Pre-Analysis Plan

Heterogeneity in Effort Provision:

Evidence from a lab-in-the-field experiment

October 8, 2021

1 Introduction

Understanding how to motivate people to provide effort is of key importance for success in many domains of life, ranging from the educational sector to the labor market. Extensive theoretical and empirical research has focused on the role of both intrinsic and extrinsic incentives to promote effort provision among workers as well as students (Prendergast, 1999; Gneezy et al., 2011; Gneezy and Rey-Biel, 2014; Cassar and Meier, 2018; Gneezy et al., 2019).

In this study, we strive to understand how personal characteristics, ability, preferences and socio-economic background shape one's performance under various incentive schemes. The idea stems from the observation that not everybody is best motivated by the same type of incentives. Some people thrive and express their best potential in competitive environments, while some instead choke under such pressure (Ariely et al., 2009). Some people are diligent and work hard regardless of the environment, while others need monetary rewards to be motivated. Scientific evidence offers surprisingly little guidance in understanding the underpinning of this heterogeneity. New evidence can provide important insights in how to improve labor market contracts to better account for individual characteristics which are crucial for one's performances. This can help to optimize output and earnings under different incentive schemes.

Here we consider a number of factors that have been proven important in understanding labor market success, such as socio-economic status (see, e.g., Heckman,

2006, 2007), conscientiousness (see Donato et al., 2017), grit (see Duckworth et al., 2007; Alan et al., 2019), economic preferences, and parenting styles (see, e.g., Bonin et al., 2007; Sutter et al., 2013; Borghans et al., 2008; Falk et al., 2016, 2018). While the link between some personal characteristics and (labor market) outcomes is quite well established, the mechanisms behind it are far less understood. In particular, it is not a priori clear if and how the above characteristics interact with specific incentive schemes.

One can think of two possible mechanisms at work. First, people with *desirable* characteristics – i.e., associated with positive outcomes, such as conscientiousness or patience – might be motivated to work hard irrespectively of the incentives. In the long run, this might give them an advantage over other students or workers as they will be persistent and productive in competitive and profitable situations as well as in situations without contingent incentives. Second, it could be that people with desirable characteristics are responsive to incentives and provide their maximal effort only when it is adequately rewarded. Considering effort as a scarce resource that can be depleted, such insights about individuals' motivation are important to maximizing earnings.[1]

A better understanding of these mechanisms can have far reaching implications for practitioners as well as for theoretical models. Unfortunately, field data on personal characteristics and productivity under different incentive contracts are scarce and a proper identification strategy is difficult to achieve because of self-selection as well as market frictions. For this reason, we conduct a controlled lab-in-the-field experimental study in a setting with limited self-selection, where we first collect a wide array of personal characteristics. This provides us with an ideal setting to investigate how people's characteristics are related to performance under different incentive schemes.

We implement two treatments. In one treatment, we exogenously assign participants to one of three different incentive schemes – fixed payment, piece-rate payment or tournament payment. We implement a tedious counting task adapted from Abeler et al. (2011) to measure effort provision. The task provides us with an objective measure to assess performance in the different incentive schemes. In the other treatment, we ask participants to self-select into one of the three incentive

[1] One could think of more nuanced mechanisms and it is not necessarily the case that all personal characteristics are relevant in understanding how different people react to incentives.

schemes. The goal of this treatment is twofold. First, it allows us to test if participants are able to correctly self-select into the incentive contract that maximizes their potential monetary earnings.[2] Second, it serves as a control to see if having agency over the type of incentive contract itself has an effect on one's performance.

We conduct our lab-in-the-field experiment in German high schools, with students from grade ten and older (16 to 20 years of age). Relying on a pool of high school students offers a number of advantages, which are crucial to our research question. First, they are highly heterogeneous in terms of socio-economic background, traits, and ability level. This is an advantage of our sample, as different personal characteristics could play a role for their performance, but also their preference for a payment scheme. Second, students are exposed to similar incentives in schools and have not yet developed a preference for different compensation schemes through experiences in the labor market, which might affect their behavior.

This document is structured as follows. In Section 2, we describe the subject pool, the experimental design and in particular the real effort task applied. In Section 3, we outline the main variables of interest.

2 Research design

Subject pool. We conduct the experiment with adolescents in schools across North Rhine-Westphalia, Germany. All students enrolled in tenth to thirteenth grade are eligible for recruitment. We target what in German is referred to as "Gesamtschule": schools that commonly comprise both low and high education tracks within the same institution. This ensures a heterogeneous sample in our study (SES; cognitive abilities, etc.). We contact schools in random order from the set of around 200 closest schools in the area of Bonn, Cologne, and Düsseldorf within the state of North Rhine-Westphalia.[3] We first inform and invite schools to participate in the study via a letter. In case of no reply, we contact the school via phone, and send them a more detailed description of the study via email.

In each school, participation in the study has to be approved by the school prin-

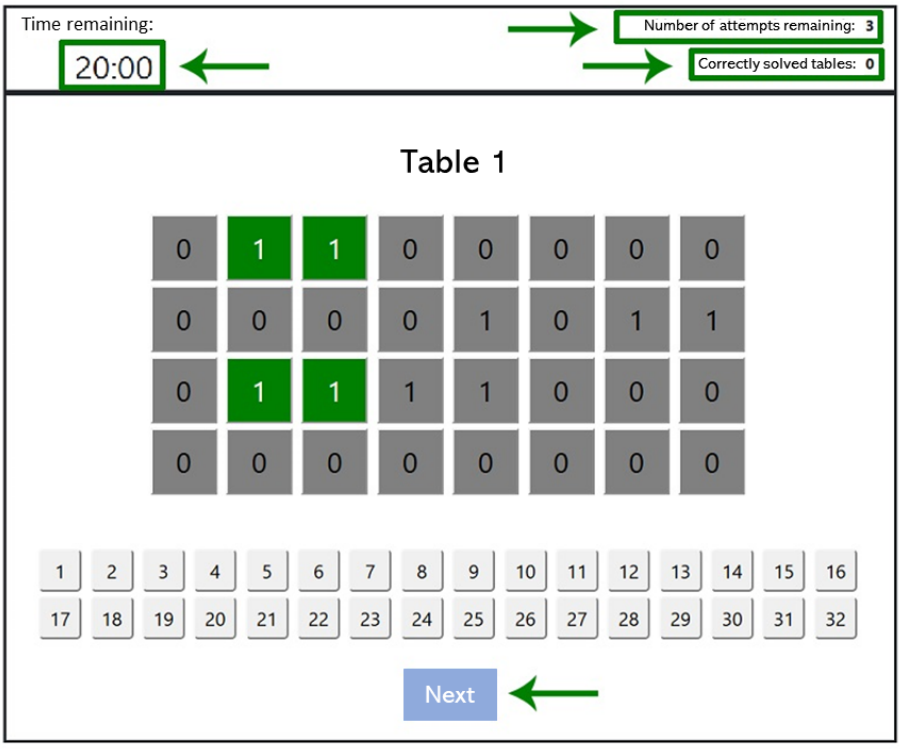
[2] Certainly, other factors than material earnings might influence the choice over different incentive schemes.

[3] Contact information are publicly available online on the webpage of the Ministry for School and Education of North Rhine-Westphalia: `https://www.schulministerium.nrw.de/BiPo/SchuleSuchen/online`.

cipal.[4] Parents are informed about the experiment and need to sign a consent form in order for a student to participate in the study.[5] Participation is voluntary and it is explicitly mentioned to the subjects that they can quit the study (or skip specific parts) at any time.

Real effort task (RET). We implement a counting task adapted from Abeler et al. (2011). Subjects are presented with a sequence of tables containing zeros and ones (Figure 1). The task consists in highlighting and counting the ones present in each table (see instructions in Appendix A). A table is correctly solved if: (i) all the ones are highlighted, (ii) none of the zeros are highlighted, and (iii) the total amount of ones is correctly reported. Subjects are given a total of three trials to solve each table. The task has a number of desirable features: it does not require any prior knowledge, performance is easy to objectively measure, learning only plays a minor role, and there is no outside value in performing the task (Abeler et al., 2011; Charness et al., 2018). The task is tedious and requires effort to be solved.

Figure 1: Real-effort task



[4]We have obtained IRB approval for our study by the ethical board of the University of Innsbruck (certificate of good standing N.o 07/2019 - 25.01.2019).

[5]Students aged 18 and older can sign the consent form themselves.

Timeline and treatments. The experiment comprises two parts to be conducted approximately two weeks apart from each other (see Table 1 for an overview over the two parts).[6] The same subjects participate in both parts of the experiment. In part 1, a broad range of socio-demographic characteristics, traits and preferences is collected from participants. In part 2, effort provision is measured under three different incentives schemes and we implement two between-subjects treatments. In the remainder of this section, we describe in detail the incentive schemes, the treatments, and the sequence of tasks in the two parts of the experiment.

Incentive schemes. We implement the following three types of incentives schemes:[7]

- Fixed payment: a flat payment (€6.5) independent of the number of correctly solved tables;
- Variable payment: subjects are paid a piece rate (€0.06) per correctly solved table;
- Tournament payment: subjects are paid either a high (€0.08) or a low (€0.04) piece rate per correctly solved table. Each participant in this treatment is matched with another participant that also chooses tournament payment; s/he will be paid the high (low) rate if they solve more (less) tables than the matched participant.[8]

Treatments. We run two between-subjects treatments: *Exogenous* and *Choice*. In the *Exogenous* treatment, participants are assigned to either a fixed, variable or tournament payment scheme. Subjects only receive information about the relevant incentive scheme they are assigned to, and are paid accordingly. In the *Choice* treatment, participants receive information about all three incentive schemes and have to choose one scheme which determines how their payment is calculated.

Part 1 – Socio-demographics, traits and preferences. This part of the study is common to all treatments and measures a number of socio-demographic characteristics, traits and preferences of the participants. We focus on four main areas: ability, family background, preferences, and personality traits (for a detailed list of included questions, see the questionnaire in Appendix A). We start by collecting a measure of ability for the real effort task (RET), where we follow Dohmen and Falk (2011)

[6] In case of deviations due to logistic reasons, we always keep the delay between the two visits below six weeks.

[7] Payment amounts for the incentive schemes were calibrated based on pilot data.

[8] Ties are solved by a random draw of the computer and participants are made aware of that.

Table 1: Timeline of the experimental tasks

Part 1 – socio-demographics, traits and preferences	**Exogenous treatment**	**Choice treatment**
Personal ID	✓	✓
RET instructions	✓	✓
Ability (RET 5 min)	✓	✓
IQ test (Raven's matrices, 5min)	✓	✓
SES questionnaire	✓	✓
Big Five (BFI-44)	✓	✓
Competitiveness (14-item)	✓	✓
Preference module	✓	✓
Positive parenting (6-item)	✓	✓
Grit (8-item)	✓	✓
Payment (€)	€4 + RET (5 mins)	€4 + RET (5 mins)
Average time	45 min	45 min
N. obs. (target)	1,029	1,029

Part 2 – Effort provision and incentives	**Exogenous treatment**	**Choice treatment**
Personal ID	✓	✓
Belief elicitation	✓	✓
RET instructions	✓	✓
Instructions (incentive schemes)	1 incentive scheme (within session randomization)	all 3 incentive schemes
Choice (incentive scheme)	-	✓
RET (20 min)	✓	✓
Payment (€)	€1 + RET (20 min) + belief	€1 + RET (20 min) + belief
Average time	45 min	45 min
N. obs. (target)	1,029[§]	1,029

Notes: [§]About 1/3 of participants were assigned to each of the three incentive schemes. See the paragraph on *Sample size and Data Collection* for information on power calculation.

in incentivizing the task. Participants are given five minutes to solve as many tables as they can, and are paid on a piece rate (€0.06) basis.[9] Next, a five minutes computerized version of a standardized non-verbal intelligence test is administered (Raven, 2000). In addition, we collect demographic information, as well as information about socio-economic status (SES). Our items are informed by three different socio-economic indices, where we also added some own questions (see next section for a detailed list). With respect to psychological measures, we collect the "Big five" (John and Srivastava, 1999), competitiveness (Newby and Klein, 2014), positive parenting style (Essau et al., 2006; Frick, 1991), and grit (Duckworth and Quinn, 2009). For all these measures, we rely on widely used psychological scales. Finally, we include a series of non-incentivized questions taken from the validated preference module by Falk et al. (2016, 2018): patience, willingness to take risk, and altruism.[10]

Part 2 – Effort provision and incentives. This part of the study captures effort provision under the different incentive schemes. The RET is the same as in part 1, but lasts for 20 minutes. We vary the incentives according to participants' payment schemes. Before the RET, we elicit participants' beliefs about their performance in part 1. More precisely, we rank all the participants present in the room based on their performance in part 1 and then ask them to guess their rank. If the guess is correct, they earn €2. If the difference between the guess and the actual ranking is at most 5 placements, they still earn €0.50. Only participants who are present in both visits can be included in the ranking (and are asked to guess). We have opted for collecting the guesses about their ranking in the second visit. We did this as we want to examine how beliefs about rankings relate to the choice of incentive scheme and to other variables collected in the second session. As students can update their beliefs between the first and the second session, we eliminate this issue by eliciting the beliefs at the beginning of the second session. Feedback about the guessing task is given only at the end of the study.

Assignment to treatments. In the *Exogenous* treatment we randomly assign participants within a session to one of the three incentive schemes based on the min MSE method developed by Schneider and Schlather (2017). Based on re-randomization, this method aims at minimizing the mean squared error of the treatment effect estimator as a function of treatment assignment. The method thus increases precision

[9] To familiarize with the task and the software, subjects are asked to solve a trial table before moving to the actual task.

[10] For patience and willingness to take risk, we include both qualitative and quantitative items (see our instructions and the questionnaire in Appendix A).

of the treatment effect estimation by choice of treatment assignment. Intuitively, the method forms comparable treatment groups considering multivariate information such as gender, SES, ability, etc. We opt for this method, as it allows us to assign three treatments in the same session while still "balancing" multivariate and continuous information in a principled way.[11] Moreover, balance with respect to the considered variables is less affected in case of attrition. To achieve balanced treatment groups, we consider pre-treatment information on the productivity in the RET, demographic information, socio-economic characteristics, psychological measures as well as preferences.[12]

Procedures. To avoid self-selection into the study, we conduct it in schools during regular school hours. Sessions are run in large lecture halls and possibly several classes take part in the experiment at the same time.[13] The number of participants in a single session can range from between 15 to 75. The experiment is conducted with up to 75 tablets and a server using oTree (Chen et al., 2016).

In each of the two sessions, subjects are randomly assigned to a desk upon arrival. They are separated by privacy screens, and communication is strictly forbidden throughout the experiment. This is enforced to avoid students comparing choices or their performance. Teachers are allowed in the classroom, but are not allowed to communicate with or observe the behavior of the participants. In the first session, the relevant instructions are read aloud, and displayed on the screens before the beginning of the real effort and IQ task. In the second session, subjects are reading the instructions displayed on the screen alone, since multiple incentive schemes are randomized within the same session in the exogenous treatment. To ensure that subjects fully understand the incentive schemes, they first have to individually answer a set of computerized control questions before proceeding with the task itself.

Since subjects take part in two separate sessions, data is matched via a personal ID created by the participants at the beginning of each part (see the instructions

[11] Previous work mostly relies on binary assignment between two different treatments.

[12] The following data is being considered during treatment assignment: Performance in the 5-minute RET, IQ, scholarly achievements (Math and German grade), intention to obtain the "Abitur", gender, an indicator for living with both parents, parental education, wealth data (consisting of the amount of pocket money, number of holidays last year, number of books at home, number of smartphones, TV sets, computers, cars, and bathrooms, and indicators for having internet access and an own room), a migration index (born outside of Germany and having at least one non-German parent), Big 5 factors (extraversion, agreeableness, conscientiousness, neuroticism, openness) separately, competitiveness and contentiousness, positive parenting style, grit, patience index, willingness to take risk/risk aversion index, and altruism.

[13] Due to logistic constraints, some session may be conducted in single classrooms.

in Appendix A). Each part lasts around 45 minutes (a regular school hour) and participants are paid anonymously and in cash at the end of each session. In part 1, participants receive a fixed payment of €4, plus the earnings for the 5 minutes RET. In part 2, participants earn €1 show-up fee, plus the earnings for the 20 minutes RET and those from the guessing task (beliefs).

Sample size and Data Collection. As we are particularly interested in heterogeneous treatment effects in order to see how, e.g., females might react differently to a given incentive system, we focus on interaction effects to determine the sample size. Based on pilot data, we aim for 80% power to detect an interaction effect of about 30% the size of a main effect (e.g., main effect +100; effect for males +115; effect for females +85). According to analytical power calculations and simulations, this can be achieved with a total sample size of about 2000 (simulations: 1992; analytical calculations: 2058). See Appendix B for more details on our power calculation.

Originally, data collection was scheduled for mid-2019 to mid-2020. Following the outbreak of the Covid-19 pandemic in January 2020, however, we have only been able to collect data of about 200 individuals in 2020 so far. Given the current Covid-situation, currently in its fourth wave, it is still impossible to predict if and how further data collection in schools will be possible. To comply with the conditions that our junior colleagues face concerning the duration of their dissertation and their contract period, we will start analyzing the data once this pre-analysis plan is preregistered, and at the same time, try to collect further data until reaching our goal.[14] At the latest in May 2022, however, we will stop data collection.

3 Empirical strategy

We start this section with an in-depth description of the explanatory variables. All variables were carefully chosen based on their potential to shape earnings and life outcomes, as reported in the literature. We give a summary of the findings in the literature on explanatory variables on sorting at the end of this section. Regarding interaction effects between explanatory variables and incentive schemes, we are only aware of few papers papers that tackle this question with respect to some of the

[14]For logistic reasons, it is impossible to aim for exactly 2058 observations: We will not refuse participation of a single class in a school only because that would mean exceeding the planned sample size; similarly, if a school that we had contacted before meeting our sample size goal offers participation, we will accept the offer, thereby also investing in a good relationship with nearby schools.

variables we elicit. Donato et al. (2017) report that higher conscientiousness and higher neuroticism predict worse response to performance incentivization among maternity care providers in India. A similar pattern between conscientiousness and reaction to incentives (albeit only for men) is found in a lab experiment by Segal (2012). Finally, Gneezy et al. (2003) and Ors et al. (2013) report evidence that women react worse to competitive incentives than men.

Socio-economic status and other relevant socio-demographic variables. SES has been shown to be strongly associated with educational outcomes and earnings (Cunha and Heckman, 2007; Heckman, 2006). Our questionnaire includes a number of proxies for SES[15], informed by three main indices:

- PISA wealth index: The PISA test provides valuable information to educational researchers and policy makers around the world by comparing countries with regard to their educational system using a variety of educational outcomes. In many ways PISA has emerged as the international benchmark in comparing educational systems (Fuchs and Wößmann, 2008; Hanushek and Woessmann, 2011; Woessmann, 2016). Their SES indicators have often been used for assessing socio-economic background with teenagers (Hanushek and Woessmann, 2011; West and Woessmann, 2010; Woessmann, 2016). We focus on the family wealth possessions index, which has been validated as a strong and reliable proxy for SES (Schulz, 2006; Rutkowski and Rutkowski, 2013).[16] It includes seven items: *(i)* Do you have a room of your own? Or do you share your room (e.g., with siblings)?; *(ii)* Do you have a link to the internet at home?; *(iii)* How many cell phones are there at your home?; *(iv)* How many televisions are there at your home?; *(v)* How many computers are there at your home?; *(vi)* How many cars are there at your home?; and *(vii)* How many rooms with a bath or shower are there at your home?

- Family Affluence Scale (FAS) score: This score is also commonly used to elicit SES among school-aged children (Andersen et al., 2008; Boyce et al., 2006; Hartley et al., 2016; Torsheim et al., 2016). The score is similar to the PISA wealth index, and three of the items are the same. It includes four items: *(i)* Do you have a room of your own? Or do you share your room (e.g. with siblings)?; *(ii)* Does your family own a car, van or truck?; *(iii)* How many times

[15]Given our sample of adolescents, elicited information about household income must be assumed to be very noisy, which is why we use several proxies.

[16]The questions are drawn from PISA tests conducted in 2015. They were accessed from `https://www.oecd-ilibrary.org/education/pisa-2015-assessment-and-analytical-framework/pisa-2015-background-questionnaires_9789264255425-8-en`.

did you and your family travel out of Germany abroad for holiday/vacation last year?; and *(iv)* How many computers does your family own?[17]

- Education and family structure: We follow Kosse et al. (2020) in considering educational and time resources available to the family as important determinants of SES. They classify a participant as low SES if at least one of the following two conditions are fulfilled: *(i)* neither parents has a college degree; *(ii)* the participant lives in a single-parent household.[18]

- We collect several other relevant variables, such as the number of books available at home, which has been found to alone be another important proxy for socio-economic status in the PISA test (Woessmann, 2016). Moreover, we add gender, age, number of siblings, zip-code, pocket money, migration background and speaking a different language than German at home (Hansson and Gustafsson, 2013; Woessmann, 2016).

Given that we elicit numerous variables to capture socio-economic status, we apply a principal component analysis (PCA) to these socio-economic variables in order to reduce dimensionality of these extensive data. Using the resulting weights from the first component, we then construct a single-item socio-economic status measure including all of the items of the three different socio-economic indices and the additional relevant variables.

Reference level of productivity and stress level. Baseline measure of performance will be captured in Part 1, where a 5 minutes RET paid on a piece-rate will be performed. This will serve as a proxy for individual productivity in playing the real effort task. The reference level of productivity has been found to be important for sorting decisions (Dohmen and Falk, 2011).[19] At the end of the RET (both 5 and 20 minute version) we follow Dohmen and Falk (2011) and elicit self-reported measures of effort, stress, and exhaustion. All three have been found to be higher in pay for performance schemes compared to fixed payment schemes.

Beliefs. Overconfidence will be computed based on incentivized beliefs. Prior to starting the 20 minutes real effort task in Part 2, we collect information about a

[17]The composite FAS score is calculated for each adolescent by adding the four items and further categorized into scores below 5, scores between 5 and 7 and scores above 8.

[18]**?**Kosse et al. (2020) consider a third dimension to define SES: household income. While we do not have the actual income, we could use the FAS score, the PISA wealth index, or possibly also the zip-code as a proxy for household income to compute an index similar to the one proposed by **?**Kosse et al. (2020).

[19]Subjects are instructed to solve as many tables as they can, and are given 0.06 cents for each correctly solved table.

participant's guessed rank in the 5 minute real effort task in Part 1. They get paid according to their guess at the end of the study.[20] Beliefs have been found to be important in sorting decisions, for example in explaining gender differences in sorting into tournaments and differences in productivity (Bordalo et al., 2019; DellaVigna and Pope, 2017; Dohmen and Falk, 2011; Larkin and Leider, 2012; Reuben et al., 2017). We can also classify participants spending a long time on their sorting decision, which arguably captures those that are marginal types, i.e., subjects who are not certain which scheme to sort into (see Dohmen and Falk, 2011).

Cognitive abilities. Cognitive ability has been found to be important for school attainment as well as future earnings (Hanushek et al., 2016; Borghans et al., 2008; Cawley et al., 2001; Segal, 2012). The main proxy for cognitive ability is the score obtained in the Raven's matrix test administered in Part 1 of the experiment (Raven, 2000). Additionally, we consider self-reported math and German grades, normalized at the grade level. All are expected to be highly correlated with productivity in the real effort task (Dohmen and Falk, 2011). Ex-ante, it is less clear whether it makes sense to combine math and German grades with the results from the Raven's matrix test for an index of cognitive abilities, as grades might also be influenced by grit and patience. If a factor analysis of these measures together with grit and patience suggests that they are rather related with cognitive abilities, we might also consider an index of cognitive abilities consisting of math and German grades, as well as the result from the Raven's matrix test instead of the measures in isolation.

Altruism, risk, and time preferences. Risk and time preferences predict labor market outcomes, educational attainment, income and wealth (Alan and Ertac, 2018; Becker et al., 2012; Bonin et al., 2007; Cadena and Keys, 2015; DellaVigna and Paserman, 2005; Dohmen et al., 2011; Golsteyn et al., 2014; Sutter et al., 2013; Von Gaudecker et al., 2011). They have also been found to be important for different sorting decisions (Bonin et al., 2007; Dohmen and Falk, 2010, 2011). Altruism, risk, and time preferences will be measured by using a subset of the global preference survey by Falk et al. (2016, 2018). For both time and risk preferences, we will collect multiple measures: a qualitative measure and a quantitative one (staircase). We will combine them as proposed by Falk et al. (2018) or using weights resulting from our own data. For altruism, we use the qualitative measure.

Big five. Personality traits, such as the big five (openness, conscientiousness, extraversion, agreeableness and neuroticism), have been shown to be stable traits in

[20] See details in the design section.

affecting performance and life outcomes (Almlund et al., 2011; Akee et al., 2018; Cubel et al., 2016; Deming, 2017; Lindqvist and Vestman, 2011; Segal, 2012). We collect data for all big-five, but will mostly focus on conscientiousness and neuroticism as they are found to be consistent predictors of performance in various settings (Borghans et al., 2008; Donato et al., 2017; Heckman and Rubinstein, 2001).

Competitiveness. A large literature documents gender differences in competitiveness. Women avoid competetive schemes, while men are competing too much (Gneezy et al., 2003; Niederle and Vesterlund, 2007, 2010). These gender differences can potentially explain differences in education and labor market outcomes (Gneezy et al., 2003; Niederle and Vesterlund, 2007, 2010). A high level of competitiveness is also a strong predictor for choosing a more prestigious academic track, controlling for ability (Buser et al., 2014, 2017; Reuben et al., 2017) as well as sorting decisions between different payment schemes (Dohmen and Falk, 2011). Competitiveness will be measured in our study on the basis of the Competitive Orientation Measure (one single composite scale; see Newby and Klein, 2014).

Parenting style. Parenting styles are important for the academic achievements and future success of children (Doepke and Zilibotti, 2017; Kosse et al., 2020). We elicit a vital component of parenting style: positive parenting (Essau et al., 2006; Frick, 1991), which indicates the use of positive stimuli and rewards by parents. Recent literature has shown that parental investments have important impact on child cognitive and non-cognitive outcomes (Carneiro and Heckman, 2003; Falk et al., 2021; Doepke et al., 2019).

Grit. Grit is defined as perseverance toward a set goal and is seen as being closely related to conscientiousness (Alan et al., 2019). Grit has been found to be predictive of success in a variety of contexts, e.g., in the case of college GPA and educational attainment (Alan et al., 2019; Duckworth et al., 2007). In their study, Alan et al. (2019) find that students participating in a grit focused intervention chose more challenging tasks and perform better in the real effort task. It will be measured in our study by the short-scale Duckworth Grit Index (Dobbie and Fryer Jr, 2015; Duckworth and Quinn, 2009). Ex-ante, it is less clear whether it makes sense to combine math and German grades with the Duckworth Grit Index for an extended scale of grit, as grades are also influenced by cognitive abilities and patience. If a factor analysis of these measures together with patience and cognitive abilities suggests that they are rather related with grit, we might also consider an extended

scale of grit consisting of math and German grades, as well as the Duckworth Grit Index instead of the Grit Index in isolation.

Table 2: Explanatory variables and sorting decisions

Paper	**Data Set**	**Female**	**Risk preferences**	**Competetiveness**	**Productivity**	**Overconfidence**	**IQ**	**Grit**	**Parenting style**	**Big 5**	**SES**
Alan et al. (2019)†	Private	▼	▲	-	-	▲	▲	▲	-	-	-
Buser et al. (2014)	Private	▼	▲	▲	-	▲	▲	-	-	-	-
Buser et al. (2017)	Private	▼	-	▲	-	-	-	-	-	-	-
Bonin et al. (2007)‡	GSOEP	-	▲	-	-	-	-	-	-	-	▲
Datta Gupta et al. (2005)	Private	▼	◓	-	-	○	-	-	-	-	-
Dohmen and Falk (2011)§	Private	○	▲	-	▲	▲	▲	-	-	○	-
Eriksson et al. (2009)	Private	○	▲	-	-	▼	-	-	-	-	-
Larkin and Leider (2012)$	Private	-	▲	-	○	▲	-	-	-	-	-
Niederle and Vesterlund (2007)	Private	▼	▲	▲	▲	▲	-	-	-	-	-
Reuben et al. (2017)¶	Private	▼	-	○	-	○	-	-	-	-	-

Notes: This table shows papers that demonstrate the role of explanatory variables on sorting decisions between piece rate or fixed and competetive environments. ▲ - Significant increase in going from piece rate or fixed to competetive payment scheme. ▼ - Significant decrease in going from piece rate or fixed to competetive payment scheme. ◓ - Mixed effects. ○ - No significant results. †- Compared sorting decisions for a more difficult risk real effort task compared to a less risky option. ‡- Investigates sorting decisions into occupations with low earnings risk. §- Compared sorting decisions for a fixed payment to piece rate or revenue sharing or tournament. $- Compared sorting decisions for a linear piece rate and a convex payout schedule that paid more per correct answer. ¶- Investigates sorting decisions into college major choices and future earnings. Data sets: GSOEP- German Socio-Economic Panel.

3.1 Research Framework I: Heterogeneity in performance across incentive schemes

We study the predictive value of relevant socio-economic and demographic characteristics, personality traits, and preferences on performance in a real effort task across exogenously assigned incentive schemes. The main objectives are:

- Study which socio-economic and demographic variables, traits and preferences are the main determinants of performance across different incentive schemes;
- Study the performance of individual types across different incentive schemes.

In the following, we provide a detailed description of the main approach we will use in this framework.

Study which socio-economic and demographic variables, traits and preferences are the main determinants of performance across different incentive schemes. In this part, we will study the effect of our main explanatory variables as determinants of performance in the 20-minute RET across three common incentive schemes which will be exogenously assigned: fixed payment, piece-rate payment and tournament payment.

The explanatory variables are socio-economic status, demographic information, cognitive abilities, economic preferences (altruism, time and risk preferences), big five personality traits, competitiveness, parenting style, and grit.[21] Moreover, we will investigate the performance in the 5-min RET as a proxy for individual productivity, and beliefs about performance in the 5-min RET (relative self-assessment and overconfidence). We might, however, have to investigate their role on incentive schemes and productivity in isolation, or add measures that are only based on these outcomes to the analyses. Such measures might be, or build on, for example, residuals from a regression of performance in the 5-min RET on the explanatory variables mentioned above. We might be unable to add these variables to the regressions (at least without additional processing), because they are most likely strongly related to performance in the 20-min RET. For example, the most overconfident student is necessarily not the best student in the classroom; most likely they are among those

[21] In this comparison, we do not take raw self-reported stress, exhaustion, and effort, as they are directly influenced by performance. We leave these three variables for potential controls and additional checks of how stress, effort and exhaustion change between incentive schemes, but also how they change within subjects, when going from the 5-min RET to 20-min RET in a particular scheme.

who performed worst in a classroom, except if a large group of participants guesses their rank correctly. Performance in the 5-min RET might be highly correlated with performance in the 20-min RET, but possibly also with all other explanatory variables.

As dependent variable we will use the total amount of correctly solved tables in the 20-minute RET. The variable represents a clean measure of one's performance.[22]

Regarding hypotheses, not much work has been done on the interaction between incentive schemes and our explanatory variables; hence, only few of our variables offer clear-cut predictions based on empirical evidence from previous research. In the domain of health care provision, Donato et al. (2017) elicit big five personality traits and report that people with high conscientiousness provide better maternal and child services, but react less to performance incentivization. People with low conscientiousness and neuroticism do better with performance incentivization. Moreover, in a lab experiment, Segal (2012) finds a similar pattern (albeit only for men) between conscientiousness and reaction to incentives.

- *Neuroticism and conscientiousness:* We expect that subjects who are more conscientious perform better overall, but exhibit a smaller increase in performance in the piece-rate and the tournament scheme in comparison to our fixed payment scheme, or for the case of the tournament scheme in comparison to the variable and fixed payment schemes, than subjects who are less conscientious.[23] Less neurotic subjects should do better in the piece-rate and the tournament scheme in comparison to our fixed payment scheme (or, for the case of the tournament scheme, in comparison to the variable and fixed payment schemes).

Furthermore, there is empirical evidence indicating that woman react worse to competitive incentive schemes in comparison to piece-rate schemes or schemes where payoff does not depend on performance, than men (Gneezy et al., 2003; Ors et al., 2013).

- *Gender:* We expect that women exhibit a smaller increase in performance in the tournament scheme in comparison to the piece-rate and the fixed payment

[22] We might construct an additional dependent variable where we divide the performance from the 20-min RET by the performance in the 5-min RET task. In this way, we would divide one's peformance by a proxy for individual productivity measured in a short RET task, where the decline in performance due to fatigue is arguably negligible, and (similarly to Dohmen and Falk, 2011) subjects are incentivized to perform in a piece-wise manner.

[23] We thus also expect performance in the fixed payment scheme to the payment schemes with stronger incentivization.

scheme, than men. We expect, however, that this is no longer true when excluding the comparison with the tournament scheme, as females, on average, are more risk averse (Sutter et al., 2019), and are less (over-)confident (Niederle and Vesterlund, 2007), which we expect to affect females' performance in the tournament scheme more than in the other two schemes. Comparing performance in the fixed vs. the piece-rate payment schemes, we expect a larger increase in performance in the piece-rate payment scheme among females, as they have been found to have more potential to improve performance when being grittier (Alan et al., 2019).

For the rest of our explanatory variables, we establish clear literature-rooted argumentation why they are relevant in our setup (see the first paragraphs of Section 3). However, the direction of how these variables might interact with incentives is less clear. We see two ways how an interaction effect can occur. On one hand, people with desirable traits and preferences — which are associated with positive life outcomes and earnings — might work hard regardless of the incentivization they are facing, i.e., their effort level might be independent of the incentivization scheme. In this way, they might gain an advantage over others in the long run by constantly working hard and being productive, irrespective of the incentives for a given task. On the other hand, people with those positive variables might be willing to work hard only when they see a sufficient benefit of investing effort. In particular, if we consider effort as a scarce resource, such people might be optimizing when exactly to exert it.

Studies have shown that socio-economic status, cognitive abilities, patience, risk tolerance, competitiveness, overconfidence and grit are all positively related to some important life-outcome variables or their proxies, e.g., educational attainment, earnings, college GPA, and standardized achievement tests (Heckman, 2006; Cunha et al., 2006; Duckworth et al., 2007; Bonin et al., 2007; Cadena and Keys, 2015; Hanushek et al., 2016; Reuben et al., 2017; Alan et al., 2019). From the remaining big five personality traits, openness was found to be positively related — while extraversion and agreeableness negatively related — to educational attainment (and to a lesser extent, standardized achievement test scores; see Almlund et al. (2011)).[24] Positive parenting was found to be highly relevant in development of cognitive and non-cognitive traits important for labor market outcomes (Falk et al., 2021). Finally, measures of individual productivity and relative self-assessment (which is tightly related to one's productivity; see Dohmen and Falk (2011)) can be indicative of

[24]Concerning conscientiousness and neuroticism, the former is positively while the latter negatively related to educational attainment.

one's performance in a task where costly effort is crucial.[25] Given these relations, we make two competing hypotheses, where one follows the argument that people with desirable traits and preferences will always invest a lot of effort, regardless of the incentivization, while the other relies on the assumption that such people will only exert effort when they see a sufficient benefit to do so.

- *Socio-economic status, cognitive abilities, patience, risk preferences, overconfidence, competitiveness, grit, openness, extraversion, agreeableness, positive parenting, relative self-asessment and individual productivity:*

    - *Hypothesis A*: We expect people with higher socio-economic status, higher cognitive abilities, higher patience, lower risk aversion, higher overconfidence, higher competitiveness, higher grit, higher openness, lower extraversion, lower agreeableness, higher positive parenting, higher relative self-assesment and higher performance in the 5-min RET (proxy of individual productivity) to increase their performance *less* as reaction to stronger incentivization, than those that hold the opposite socio-demographics, preferences and traits. With stronger incentivization we mean the tournament scheme in comparison to the piece-rate scheme and the fixed payment scheme, or the piece-rate scheme and the tournament scheme in comparison to the fixed payment scheme.
    - *Hypothesis B*: We expect people with higher socio-economic status, higher cognitive abilities, higher patience, lower risk aversion, higher overconfidence, higher competitiveness, higher grit, higher openness, lower extraversion, lower agreeableness, higher positive parenting, higher relative self-assesment and higher performance in the 5-min RET (proxy of individual productivity) to increase their performance *more* as a reaction to stronger incentivization, than those that hold the opposite socio-demographics, preferences and traits.

To investigate our conjectures and find out which of the explanatory variables is indeed relevant in our setting (and which lack importance), we test how the variables predict performance in reaction to different incentive schemes. To this end, we perform OLS regressions of performance, where we use the explanatory variables as discussed above as well as their interactions with payment schemes as

[25] Following similar argumentation as with other variables, individual productivity and relative self-assessment on a given task — although not necessarily related to long-term outcomes — might also interact with incentives in two opposing ways. Those who are more productive or have positive beliefs about their relative productivity could be willing to always work hard, or on the other hand, they could choose to work only if the benefit of exerting effort is sufficient.

independent variables. As interaction effects might be non-linear, we additionally might investigate interaction effects when replacing the explanatory variables with binary variables indicating whether or not a subject's given characteristic, trait or preference falls above or below a certain quantile. We consider the 25%, the 50% and the 75% quantiles. Using this approach, we will investigate the possibly heterogeneous effects on performance of the tournament scheme compared to the fix and pice-rate payment schemes (pair-wise and groupwise comparison), as well as those of the piece-rate scheme compared to the fixed payment scheme, and those of the fixed scheme compared to the piece-rate and tournament scheme (groupwise comparison).

Besides focusing on significantly estimated variables in these regressions, we might also use a LASSO approach to reduce the set of included variables to the most relevant ones, and will rely on methods such as Belloni and Chernozhukov (2013) or Belloni et al. (2013) for inference after variable selection via the LASSO. We might also apply LASSOplus (Ratkovic and Tingley, 2017), a Bayesian approach to variable selection, which yields approximate confidence intervals with nominal coverage for the estimated posterior. Using LASSO for variable selection (with appropriate inference) also offers the possibility to perform an additional analysis where we consider all the variables that we collect to measure socio-economic status and demographic characteristics in isolation without relying on indices. Moreover, it might be possible to investigate possibly heterogeneous effects of payment schemes accross all payment schemes.

Besides these rather classical variable selection approaches to identifying subgroups with respect to treatment effects, we might also apply a tree-based approach that combines classical inference with random forests (Wager and Athey, 2018; Athey et al., 2019), as well as a generic machine learning inference approach that may be combined with predictions based on support vector machines, neural networks, random forests, and several other machine learning techniques alike (Chernozhukov et al., 2018).

Study the performance of individual types across different incentive schemes. Based on our knowledge of subjects' socio-demographics, traits and preferences, we will build types using cluster analyses. The exercise will give us the tools to engage in a complementary analysis of our research question. In particular, we will investigate whether there are specific types of individuals that increase their performance in certain incentive schemes. We use the k-medoids clustering algorithm (Kaufman and Rousseeuw, 1990), also known as partitioning around medoids (PAM)

clustering, which is more robust to outlier and noise than the well-known k-means approach. We determine the number k of clusters such that the reduction in average silhouette width (Rousseeuw, 1987) or in the Calinski-Harabasz statistic (Calinski and Harabasz, 1974) is maximal ("kink-method"). In case of ambiguity, we will opt for the number of clusters allowing for a more intuitive interpretation of types. Ex-ante, the interpretation of clusters cannot be specified, as this is a data driven approach. Once clusters are formed, we will regress performance on these types, interacted with payment schemes.

3.2 Research Framework II: Sorting and incentive schemes

Here we focus on the predictive value of socio-economic and demographic variables, traits, and preferences on sorting into three common incentive schemes. The main objectives are:

- Study which socio-demographics, traits and preferences predict sorting into different incentive schemes;
- Study whether subjects sort according to socio-demographics, traits and preferences that are predictive of their performance;
- Study whether the possibility to sort influences the predictiveness of socio-demographics, traits and preferences on performance identified in the exogenous treatment.

In the following we provide a more detailed description of our approach in this framework.

Study which socio-demographics, traits and preferences predict sorting into different incentive schemes. Sorting is a key aspect of functioning labor markets. We focus on sorting into one of our three incentive schemes — fixed payment, piece-rate payment and tournament payment — which will serve as our dependent variable(s). For independent variables, we will use the same variables as in Research Framework I.

When it comes to sorting, there are multiple studies that focus on sorting between piece-rate and competitive schemes, and only few that take into account fixed payment schemes. So far, one of the most robust findings in the literature is the effect of gender and risk preferences. It was shown that women and people that are more risk averse are less likely to sort into competitive or generally more risky

incentive schemes (Niederle and Vesterlund, 2007; Dohmen and Falk, 2011; Buser et al., 2014, 2017; Alan et al., 2019).

- *Risk preferences and gender:* We expect that women as well as subjects with higher risk aversion will be less likely to sort into stronger incentivized payment schemes in comparison to men and subjects with lower risk aversion. Stronger incentivization implies the tournament scheme in comparison to the piece-rate and the fixed payment scheme, or the tournament and the piece-rate scheme in comparison to the fixed payment scheme.

It was also found that being very productive leads to choosing incentivized schemes with variable pay, while being overconfident or simply having high relative self-assesment encourages subjects to be more likely to opt for a competitive scheme (Buser et al., 2014; Dohmen and Falk, 2011; Niederle and Vesterlund, 2007).

- *Individual productivity, relative self-assessment and overconfidence:* We expect that having higher performance in the 5-min RET (proxy for individual productivity) will lead to more sorting into incentive schemes with stronger incentivization in comparison to having lower performance. Moreover, we expect that being overconfident as well as having high relative self-assessment will lead to more sorting from piece-rate or fixed into tournament scheme, in comparison to being less overconfident or having lower relative self-assessment.

There also exists evidence that cognitive as well as non-cognitive skills are relevant when it comes to sorting into competitive schemes. Being grittier and smarter was found to be positively related with sorting into more competitive and generally more risky payment schemes. Moreover, it was shown that being more competitive causes higher sorting into competition (Alan et al., 2019; Buser et al., 2014, 2017; Niederle and Vesterlund, 2007; Dohmen and Falk, 2010).

- *Grit, cognitive abilities and competitiveness:* We expect that subjects with higher grit and higher cognitive abilities will be more likely to sort into payment schemes with stronger incentivization in comparison to subjects that have lower grit and lower cognitive abilities. Moreover, we expect that subjects that are more competitive will be more likely to sort into stronger incentivized payment schemes.

For the rest of the explanatory variables, the predictions are less clear when looking at the existing literature. Continuing on our argumentation of how socio-demographics, preferences and traits relate to positive life outcomes, one could argue that people with characteristics that relate to positive life outcomes are not only

more productive, but given their productivity, might be more willing to opt for incentive schemes where this work pays out. Hence, such individuals could be more willing to enter variable schemes, where they will have a higher expected return on their invested effort.[26] Following this argumentation, and taking into account the relations between socio-demographics, preferences, traits on one side and positive life outcomes on the other (see Research Framework I), we state the following hypothesis.

- *Socio-economic status, patience, opennes, extraversion, agreeableness, neuroticism, conscientiousness, positive parenting:* We expect subjects with higher (or lower) socio-economic status, higher patience, higher openness, lower extraversion, lower agreeableness, higher conscientiousness, lower neuroticism and higher positive parenting will be more likely to sort into stronger incentivized payment schemes, than those who do not possess these characteristics, traits and preferences.

To test our conjectures and investigate which of the explanatory variables is indeed relevant for sorting into different payment schemes, we use OLS and GLM regression (probit/logit as well as ordered logit) to explain the choice of a payment scheme by our explanatory variables. Besides reporting significantly estimated variables, we might also use a LASSO approach to reduce the set of included variables to the most relevant ones, and will rely on methods such as Belloni and Chernozhukov (2013) for inference after variable selection via the LASSO. Using LASSO for variable selection (with appropriate inference) also offers the possibility to perform an additional analysis where we consider all the variables that we collect to measure socio-economic status and demographic characteristics in isolation without relying on incides. We will run these analyses (except for the ordered logit) for pair-wise comparisons of incentive schemes, but also for grouped comparisons: For example, choosing a risky/competitive payment scheme (piece-rate or tournament) compared to choosing the fixed payment scheme; or choosing the high-risk/highly competitive payment scheme compared to choosing a scheme that will less likely result in big surprises (piece-rate or fixed payment scheme).

Study whether subjects sort according to socio-economic and demographic characteristics, traits and preferences that are predictive of their performance. One of the key questions to understand when it comes to sorting, is whether

[26]Note that the same argument can apply both if subjects with those desirable preferences and traits only exert effort when there is sufficient benefit, or if they are always willing to exert effort. If it is the former, they can choose a variable payment scheme that will in expectation offer a sufficient benefit for them to exert effort, while if it is the latter, they will anway invest high effort; hence, choosing a variable incentive scheme will only increase their expected profit.

subjects actually understand their performance potential and how it might change across incentive schemes. Here, we want to investigate whether socio-demographics, traits and preferences which we identify as predictive of one's performance in Research Framework I are also decisive when subjects actually choose an incentive scheme. It could as well be that subjects do not always sort according to what might increase their performance, and there could be plenty of reasons for it. For instance, they might not realize they would be good in a specific incentive scheme, they could shy away from a particular scheme (e.g., being in competition could cause higher stress), or there could be another personal characteristic that dominates the one predicting high performance (e.g., if women do not favor tournament scheme, having high beliefs about their performance might not suffice to choose it). Furthermore, if subjects sort according to the explanatory variables that predict their performance it does not automatically imply that they successfully maximize their payoff. Hence, we will also investigate which subjects actually choose their payoff maximizing incentive scheme.

- *Predictive Variables for Performance in Research Framework I:* If given the choice, we expect subjects with characteristics, traits and preferences that are associated with higher performance under a given incentive scheme will be more likely to sort into the corresponding payment scheme than those who do not possess these characteristics, traits and preferences.

To investigate this hypothesis, we will first check which explanatory variables predict sorting in the endogenous treatment as outlined above, and see whether those are the same variables predicting higher performance in the corresponding incentive schemes in the exogenous treatment (identified in Research Framework I).

Moreover, we will check whether subjects actually sort according to their payoff-maximizing incentive scheme, and compute the potentially resulting welfare-loss. To this end, we will compute the payoff not only under the chosen incentive scheme, but also under the alternative schemes. For all subjects who did not choose the tournament scheme, we compute their expected payment under the tournament scheme by averaging over all possible pairings in their session. Analogously to the analyses above, we will identify whether certain characteristics, traits or preferences are predictive for the payoff foregone by chosing a suboptimal payoff scheme.

Study whether the possibility to sort influences the predictiveness of socio-economic and demographic characteristics, traits and preferences on performance identified in the exogenous treatment. Here, we will investigate whether the observed relations from the exogenous treatment between

explanatory variables and performance also hold in the endogenous treatment. We expect that the option of sorting affects at least some of the observed relations, for example, because some characteristics describe individuals that might be particularly motivated when having the freedom to choose. If that occurs, we have to ensure that the change in relations does not come from other reasons, e.g., from self-selection. In particular, as subjects choose the incentive scheme they prefer, we will likely have a different distribution of subjects across incentive schemes between exogenous and endogenous treatment, which might influence the relations we observe.

To analyze the effect of sorting on performance, we will thus, for every individual, first predict the performance in the chosen incentive scheme based on their socio-economic and demographic characteristics, traits and preferences using insights from the *exogenous treatment*: These predictions will be obtained either using machine learning approaches (e.g., random forest) or conventional regression models trained/fitted on the data from the exogenous treatment. The model for prediction will be selected based on the out-of-sample mean squared prediction error in the sample with the exogenous treatment using cross-validation. In this way, we can predict the performance *in absence of self-selection* for a certain individual under an incentive scheme given their characteristics, traits and preferences. The difference between *predicted* and *actual* performance can thus be attributed to having the choice to decide for an incentive scheme. Analogously to the analyses above, we will identify characteristics, traits or preferences that are predictive for an increased (decreased) performance due to having had the option to choose.

3.3 Research Framework III: Additional analyses

Our main interest lies in Research Framework I and II, but of course, we are also interested in

- whether stronger incentivized payment schemes lead to higher productivity in the *Exogenous* treatment (on average and over the whole distribution);
- the distribution of chosen incentive schemes in the *Choice* treatment, in particular, if it deviates from a uniform distribution;
- whether we can detect productivity sorting, meaning that
  - those with higher performance in Part 1 are more likely to choose the tournament scheme and less likely to choose the fixed payment scheme;

– the higher incentivized payment schemes lead to higher performance in the *Choice* treatment in Part 2, above and beyond what we observe for the *Exogenous* treatment;

References

Abeler, J., A. Falk, L. Goette, and D. Huffman (2011). Reference points and effort provision. *American Economic Review 101*(2), 470–92.

Akee, R., W. Copeland, E. J. Costello, and E. Simeonova (2018). How does household income affect child personality traits and behaviors? *American Economic Review 108*(3), 775–827.

Alan, S., T. Boneva, and S. Ertac (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics 134*(3), 1121–1162.

Alan, S. and S. Ertac (2018). Fostering patience in the classroom: Results from randomized educational intervention. *Journal of Political Economy 126*(5), 1865–1911.

Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011). Personality psychology and economics. In *Handbook of the Economics of Education*, Volume 4, pp. 1–181. Elsevier.

Andersen, A., R. Krolner, C. Currie, L. Dallago, P. Due, M. Richter, Á. Örkényi, and B. Holstein (2008). High agreement on family affluence between children's and parents' reports: International study of 11-yearolds. *Journal of Epidemiology & Community Health*.

Ariely, D., U. Gneezy, G. Loewenstein, and N. Mazar (2009). Large stakes and big mistakes. *The Review of Economic Studies 76*(2), 451–469.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2).

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.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2).

Belloni, A., V. Chernozhukov, and C. Hansen (2013). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Bonin, H., T. Dohmen, A. Falk, D. Huffman, and U. Sunde (2007). Cross-sectional earnings risk and occupational sorting: The role of risk attitudes. *Labour Economics 14*(6), 926–937.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2019). Beliefs about gender. *American Economic Review 109*(3), 739–73.

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

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.

Boyce, W., T. Torsheim, C. Currie, and A. Zambon (2006). The family affluence scale as a measure of national wealth: Validation of an adolescent selfreport measure. *Social Indicators Research 78*(3), 473–487.

Buser, T., M. Niederle, and H. Oosterbeek (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics 129*(3), 1409–1447.

Buser, T., N. Peter, and S. C. Wolter (2017). Gender, competitiveness, and study choices in high school: Evidence from switzerland. *American Economic Review 107*(5), 125–30.

Cadena, B. C. and B. J. Keys (2015). Human capital and the lifetime costs of impatience. *American Economic Journal: Economic Policy 7*(3), 126–53.

Calinski, T. and J. Harabasz (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods 3*(1), 1–27.

Carneiro, P. M. and J. J. Heckman (2003). Human capital policy. *IZA Discussion Paper* (821).

Cassar, L. and S. Meier (2018). Nonmonetary incentives and the implications of work as a source of meaning. *Journal of Economic Perspectives 32*(3), 215–38.

Cawley, J., J. Heckman, and E. Vytlacil (2001). Three observations on wages and measured cognitive ability. *Labour Economics 8*(4), 419–442.

Charness, G., U. Gneezy, and A. Henderson (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization 149*(C), 74–87.

Chen, D. L., M. Schonger, and C. Wickens (2016). otree- an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance 9*, 88–97.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2018). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report.

Cubel, M., A. Nuevo-Chiquero, S. Sanchez-Pages, and M. Vidal-Fernandez (2016). Do personality traits affect productivity? evidence from the laboratory. *The Economic Journal 126*(592), 654–681.

Cunha, F. and J. Heckman (2007). The technology of skill formation. *American Economic Review 97*(2), 31–47.

Cunha, F., J. J. Heckman, L. Lochner, and D. V. Masterov (2006). Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education 1*, 697–812.

Datta Gupta, N., A. Poulsen, and M. C. Villeval (2005). Male and female competitive behavior - experimental evidence. *GATE Working Paper* (05-12).

DellaVigna, S. and M. D. Paserman (2005). Job search and impatience. *Journal of Labor Economics 23*(3), 527–588.

DellaVigna, S. and D. Pope (2017). What motivates effort? evidence and expert forecasts. *The Review of Economic Studies 85*(2), 1029–1069.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics 132*(4), 1593–1640.

Dobbie, W. and R. G. Fryer Jr (2015). The medium-term impacts of high-achieving charter schools. *Journal of Political Economy 123*(5), 985–1037.

Doepke, M., G. Sorrenti, and F. Zilibotti (2019). The economics of parenting. *Annual Review of Economics 11*.

Doepke, M. and F. Zilibotti (2017). Parenting with style: Altruism and paternalism in intergenerational preference transmission. *Econometrica 85*(5), 1331–1371.

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), 556–90.

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*(3), 522–550.

Donato, K., G. Miller, M. Mohanan, Y. Truskinovsky, and M. Vera-Hernández (2017). Personality traits and performance contracts: Evidence from a field experiment among maternity care providers in india. *American Economic Review 107*(5), 506–10.

Duckworth, A. L., C. Peterson, M. D. Matthews, and D. R. Kelly (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology 92*(6), 1087.

Duckworth, A. L. and P. D. Quinn (2009). Development and validation of the short grit scale (grit–s). *Journal of Personality Assessment 91*(2), 166–174.

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

Essau, C. A., S. Sasagawa, and P. J. Frick (2006). Psychometric properties of the alabama parenting questionnaire. *Journal of Child and Family Studies 15*(5), 595–614.

Falk, A., A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics 133*(4), 1645–1692.

Falk, A., A. Becker, T. Dohmen, D. Huffman, and U. Sunde (2016). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Netspar Discussion Paper* (01).

Falk, A., F. Kosse, P. R. Pinger, H. Schildberg-Hörisch, and T. Deckers (2021). Socio-economic status and inequalities in children's iq and economic preferences. *Journal of Political Economy, forthcoming*.

Frick, P. J. (1991). The alabama parenting questionnaire. *University of Alabama*.

Fuchs, T. and L. Wößmann (2008). What accounts for international differences in student prformance? a re-examination using pisa data. In *The Economics of Education and Training*, pp. 209–240. Springer.

Gneezy, U., J. A. List, J. A. Livingston, X. Qin, S. Sadoff, and Y. Xu (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights 1*(3), 291–308.

Gneezy, U., S. Meier, and P. Rey-Biel (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives 25*(4), 191–210.

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

Gneezy, U. and P. Rey-Biel (2014). On the relative efficiency of performance pay and noncontingent incentives. *Journal of the European Economic Association 12*(1), 62–72.

Golsteyn, B. H., H. Grönqvist, and L. Lindahl (2014). Adolescent time preferences predict lifetime outcomes. *The Economic Journal 124*(580), F739–F761.

Hansson, Å. and J.-E. Gustafsson (2013). Measurement invariance of socioeconomic status across migrational background. *Scandinavian Journal of Educational Research 57*(2), 148–166.

Hanushek, E. A., S. J. Machin, and L. Woessmann (2016). *Handbook of the Economics of Education*. Elsevier.

Hanushek, E. A. and L. Woessmann (2011). The economics of international differences in educational achievement. In *Handbook of the Economics of Education*, Volume 3, pp. 89–200. Elsevier.

Hartley, J. E., K. Levin, and C. Currie (2016). A new version of the hbsc family affluence scale-fas iii: Scottish qualitative findings from the international fas development study. *Child Indicators Research 9*(1), 233–245.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science 312*(5782), 1900–1902.

Heckman, J. J. (2007). The economics, technology, and neuroscience of human capability formation. *Proceedings of the National Academy of Sciences 104*(33), 13250–13255.

Heckman, J. J. and Y. Rubinstein (2001). The importance of noncognitive skills: Lessons from the ged testing program. *American Economic Review 91*(2), 145–149.

John, O. P. and S. Srivastava (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research 2*(1999), 102–138.

Kaufman, L. and P. J. Rousseeuw (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.

Kosse, F., T. Deckers, P. Pinger, H. Schildberg-Hörisch, and A. Falk (2020). The formation of prosociality: Causal evidence on the role of social environment. *Journal of Political Economy 128*(2), 434–467.

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

Lindqvist, E. and R. Vestman (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the swedish enlistment. *American Economic Journal: Applied Economics 3*(1), 101–28.

Newby, J. L. and R. G. Klein (2014). Competitiveness reconceptualized: Psychometric development of the competitiveness orientation measure as a unified measure of trait competitiveness. *The Psychological Record 64*(4), 879–895.

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.

Niederle, M. and L. Vesterlund (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives 24*(2), 129–44.

Ors, E., F. Palomino, and E. Peyrache (2013). Performance gender gap: does competition matter? *Journal of Labor Economics 31*(3), 443–499.

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

Ratkovic, M. and D. Tingley (2017). Sparse estimation and uncertainty with application to subgroup analysis. *Political Analysis 25*(1), 1–40.

Raven, J. (2000). The raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology 41*(1), 1–48.

Reuben, E., M. Wiswall, and B. Zafar (2017). Preferences and biases in educational choices and labour market expectations: Shrinking the black box of gender. *The Economic Journal 127*(604), 2153–2186.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*, 53–65.

Rutkowski, D. and L. Rutkowski (2013). Measuring socioeconomic background in pisa: One size might not fit all. *Research in Comparative and International Education 8*(3), 259–278.

Schneider, S. O. and M. Schlather (2017). A new approach to treatment assignment for one and multiple treatment groups. Technical Report 228.

Schulz, W. (2006). Measuring the socio-economic background of students and its effect on achievement on pisa 2000 and pisa 2003. *Annual Meetings of the American Educational Research Association*.

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

Sutter, M., M. G. Kocher, D. Gältzle-Rützler, and S. T. Trautmann (2013). Impatience and uncertainty: Experimental decisions predict adolescents' field behavior. *American Economic Review 103*(1), 510–31.

Sutter, M., C. Zoller, and D. Glätzle-Rützler (2019). Economic behavior of children and adolescents–a first survey of experimental economics results. *European Economic Review 111*, 98–121.

Torsheim, T., F. Cavallo, K. A. Levin, C. Schnohr, J. Mazur, B. Niclasen, and C. Currie (2016). Psychometric validation of the revised family affluence scale: A latent variable approach. *Child Indicators Research 9*(3), 771–784.

Von Gaudecker, H.-M., A. Van Soest, and E. Wengstrom (2011). Heterogeneity in risky choice behavior in a broad population. *American Economic Review 101*(2), 664–94.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

West, M. R. and L. Woessmann (2010). 'every catholic child in a catholic school': Historical resistance to state schooling, contemporary private competition and student achievement across countries. *The Economic Journal 120*(546), F229–F255.

Woessmann, L. (2016). The importance of school systems: Evidence from international differences in student achievement. *Journal of Economic Perspectives 30*(3), 3–32.

A Appendix: Instructions and Questionnaire

Choice & Exogenous Treatments

Instructions for Part 1

Create your ID

Experimenter reads aloud: *[Welcome to the study. This study consists of two sessions: this session today, and another session in which you will participate in the near future. In both sessions you will earn money, please listen carefully to the instructions.* ***For today, you will receive a fixed payment of 4€ if you complete the session. You can also earn additional money depending on your performance in a task that I will explain later.*** *During the session you cannot talk to the other students in this room. This is a very important rule, and if you break it, you will not receive the money that you earned. On the first page, you will be asked to enter your ID as explained on the screen. Your name will never be used during the entire study. Whatever you do, and all the answers you give will only be recorded under your ID. That means that everything you do in the study is going to be anonymous. When you will participate in the second session, you will also be identified via the same anonymous ID code. If you have any questions, please raise you hand and one of us will come to your desk to answer it in private. Please remember that your participation on this study is fully voluntary, and you can decide to quit at any time. If you decide to quit before finishing the study, you are not allowed to leave the room, and you are still required to stay seated at your desk.]*

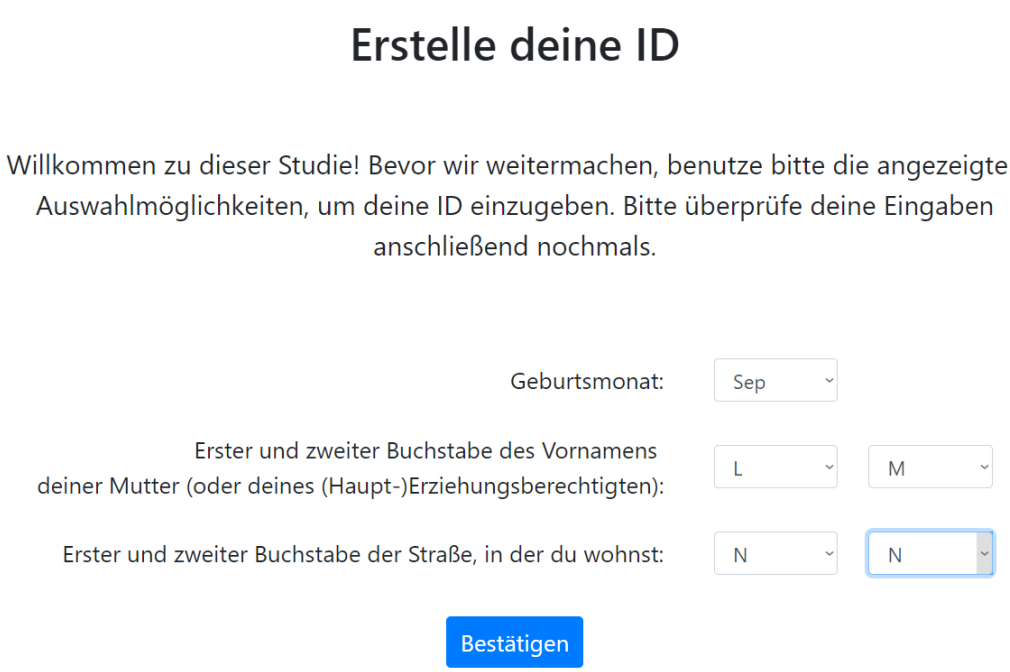
$\Longrightarrow$ ——— *Enter ID (first trial)* ——— $\Longleftarrow$

Welcome to this study. Before we proceed, use the drop-downs to enter your ID. Please double check all your entries before proceeding, as it is very important that your ID is specified correctly.

- Month of birth *[drop down – Jan, Feb, Mar, ...]*

- First and second letter of your mother's first name (or your legal guardian's first name)
- First and second letter of the street where you live

CONFIRM



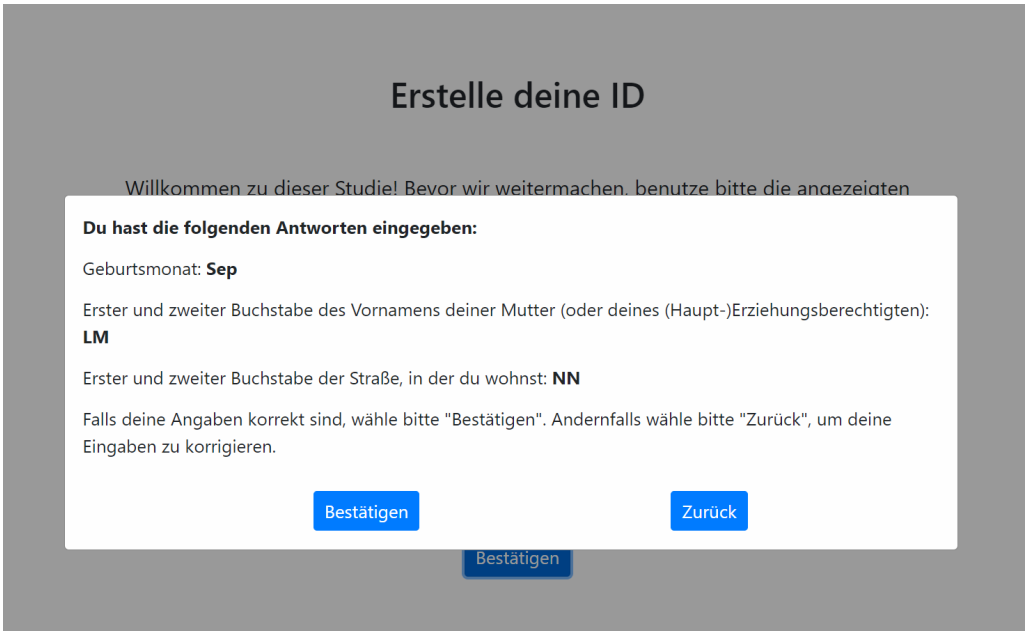
$\Longrightarrow$ ——— *pop-up* ——— $\Longleftarrow$

You provided the following answers:

- Month of birth: _____
- First and second letter of your mother's first name (or your legal guardian's first name): _____
- First and second letter of the street where you live: _____

If your answer is correct please press CONFIRM otherwise press BACK to revise your entries.

CONFIRM BACK



$\Longrightarrow$ ——— *wait for all & new screen for double IDs* ——— $\Longleftarrow$

Your ID is the same of someone else in this room. We hence ask you to answer an additional question:

- Last two letter of your first name

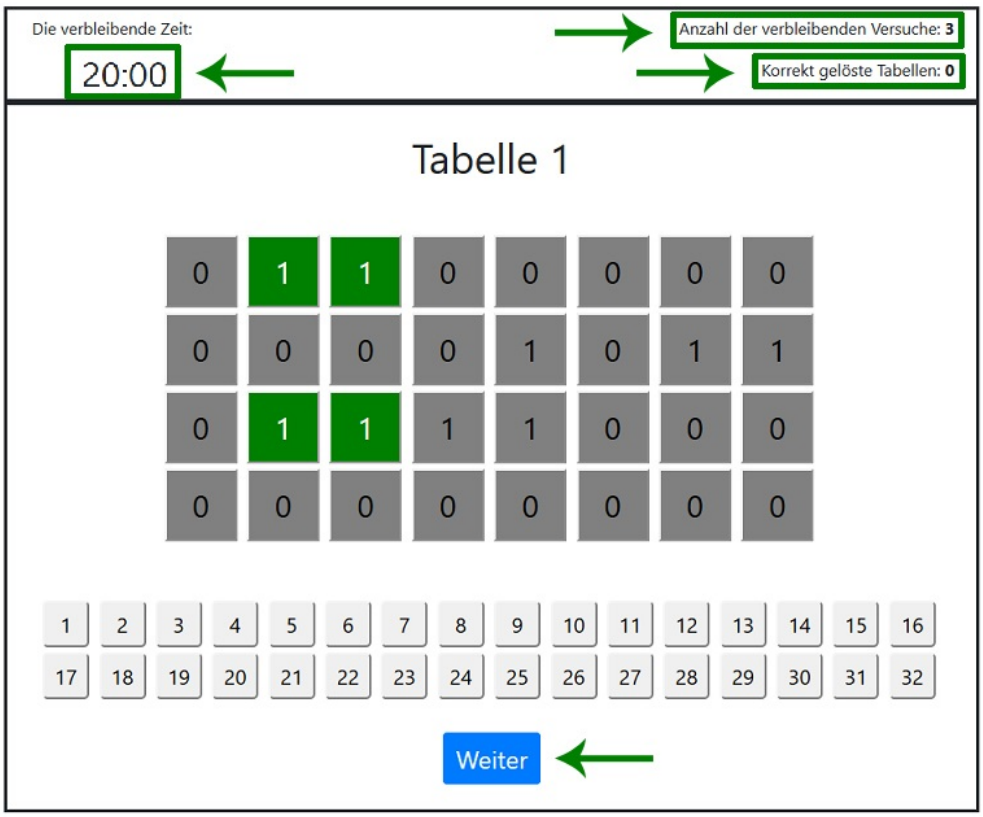
CONFIRM

$\Longrightarrow$ ——— *wait for all & new screen* ——— $\Longleftarrow$

Your Task

Experimenter reads aloud:

You will be shown a set of tables with 1s and 0s, like the one reported below. Your task is to correctly solve as many tables as you can.



To correctly solve a table, you have to:

1. **Tap** on all the cells containing a 1, which will highlight them in a dark green color

2. **Count** the correct amount of 1s that you see in the table, and report this amount in the number pad underneath the table.

Be aware, you are **not** allowed to highlight the 0s! If you accidentally highlight a 0, you can tap on the cell again to change it back to grey.

Once you are done with the tapping and you have reported the number, press CONFIRM. You get three tries to solve a table correctly. You will see the amount of remaining tries in the upper-right corner. If you do not manage to solve a table within the three tries, the next table will be shown on your screen. There are no penalties for not solving a table. You can see the amount of correctly solved tables in the upper right corner at any point during the task.

You have a total of 5 minutes to solve as many tables as you can. You will be **paid 0.06€ for each table you solved correctly.** For instance, if by the end of the 5 minutes you solve 1 table correctly, you will earn 0.06€ . If by the end of the 5 minutes you solve 10 tables correctly, you will earn 10 times 0.06€, so you will earn 0.6€. Or for instance, if by the end of the 5 minutes you solve 100 tables correctly, you will earn 100 times 0.06€, so you will earn 6€.

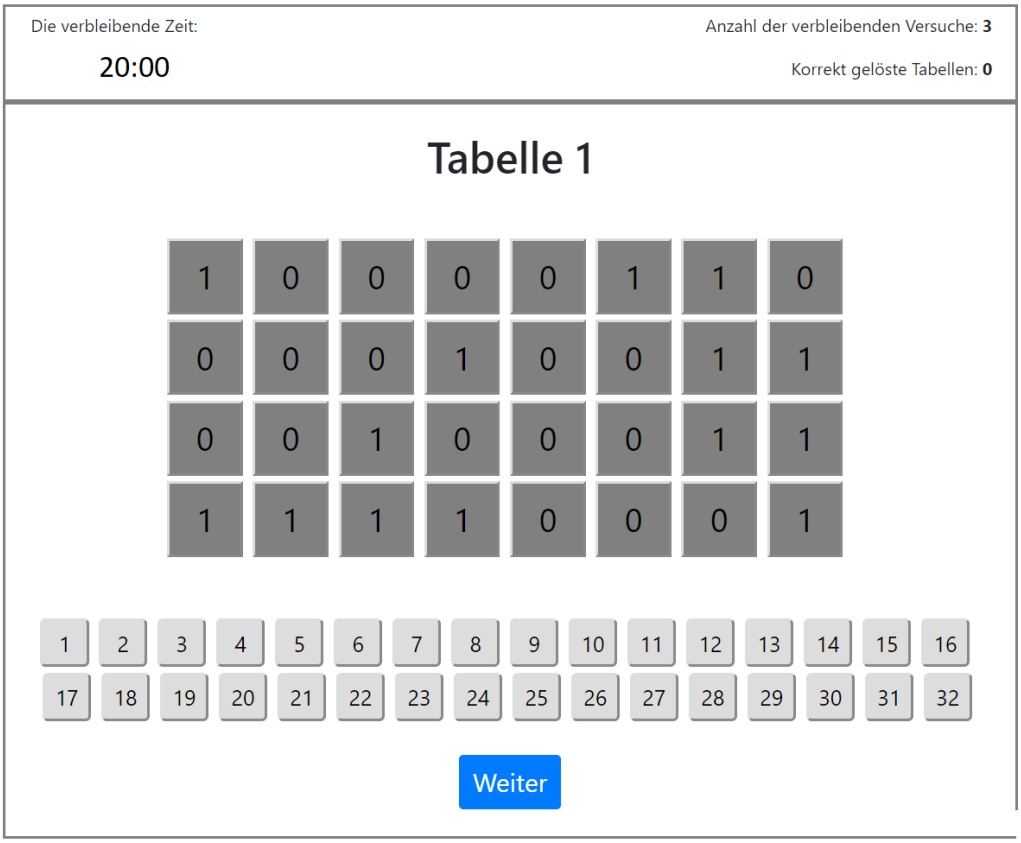
Before you start with the task you will have one trial round. That means that the first table you solve will not count for money, but will help you get acquainted with the task. After you correctly solve the first table, the 5 minute period will start.

Remember that you are not allowed to talk to the other participants in this room. If you have any question, please rise you hand and one of us will come to your desk to answer it privately. [1]

[1]The program is advanced by the experimenter after about 2-3 min (A "continue" button is displayed for the subjects once the experimenter advances the program) and participants are told to click "Continue" once they are ready to continue the experiment.

$\Longrightarrow$ ——— *wait for all & new screen*[2] ——— $\Longleftarrow$

The real effort task



[2]A similar table with "Trial round" is displayed. The picture is the same without the remaining time and correctly solved tables. After subjects correctly solve the trial round, they enter a waiting screen which lasts until everybody solves the trial round. Then a new screen appears with a 5 second countdown: "The 5 minute period for solving the task will start in 5, 4, 3... " Following that, Table 1 is displayed and the 5 minute period starts.

$\Longrightarrow$ ——— *new screen* ——— $\Longleftarrow$

The following questions are related to the task you completed. Please answer the following questions referring to the task you just solved. Please indicate your answers on a 7 point scale, where 1 means "not at all" and 7 means "very much":

	1	2	3	4	5	6	7
How much effort did you exert?	◯	◯	◯	◯	◯	◯	◯
How stressed did you feel?	◯	◯	◯	◯	◯	◯	◯
How exhausted did you get?	◯	◯	◯	◯	◯	◯	◯

CONTINUE

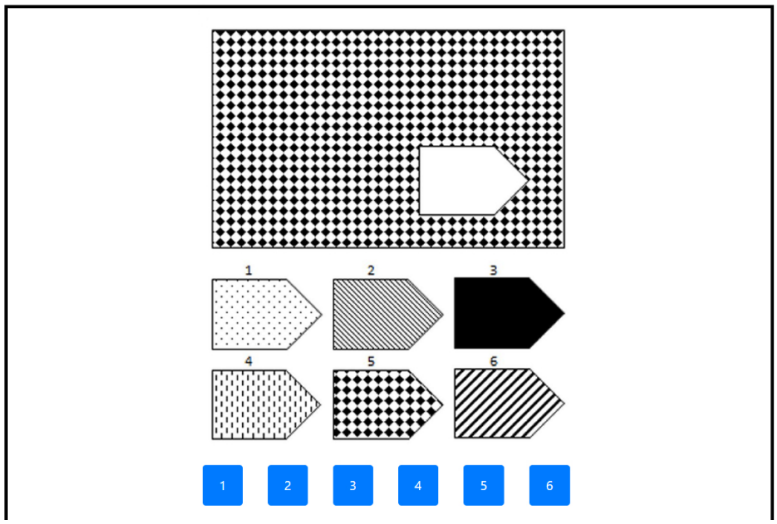
$\Longrightarrow$ ——— *new screen* ——— $\Longleftarrow$

For the following tasks you have to look at the picture, and find the missing piece of the picture. Once you find it, you need to circle it, as it is shown in the example below. Your goal is to solve as many tasks within 5 minutes as possible. If you have any questions, please raise your hand. If everything is clear, you can start immediately.[3]

Rätselaufgabe

In der nächsten Aufgabe werden ihnen einige Bilder wie unten abgebildet gezeigt. Sie müssen sich das jeweils gezeigte Bild genau anschauen und das fehlende Teil finden, welches das Bild vervollständigt.

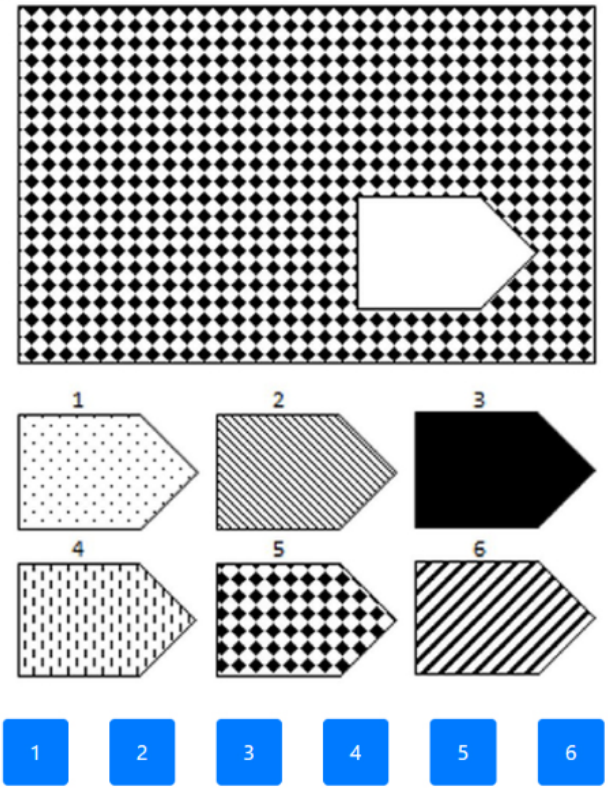
Sobald sie das fehlende Stück gefunden haben, tippen sie auf die entsprechende Nummer im unteren Teil des Bildschirms und bestätigen ihre Auswahl. Ihr Ziel ist es, möglichst viele Bilder innerhalb von 5 Minuten korrekt zu lösen. Sollten sie eine Frage haben, heben sie bitte ihre Hand und ein Mitarbeiter wird zur Klärung zu ihnen kommen.



NEXT

[3]Instructions are given with an example to make the task clear to the participants. The program is then advanced by the experimenter after they are done with reading the instructions (A "continue" button is displayed for the subjects once the experimenter advances the program) and participants are told to click "Continue" once they are ready to continue with the task.

$\Longrightarrow$ ——— *new screens* – ***matrices***[45] ——— $\Longleftarrow$

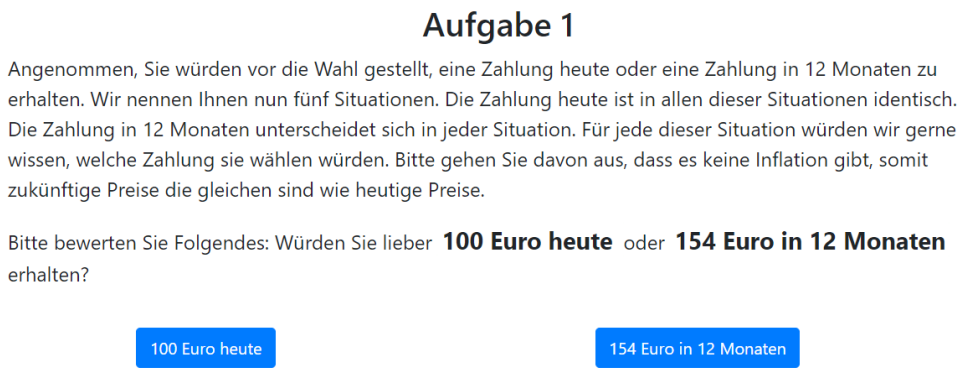


[4]The matrices are shown and the 5 minutes count-down starts to solve the 10 different matrices. Participants cannot go back once they have submitted an answer for a given task or skip between pictures. If a participant completes the tasks before the 5 minutes, they would have to wait for the other participants to finish.

[5]When subjects have answered all the matrices in the IQ task, they are told to proceed with the remainder of the tasks by themselves.

$\Longrightarrow$ ——— *new screen* – ***staircase for time preferences*** ——— $\Longleftarrow$

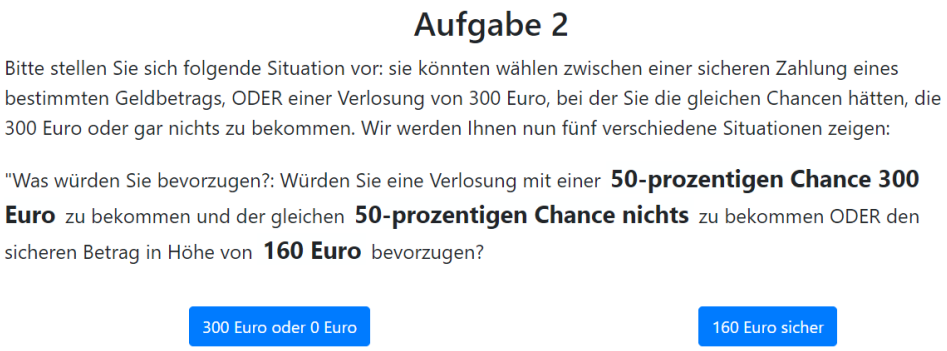
Suppose you were given the choice between the following: receiving a payment today or a payment in 12 months. We will now present to you five situations. The payment today is the same in each of these situations. The payment in 12 months is different in every situation. For each of these situations we would like to know which you would choose.[6]



[6]The two different options are shown with the corresponding amounts and for all 5 pages and the different amounts are highlighted to make them salient. Subjects advance to the next page when they click on one of the two buttons. The staircase approach is taken from Falk et al. (2016, 2018).

$\Longrightarrow$ ——— *new screen* – ***staircase for risk preferences*** ——— $\Longleftarrow$

Please imagine the following situation: You can choose between a sure payment and a lottery. The lottery gives you a 50 percent chance of receiving 300 Euro. With an equally high chance you receive nothing. Now imagine you had to choose between the lottery and a sure payment. We will present to you five different situations. The lottery is the same in all situations. The sure payment is different in every situation.[7]



[7]The two different options are shown with the corresponding amounts and for all 5 pages and the different amount are highlighted to make them salient. Subjects advance to the next page when they click on one of the two buttons. The staircase approach is taken from Falk et al. (2016, 2018)

$\Longrightarrow$ ——— *new screen* ——— $\Longleftarrow$

$\Longrightarrow$ ——— ***Questionnaire*** ——— $\Longleftarrow$

1. Are you female or male?	O Female	O Male	O I don't want to comment
2. What is your zip code?			
3. Where are you born?	O In Germany O In another EU country O In a European country outside of EU O In an Asian country O In an African country O In a South American country O In a North American country O In Australia		
4. What grade are you in?	O Grade 10 O Grade 11 O Grade 12 O Grade 13		
5. Year of birth?			
6. Grade in math? (final grade for last school year)	O 1 O 2 O 3 O 4 O 5 O 6		
7. Grade in German? (final grade for last school year)	O 1 O 2 O 3 O 4 O 5 O 6		
8. If everything goes as planned, when do you plan to finish the Abitur? (If you don't plan to finish the Abitur, please answer "No plans about finishing the Abitur")	O 2019 O 2020 O 2021 O 2022 O 2023 O No plans about finishing the Abitur		
9. How much pocket money/allowance do you get per week?	0-95 euros per week		
10. Do you have a mother/father born outside of Germany?	O Mother born outside of Germany O Father born outside of Germany O Both parents born outside Germany O Both parents born in Germany		

11. Do you live together with one or two parents (legal guardians)? (If you live with one parent and his/her partner, please answer: Two parents)	O One parent	O Two parents	O Neither
12. What is the highest education level of your mother?	O University or similar O High school O Middle school or lower O No schooling O I don't know		
13. What is the highest education level of your father?	O University or similar O High school O Middle school or lower O No schooling O I don't know		
14. What do you plan to do after you finish high school?	O University degree in STEM (Science, Technology, Engineering and Mathematics) O University degree outside of STEM O Vocational training (Ausbildung) O I want to find a job O I want to take some time off O Voluntary military service O I don't know		
15. Do you have any siblings?	O 0 O 1 O 2 O 3 O 4 or more		
16. How many books are there in your home?	O 0-10 books O 11-25 books O 26-100 books O 101-200 books O 201-500 books O More than 500 books		
17. What languages do you speak at home most of the time?	O German O English O Turkish O Spanish		O Italian O French O Arabic O Other

18. How many times did you and your family travel out of Germany abroad for holiday/vacation last year?	O None O Once O Twice O More than twice		
Which of the following are in your home?			
19. A room of your own?	O Yes	O No	
20. A link to the Internet?	O Yes	O No	
How many of these are there at your home?			
21. Cell phones?	O None O One O Two O Three or more		
22. Televisions?	O None O One O Two O Three or more		
23. Computers/PCs?	O None O One O Two O Three or more		
24. Cars?	O None O One O Two O Three or more		
25. Rooms with a bath or shower?	O None O One O Two O Three or more		

$\Longrightarrow$ ——— *new screen – from preference module* ——— $\Longleftarrow$

Please tell me, in general, how willing or unwilling you are to take risks. Please use a scale from 0 to 10, where 0 means you are "completely unwilling to take risks" and a 10 means you are "very willing to take risks". You can also use any numbers between 0 and 10 to indicate where you fall on the scale, like 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

0 = Completely unwilling to take risks *Very willing to take risks = 10*

◯ ◯ ◯ ◯ ◯ ◯ ◯ ◯ ◯ ◯ ◯

0 1 2 3 4 5 6 7 8 9 10

We now ask for your willingness to act in a certain way in four different areas. Please again indicate your answer on a scale from 0 to 10, where 0 means you are "completely unwilling to do so" and a 10 means you are "very willing to do so". You can also use any numbers between 0 and 10 to indicate where you fall on the scale, like 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

	Completely **unwilling** to do so										Completely **willing** to do so
How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯
	0	1	2	3	4	5	6	7	8	9	10
How willing are you to give to good causes without expecting anything in return?	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯
	0	1	2	3	4	5	6	7	8	9	10

$\Longrightarrow$ ——— *new screen – BFI-44* ——— $\Longleftarrow$

Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please select a number next to each statement to indicate the extent to which you agree or disagree with that statement.

I see Myself as Someone Who...	1. Disagree strongly	2. Disagree a little	3. Neither agree nor disagree	4. Agree a little	5. Agree strongly
1. Is talkative	◯	◯	◯	◯	◯
2. Tends to find fault with others	◯	◯	◯	◯	◯
3. Does a thorough job	◯	◯	◯	◯	◯
4. Is depressed, blue	◯	◯	◯	◯	◯
5. Is original, comes up with new ideas	◯	◯	◯	◯	◯
6. Is reserved	◯	◯	◯	◯	◯
7. Is helpful and unselfish with others	◯	◯	◯	◯	◯
8. Can be somewhat careless	◯	◯	◯	◯	◯
9. Is relaxed, handles stress well	◯	◯	◯	◯	◯
10. Is curious about many different things	◯	◯	◯	◯	◯
11. Is full of energy	◯	◯	◯	◯	◯
12. Starts quarrels with others	◯	◯	◯	◯	◯
13. Is a reliable worker	◯	◯	◯	◯	◯
14. Can be tense	◯	◯	◯	◯	◯
15. Is ingenious, a deep thinker	◯	◯	◯	◯	◯
16. Generates a lot of enthusiasm	◯	◯	◯	◯	◯
17. Has a forgiving nature	◯	◯	◯	◯	◯
18. Tends to be disorganized	◯	◯	◯	◯	◯
19. Worries a lot	◯	◯	◯	◯	◯
20. Has an active imagination	◯	◯	◯	◯	◯
21. Tends to be quiet	◯	◯	◯	◯	◯

I see Myself as Someone Who...	1. Disagree strongly	2. Disagree a little	3. Neither agree nor disagree	4. Agree a little	5. Agree strongly
23. Tends to be lazy	◯	◯	◯	◯	◯
24. Is emotionally stable, not easily upset	◯	◯	◯	◯	◯
25. Is inventive	◯	◯	◯	◯	◯
26. Has an assertive personality	◯	◯	◯	◯	◯
27. Can be cold and aloof	◯	◯	◯	◯	◯
28. Perseveres until the task is finished	◯	◯	◯	◯	◯
29. Can be moody	◯	◯	◯	◯	◯
30. Values artistic, aesthetic experiences	◯	◯	◯	◯	◯
31. Is sometimes shy, inhibited	◯	◯	◯	◯	◯
32. Is considerate and kind to almost everyone	◯	◯	◯	◯	◯
33. Does things efficiently	◯	◯	◯	◯	◯
34. Remains calm in tense situations	◯	◯	◯	◯	◯
35. Prefers work that is routine	◯	◯	◯	◯	◯
36. Is outgoing, sociable	◯	◯	◯	◯	◯
37. Is sometimes rude to others	◯	◯	◯	◯	◯
38. Makes plans, & follows through with them	◯	◯	◯	◯	◯
39. Gets nervous easily	◯	◯	◯	◯	◯
40. Likes to reflect, play with ideas	◯	◯	◯	◯	◯
41. Has few artistic interests	◯	◯	◯	◯	◯
42. Likes to cooperate with others	◯	◯	◯	◯	◯
43. Is easily distracted	◯	◯	◯	◯	◯
44. Is sophisticated in art, music, or literature	◯	◯	◯	◯	◯

$\Longrightarrow$ ——— *new screen* – ***Revised Competitiveness Index*** ——— $\Longleftarrow$

The following scale measures aspects of competitiveness. Please read each question carefully and try to answer as honestly as possible. Do not spend too much time on any one item; if trying to decide between two responses, choose the one that first comes to mind.

	1. Strongly disagree	2. Slightly disagree	3. Neither agree nor disagree	4. Slightly agree	5. Strongly agree
1. I like competition.	○	○	○	○	○
2. I am a competitive individual.	○	○	○	○	○
3. I enjoy competing against an opponent.	○	○	○	○	○
4. I don't like competing against other people.	○	○	○	○	○
5. I get satisfaction from competing with others.	○	○	○	○	○
6. I find competitive situations unpleasant.	○	○	○	○	○
7. I dread competing against other people.	○	○	○	○	○
8. I try to avoid competing with others.	○	○	○	○	○
9. I often try to outperform others.	○	○	○	○	○
10. I try to avoid arguments.	○	○	○	○	○
11. I will do almost anything to avoid an argument.	○	○	○	○	○
12. I often remain quiet rather than risk hurting another person.	○	○	○	○	○
13. I don't enjoy challenging others even when I think they are wrong.	○	○	○	○	○
14. In general, I will go along with the group rather than create conflict.	○	○	○	○	○

$\Longrightarrow$ ——— *new screen* – ***Alabama Parenting Style (positive parenting)*** ——— $\Longleftarrow$

The following are statements about your family. Please rate each item and how often it TYPICALLY occurs in your home.

		1. Never	2. Almost Never	3. Sometimes	4. Often	5. Always
1	Your parents tells you that you are doing a good job.	◯	◯	◯	◯	◯
2	Your parents reward you or give you something extra to you for behaving well.	◯	◯	◯	◯	◯
3	Your parents compliment yuo when you have done something well.	◯	◯	◯	◯	◯
4	Your parents prise you for behaving well.	◯	◯	◯	◯	◯
5	Your parents hug or kiss you when you done something well.	◯	◯	◯	◯	◯
6	Your parents tell you that they like it when you help out around the house.	◯	◯	◯	◯	◯

$\Longrightarrow$ ——— *new screen* – ***short grit scale*** ——— $\Longleftarrow$

Please respond to the following 8 items. Be honest – there are no right or wrong answers!

	1. Not like me at all	2. Not much like me	3. Somewhat like me	4. Mostly like me	5. Very much like me
1. New ideas and projects sometimes distract me from previous ones.	◯	◯	◯	◯	◯
2. Setbacks don't discourage me.	◯	◯	◯	◯	◯
3. I have been obsessed with a certain idea or project for a short time but later lost interest.	◯	◯	◯	◯	◯
4. I am a hard worker.	◯	◯	◯	◯	◯
5. I often set a goal but later choose to pursue a different one.	◯	◯	◯	◯	◯
6. I have difficulty maintaining my focus on projects that take more than a few months to complete.	◯	◯	◯	◯	◯
7. I finish whatever I begin.	◯	◯	◯	◯	◯
8. I am diligent	◯	◯	◯	◯	◯

$\Longrightarrow$ ——— *new screen at the end of part 1* ——— $\Longleftarrow$

Thanks for taking part in the study.

In the first task you solved ____ tables correctly.

You earnings for this task are: ____€ (rounded up at the 10 cents)

In addition, you earned a 4€ fee for taking part in the study.

Your total earnings for today are: ____€

Please remain seated and remember that you are not allowed to talk to the other participants. One of the experimenters will come to your desk to give you your earnings.

Choice Treatment[8]
Instructions for Part 2[9]

Create your ID

Experimenter reads aloud: *[In this study you will earn money, so please listen carefully to the instructions. During the study you cannot talk to the other students in this room. This is a very important rule, and if you break it, you will not receive the money that you earned during the study. On the first page, you will be asked to enter your ID as explained on the screen. Your name will never be used during the study. Whatever you do, and all the answers you give will only be recorded under your ID. That means that everything you do in the study is going to be anonymous! If you have any questions, please raise you hand and one of us will come to your desk to answer it in private. Please remember that your participation on this study is fully voluntary, and you can decide to quit at any time. If you decide to quit before finishing the study, you are not allowed to leave the room, and you are still required to stay seated at your desk.]*

Welcome to this study! Before we proceed, use the drop-downs to enter your ID. Please double check all your entries before proceeding, as it is very important that your ID is specified correctly.

- Month of birth *[drop down – Jan, Feb, Mar, ...]*
- First and second letter of your mother's first name (or your legal guardian's first name)
- First and second letter of the street where you live
- Last two letter of your first name [10]

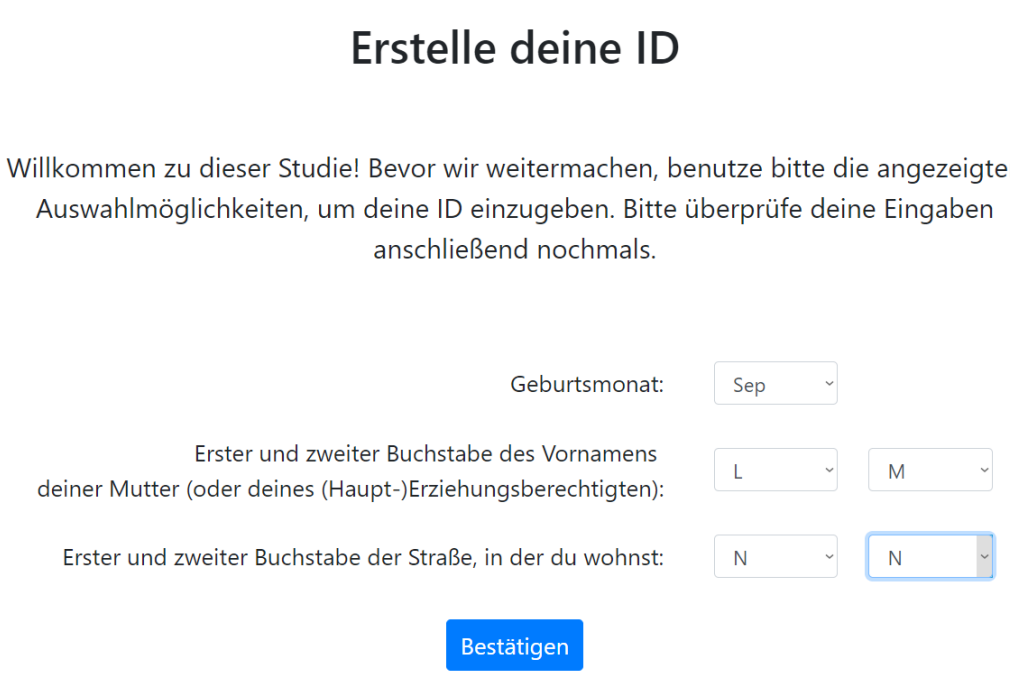
CONFIRM

[8] Same instructions as for exogenous, except the subjects face no choice screen, and only information about the relevant payment scheme is diplayed.

[9] The original German instructions and available upon request from the authors.

[10] Extra question in case of double ID.

$\Longrightarrow$ ——— *Enter ID (first trial)* ——— $\Longleftarrow$



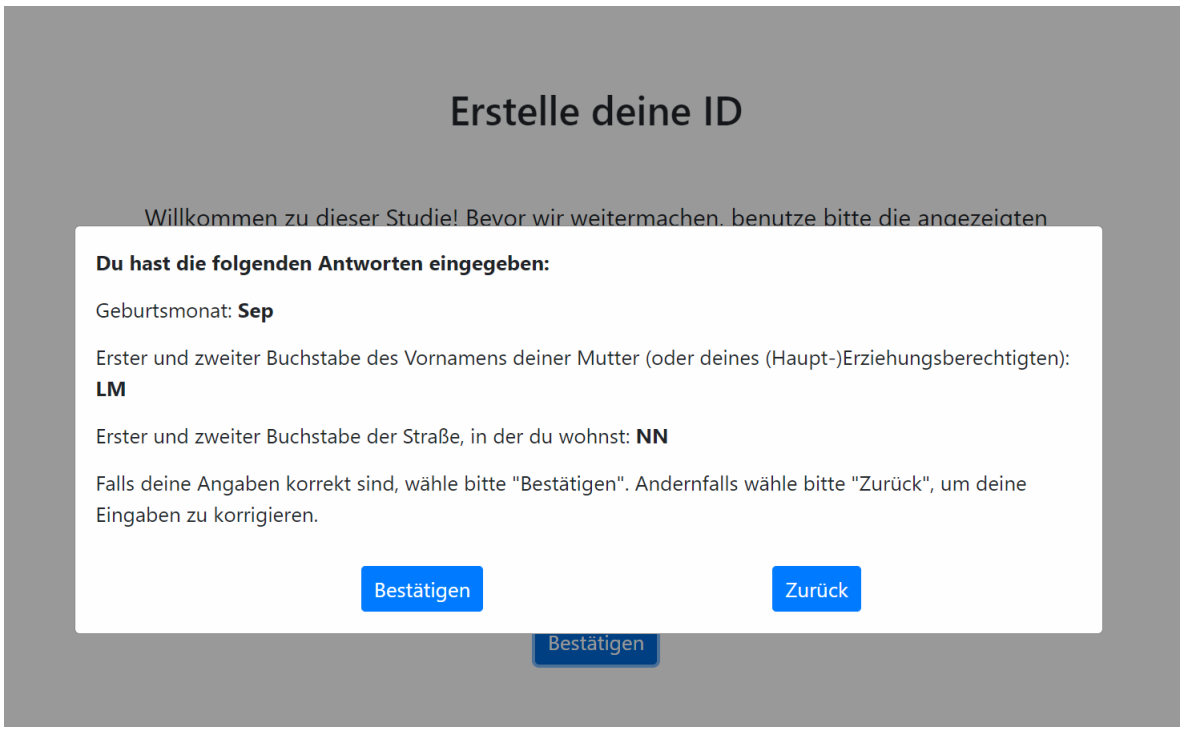
$\Longrightarrow$ ——— *pop-up* ——— $\Longleftarrow$

You provided the following answers:

- Month of birth: _____
- First and second letter of your mother's first name (or your legal guardian's first name): _____
- First and second letter of the street where you live: _____

If your answer is correct please press CONFIRM otherwise press BACK to revise your entries.

CONFIRM BACK



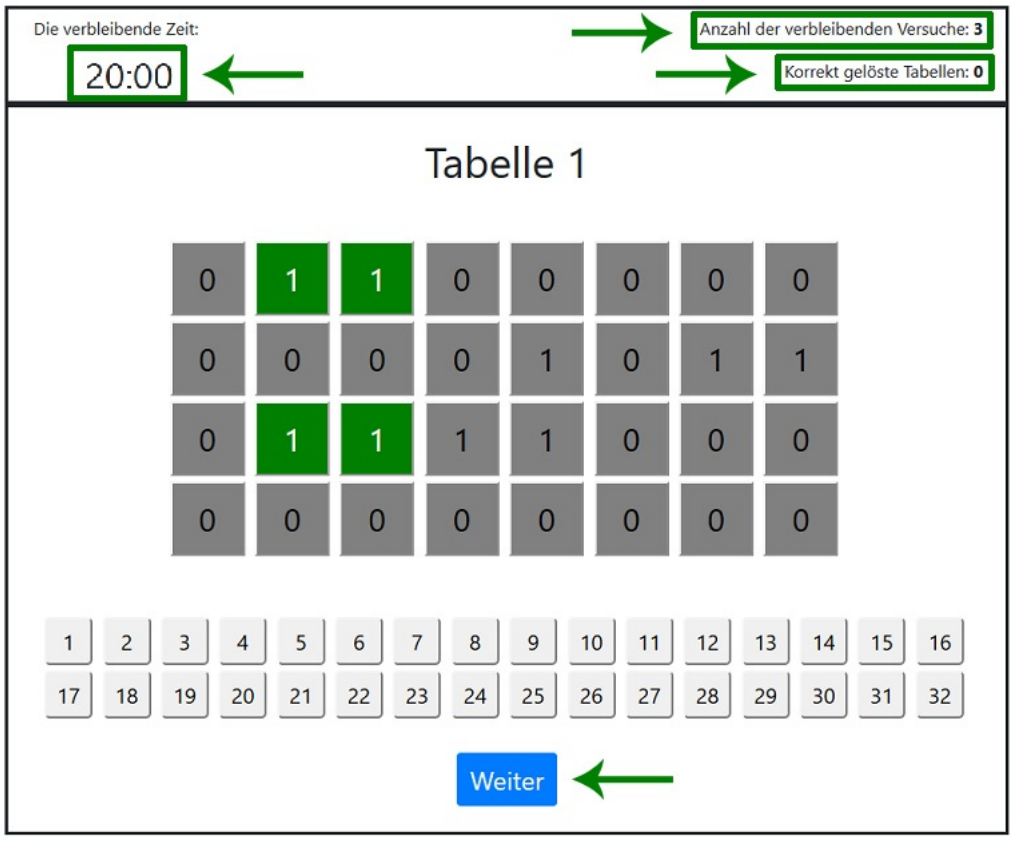
$\Longrightarrow$ ——— *new screen* ——— $\Longleftarrow$

$\Longrightarrow$ ——— *wait for all & new screen* ——— $\Longleftarrow$

Your Task

Experimenter reads aloud: *[I will now explain you the task in which you can earn money. Some of you have already seen the task as you did it the first time we came to your class. But some of you were not here; to be certain that you all know the task, I will explain it in detail again. Please follow the instructions carefully.]*

You will be shown a set of tables with 1s and 0s, like the one reported below. Your task is to correctly solve as many tables as you can.



To correctly solve a table, you have to:

1. **Tap** on all the cells containing a 1, which will highlight them in a dark green color;

2. **Count** the correct amount of 1s that you see in the table, and report this amount in the number pad underneath the table.

Be aware, you are **not** allowed to highlight the 0s! If you accidentally highlight a 0, you can tap on the cell again to change it back to grey.

Once you are done with the tapping and you have reported the number, press CONFIRM. You get three tries to solve a table correctly. You will see the amount of remaining tries in the upper-right corner. If you do not manage to solve a table within the three tries, the next table will be shown on your screen. There are no penalties for not solving a table. You can see the amount of correctly solved tables in the upper right corner at any point during the task. **You have a total of 20 minutes for solving the task.**

Remember that you are not allowed to talk to the other participants in this room. If you have any question, please raise you hand and one of us will come to your desk to answer it privately.[11]

[11]The program is advanced by the experimenter after about 2-3 min (A "continue" button is displayed for the subjects once the experimenter advances the program) and participants are told to click "Continue" once they are ready to continue the experiment

$\Longrightarrow$ ——— *new screen*[12] ——— $\Longleftarrow$

Guessing Task

Before explaining how you will be paid for the task, you have a chance to earn some additional money.

In this room, there are XY students (including you) that were present also during our previous visit. You all performed the task for 5 minutes the other time.

In the 5-minute version of the task, **you correctly solved X tables.**

We ranked you and the other participants present in the previous visit.[13] You were all ranked based on the number of tables correctly solved in 5 minutes. For example, position number 1 is for the one who solved the most tables, position number 2 is for the one who solved the second most tables, and so on, with the last position XY for the one who solved the least.

We would now like you to guess your position in the ranking.

If you were to guess the correct number, you earn 2€. If you come within up to 5 positions (higher or lower), you will earn 0.50€ that will be added to your total earnings for today's session.

\- ——————————————————————— +
1 — — — — — — — — — — — — — — — — — — — XY

I think I ranked number .....

CONFIRM[14]

[12]Screen only appears for subjects that are present in both sessions.

[13]Participants that are present in both sessions are ranked by standard competition ranking.

[14]Participants need to touch the slider to activate it. They can adjust the number either by touching the slider or clicking on the + and - signs at the ends of the slider. Absolute numbers of the different options for the ranking appear after the participant click on the slider. "I think I ranked number ...." only appears when participants click on the slider with the number of the ranking clicked on.

$\Longrightarrow$ ——— *new screen* ——— $\Longleftarrow$

Your Earnings[15]

You can determine the payment mode yourself. In particular, you can choose between three alternative payment modes.

Fixed Payment. When the 20 minutes are up, you will receive 6.5€, independent of the number of tables you solved correctly.

Variable Payment. When the 20 minutes are up, you will be paid 0.06€ for each table you solved correctly.

Tournament. When the 20 minutes are up, you will be paid either 0.08€ or 0.04€ *for each table* you solved correctly. To establish whether you will be paid 0.08€ or 0.04€ per correct table, your performance will be compared with one other student in this room, whose payment will also be determined in the same way. At the end of the 20 minutes, if you solved more tables than this other student matched with you, you will get 0.08€ per correct table. If instead you solved less tables than this other participant matched with you, you will get 0.04€ per correct table. If you and this other participant solved the same number of tables, the computer will randomly determine if you are paid 0.08€ or 0.04€ per correct table.

For your information, in the first visit you have solved ____ in ____ minutes. [16]

Before choosing your payment mode, please answer a few control questions designed to make sure you understood how the earnings are computed. If you have any questions, please raise your hand and wait for an experimenter to come to your desk. [17]

1. In the **fixed payment** alternative, if you solve 10 tables correctly by the end of the 20 minutes, how many Euros will you get?

   a I will get 0.60€

   b I will get 6.50€

   c I will get 80.00€

[15] In the exogenous treatment, subjects would only be able to see the paragraph explaining the payment scheme that they were assigned, and would only receive control questions referring to that payment scheme.

[16] This info was displayed only if the ID was present in first study and it is unique in second study.

[17] The correct answers are marked here in bold for display.

2. In the **fixed payment** alternative, if you solve 1000 tables correctly by the end of the 20 minutes, how many Euros will you get?

   a I will get 0.60€

   b I will get 6.50€

   c I will get 80.00€

3. In the **variable payment** alternative, if you solve 10 tables correctly by the end of the 20 minutes, how many Euros will you get?

   a I will get 0.60€

   b I will get 6.50€

   c I will get 90.00€

4. In the **variable payment** alternative, if you solve 1000 tables correctly by the end of 20 minutes, how many Euros will you get?

   a I will get 0.60€

   b I will get 6.50€

   c I will get 60.00€

5. In the **tournament payment** alternative, if you solve 1000 tables correctly, and the student matched with you solves 10 tables correctly by the end of the 20 minutes, how many Euros will you get?

   a I will get 0.40€

   b I will get 6.50€

   c I will get 80.00€

6. In the **tournament payment** alternative, if you solve 10 tables correctly, and the student matched with you solves 1000 tables correctly by the end of the 20 minutes, how many Euros will you get?

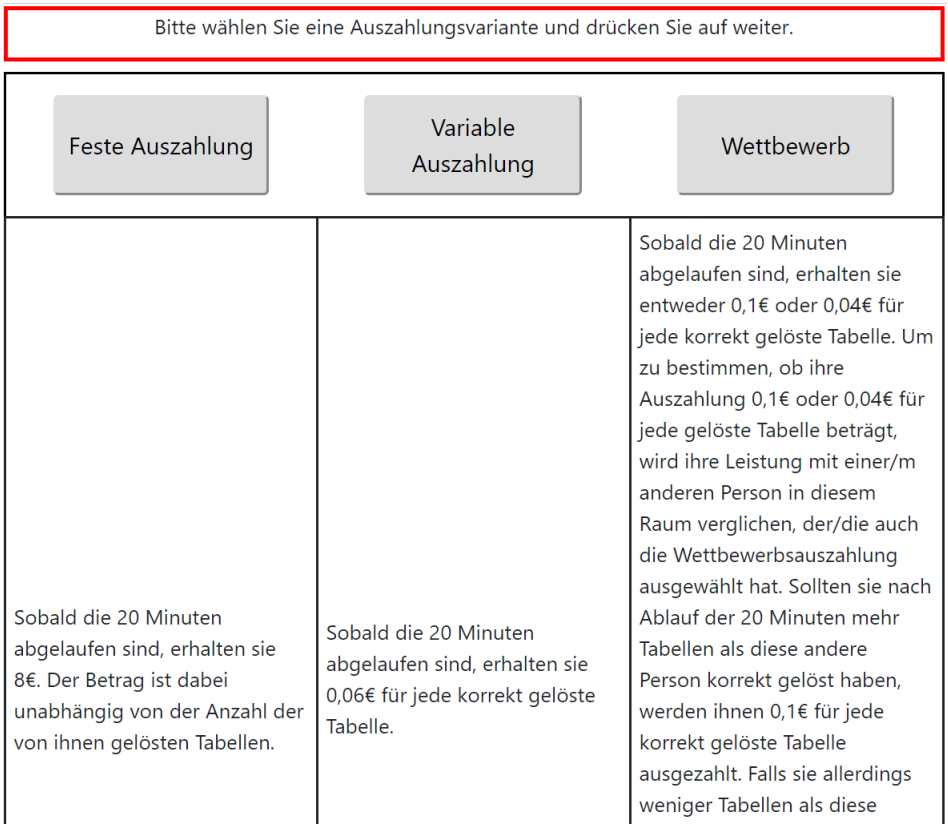
   a I will get 0.40€

   b I will get 6.50€

   c I will get 80.00€

$\Longrightarrow$ ——— *new screen* ——— $\Longleftarrow$

Choice of Payment Mode



$\Longrightarrow$ ——— *wait for all & new screen* ——— $\Longleftarrow$

Show if only one person chose tournament

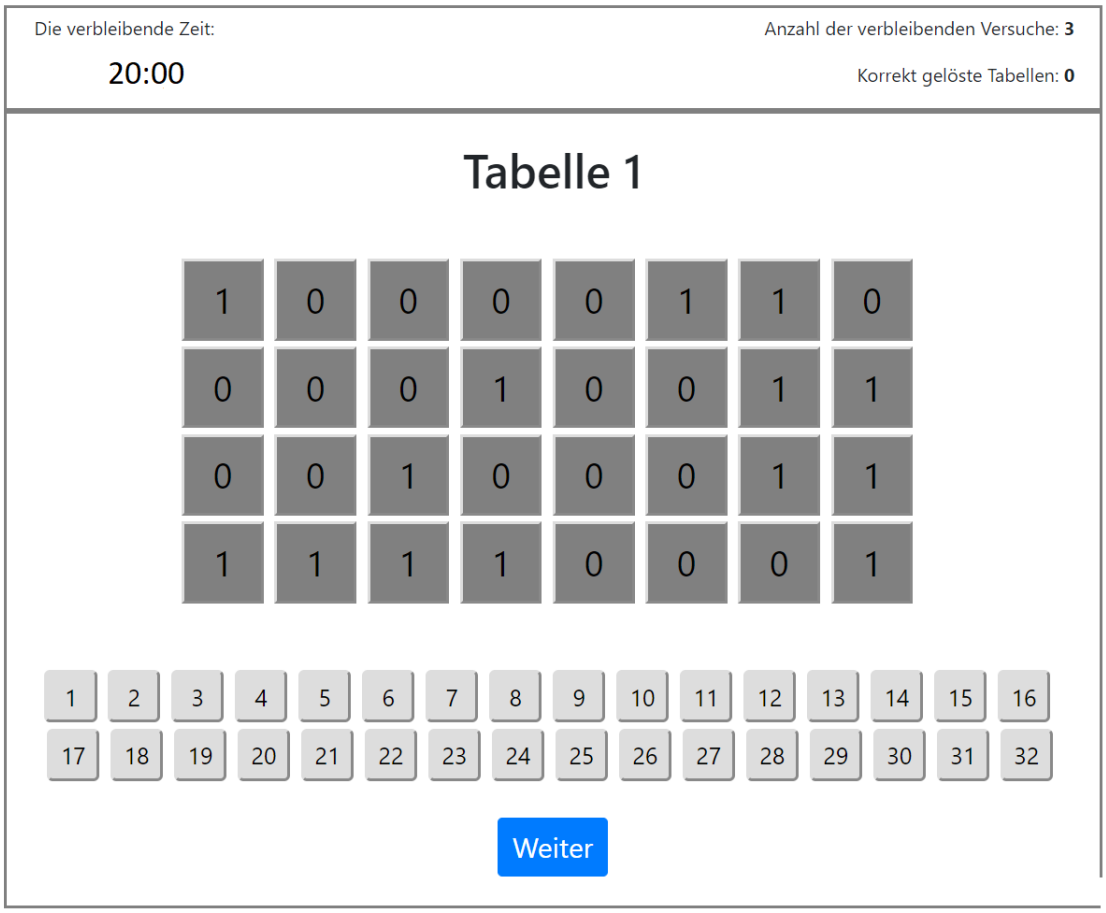
You are the only one who chose tournament. Unfortunately, it is not possible to match you with another student in this room. Please choose again, this time between fixed and variable payment.[18]

$\Longrightarrow$ ——— *wait for all & new screen*——— $\Longleftarrow$

Countdown. The task will start in 10, 9, 8...seconds.

[18] A menu with the two possible choices are shown to the participant.

$\Longrightarrow$ ——— *new screen – the real effort task* ——— $\Longleftarrow$



$\Longrightarrow$ ——— *new screen* ——— $\Longleftarrow$

Before proceeding to the final payments, please answer the following questions referring to the task you just solved. Please indicate your answers on a 7 point scale, where 1 means "not at all" and 7 means "very much":

	1	2	3	4	5	6	7
How much effort did you exert?	○	○	○	○	○	○	○
How stressed did you feel?	○	○	○	○	○	○	○
How exhausted did you get?	○	○	○	○	○	○	○

NEXT

$\Longrightarrow$ ——— *new screen at the end of the task* ——— $\Longleftarrow$

Thank you for taking part in our study. The task is now over.

You solved ____ tables correctly.

You earnings for this task are: ____€

In addition, you earned a 1€ fee for taking part in the study.

In the guessing task you earned ..... €

[**Only for Tournament.**] You solved more/less/the same number of tables than/as the student you are compared to.
[**Only in case of tie.**] The computer randomly determined that you earn 0.08/0.04 € per solved table.

Your earnings for the task are: ____ €
(rounded up at the 10 cents)

Please remain seated and remember that you are not allowed to talk to the other participants. One of the experimenters will come to your desk to give you your earnings.

B Appendix: Power Calculations

Based on pilot data, we assume that individuals solve 105 tables on average under the fixed payment scheme, and 140 under either of the alternative schemes with a pooled SD of 23. That is, we assume a treatment effect of +35 tables solved. Moreover, again informed by pilot data, we assume that the effect for males is +40, while it is +30 for females. The interaction of being female (−10) is thus about 30% of the main effect.

Analytical Calculations

Assuming equal numbers of observations per payment scheme, for a desired significance level of 95% and 80% power, we need a sample size of 14 to detect the assumed main treatment effect. Assuming that the SD of the interaction is about twice as large as that of the main effect,[19] we need to multiply the sample size for the main treatment effect by 49. Thus, to detect the interaction effect with 80% power at a significance level of 95%, wee need a sample of 686 for a treatment comparison. Having two payment schemes in addition to the base scheme, and, in addition, having the same setup for the *Choice* treatment, we need $686/2 \cdot 3 \cdot 2 = 2058$ observations.
The following R Script replicates the calculations.

```
## Sample size for main effect (from http://powerandsamplesize.com/)
muA=105 # outcome group A
muB=140 # outcome group B
kappa=1 # ratio nA/nB, where nA number of observations in group A
sd=23
alpha=0.05 # significance level
beta=0.20 # 1-beta = power
(nB=(1+1/kappa)*(sd*(qnorm(1-alpha/2)+qnorm(1-beta))/(muA-muB))^2)
ceiling(nB)

## Sample size for interaction and total sample size
interaction = 10/35
(scale = (2/interaction)^2)
(N = ceiling(nB)*2*scale)
(N_total = N/2*3*2)
```

Simulation

Using a simulation based approach with the script below, the sample size amounts to 1992 observations. (Note that (indented) linebreaks might have to be removed before execution

[19] $\mathbb{V}[\Delta_1 - \Delta_2] = \mathbb{V}[\Delta_1] + \mathbb{V}[\Delta_2]$ for uncorrelated random variables Δ_1 and Δ_2.

of the script.)

```
power_interaction = function(N = 1000) {
  sigma <- rep(23,N)
  treatment <- c(rep(0,N/2),rep(1,N/2))
  female <- c(rep(0,N/4),rep(1,N/4),rep(0,N/4),rep(1,N/4))
  means <- 105 + treatment*(145-105) + (115-105)*female +
      ((145-115)-(145-105))*(female*treatment)
  y <- rnorm(N, means, sigma)
  reject <- ifelse(coef(summary(lm(y ~ treatment + female +
      treatment:female)))[4,4] < .05,1,0)
  return(reject)
}

x <- seq(from=640, to = 680, by = 4)
y <- NULL
for (i in 1:length(x)) {
  y <- c(y,mean(replicate(10000, power_interaction(N = x[i]))))
}
plot(x,y)

rbind(sample_size = x,power = y)
664/2*3*2
```