DISCUSSION PAPER SERIES

IZA DP No. 7182

The Importance of Intrinsic and Extrinsic Motivation for Measuring IQ

Lex Borghans
Huub Meijers
Bas ter Weel

January 2013

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

The Importance of Intrinsic and Extrinsic Motivation for Measuring IQ

Lex Borghans
ROA, Maastricht University and IZA

Huub Meijers
UNU-MERIT, Maastricht University

Bas ter Weel
CPB Netherlands Bureau for Economic Policy Analysis, Maastricht University and IZA

Discussion Paper No. 7182
January 2013

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 7182
January 2013

ABSTRACT

The Importance of Intrinsic and Extrinsic Motivation for Measuring IQ*

This research provides an economic model of the way people behave during an IQ test. We distinguish a technology that describes how time investment improves performance from preferences that determine how much time people invest in each question. We disentangle these two elements empirically using data from a laboratory experiment. The main findings is that both intrinsic (questions that people like to work on) and extrinsic motivation (incentive payments) increase time investments and as a result performance. The presence of incentive payments seems to be more important than the size of the reward. Intrinsic and extrinsic motivation turn out to be complements.

JEL Classification: J20, J24

Keywords: incentives, cognitive test scores

Corresponding author:

Bas ter Weel
CPB Netherlands Bureau for Economic Policy Analysis
P.O. Box 80510
2508 GM The Hague
The Netherlands
E-mail: b.ter.weel@cpb.nl

* We would like to thank Anton de Vries for advice about the use of cognitive tests. Flavio Cunha and James Heckman have contributed to this work through very helpful discussions. The thoughtful comments of the referee of this journal have improved the analysis further. We also acknowledge the useful comments of participants at the 2008 Meeting of the European Association of Labour Economists in Amsterdam and seminar participants at Maastricht University.

1. Introduction

It has been well-documented that individual test scores on achievement and IQ tests are sensitive to incentives (e.g., Borghans, Duckworth, Heckman & Ter Weel, 2008). Students perform better on high-stakes tests and if they are paid for their performance (e.g., Duckworth & Seligman, 2005 and Sackett, Borneman & Connelly, 2008). The role of incentives in stimulating the performance during achievement tests implies that the expressed effort can be seen as an economic decision. Hence, an economic perspective sheds light on how IQ tests are influenced by changing circumstances during the test.

The aim of this research is to analyse people's performance on an IQ test from an economic point of view. Economists are increasingly interested in including psychological measures, such as IQ, in their analyses explaining differences across a range of individual outcomes. However, these measures have been developed to answer psychological questions and careful attention is needed in applying them to economic analyses. To answer economic questions by using these measures it is important to apply an economic perspective on what is measured and consider how circumstances and incentives affect scores. We present an economic model and estimate its components. When answering questions on an IQ test, performance improves when a participant takes more time. A participant has to decide how much time to invest in a question. Outcomes on each question depend on (i) the technology described by a production function in which time investment determines the probability of a right answer; and (ii) preferences for a right answer relative to the time invested. Preferences and technology together determine the optimal time investment and hence the score on each of the questions. It is important to note that this research is concerned with individual behaviour during an IQ test, not with measuring an individual's IQ per se.

We empirically investigate this model by applying data from a laboratory experiment, described in Borghans, Meijers & Ter Weel (2008), in which students answered several types

of questions, common in standard IQ tests, with varying monetary incentives and time constraints. The empirical part of this paper is based on the data obtained in that experiment. In the previous paper we investigate how people with different personalities respond to incentives. In this paper we use the same data to disentangle the production function and preferences in answering questions, to understand how people perform on IQ tests.

Our most important findings can be summarised as follows. Participants have higher intrinsic motivation on some questions (especially the most difficult ones) relative to others. In addition, participants invest more time in answering questions when incentive payments are introduced. However, the preference for correctly answering a question with incentives is not proportional to the size of the incentive. An incentive as such seems to be more important than its monetary value. Extrinsic monetary incentives have a bigger impact on the time invested into questions for which students have a higher intrinsic motivation. This suggests complementarity between intrinsic and extrinsic motivation. Indeed, when estimating a CES production function with intrinsic and extrinsic motivation as the two inputs, we obtain complementarity.

The analysis in this research contributes to our understanding of how the circumstances of a test affect the performance of people during the test. To better understand the relationship between measures such as IQ and achievement and economic outcomes, these economic factors influencing tests scores have to be taken into account. The model also offers a framework to analyse differences in the way people perform on tests, which contributes to our understanding of variation in test scores beyond the pure variation in intelligence. The approach can be used in the same way to analyse achievement tests because it is conceptually equivalent to the process during an IQ test. A large body of literature linking intelligence to future outcomes uses data obtained from achievement tests, because these measures are available in large longitudinal surveys. A well-known measure of achievement is the AFQT in

the NLSY.

This paper is related to the literature about the effects on incentives on IQ scores and individual performance on achievement tests. In the economic literature there is a body of work studying the determinants of cognitive achievement. This work focuses both on parental inputs and youth environment and on the relationship between schooling inputs and cognitive test scores (e.g., Todd & Wolpin, 2003 for an elaborate and excellent review). In general, a positive correlation between inputs and cognitive test scores is obtained. The present research adds the importance of intrinsic and extrinsic motivation in determining test scores, which influences how well students perform on these cognitive tests. IQ and achievement during childhood (typically measured by psychologists), is very predictive for a wide variety of later outcomes. An important question is whether this predictive power is due to intelligence or can be explained by other factors that determine test scores. Recent papers have pointed at the interrelationship between cognitive and noncognitive skills (e.g., Heckman, Stixrud & Urzua, 2006, Cunha, Heckman & Schennach, 2010, Heckman, Humphries & Mader, 2011, Moffitt et al., 2011 and Prevoo & Ter Weel, 2012). They demonstrate that self-discipline or self-control, conscientiousness and determination are equally important in explaining a variety of economic outcomes in the sense that movements from the bottom to the top of such noncognitive distributions have comparable effects on many outcome measures relative to cognitive skills. If IQ and achievement tests are affected by noncognitive factors, it could explain the high predictive power of these tests. This calls for an economic framework to disentangle the various influences (e.g., Borghans, Golsteyn, Heckman & Humphries, 2011). This paper offers such a framework. The relationship between motivation and test scores is also subject to investigation for a long time (e.g., Borghans, Duckworth, Heckman & Ter Weel, 2008, for a discussion and overview).

Psychologists have been worried about motivation interfering with intelligence and

economists have shown that test scores can be improved by incentive payments. For example, Gneezy & Rustichini (2000), Angrist & Lavy (2009) and Kremer, Miguel & Thornton (2009) show that high enough payments provide incentives to people to work harder. By contrast, Fryer (2011) only finds moderate effects for rewarding students to read books or for other desirable behaviours on performance across several experiments in large US cities. However, his experiments do not involve direct incentive payment during achievement tests. In another experiment he shows that aligning incentives between students, parents and schools by financial rewards improves students' test scores (Fryer, 2012). Bettinger (2010) also reports that providing financial incentives (in elementary school) for getting better test scores or grades yields little to no effects on student achievement. Angrist, Lang & Oreopoulos (2009) report on an experimental evaluation of strategies to improve student performance in a Canadian university. They find that particularly women improve study habits and obtain higher grades when they are offered academic support services and financial incentives. Winters, Trivitt & Greene (2010) and Liu & Neilson (2011) investigate whether schools with an incentive to focus on those subjects that play a role in the accountability system decrease attention to subjects that are not part of such a programme. Our paper adds to these papers by estimating the technology of answering questions during an IQ test, which helps us to understand the relationship between extrinsic and intrinsic motivation.

Finally, our work is most closely related to the contribution of Segal (2012). She investigates whether the most motivated subjects are the most cognitive able ones. She finds that test scores relate to economic success not only because of cognitive ability but also because of favourable personality traits, which is consistent with our findings in Borghans, Meijers & Ter Weel (2008). The present paper is unique in the sense that we are able to distinguish the degrees of difficulty of the questions.

This paper proceeds as follows. Section 2 builds the theoretical background and

presents the empirical strategy. Section 3 presents the data and documents a number of descriptive results. Section 4 shows the estimation results, and Section 5 concludes.

2. Theory and strategy

This section presents an economic approach to answering questions on a cognitive test. Conducting a cognitive test implies making decisions. This decision-making process is considered to be an economic activity in the same way as other behavioural outcomes such as school and job performance.

The main input in the performance of a cognitive test is time. By thinking longer the probability of finding the right answer can be increased. The technology of answering a question is expected to be an upward sloping and concave function of time. Figure 1 shows this relationship. On the horizontal axis we plot time and on the vertical axis the probability of giving the correct answer. The concave relationship is the probability of submitting the correct answer, which is increasing in time. The participant's decision is when to stop thinking and to submit the answer. The decision to stop and submit depends on the technology (the expected increase in performance when thinking longer), the preference for submitting the right answer and the disutility of time. These preferences are shown by the line β.

The utility function of a person conducting on an achievement test can be written as

$$U = \beta(i_q, e)p_q(t) - t \quad (1)$$

in which β represents the value of a good answer relative to the value of time (t). In equation (1) $p_q(t)$ is the probability of a right answer on question q conditional on time, and i_q and e are variables to capture intrinsic and extrinsic motivation, respectively. The intrinsic motivation can vary between questions. The first-order condition with respect to time determines optimal time investment: $\frac{dU}{dt} - \beta(i_q, e)\frac{dp_q(t)}{dt} - 1 = 0$.

We analyse performance as the optimisation of preferences conditional on technology.

This implies that we separate ability, as revealed by the technology used to answer a question, from preferences. Together ability and preferences determine choices and outcomes on the test. Technology is identified by exogenously varying the time people use to answer a question. Our data (which will be described in Section 3) contain both variation in the time constraint a participant faces and in monetary incentives. Together, time constraints and incentives determine the time people invest to answer a question. Different combinations of time constraints and incentives enable us to measure the technology function. If participants are able to improve test scores not only by investing more time but also by exerting more effort, variation in incentives will not only affect the time invested but also the effort exerted. Our strategy is to use the variation in monetary incentives to investigate whether or not the time invested to answer a question sufficiently describes what people do to improve their scores.

Once the production function representing the technology is known, we can investigate β, which is determined by the technology assuming, that people decide optimally about the time they invest in a question:

$$\beta\left(i_q, e\right) = \frac{1}{\left(\frac{dp_{q(t)}}{dt}\right)} \tag{2}$$

By estimating the technology and measuring the time people invest to answer a question, we are able to disentangle how intrinsic and extrinsic motivation affect the value of answering a question rightly. Our strategy is to estimate β under different circumstances to investigate how the value of answering a question rightly varies.

A number of possibilities is investigated. First, if only the expected monetary benefits would matter for participants, β would be proportional to the monetary incentives provided. Since we observe that participants also invest time in answering questions when there are no rewards, there appears to be an intrinsic value of answering questions. It seems natural to add intrinsic motivation to the model additively. This would lead to a linear relationship between

β and the monetary incentives provided: $\beta(i_q, e) = i_q + \alpha e$, where i_q and e are again variables to capture intrinsic and extrinsic motivation (incentives).

Our second approach is to investigate possible complementarities between intrinsic and extrinsic motivation. We model this as a CES-function: $\beta(i_q, e) = \left({i_q}^{\rho} + \alpha e^{\rho}\right)^{1/\rho}$. The target is to estimate the value of the elasticity of substitution $\frac{1}{1-\rho}$, which should be equal to 1 in case of complementarity.

If a full non-parametric estimation of the technology function would be available, we would be able to estimate β by one over the derivatives of the technology function evaluated at the time people invested, given the circumstances. Since our data only allows us to determine the technology function at a few points on the time-axis, this approach leads to an upper and a lower bound for β. Figure 2 shows how we can determine β using the properties of the technology function, which is in this example based on two observations combining the amount of time used answering questions (t_1 and t_2) and the probability of a correct answer ($p(t_1)$ and $p(t_2)$). Due to the concavity of the production function, β has to be larger than $\frac{t_1}{p(t_1)}$ and smaller than $\frac{t_2 - t_1}{p(t_2) - p(t_1)}$. Finally, to further investigate how people answer questions a parametric specification of the technology function is needed. The functional form we apply in this paper is: $p(t) = 1 - \delta t^{\alpha}$, which can be linearly estimated by $\ln(1 - p(t)) = \ln(\delta) + \alpha \ln(t)$. Assuming optimal behaviour, it can be derived that $\beta = -t^{1-\alpha}/\alpha\delta$.

In the empirical application of this model we will pool questions together and estimate this equation for groups of questions separately.

3. Data

To estimate the model we use data from an experiment we reported about elsewhere (Borghans, Meijers and Ter Weel, 2008). In this experiment students have to answer

questions from several different IQ tests with different time constraints and financial incentives. 128 students participated in the experiment. They were all Dutch students from Maastricht University and the experiment was conducted in Dutch. The experiment was conducted in thirteen sessions in one week during the spring of 2006. We do not obtain differences in test scores between groups of students, which makes us confident that contamination and sharing answers with others is not biasing our results. We present the most salient details here and have put other details and information about the different types of questions in an appendix at the end of the paper. Our previous paper reports in greater detail about the setup of the total experiment.

The data we use here are from the part of the experiment in which seven sets of ten questions had to be answered. In each set there was a possible time constraint (no time constraint, 60 seconds or 30 seconds) and incentive payment (no payment, €0.10, €0.40 or €1.00 for each correct answer). Subjects always had to complete one set of questions without incentive payment and two sets of questions under each incentive payment regime. The maximum earnings are €30.00. The average earnings were €16.53 (standard deviation €3.44). All respondents had to answer the full set of questions, but we randomized the order to separate the effect of tiredness and experience with the questions from the difficulty of the question.

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 (e.g. Epstein, 1994). The questions we have applied in our experiment refer to the former, with the exception of the cognitive reflection test (CRT). Similar to the questions that can be executed relatively quickly, the CRT questions have a more or less spontaneously answer, but this is often the wrong answer. Frederick (2005) provides a number of examples of such questions. Appendix A.2 presents an example of such a question and examples of each of the other six

different types of questions. The questions we use in the experiment are often used in IQ tests and differ in the degree of difficulty.

After each block of ten questions, there was a one minute break during which subjects could recover but were not allowed to do anything else then sit still. After these seven sets of questions this part of the experiment ended.

4. Results

This section documents a set of estimation results in which we estimate the technology to obtain a better understanding of how participants behave during our experiment.

4.1. Basic results

The experimental variation in the time limits to answer questions on the test combined with the different payments assigned to a block of questions induce exogenous variation in the time subjects think about answering a question. In this section we explore this variation.

Figure 3 presents the experimental equivalent of Figure 1. The dots in the figure represent the average time that is invested to answer a question and the average scores for each of these circumstances. The coding of the dots in the figure is the following. The first digit represents the time limit imposed: 1 is a 30 second time limit, 2 is a 60 second time limit and 3 is no time limit imposed. The second digit represents the incentive pay: 0 is no pay, 1 is €0.10, 2 is €0.40 and 3 is €1.00 for submitting the correct answer on a question. The curve has been fitted by $\ln\big(1-p(t)\big) = C + \alpha_1 \ln(t)$, in which $p(t)$ and $\ln(t)$ are population and question averages for the different circumstances. Since questions have been pooled, we omit the index q from now on. Overall, Figure 3 shows a concave pattern of the relationship between the probability of submitting a correct answer and time investment, which is consistent with the theory plotted in Figure 1.

An empirical question is whether time sufficiently describes the effort people put in to answer a question. To investigate this we make use of the points in Figure 3 to estimate the technology controlling for time investments and incentive payments. The idea is that if participants are not only able to vary the time they invest in answering a question but also improve their scores by thinking harder, we expect that questions with a high reward will be answered better, even when conditioning for time investments. Table 1 displays the results of estimating several versions of the relationship between $\ln\big(1-p(t)\big)$ and time investments and incentives payments. In the first column we only include time investments to explain $\ln\big(1-p(t)\big)$. This exercise returns a significant coefficient, suggesting that the longer people think about answering a question the higher is the probability of submitting the correct answer. When we add a dummy for incentive payments, it turns out that incentive payments do not explain $p(t)$. In the third column we show the results of adding the different incentive schemes. This leads to similar conclusions. In the results presented in the fourth column we leave out time investments and only include incentive payments to explain $\ln\big(1-p(t)\big)$. The estimates in the fourth column of Table 1 show that the value of the question indeed leads to better scores, especially in the case of no time restrictions. However, when we add time investments (see column (5)) this effect becomes small and insignificant. The estimation results presented in Table 1 suggest that time investments are the main channel through which subjects are able to invest in higher test scores. Thinking harder as a result of incentive payments does not improve scores. The line in Figure 3 provides the prediction of this specification.

4.2. Exploring heterogeneity between questions

Not all questions are equally difficult to answer. Some questions turn out to be easier and others are extremely hard for our population. We take advantage of this heterogeneity to

explore the data further. To take into account the heterogeneity across different questions, we split the sample into four quartiles of varying difficulty. To determine the difficulty of a question we use the average scores in the case when a time limit of 30 seconds was applied. Under these circumstances there is not much scope to vary time investment, which provides us with approximately the scores conditional on the time investment.

Figure 4 shows the relationship between time investments and the probability of submitting a correct answer for the four quartiles. The darkest dots present the easiest questions, the lightest dots the hardest ones. Again, we make use of the exogenous variation provided by the experiment. We observe that the four levels of difficulty yield differences in the technology of answering these questions. For harder questions much more time is invested relative to easier questions.

Table 2 provides a set of regression results for this technology function. We estimate models in which we interact time investments with dummies for the different quartiles. The easiest questions serve as the reference group. The results displayed in the first column of Table 2 confirm the estimates in Table 1 by pointing towards a strong role for time investments in explaining the probability of a correct answer. Also, as columns (2) and (3) suggest, incentive payments only change this picture when we neglect time investments (column (2)). More precisely, as column (3) reveals, incentive payments do not seem to lead to higher scores when we control for time investments. This is consistent with the set of estimates presented in Table 1.

4.3. Explaining intrinsic and extrinsic behaviour

Figure 2 has shown how we can use the different time limits to determine the functional form of the relationship between the probability of submitting a correct answer and time investments. Using these data points from our sample from the experiment, we can now

determine the upper and lower bounds of the value of time investments when answering a question. We again split the data according to the four levels of difficulty and in questions with and without incentive payments.

Table 3 provides an overview of the results. We only use the answers to questions for which there was no time limit because time limits may interfere with optimal levels of time investments. This has the additional advantage that the results are independent from the determination of the quartiles in level of difficulty, which is based on questions with a 30 second time limit. In the first two columns we document the time investment and average scores without incentive payments and in the next two columns the same with incentive payments. The rows show the level of difficulty with d_1 being the easiest quartile and d_4 the hardest questions. As before, we obtain that time investments are higher when incentives are provided. Also, the scores are higher. From this information lower and upper bounds of the value of answering a question correctly can be derived, following the procedure sketched in Section 2. The final two columns in Table 3 show the results. The standard errors of this approach are large. However, even at this level of uncertainty, the value of answering the hardest questions correctly is significantly higher than the value of answering the easiest question in a correct way. This suggests that people are not only triggered by incentive payments, but seem to like to invest more time in the harder questions relative to the easier questions.

When we apply the parametric model to the data, more precise estimates are obtained. Table 4 provides estimates for β using the same specifications as shown in Table 1, in which all questions are pooled. The assumption is that time investments are done optimally. For questions with time limits, optimal time investment might not be feasible because it could be the case to subjects would have liked to invest more time than 30 or 60 seconds. The coefficients in Table 4 indeed show that the implied value of answering a question is rather

constant across the different incentive schemes for the questions with a time limit of 30 or 60 seconds. In case the subjects have unlimited time to answer a question (shown in the final row of Table 4), the increase in the value of submitting the correct answer is much larger. The increase in value is relatively large moving from no incentive payment to a payment of €0.10. This suggests that the distinction between no incentive payments and any incentive is relatively large compared to the effect of the monetary reward as such.

Again we can do the same for the four quartiles. Table 5 provides the results for the questions without time limits. Comparing this parametric approach with the nonparametric upper and lower bounds in Table 3 shows that the parametric estimates for the easy questions are above the upper bound but still within the 95% confidence interval of the nonparametric estimates. This suggests that the parametric specification we applied is flatter for low values of t than the data suggests. A slightly more curved function might fit the data better. The results in the table reveal several things. Without incentive payments subjects attach a higher value to harder questions relative to easier questions as is shown by the first column. Moving to columns two to four incentive payments increase the value of time for answering questions. The difference between easier and harder questions becomes larger when incentive payments are introduced. Together these results suggest that the effect of intrinsic motivation (revealed by the value attached to harder questions) and extrinsic motivation (the incentive payments) is not an additive process.

To show this, we fit a CES production function to the coefficients displayed in Table 5. We include dummy variables for each level of difficulty (d_i) and for each incentive payment scheme (d_e). We estimate $\beta(i,e) = (d_i + d_e)^{1/(1+\rho)}$. A value of $\rho = 0$ indicates linearity which would imply that the value of submitting a correct answer is an additive function of intrinsic and extrinsic motivation. Note that since each value of i and e is represented by a dummy variable, the power of ρ on the terms within brackets has no

meaning and is not displayed here.

Table 6 provides the results of estimating the CES production function. We apply two different specifications: a sample using the four quartiles and a sample using eight levels of difficulty. In both specifications we obtain an estimated ρ close to –1. This suggests that an additive specification of intrinsic and extrinsic motivation does not fit the data and the actual model is close to multiplicative. The conclusion we draw from this finding is that both intrinsic and extrinsic motivation in answering questions during IQ tests are relevant and that both types of motivation are not independent from each other. This suggests that extrinsic motivation, such as in high-stakes tests, is likely to only partially explain student performance and that intrinsic motivation is relevant too.

Together, our findings suggest that performance on an IQ test is not just a matter of intelligence, but is also in an interactive way influenced by intrinsic and extrinsic motivation. Participants perform best when they are challenged by the type of questions and are rewarded for good performance.

5. Conclusion

This research presents an economic framework of how participants behave during an IQ test. We have built a theory in which we distinguish a technology that shows how performance can be improved by investing more time into answering questions from individual preferences to exert effort when confronted with different types of questions. Our main findings are that the individual scores on IQ tests are not only the result of intelligence, but also of time investments and preferences. In particular our experimental data allow us to distinguish between intrinsic and extrinsic motivation. By providing incentive payments, people invest more time in answering a question and this complements their intrinsic motives to show their willingness to do well on the test. The alternative theory that extrinsic and

intrinsic motivation are additive is not confirmed in our data.

This paper is a first step to better understand how IQ scores are determined by different circumstances and how they differ not only based on intelligence but also on differences in preferences across individuals. This notion is potentially important for applied research in both economics and psychology. Economists are increasingly interested in psychological variables concerning intelligence, achievement and personality to understand investments in human capital, allocation in the labour market and economic outcomes of behaviour. Psychologists have well-developed instruments, such as IQ tests and personality tests, that are targeted at psychological questions. Applying these instruments for questions that typically interest economists can easily lead to biases, as shown by the literature on the effects of incentives on IQ scores (e.g., Gneezy & Rustichini, 2000). For these purposes these psychological measures have to be interpreted in an economic perspective.

This paper is a first step to understand how an IQ test has to be understood from an economic point of view by explicitly taking into account how the ability of a participant and his preferences (for a high score and time investments) determine the scores. Such a framework allows us to disentangle the effects of ability and various aspects of preferences, which is necessary for our understanding of why some people do well on IQ tests and why IQ tests predict later outcomes.

Appendix

This appendix provides details about the setup of the experiment and provides examples of each of the seven types of IQ questions asked to subjects. The instructions and the computer programs used to conduct the experiment are available upon request.

A.1. Experimental setup

128 subjects participated in the experiment. They were all students from Maastricht University recruited by email through the communication office of the university. The email contained a hyperlink referring to a webpage through which people could register. Upon registration we asked questions about gender, date and place of birth, highest level of education of both parents, and college major.

The experiment was run in the week of 15-19 May 2006 in the experimental laboratory of the Faculty of Economics and Business Administration at Maastricht University. There were thirteen sessions: Three on Monday, Wednesday, Thursday and Friday and one on Tuesday morning. The sessions lasted for almost 2.5 hours. The morning session started at 8.30 hrs, the early afternoon session at 12.00 hrs. and the late afternoon session at 15.30 hrs. During initial registration we randomly selected subjects into groups of 10-15 subjects and assigned them to sessions. All subjects received an invitation by email. Upon arrival subjects had to wait in front of the laboratory until everybody arrived. There are no differences between outcomes for different groups.

The laboratory consists of two rooms separated by a slide door. In both rooms there are twelve computers available separated by screens, so people cannot see each other. Every subject was assigned to a computer number and login name and password upon arrival. We experimented with rooms consisting of females and males only and with rooms where males sat next to females only and females next to males. There are no significant differences between females sitting next to males and females sitting in a room with females only. The same goes for males. One room was equipped with an air-conditioning system, but this too does not show up in the results. There are also no significant differences between sessions at different times of the day, supervisor etc.

The supervision during the experiment was always conducted by two persons: One

professional and one of the authors. The professional made sure people were not talking and looking at other people's answers (which made no sense because all questions were assigned randomly, so nobody had the same question at the same time). He also guided people to the exit when they completed the experiment. We controlled the progress of the experiment on a master computer. On this computer the progress and cumulative earnings of every participant were followed. After a subject completed the experiment, he had to leave the room to receive his total earnings in cash in a separate room.

Before the sessions started, one of the authors read the instructions to all participants. All subjects were entitled to receiving a show up fee of €5.00, which was paid after having finished the entire test. During the cognitive tests they could earn €30.00 when they answered all questions correctly. The average earnings during the cognitive test were €16.53.

The experiment was programmed in PHP/MySQL and subjects completed the experiment using the Microsoft Internet Explorer. All computers showed the login screen upon arrival and when subjects logged on to the experiment they could start. It was not possible to go back and forth in the program so once an answer had been given and the subject had pressed "continue" or once time ran out during questions with a time constraint, the next question appeared on the screen.

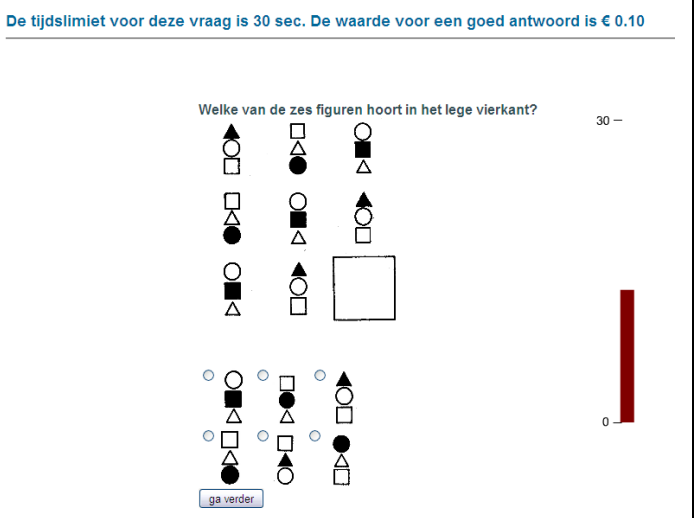
The data were collected on a server. The investment in answering the cognitive test questions is measured in milliseconds. Because server time can be longer when the network is in heavy use, we checked delays. The average delay is about 2 seconds, with no differences between the different sessions.

A.2. Types of questions

We now document representative examples of the seven types of different IQ questions subjects had to answer during the experiment. For each type of IQ questions an

example is given here by a screen-dump of that specific question, followed by a translation. The button “ga verder”, which is present in all examples, means “continue”. In case of time limits the bar at the right start at the marks 30 or 60 and becomes smaller and smaller until it reaches 0. If the time limit is reached the experiment continues with the next question. Awards for questions and time limits are randomly chosen in the experiment and here added to some question by means of example

Raven matrix



The time limit for this question is 30 sec. The value for a correct answer is €0.10

Which of the six figures should be placed in the empty square?

Cognitive Reflection Test

The time limit for this question is 60 sec. The value for a correct answer is €1

If 5 machines need 5 minutes to produce 5 widgets, how long need 100 machines to produce 100 widgets?

…minutes

Fill out the number and click on ‘continue’

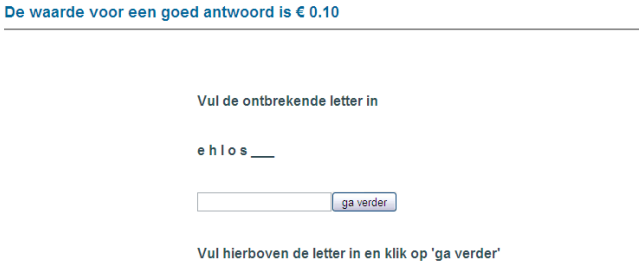
Anagram

	Choose the word containing the characters of which no car brand can be made Select the correct answer and click subsequently on 'continue'

Sequences or matrices of numbers

	The value for a correct answer is €0.40 Fill out the missing number Fill out the number and click on 'continue'

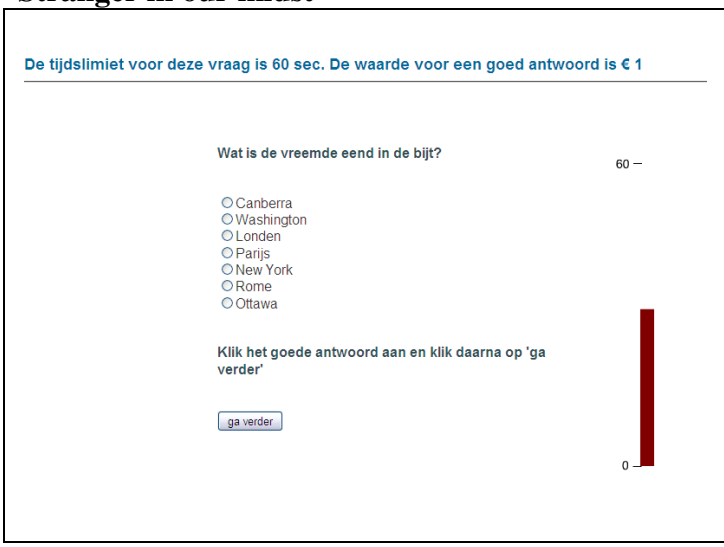
Sequence or matrix of characters

	The value for a correct answer is €0.10 Fill out the missing character Fill out the character and click on 'continue'

Filling in linking words

	The time limit for this question is 30 sec. Fill out the word that forms the last characters of the first word and the first characters of the last word Fill out the word and click on 'continue'

Stranger in our midst



The time limit for this question is 60 sec. The value of a correct answer is €1

What is the stranger in our midst?

Tick the correct answer and click subsequently on 'continue'

A.3. Instructions

Before the experiment started students were assigned a chair and a computer. They were read a set of rules of conduct and given a username and log in code. Once they were logged in, they had to read instructions from the screen before they could start with the experiment. All instructions were in Dutch and below we provide a translation. The translation is in *italics*.

Screen 1

This research consists of a number of parts. First, a set of questions on personality traits will be asked. There are four blocks with statements. You have to state to what extent they are applicable to you. Everybody is different, so there are no good or bad answers. Fill in the answer that you think suits you best.
After this block you will have to answer a set of 10 questions from an IQ test. Only one possible answer is the right one. Questions differ in terms of difficulty and type. Some questions are easy, while others are almost impossible to answer correctly.
The first two parts of the experiment take about 30 minutes.
After this, we present 7 sets of 10 questions where you can earn money. Everybody receives a payment of €5 for showing up, but most participants are likely to add €15-€25 to this amount. This can be more if you perform very well.

This part of the experiment will take about 90 minutes.
After you have completed the questions you are allowed to leave. On the screen you will see in what room you can collect your money. If you leave earlier or do not complete the experiment, we cannot pay you any money.

Screen 2
Instructions for personality traits measurement. These are irrelevant for the present paper.
After the personality questions are done the following instruction screen pops up.

Screen 3
You are now about to start the IQ test. The first set of 10 questions will be presented to you now. Questions differ in terms of difficulty and type. Some questions are easy, while others are almost impossible to answer correctly.
After this set of 10 questions the following instruction appeared on the screen.

Screen 4
We continue with 7 blocks of 10 questions for which you can earn money. You will receive similar types of questions as in the first block.
Two things are going to change.
First, during some blocks you will receive money for submitting the correct answer.
Second, sometimes we include a time limit. If time has run out the following question appears.
At the beginning of each block of 10 questions the one of the following screens appeared.

Screen 5a
For the next block of ten questions you will receive € … for submitting the correct answer. You are allowed to spend … seconds on the question. When time runs out the next question appears on the screen.

Screen 5b
For the next block of ten questions you will receive € … for submitting the correct answer. There is no time limit for this block.

Screen 5c
For the next block of ten questions you will receive no pay for submitting the correct answer. You are allowed to spend … seconds on the question. When time runs out the next question appears on the screen.

Screen 5d
For the next block of ten questions you will receive no pay for submitting the correct answer. There is no time limit for this block.

References

Angrist, J.D., Lang, D. & Oreopoulos, P. (2009). Incentives and services for college achievement: evidence from a randomized trial. *American Economic Journal: Applied Economics, 1* (1), 136-163.

Angrist, J.D. & Lavy, V. (2009). The effect of high stakes high school achievement awards: Evidence from a school-centred randomized trial. *American Economic Review, 99* (4), 1384-1414.

Bettinger, E. (2010). Paying to learn: the effect of financial incentives on elementary school test scores. NBER Working Paper No. 16333.

Borghans, L., Duckworth, A.L., Heckman, J.J. & Ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of Human Resources, 43* (4), 974-1061.

Borghans, L., Golsteyn, B.H.H., Heckman, J.J. & Humphries, J.E. (2011). Identification problems in personality psychology. *Personality and Individual Differences, 51*, 315-320.

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

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

Duckworth, A.L. & Seligman, M.E.P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science, 16* (12), 939-944.

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

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19* (1), 24-42.

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

Fryer, R.G. (2012), Aligning student, parent and teacher incentives: evidence from Houston public schools. NBER Working Paper No. 17752.

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

Heckman, J.J., Stixrud, J. & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics, 24* (2), 411-482.

Heckman, J.J., Humphries, J.E. & Mader, N.S. (2011). The GED. In E.A. Hanushek, S. Machin & L. Woessmann, *Handbook of the Economics of Education* (pp. 423-484). Amsterdam: North-Holland.

Kremer, M., Miguel, E. & Thorton, R. (2009). Incentives to learn. *Review of Economics and Statistics, 91* (3), 437-456.

Liu, L. & Neilson, W.S. (2011). High scores but low skills. *Economics of Education Review, 30* (3), 507-516.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., et al. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America, 108* (7), 2693-2698.

Prevoo, T. & Ter Weel, B. (2012), The importance of early conscientiousness for socio-economic outcomes. Working paper.

Sackett, P.R., Borneman, M.J. & Connelly, B.S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*, 215–227.

Segal, C. (2011). Working when no one is watching: Motivation test scores, and economic success, *Management Science, 58* (8), 1438-1457.

Todd, P. & Wolpin, K.I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal, 113*, F3-F33.

Winters, M.A., Trivitt, J.R. & Greene, J.P. (2010). The impact of high-stakes testing on student proficiency in low-stakes subjects: Evidence from Florida's elementary science exam. *Economics of Education Review, 29* (1), 138-146.

Figure 1
The theoretical relationship between the probability of submitting the correct answer and time

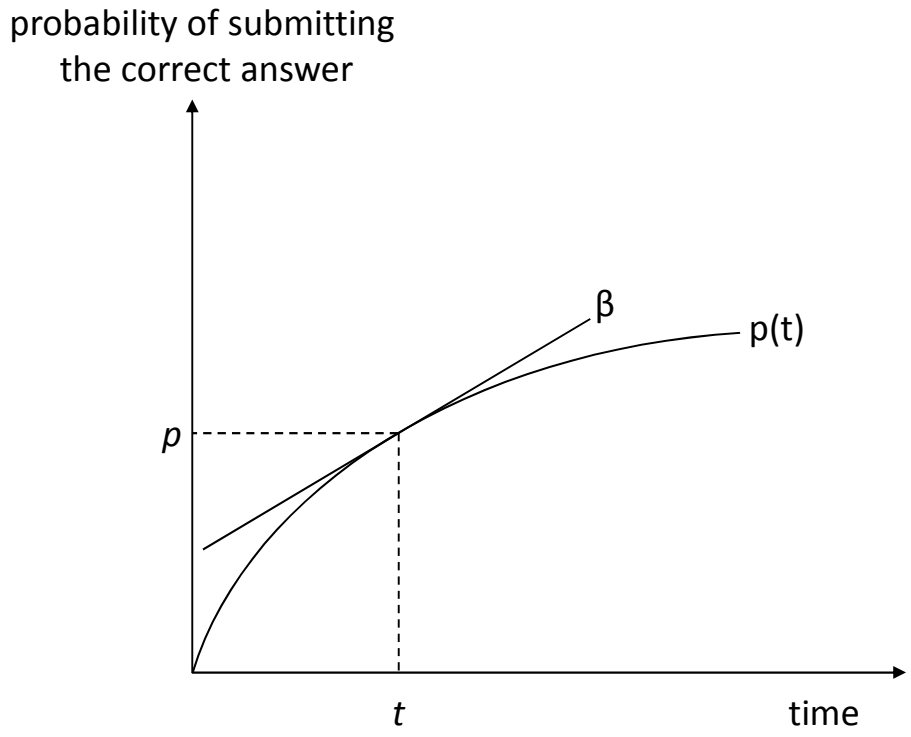


Figure 2
Constraints on the slope of the tangent

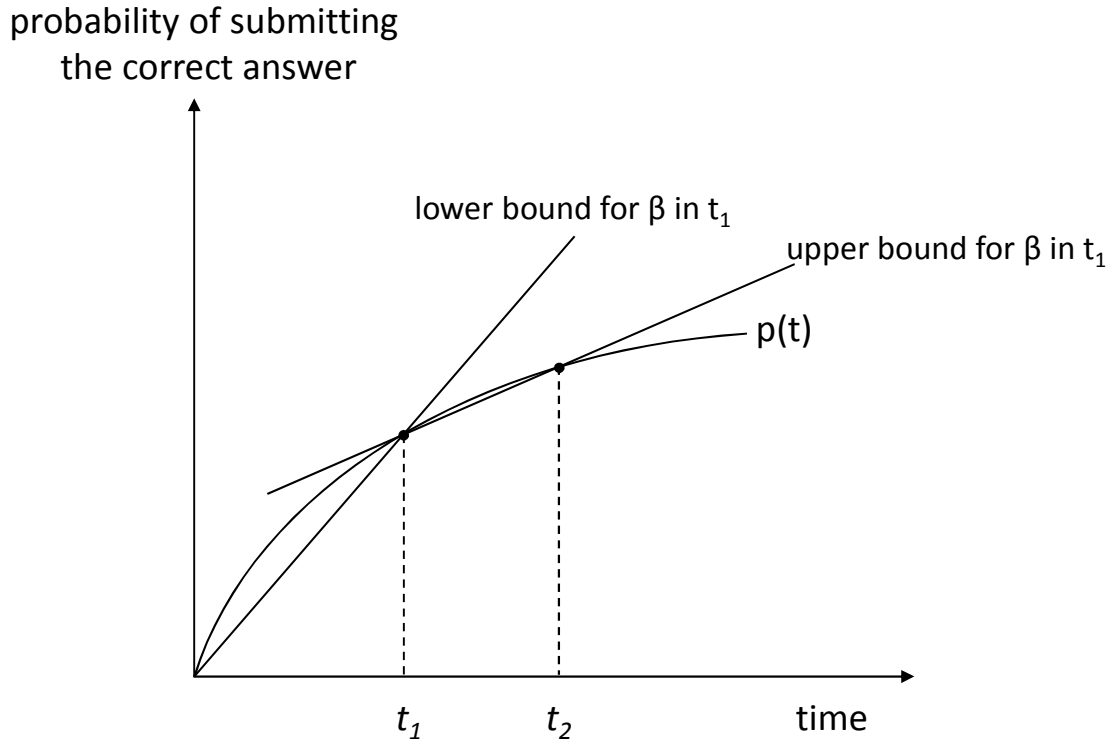
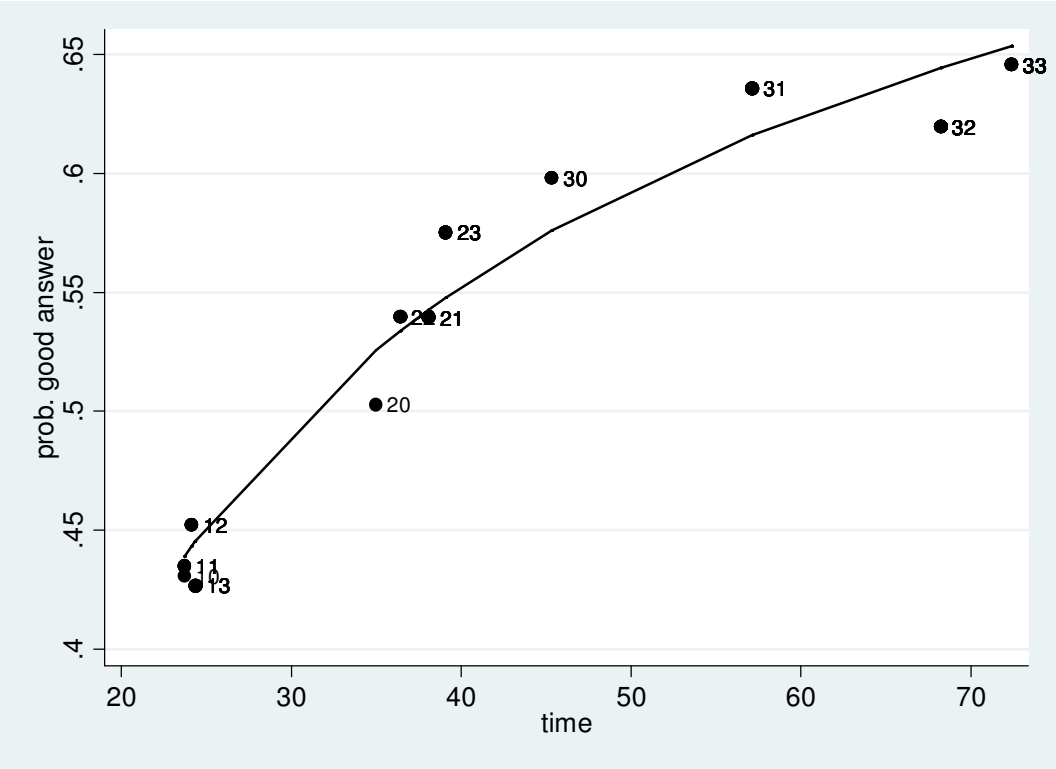


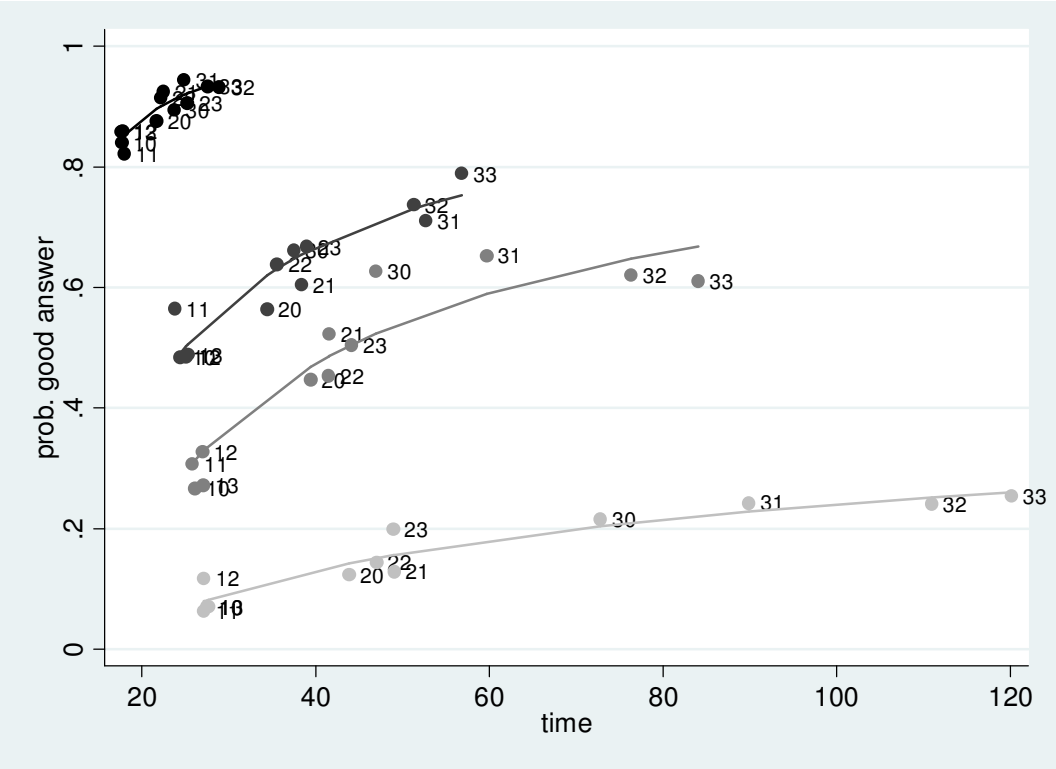
Figure 3
The pattern obtained from the experiment between the probability of submitting the correct answer and time investment



Note: The coding of the dots is the following. The first digit represents the time limit imposed: 1 is a 30 second time limit, 2 is a 60 second time limit and 3 is no time limit imposed. The second digit represents the incentive pay: 0 is no pay, 1 is €0.10, 2 is €0.40 and 3 is €1.00 for submitting the correct answer on a question.

Figure 4

The pattern obtained from the experiment between the probability of submitting the correct answer and time investment for different quartiles of the distribution



Note: Questions are selected into four different quartiles based on the average scores. The darkest dots are the questions that were the easiest, the lightest dots are the questions that turned out to be the hardest ones. The coding of the dots is the following. The first digit represents the time limit imposed: 1 is a 30 second time limit, 2 is a 60 second time limit and 3 is no time limit imposed. The second digit represents the incentive pay: 0 is no pay, 1 is €0.10, 2 is €0.40 and 3 is €1.00 for submitting the correct answer on a question.

Table 1
Estimation results of the technology function
(dependent variable: probability of submitting a correct answer)

		(1)	(2)	(3)	(4)	(5)
time investments	$\ln(t)$	-0.43 (0.03)	-0.43 (0.03)	-0.43 (0.04)		-0.48 (0.06)
incentive payments	dummy		0.00 (0.03)		0.04 (0.09)	-0.02 (0.03)
	€0.10			-0.02 (0.04)		
	€0.40			0.01 (0.04)		
	€1.00			-0.01 (0.04)		
Combined	dummy*no limit				-0.32 (0.09)	0.06 (0.06)
R^2		0.95	0.96	0.96	0.59	0.95

Note: Clustered standard errors in parentheses.

Table 2

Regression results for the technology function in which time investments interact with dummies for four different quartiles (dependent variable: probability of submitting the correct answer)

		(1)	(2)	(3)
time investments	$\ln(t)$	-1.87 (0.21)		-1.76 (0.25)
	$\ln(t) * d_2$	1.02 (0.24)		0.98 (0.25)
	$\ln(t) * d_3$	1.25 (0.23)		1.19 (0.24)
	$\ln(t) * d_4$	1.73 (0.22)		1.65 (0.24)
incentive payments	dummy		0.00 (0.07)	-0.03 (0.04)
combined	dummy*no limit		-0.43 (0.07)	-0.06 (0.07)
dummies	d_2	-1.44 (0.79)	1.30 (0.08)	-1.35 (0.80)
	d_3	-1.85 (0.74)	1.63 (0.08)	-1.70 (0.76)
	d_4	-3.09 (0.71)	2.12 (0.08)	-2.89 (0.74)
R^2		0.98	0.95	0.98

Note: Clustered standard errors in parentheses. d_2, d_3, and d_4 are dummies for second, third and fourth quartiles in difficulty, respectively. The first quartile, the easiest questions, serves as a reference group.

Table 3

Upper and lower bounds based on four levels of difficulty with and without incentive payments

	Without incentive payments		With incentive payments		Bounds	
	t	*p*	*t*	*p*	Lower bound	Upper bound
d_1	21.12	0.87	22.54	0.90	24.25 (1.32)	52.20 (60.72)
d_2	32.96	0.58	37.89	0.62	56.40 (3.57)	122.08 (98.39)
d_3	38.45	0.47	47.63	0.48	82.47 (5.63)	783.35 (1964.49)
d_4	51.78	0.15	59.38	0.16	349.52 (52.20)	703.39 (1478.07)

Note: *t* is measured in seconds.

Table 4

The value of answering a question in different incentive schemes obtained from assuming optimal investment behaviour

		Incentive payments			
		No incentive payment	€0.10	€0.40	€1.00
Time limit	30 seconds	97.87 (7.69)	97.89 (7.45)	100.53 (7.58)	101.68 (7.63)
	60 seconds	170.80 (12.93)	192.98 (12.57)	180.86 (12.05)	200.31 (12.88)
	No limit	247.76 (20.06)	344.90 (24.54)	444.68 (30.51)	484.07 (34.35)

Note: Clustered standard errors in parentheses.

Table 5

The value of answering a question in different incentive schemes as obtained from assuming optimal investment behaviour without time limits

	Incentive payments			
	No incentive payment	€0.10	€0.40	€1.00
d_1	144.53 (35.31)	165.34 (34.77)	254.73 (56.13)	221.43 (41.86)
d_2	124.42 (22.23)	233.95 (32.81)	222.18 (33.52)	268.83 (38.23)
d_3	158.50 (23.52)	233.62 (29.70)	348.46 (44.78)	407.26 (45.77)
d_4	619.86 (66.44)	789.54 (78.34)	1004.86 (90.09)	1101.23 (114.08)

Note: Clustered standard errors in parentheses.

Table 6

Estimation results of the CES function with intrinsic and extrinsic motivation, based on four and eight levels of difficulty

		(1)	(2)
	ρ	-0.93 (0.23)	-0.88 (0.19)
difficulty of questions	d_2	1.01 (0.03)	1.01 (0.03)
	d_2	1.04 (0.16)	0.93 (0.17)
	d_4	1.18 (0.78)	0.99 (0.05)
	d_5		0.96 (0.09)
	d_6		1.06 (0.17)
	d_7		1.29 (0.76)
	d_8		1.36 (0.97)
incentive payments	no incentive payments	0.44 (1.58)	0.86 (1.76)
	€0.10	0.48 (1.75)	0.95 (1.99)
	€0.40	0.51 (1.88)	1.01 (2.14)
	€1.00	0.51 (1.91)	1.04 (2.21)

Note: Clustered standard errors in parentheses.