg (1992)) In Table C.1 we present the items which were used to measure each of the traits. The rows below that show measures for their reliability. First we document Cronbachs alpha which is a measure for the consistency of the items. All values are above .6 which indicates a reasonable degree of consistency. Factor analysis extracts one latent factor per measured trait using Kaisers criterion. Next, we document the proportion of the eigenvalue of the first factor. Lastly, Figure C.10 shows a dendogram of the clustering using all 22 items of the personality questionnaire. The number in the graphs correspond to the respective items in Table C.1. Except for the second item of the trait agreeableness the grouping is as expected.

Table C.1: Items & Reliability of the Big Five personality traits

	Openness to Experience	Item No.	**Agreeableness**	Item No.
	I use difficult words	1	I try to help people	1
	I am full of ideas	2	I am interested in others	2
	I am quick to understand things	3	I sympathize with others' feelings	3
	I do not have a good imagination (R)	4	I am friendly	4
	I have a rich vocabulary	5		
Cronbach's alpha	0.6165		0.7864	
Proportion	120.60%		118.68%	
	Conscientiousness		**Neuroticism**	
	I do chores right away	1	I get upset easily	1
	I leave my things lying around (R)	2	I get stressed out easily	2
	I keep appointments	3	I have frequent mood swings	3
	I sometimes forget that I need to do something. (R)	4	I often feel blue.	4
	I am accurate in my work	5		
Cronbach's alpha	0.6827		0.7667	
Proportion	120.02%		122.96%	
	Extraversion			
	I talk a lot.	1		
	I am quiet around strangers. (R)	2		
	I am the life of the party.	3		
	I find it nice to be around people	4		
Cronbach's alpha	0.6863			
Proportion	137.23%			

Note. The table shows the translated items of the personality questionnaire. An (R) indicates that this item is a reversed measure of the trait. We obtained one factor per personality trait using Kaisers criterion to drop all components with an eigenvalue lower than 1. Proportion" is the proportion of the variance explained by this one factor.

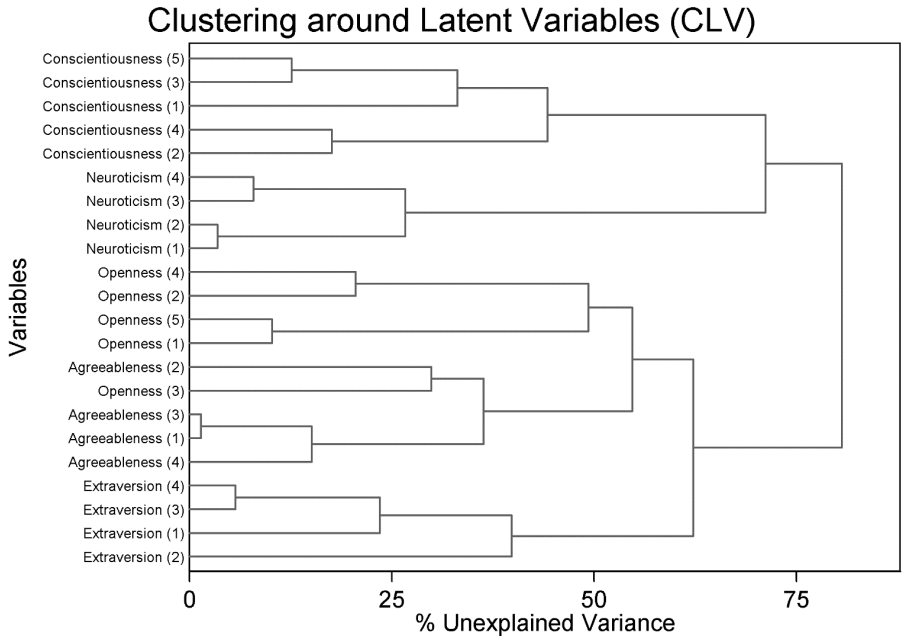


Figure C.10: Dendogram of the Big Five personality traits.

Appendix D

Appendix to Chapter 5

D.1 Screenshots

This section contains the two most important screen shots of each of the two treatments and a screenshot of the calculation task. All the other screenshots can be found in the online appendix of Dohmen and Falk (2011).

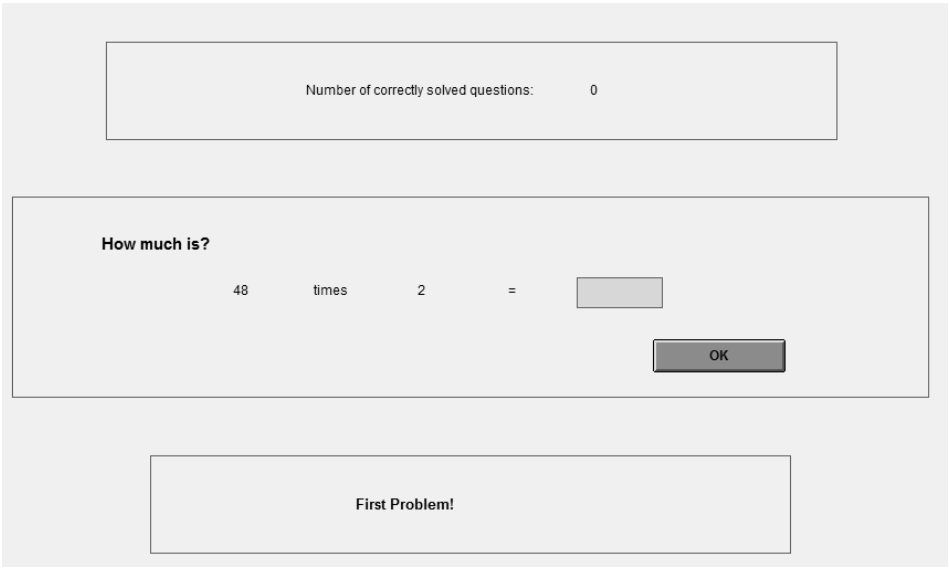


Figure D.1: The calculation task.

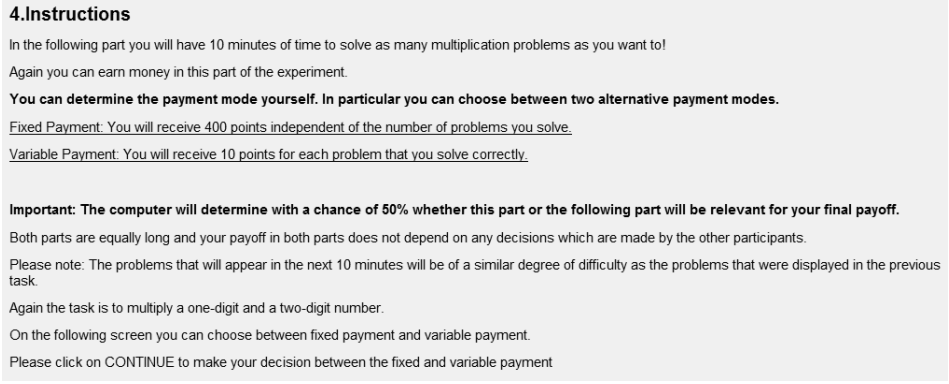


Figure D.2: Instructions of the ENDO treatment.

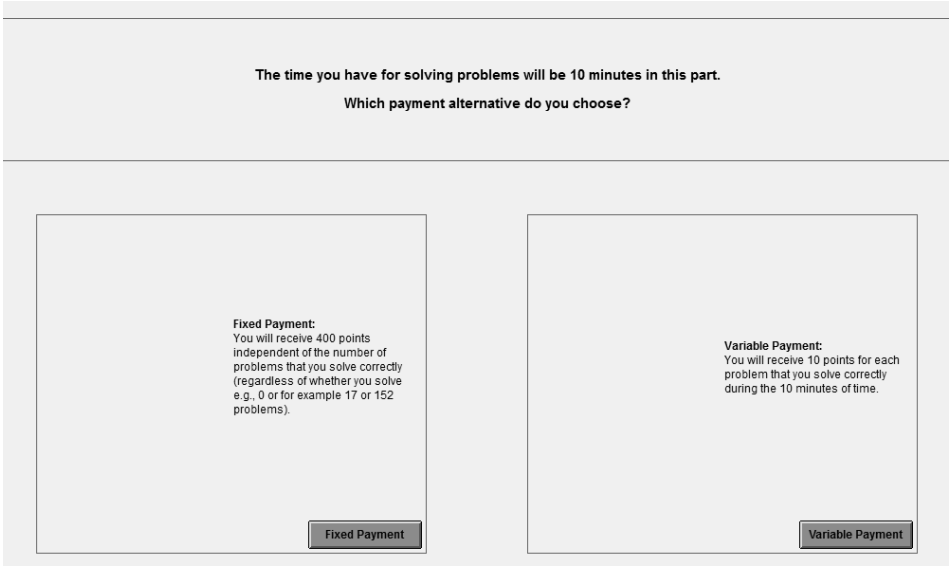


Figure D.3: Decision screen of the ENDO treatment.

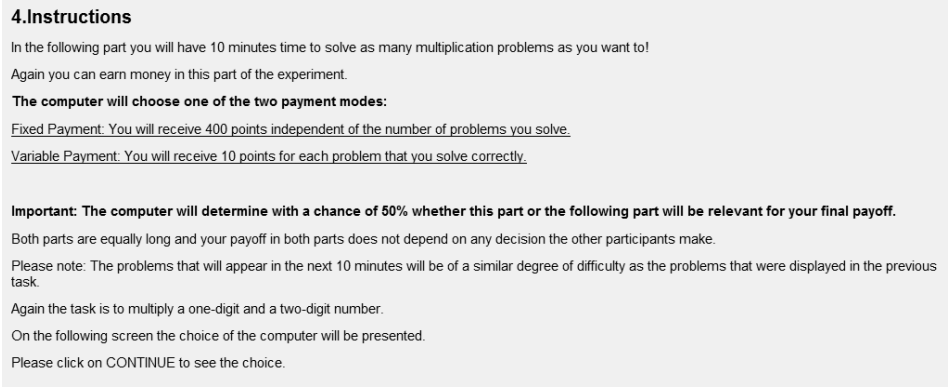


Figure D.4: Instructions of the EXO treatment.

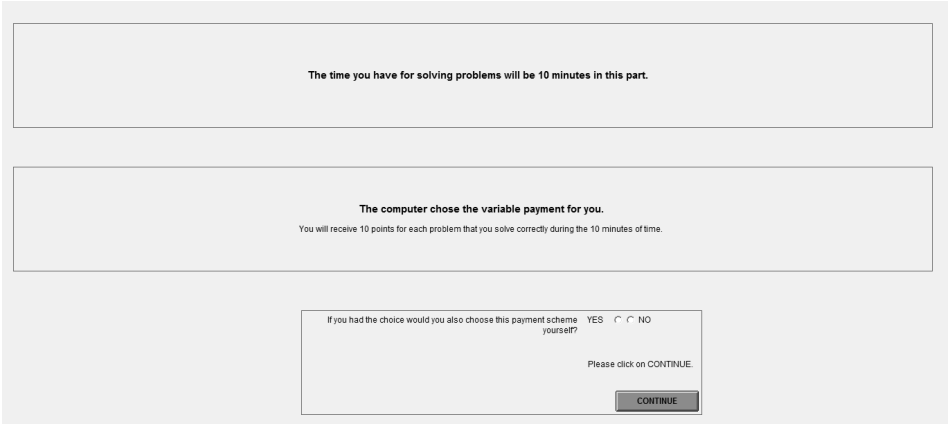


Figure D.5: Information screen of the EXO treatment.

D.2 Additional Results

This section presents additional results of the experiment. Figure D.6 is the equivalent to Figure 5.5 but for the EXO treatment. Figure D.7 shows a randomization check.

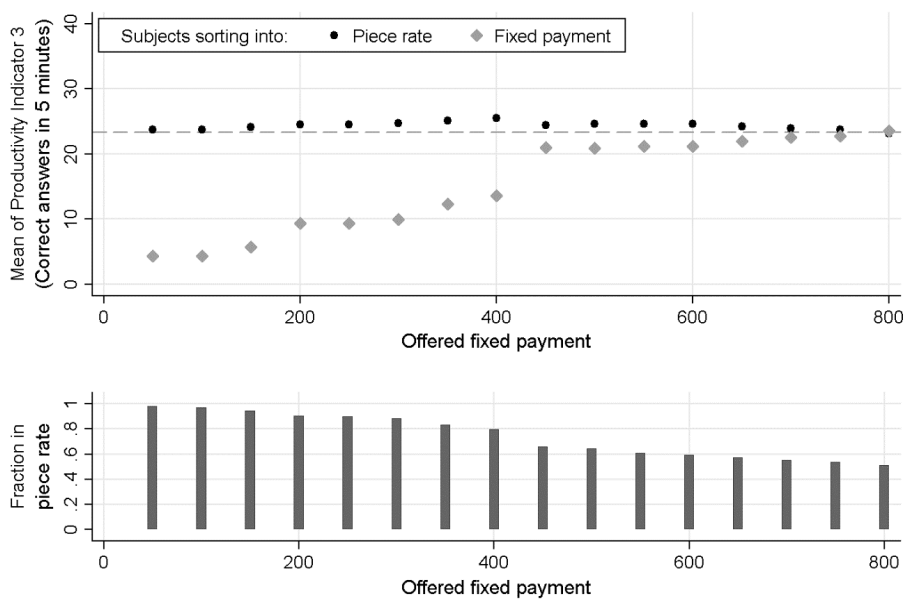


Figure D.6: Sorting choices depending on productivity and offered fixed payment in the EXO treatment (c.f. Dohmen and Falk (2011))

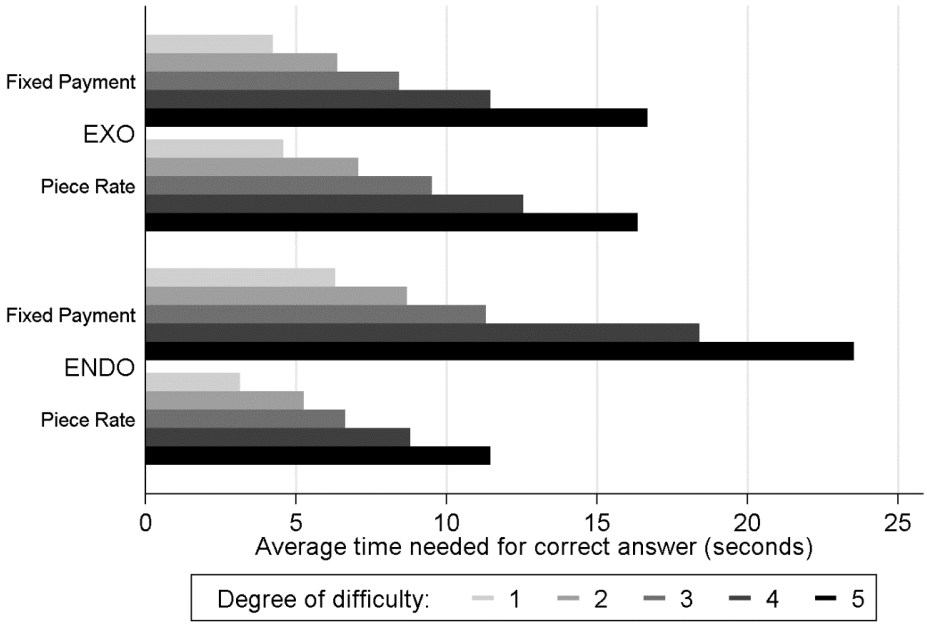


Figure D.7: Average calculation time and treatment.

Table D.1: Determinants of choosing a piece rate remuneration in the EXO treatment

	(1)	(2)	(3)	(4)	(5)
Productivity	-0.0002	0.0001	0.002	0.003	0.0041
	(.0034)	(.0035)	(.0035)	(.0036)	(.0032)
1 if Female		0.0548	0.0831	0.1282	0.1212
		(.0722)	(.0698)	(.0865)	(.1012)
Risk Attitude			-0.0456	-0.0213	-0.0219
			(.0409)	(.0441)	(.0466)
Trust			0.0378***	0.0209**	0.0194**
			(.0102)	(.0090)	(.0082)
Reciprocity			0.0271	0.0375	0.0398
			(.0231)	(.0273)	(.0293)
Cognitive Ability				0.0483	0.0434
				(.0336)	(.0382)
Observations	165	165	165	145	145
Controls	YES	YES	YES	YES	YES
Big Five	NO	NO	NO	NO	YES

Note: The table shows marginal effects of a probit estimation evaluated at the mean. The dependent variable takes the value one if a piece rate was chosen by the computer. All regressions contain session dummies as controls. Robust standard errors clustered on the session level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

D.3 Pictures of the Experimental Settings at the different Locations

Figure D.8: Experimental Location at Maastricht University.

Figure D.9: Example of Experimental Location at the University of Applied Sciences.

Nederlandse samenvatting

Een van de belangrijkste vraagstukken in de economische wetenschap is het begrijpen van de manier waarop mensen beslissingen nemen. Hiertoe is het traditionele theoretische raamwerk van nutsmaximalisatie opgesteld (bijv. Edgeworth (1879)) en worden ook elementen als voorkeuren en prikkels onderzocht. Voorkeuren en prikkels bepalen de keuzes van een individu. Deze keuzes hangen af van de situatie waarin het individu zich bevindt en manifesteren zich in (geobserveerd) gedrag.

In de afgelopen decennia zijn theorieën over besluitvorming van mensen getoetst op basis van onderzoek in het veld zowel als labexperimenten (Harrison and List, 2004; Falk and Heckman, 2009). Het labexperiment is een onderzoeksmethode die veel gebruikt wordt in de economie. Besluitvorming van mensen kan niet goed worden voorspeld wordt door de klassieke ingrediënten van de nutsmaximalisatie. Nutsmaximalisatie veronderstelt onder andere dat mensen rationale beslissingen nemen. Echter, er zijn dankzij gedragseconomisch onderzoek inmiddels vele voorbeelden waarin gedrag van mensen afwijkt van dit rationele patroon en dus niet in overeenstemming is met de verwachte nutsmaximalisatie (bijv. Kahneman and Tversky (1979)), verliesaversie (bijv. Tversky and Kahneman (1991)), sociale voorkeuren (bijv. Fehr and Schmidt (1999)) en hyperbolische verdisconter-ing (bijv. Laibson (1997)). Inzichten uit de gedragseconomie blijven economische modellen verbeteren en leggen de basis voor nieuw empirisch onderzoek in het veld en het laboratorium (Dohmen, 2014).

Een mooi voorbeeld van zulke nieuwe inzichten uit labexperimenten is onderzoek dat is uitgevoerd naar de verschillende typen vaardigheden die meespelen in besluitvorming van mensen. Onlangs hebben arbeids- en onderwijseconomen het onderscheid tussen niet-cognitieve en cognitieve vaardigheden aangetoond (bijv.

Borghans et al. (2008) voor een uitgebreid overzicht). Economen meten aan de hand van methodes uit de psychologie de cognitieve en niet-cognitieve vaardigheden. Het blijkt dat deze verschillende typen vaardigheden op verschillende manieren de individuele besluitvorming beïnvloeden. Deze bevindingen kunnen ze vervolgens in hun empirische en theoretische analyse meenemen (zie bijv. Dohmen (2014)). Het is nog niet bekend wat de toetsen van cognitieve en niet-cognitieve vaardigheden precies meten en hoe cognitieve vaardigheden met niet-cognitieve vaardigheden, prikkels en inspanning bij individuele besluitvorming interacteren.

Een ander voorbeeld illustreert de ontwikkeling van de gedragseconomie. Psycholoog Walter Mischel en zijn co-auteurs (Mischel et al., 1972) hebben een eenvoudig experiment uitgevoerd. Kinderen van 3 tot 5 jaar oud werden alleen in een kamer geplaatst met op de tafel voor hen een marshmallow. De onderzoekers hadden de kinderen verteld dat ze de marshmallow nu mochten opeten of een tijdje moesten wachten, waarna ze een tweede marshmallow zouden krijgen. Het onderzoek volgt de kinderen gedurende hun jeugd en laat zien dat de kinderen die ervoor kozen op een tweede marshmallow te wachtten, later op school beter presteerden en bijvoorbeeld beter met stressvolle situaties om konden gaan (Mischel et al., 1989). Dit resultaat leidde tot discussie onder onderzoekers en beleidsmakers. Economen en psychologen raakten geïnteresseerd in de mechanismen achter deze resultaten. Waarom bleven sommige kinderen wachten op de marshmallow en waarom deden de andere kinderen dit niet? Waren sommige kinderen simpelweg slimmer en begrepen zij dat het beter was om te wachten? Of verschilden zij van persoonlijkheid? Is het waargenomen gedrag typisch voor kinderen of is het resultaat alleen maar toepasbaar in deze specifieke context?

Bovenstaande onderzoeken illustreren de hoofdgedachte van dit proefschrift. Het hoofddoel van mijn onderzoek is het identificeren van determinanten van besluitvorming vanuit een empirische invalshoek. In vier studies onderzoek ik de factoren die individuele besluitvorming bepalen en probeer ik de hieraan onderliggende mechanismen te begrijpen.

In hoofdstuk 2, mede geschreven door Lex Borghans, Huub Meijers en Bas ter Weel, onderzoek ik het gedrag van deelnemers die een cognitieve toets maken. De toets vindt plaats in het BEElab, een laboratorium van de Universiteit Maastricht. De prestaties op en uitkomsten van een dergelijke cognitieve toets kunnen als economisch gedrag worden gezien. De score die de deelnemers behalen hangt af van hun vaardigheden en de manier waarop ze met tijdsdruk omgaan. Een resultaat op een toets reflecteert niet alleen die specifieke vaardigheden of kennis

die in de toets gevraagd worden, maar ook de vaardigheden van een persoon om een toets zelf te maken. Anders gezegd: hoe een individu met de toetssituatie omgaat. Eerder onderzoek in de psychologie en de economie geeft aan dat mensen reageren op prikkels wanneer zij een cognitieve toets maken (Edlund, 1972; Borghans et al., 2008). Om deze economie van het toetsen verder te onderzoeken voeren wij een labexperiment uit. We kunnen gedurende het experiment het moment vaststellen waarop de deelnemer het antwoord op een bepaalde vraag weet, maar ook de tijd die het kost voordat hij de vraag echt invult. We veranderen de toetsomgeving door het variëren van de tijdsdruk en de monetaire prikkels om een vraag (snel) te beantwoorden. Op deze manier kunnen wij de intensiteit van het denken en de reactie op verschillende prikkels van elkaar onderscheiden. Het onderzoek geeft drie hoofdresultaten. Ten eerste, deelnemers komen niet sneller op het antwoord van een vraag bij sterkere financiële prikkels of hogere tijdsdruk. Ten tweede, de verandering in het gedrag van de deelnemers komt overeen met de economische theorie. Ten slotte, de veranderingen in het antwoordgedrag van de deelnemers zijn verassend klein. Blijkbaar tonen de kleine veranderingen in het antwoordgedrag aan dat de deelnemers van het experiment al sterk gemotiveerd zijn om goed te presteren en niet extra gemotiveerd worden door een hogere betaling.

Naast prikkels zijn voorkeuren en persoonlijkheidsvaardigheden belangrijke determinanten van toetsresultaten (zie bijv. Borghans and Schils (2013)). In de volgende stap, hoofdstuk 3, onderzoek ik hoe deze persoonlijkheidsvaardigheden en voorkeuren een toetsresultaat beïnvloeden. Cognitieve toetsen worden gebruikt om cognitieve vaardigheden te meten. Echter zij meten meer dan dat alleen: de toetsscore is het resultaat van een mentaal proces dat meer omvat dan alleen cognitieve vaardigheden. In hetzelfde experiment zoals in hoofdstuk 2 beschreven hebben wij ook persoonlijkheidsvaardigheden zoals de Big Five (Goldberg (1990)) en economische voorkeuren zoals risico voorkeur en tijd voorkeur gemeten. Wij onderzoeken het besluitvormingsproces tijdens een cognitieve toets met hulp van een labexperiment waarin studenten Raven matrices moesten oplossen. Met behulp van het experimentele design weten wij wanneer een individu het antwoord op een vraag weet en wanneer een individu een vraag beantwoordt. De resultaten van het onderzoek tonen aan dat autonomie, emotionele stabiliteit en de risicovoorkeur van een individu de snelheid van het denken beïnvloed. Alleen de verdisconteringsvoet van een individu bepaalt de timing van een antwoord. De resultaten bevestigen eerdere bevindingen in de literatuur (bijv. Burks et al. (2009); Duckworth et al. (2011)) maar belichten ook het mechanisme achter de totstandkoming van

een toetsscore. Onze bevindingen hebben implicaties voor toetsresultaten omdat het mogelijk is om de prestatie op toetsen van individuen te verbeteren door het veranderen van hun gedrag.

Hoofdstuk 4 geeft aanvullende resultaten van veldonderzoeken die de resultaten van het labexperiment ondersteunen. In dit hoofdstuk onderzoek ik de relatie tussen het geduld van een individu en het toetsresultaat. Deze studie lijkt dus op het marshmallowexperiment dat hierboven als voorbeeld is aangehaald. Uit dat experiment bleekt dat kinderen die geduldiger zijn, betere resultaten op school boeken. In deze studie maak ik een onderscheid tussen high-stake en low-stake toetsen. De hoofdvraag luidt: in hoeverre scoren meer geduldige leerlingen beter op hun toets? Ik meet het geduld van leerlingen in een steekproef van 15-jarigen door middel van een experimenteel gevalideerde maat van het interne rendement. De schattingsresultaten laten een sterk en significant verband tussen geduld en het toetsresultaat zien bij een high-stake toets. Leerlingen die een standaarddeviatie hoger scoren op de maat van geduld, scoren 16 procent hoger op hun high-stake toets. Deze samenhang is kleiner en niet statistisch significant bij een low-stake toets. Het onderzoek vindt ook verschillende niveaus van geduld bij leerlingen van verschillende onderwijsniveaus.

Het begrijpen van de manier waarop beloningsvormen uitkomsten en gedrag beïnvloeden is cruciaal voor het begrijpen van moderne economieën. In hoofdstuk 5, samen geschreven met Trudie Schils en Bas ter Weel, doen we experimenteel onderzoek naar het effect van vaste en variabele beloningsvormen op productie in een 'real-effort' taak. We voeren een aantal labexperimenten uit onder studenten van de Universiteit Maastricht en de Hogeschool Zuyd in Sittard en Heerlen. De vraag die we in dit onderzoek willen beantwoorden is of veranderingen in beloningsvormen leiden tot andere uitkomsten. In het onderwijs is bijvoorbeeld discussie over het invoeren van prestatiebeloning voor leerkrachten. Het is echter ingewikkeld om de effectiviteit van een nieuwe beloningsvorm aan te tonen, omdat de huidige populatie leerkrachten heeft gekozen voor een baan waarin prestatiebeloning geen rol speelt. Invoering hiervan zou ze wellicht anders hebben doen besluiten bij hun studiekeuze. De allocatie van verschillende mensen over beroepen met verschillende beloningsvormen is niet toevallig zo tot stand gekomen. Deze allocatie onder twee verschillende beloningsregimes is het onderwerp van studie in dit hoofdstuk. Een eerste belangrijke observatie in de data is dat mensen op basis van hun productiviteit en geslacht een betalingsvorm kiezen wanneer zij die vrije keuze hebben (zie ook Dohmen and Falk (2011). De relatief hoogproductieve mensen kiezen voor

prestatiebeloning, terwijl de relatief laagproductieve mensen kiezen voor een vaste beloning). Als wij zelf de mensen een betalingsvorm opleggen, verandert hun productiviteit echter niet. Wel rapporteren zij een hoger stressniveau, meer uitputting en hogere inspanning. Bij een hogere vaste betaling en constant stukloon kiezen ook productievere individuen voor de vaste beloningsvorm (zie hier ook Dohmen and Falk (2011)). Het opleggen van een andere betalingsvorm lijkt dus ineffectief te zijn om de prestatie te veranderen omdat individuen naar productiviteit selecteren. De meest productieve individuen kiezen onder alle omstandigheden de variabele betalingsvorm. Onze resultaten laten zien dat de meerderheid van de individuen al maximaal presteert ook als hun betaling niet direct met hun prestatie verbonden is. In het algemeen is dit in lijn met wat we vaak op de werkvloer zien. Veel contracten zijn onvolledig en agenten leveren veel inspanning ook als zij weten dat de inspanning niet noodzakelijk door hun leidinggevende afgedwongen kan worden. Eerdere resultaten uit het laboratorium (bijv. Fehr et al. (1998)) en het veld (bijv. Kube et al. (2012)) interpreteren dit gedrag als geschenkuitwisseling. De agent beantwoordt de goedwillendheid van de principaal met een hoger niveau van inspanning dan het 'rationele' niveau (Akerlof, 1982). Bovendien, vinden we dat prestatie afhankelijke beloning de productiviteit van relatief onproductieve individuen niet verhoogd. De oorzaak hiervan is waarschijnlijk dat de allocatie van mensen zo is dat de minst productieve werknemers of de werknemers die het minst gevoelig zijn voor financiële prikkels in banen terechtkomen waar de beloning niet een-op-een afhankelijk is van de prestaties.

Valorization

The National Valorization Commitee refers to valorization as the "process of creating value from knowledge, by making knowledge suitable and [...] available for social and [...] economic use and by making knowledge suitable for translation into competitive products, services, processes and new commercial activities". The goal of this part is to put the results obtained in this dissertation in a broader perspective. I will further elaborate why they might be relevant for society. It is important to keep in mind that the interpretation and possible application of the results to real world settings has to be done extremely carefully. The reason for that is the studies in this dissertation use for instance samples which are subsamples of the Dutch population. In addition, the results in Chapters 2, 3 and 5 are based on lab experiments. There is a lively debate in economics about the application of such results into real life settings (see for instance Harrison and List (2004); Falk and Heckman (2009)). Nonetheless, the new empirical evidence documented in this thesis adds to the body of knowledge in the fields of education economics and behavioral economics and is in all likelihood useful in combination with the existing literature for a better understanding in public policy making or decision-making in companies.

One core subject of economic research has always been to better understand the nature of decision-making and human motivation. This is important since theoretical and empirical methods used by economists serve as key elements to steer individuals in companies and base policy decisions in entire countries. Hereby, assumptions and theoretical models about how individuals decide in certain situations are used to introduce certain payment schemes in companies. Many models assume that correctly contracted monetary incentives lead to better outcomes,

since they enhance individual performance. This neoclassical view has always been challenged but more recently the field of behavioral economics has established a couple of deviations of the framework how economists (used to) think about the determinants of human decision-making (see for instance Dohmen (2014); Koszegi (2014)). This involves the ingredients of the decision-making process as well as the measurement of the ingredients. Early models do for instance not capture personality traits. Many studies, however, do show that personality traits matter for important outcomes in life (see for instance Cunha and Heckman (2009)). Almlund et al. (2011) introduce a utility maximization framework which allows personality traits to matter in the outcome of the decision-making process. These additions to the early and less elaborate models better take into account heterogeneity across individuals to explain a variety of socio-economic outcomes. It is important to stress that these relative recent studies in the field that has become known as 'behavioral economics' are additions to the traditional neoclassical workhorse that help explain outcomes.

This thesis adds to the research in economics and psychology which links decision-making outcomes to measures of preferences defined by economists and personality traits defined by psychologists. It further adds to the literature on incentives and to what extent they can act as sorting devices as well as performance boosters. These are important topics in modern societies which base their politics on evidence-based economic research. Only recently data sources and methods have been developed to test the validity of economic theories with rich new data sets in the field and experimental methods. By doing so, relationships which have been formulated thus far only theoretically can be tested and improved with real world data. This thesis contributes to this literature by investigating the effects of incentives on complex problem solving (Chapter 2). The key result of this study is that monetary incentives are useful to trigger individuals to exert effort, but that as of a certain level, monetary incentives do not enhance the outcome of a complex problem solving task. If we change the conditions of the problem solving environment individuals change their behavior according to a simple economic model. The results from Chapter 2 are complemented with those from Chapter 3. I investigate the mechanisms behind personality traits, preferences and incentives in problem solving. I find that personality traits and economic preference parameters are associated with the results on a problem solving task. One of the key findings is that patient individuals have better problem solving skills because they just take more time to think about a problem compared to their rather impatient

counterparts. At the same time both types of individuals have the same technology to solve a problem. This means that more patient individuals only seem to obtain better results because they wait longer until they make a decision. The relationships which are found in laboratory data are tested and confirmed with field data in Chapter 4. I investigate the determinants of achievement test scores which are important for further education careers of Dutch students. I show that not only those students who are equipped with better cognitive skills have better achievement test results but also those students who exhibit greater patience show better test results.

The findings from Chapters 2, 3 and 4 can have implications for a variety of settings in companies, other institutions or for instance in the education sector. The results in Chapter 2 could imply that, when it comes to complex problem solving, monetary incentives are important performance boosters but that the marginal effects of additional monetary incentives are not always effective. This can be relevant for the design of incentive schemes in companies or institutions such as research institutes which require complex problem solving skills. Higher monetary incentives and more time pressure do not necessarily yield faster and better solutions. Moreover, the results which are obtained in Chapter 2, can help understanding the effect of policies that aim at paying pupils for their test performance (Bettinger, 2011). We find that at least during the test increasing pressure and monetary incentives does not improve the number of correct answers obtained by our subjects.

Chapter 3 and 4 add to the debate in policy to what extent personality traits and preferences matter for outcomes. If we know which traits matter for problem solving skills and other important outcomes in life this can be helpful to design interventions in politics which aim at the enhancement of these skills. Recent studies have shown that personality skills start to develop in early childhood (Borghans et al., 2008; Cunha et al., 2010) and are still malleable at certain stages in life (Prevoo, 2013). Many other studies such as Moffitt et al. (2011) and Golsteyn et al. (2014) show that the ability to delay immediate gratification at young ages is associated with favorable outcomes such as higher levels of education, health and more wealth. One potential conclusion could be to train patience or and delay of immediate gratification to enhance the quality of decisions and thus later outcomes in life. Another relevant finding from this research is that achievement tests capture many other skills besides the ones they are measuring. Policy makers want to assess for instance the skills of students in a certain subject. They make

use of these skill measures to compare the education level of entire countries and trigger big debates about educational reforms. The findings from Chapter 3 and 4, however, show that good test performance is associated with more than just the skill which should be measured by the test scores. Favorable personality traits such as emotional stability, conscientiousness and patience go along with higher scores on these tests. This implies that policy makers should also focus on methods that aim at the measurement of theses skills (Kautz et al., 2014). Lastly, it is important to investigate how these skills develop. If this is known, training these skills could help to improve lifetime outcomes such as educational attainment.

Lastly, Chapter 5 investigates the role of incentives on sorting decisions, as well as the role of imposed incentive schemes for output and stress levels. Our findings have implications for the change of payment schemes in companies or other institutions (see for instance Delfgaauw and Dur (2008); Dur and Zoutenbier (2014); Zoutenbier (2015)). We find that different incentive schemes attract different individuals. More productive individuals are more likely to select themselves into performance dependent payment schemes. Incentive schemes which are tightly tailored to performance induce higher stress levels. Most importantly, a performance depended incentive scheme does not increase the output of less productive individuals but only seems to induce higher stress levels. These results confirm the results by Dohmen and Falk (2011) and can help to better understand potential mechanisms in the debate of changing payment schemes for certain occupations.

The methods used in the lab experiments and the field study can for instance be applied in assessment centers of companies or schools to learn more about an individual's skill package. The (incentivized) elicitation of time preferences could serve as complementary measure to psychological traits since it yielded more reliable results than some of the psychological measures. Moreover, Dohmen and Falk (2011) already suggest to use the experimental sorting decisions to assess individual preferences of payment schemes. They argue that, this could help to optimize payment schemes and select individuals in the right career paths. Our study adds to their finding. If individuals are in the 'wrong' payment scheme according to their preferences this might have severe consequences for their health if stress levels indeed remain on such a high level. Hence the experimental methods can for instance be helpful to improve HR policies in organizations. Firms could use the experimental design, to assess productivity of individuals, their payment scheme preference and their stress resistance. However, these results might need more long-term studies as additional support.

The studies in this dissertation also provide a fundament for new research. This is also relevant for society since additional research can eventually help to further develop better policies and institutional setups. First, it is important to further investigate which incentives besides monetary incentives influence performance. The relationship between performance and non-monetary incentives such as social approval could be tested in further research projects (see for instance Fehr and Falk (2002)). One way to pursue this research could be to test problem solving in tournament settings. Instead of only rewarding good performance with monetary rewards the results could also be revealed in public such that individuals have the possibility to compare themselves with their competitors (Gerhards and Siemer, 2014). This would come closer to settings in schools or companies. A second promising part of the research agenda is to investigate the origins and mechanisms which enables an individual to delay gratification. The investigation of these channels can induce a better understanding of why some children perform better or worse in certain environments. Another important avenue for research is the investigation of consequences of performance dependent payment schemes. A potential follow-up to Chapter 5 would be a more precise measure of actual stress levels with bio-markers. In addition it is important to investigate the long-term effects of imposed payment schemes on performance and other outcomes such as health. Here it might be useful to conduct field experiments to circumvent potential biases arising from self-selection of individuals. Last but not least, there is still need for research to establish well-tested measures of individual preferences. In this dissertation I make use of experimentally validated measures. However, more studies are needed to further stress test the validity of these measures.

The different chapters of this dissertation were presented at various international conferences which aim at the distribution and application of knowledge. The conferences were not only purely scientific but also suited for policy makers such as the ORD or EALE Conference. The ORD ('Onderwijs Research Dagen' which stands for Education Research Days) is a Dutch conference for policy makers and practitioners in the field of education. The European Association of Labour Economists (EALE) Conference is a conference which brings together researchers and policy makers in the field of labour economics.

Curriculum Vitae

Benedikt (1986) is ne rheinische Jung. He graduated from high school in Bonn in 2005. Afterwards he started his studies in economics at the University of Bonn. He also enjoyed is student life by being engaged in various positions in the students council of the faculty of economics and by singing in the youth choir of the opera in Bonn. He spent his studies abroad at Charles University in Prague. Before and afterwards he worked as a student assistant at the Institute for Applied Microeconomics and the Center for Economics and Neuroscience. In 2011 he obtained his Diploma degree in economics and started his PhD at Maastricht University. He presented his work at various international conferences such as the EALE conference. Since September 2014 he has been working as a researcher at the CPB Netherlands Bureau for Economic Policy Analysis in the Hague.

Acknowledgements

The start and completion of this dissertation would not have been possible without the support of many people. At the risk of making the (wrong) impression that I spent more time on everything else but writing a dissertation, I did my best to collect a more or less complete list of those people who supported me and made my life beautiful in the last three years.

First of all, I would like to show gratitude to Bas ter Weel and Trudie Schils. I see it as a great privilege that you gave me the opportunity to think more in depth about economic problems by hiring me as your PhD student. You advised me in asking the right questions and supported me during my analysis and writing. Most importantly, as good principals you kept up my intrinsic motivation for the projects and pushed me in phases of sophisticated procrastination. Thank you very much!

Next, I would like to thank the members of the assessment committee, Andries de Grip, Thomas Dohmen, Robert Dur, Dinand Webbink and Inge de Wolf. Moreover, many people at Maastricht University gave useful input during different stages of the research projects. The participants of the Economics of Education lunch seminar were always open to listen to presentations of my work and gave very helpful input. I would like to thank Paul Jungbluth who made it possible that I could use the KAANS data for one of my projects. I would especially like to express my thanks to Lex Borghans and Huub Meijers for interesting and inspiring discussions. I am happy to have you as co-authors of one of the chapters. I am thankful to Bart Golsteyn for patient discussions about patience. I thank Olivier Marie and Martin Strobel for inspiring discussions. I would also like to thank the participants of the 'kids' seminar and particularly Nico, Maria and Jan

for questions and comments on different projects.

Silvana de Sanctis, Fleur Keune and Sylvia Behnen were very supportive and patient with all organizational questions, regarding the funding of the lab experiments, teaching hours and conference trips, thank you!

Eva, I thank you for being such an inspiring office mate. Many thanks also go to all the people from our PhD cohort. I would especially like to thank Andreea, Dennis, Gabri, Len, Matteo, Seher, Thomas and Tyas for inspiring talks as well as decent trash-talks which gave me new motivation for the next endogeneity problem.

Many thanks go out to the people who accompanied me on various sport events. I enjoyed the cycling trips through 'Heuvelland' with Ruud and the accounting racing team. I loved doing the Triathlons and various city runs with Matze.

During the time abroad you learn to appreciate how important friendships back home are. I enjoyed our trips to Berlin, Essen, Eschwege, Hamburg, Leuwen, Mannheim, Nürburgring, Prague, Regensburg, and the fun nights we had in Bonn. I would also like to thank the 'WG' in Aachen. Dominik, I still have a laugh when I think about 'the golden pig' and the brunches we had together.

Nina, Eva and Matt, thank you very much for sharing your excellent culinary skills on various dinners, barbecues, and, of course, eggs Benedict.

Dear Anke, Franzi and Ulf. I probably would not have started this PhD project without the inspiration, motivation, fun and encouragement I got during my times in Bonn. Thank you, AFU! Ulf and Franzi, thank you for being mental and professional support during our time together in Maastricht. I loved discussing research with you and it always gave me a better understanding what we are actually doing. I enjoyed my first three months in Maastricht behind your couch, as well as the nights and weekends we spent barbecuing and enjoying Maastricht. Thank you for being my paranymphs!

Last, but definitely not least, I am most grateful to my family. I would like to thank my parents for their unconditional support and patience in the things I have done. Understanding a little bit better how cognitive and non-cognitive skills are formed and how they matter for the things we do, I know that you set the fundament with your investments for all of this. Christine, Thomas and Moritz, thank you very much for being the best brother and sister and friends I could ever have.