Essays on the Determinants of Student Achievement: Education Policies, Confidence and Effort

James Robert Merewood

Submitted for the Award of

DOCTOR OF PHILOSOPHY

in Economics

School of Economics

University of East Anglia

Norwich – September 2022

Abstract

In this thesis, I examine the role of effort, and school management as determinants of student achievement. In addition, I present evidence demonstrating the role of task complexity on confidence.

In chapter 1, I study the impact of a nationwide merit-based selection policy implemented in Romanian high schools. This policy saw high school principal posts opened for applications; the highest scoring candidate across a series of tests was selected for each post. Using a staggered difference-in-difference design, I identify the impact of the policy on student outcomes. I find that principals who move between posts (compared to those who retain their position) improve outcomes in low-to-average schools two years on. Improvements are related to the selection of students into school leaving exams.

In chapter 2, I examine the impact of the complexity of real effort tasks on subjects' beliefs about performance. Here, I note that the choice of effort tasks used in many lab experiments is not trivial. Some evidence suggests that task complexity influences subject beliefs about performance, however little is known about this interaction when using standardised real effort tasks. I conduct an experiment to test the interaction between task complexity and beliefs about performance. I find that subjects are more confident about their relative performance, and make more accurate predictions about performance, when facing complex tasks. The findings of this chapter have ramifications for real-effort task choice within experimental economics.

In chapter 3, I explore the impact of ability tracking systems on effort using an online lab experiment. Evidence suggests that ability tracked classrooms allow teachers to better target teaching, however low ability students often suffer from studying alongside low ability peers. The mechanism of peer effects is well established in literature, but little is known about the impact of tracking on effort. Using an experiment, I find that when ability tracking is implemented so that subjects can frequently move between ability groups, effort increases overall. I show that increases in effort are due to high ability subjects increasing their effort, and further that low ability subjects do not reduce effort. I go further to show that differences are not driven by group composition or differences in ability.

Contents

List of Figures

List of Tables

Acknowledgements

Firstly, I would like to thank my supervisors Oana Borcan and Sheheryar Banuri. Your constant encouragement, care and wisdom have made this research possible. The lessons you have imparted on me throughout our time working together will stay with me for life.

Second I must thank my colleagues in the School of Economics. I am grateful for the helpful comments and personal support you have provided during my time at the school. Thanks also go to the Centre for Behavioural and Experimental Social Sciences, for funding a large portion of my experimental research.

Third, I thank my friends within my PhD cohort. Your companionship across the past 4 years has provided me with a great source of inspiration. I know that I will continue to be good friends in the years to come.

My eternal gratitude goes to my family for enabling me to follow my dreams. Particularly, I thank my parents, who have endured many phone calls in which I complain endlessly. Your calming presence and constant support has guided me through countless situations.

Finally, I thank Joe. I thank you for your loving and encouraging nature, for your motivation and perseverance, and for the endless cups of tea at ungodly hours. Witnessing the dedication you have for your own work continues to inspire me. Your compassion underpins this work, without you it would not have been possible.

In loving memory of

Sarah Merewood

(1972 – 2023)

Introduction

This thesis examines issues related to education policy and is comprised of three independent essays. We utilise both empirical and experimental methodologies to analyse several determinants of student achievement. Initially, using a staggered difference-in-difference strategy we examine the role of school management on student outcomes. Secondly, we study the effect of task complexity on subjects' ability to predict their performance in a laboratory setting. Finally, we shed light on the impact of different types of ability-tracking systems on subject effort. Taken together, the results from this thesis form the basis for future education policy suggestions.

In chapter 1, we study the impact of the merit-based selection of Romanian high school principals. As gatekeepers of education, high school principals often manage the allocation of resources in an attempt to provide high quality learning and student performance. Relatively few studies provide evidence that principal management practises are linked to student outcomes (Di Liberto *et al.*, 2015 Coelli and Green, 2011). Those that do, suggest that high quality principals improve outcomes (Bohlmark *et al.*, 2016). As a result of this evidence, it is important to study the selection process of principals; attempts should be made to have the highest quality principals possible.

Outside of high-income countries, hiring decisions are often made based on personal and political affiliations. In the Romanian university context, evidence suggests that the successful candidate is often known in advance of the application process (PEIS, 2007). In an attempt to remove the existing political influence in high school principals, a nationwide meritocratic selection policy was implemented in Romanian high schools in 2016/17. As a result of the policy implementation, all high school principal posts were opened, and candidates were required to undertake a series of exams, interviews and CV assessments; the highest scoring candidate for each post was appointed. We test empirically the impact of the meritocratic selection policy on student outcomes, using a staggered difference-in-difference strategy.

We study the impact of competitively selected principals (compared to those appointed), and further the impact of new managers (compared to legacy managers who retain their positions). The average treatment effect is small and insignificant immediately after the policy. However, we provide evidence that new managers in low-to-average schools begin to improve outcomes two years on. The improvement is related to the selection of students into sitting the school leaving exam, but additional survey evidence also suggests that policy selected managers have more motivation towards their job. The evidence presented in this chapter suggests that merit-based selection policies have the scope to reduce the inequality in education.

In chapter 2, we examine the impact of real-effort task complexity on the ability of subjects to predict their performance. Real-effort tasks are frequently used in experimental economic literature, and require subjects to exert a costly effort in the form of a working task. Examples of frequently used real-effort tasks include an arithmetic task in which subjects add 2-digit numbers together (Niederle and Vesterlund, 2007) and a matrix task which requires subjects to count the number of 0's in a grid (Abeler *et al.*, 2011). Typically, real-effort tasks used in the literature are simple, as this allows subjects to learn the task quickly, preventing changes in performance over time (Benndorf *et al.*, 2019). However, evidence from psychology literature suggests that the confidence of a subject about their own performance is impacted by the complexity of the task (Moore & Healy, 2008).

Despite this, little is known about the impact of increasing task complexity for standard experimental economic real-effort tasks. The potential confidence related biases which may exist when subjects face simple tasks may impact behaviour which we measure in the lab. One important example is experiments in which subjects must choose between different incentive schemes; for example, Niederle and Vesterlund (2007) require subjects to pick between a piece rate and a tournament payment scheme. Here, the decision to enter a competitive incentive scheme is driven the ability of a subject to accurately predict their performance; overconfident subjects are shown to place stronger weight on success contingent payments (de la Rosa, 2011). If subjects are better able to predict their performance when facing more complex tasks, then they may make sub-optimal incentive scheme choices when they face simpler tasks.

We conduct an experiment to test the interaction between task complexity and the accuracy of predictions which subjects make about their performance. In our baseline, subjects are tasked with a simple arithmetic task in which they must add together three 2-digit numbers. Our treatments manipulate task complexity by introducing an additional grid search component. Before each round of the task, we ask subjects to predict their upcoming performance, and use this prediction to calculate a confidence measure (confidence = prediction – performance). We find that subjects are, on average, underconfident; their predictions are lower than their true performance. Despite this, our results show that subjects are better at predicting their performance when faced with more complex tasks. We provide further evidence that this effect persists across many rounds of the real-effort task. The findings of this chapter have ramifications for real-effort task choice within experimental economic literature, particularly where beliefs about performance play an important role.

In chapter 3, we explore the impact of ability tracking systems on effort using an online lab experiment. Around the world, students are placed into ability tracked classrooms based on their performance in past exams. In some countries, students are placed into different schools and are limited in the subjects that they can continue to study. In the UK, students are placed into high- and low-ability classrooms for core subjects such as Mathematics and English. The use of ability-tracking is widespread, and is used in more than 95% of US schools (Fu and Mehta, 2018). Ability tracking enables teachers to better target their material to similar ability students (Duflo *et al.*, 2011), and high-ability students are shown to benefit from studying alongside high-ability peers (McEwan, 2003). Despite this, low-ability students often become discouraged, experience a drop in self-confidence (Francis *et al.*, 2019), and exert less effort (Jagacinski and Nicholls, 1990).

Existing literature focuses heavily on the impact of ability tracking systems on student outcomes, and has spent little time addressing student effort as an important channel through which ability tracking impacts students. Often, ability tracking systems are very restrictive. Subjects move between tracks infrequently, and often through informal channels. When ability tracks are fixed, low-ability students have reduced incentives to continue to exert effort; despite how hard they work, they remain in the low-ability class. We propose an alternative ability tracking system, which allows students to frequently move between ability tracked groups. We define this ability tracking system as 'retracking'. Theoretically, 'retracking' provides an incentive for high-ability students to exert effort to remain in the high-ability group, and also provides incentives for low-ability students to exert effort in a bid to move into the higher ability track.

In this paper, we implement an online laboratory experiment in which treatments manipulate the grouping strategy: mixed-ability groups; ability tracked groups; ability retracked groups. We find that ability tracking, when implemented such that individuals can move between tracks, increases effort overall. Total output is significantly greater under our new retracking system than under a random ability grouping, and further when compared to ability tracking. We show that this increase is due to high-ability subjects increasing their effort, suggesting that retracking might benefit high-ability students. Importantly, we do not find evidence that low-ability subjects reduce their effort provision, highlighting the idea that retracking may not lead to inequality. We go further to show that differences are not driven by group composition or differences in ability.

Chapter 1

Positive Disruption: Meritocratic Principal Selection and Student Achievement

1.1 Introduction

Principals are the gatekeepers of education, exercising considerable authority in the allocation of material and human resources to ensure quality learning, equity and student performance. In 2017, pre-university education expenditure was, on average, 3.5% of the GDP in OECD countries (OECD, 2020). School principals manage a large share of the national education budgets; it is therefore critical that principals are able to ensure efficient use of resources and to support student attainment. A small number of recent studies show that principals' management practises are linked to student outcomes (e.g. Jacob *et al.*, 2018; Di Liberto *et al.*, 2015; Coelli and Green, 2011); for example, Bohlmark *et al.* (2016) demonstrate that students in schools with higher quality principals have improved outcomes.

As a result, the selection process of school principals is vitally important in order to ensure that capable and motivated individuals are entrusted with leadership roles. Despite this, many education systems outside high-income countries rely on discretionary appointments, either based on evaluative interviews (which can result in subjective and inconsistent assessments), or based on criteria such as nepotism and political affiliation, adversely affecting management performance.[1][2] Standardised and objective principal selection processes are an underexplored avenue for improving student learning and test scores. This is partly because meritocratic selection policies for principals utilising standardised testing are relatively rare. Moreover, national or state policies on principal (or more generally public sector) recruitment do not typically leave room for experimental variation in meritocratic or competitive selection processes.

In our study, we examine the impact of a gradually introduced meritocratic selection of principals in Romanian secondary schools on student outcomes. In the economics literature, there is emerging evidence linking objective rule-based selection of teaching staff to improved job performance (Estrada, 2019). In the private sector, employee recruitment is known to deliver better quality hires when utilising objective meritocratic selection, than when using discretionary appointments (Hoffman *et al.*, 2018). However, we are not aware of any studies causally linking the merit-based selection of school principals to management performance or student outcomes. Whilst some studies in education and public administration utilise survey techniques to provide descriptive evidence, to the best of our knowledge we are the first study to link merit-based selection of principals to student outcomes utilising a quasi-experimental design.

Romania provides a relevant context because its post-communist education landscape has been marred by corruption and nepotistic appointments. For instance, in Romanian universities, 23% of staff say that the person obtaining a job at the institution is often known before interviews are held (PEIS, 2007). Against this backdrop, a series of anti-corruption policies have been implemented since 2011, beginning with the introduction of CCTV monitoring of exams and tougher punishments for teachers and students (Borcan *et al.*, 2017). Another recent reform was a nationwide meritocratic manager selection policy which ran between 2016 and 2017. The policy

[1] Whilst not the direct focus of this paper, for more information on patronage or nepotistic appointments see Colonnelli *et al.* (2018), Scoppa (2009), Huseyin and Mustafa (2008).

[2] For example, Gurmu (2020) documents that principalship in Ethiopia is often denied to those who have the relevant level of education, and suggests that political affiliation is the main criteria used for selecting principals into positions. Similarly, Walker and Kwan (2012) highlight the role of cultural and religious affiliation in the selection of principals in Hong Kong; despite this, prospective principals in Hong Kong are required to provide proof of a pre-principal certification and must undertake an interview (Walker and Kwan, 2012). In many districts of the US, evidence suggests that most principals perceive merit-based selection to be the main practise (Palmer and Mullooly, 2015); however, several participants still perceive inequality in the selection process (Palmer and Mullooly, 2015).

was a hurried attempt by a provisional technocrat government to purge schools of political influence and came as shock. In September 2016 *all* manager posts across the country were vacated and a national competition including standardised cognitive and management competency tests was open from October 2016. National elections in December saw a new political party in government, which halted the exam process before all schools were assigned principals, and postponed the competition over remaining schools until July 2017. A vast majority of principals who were eventually confirmed in a post in 2016 and 2017 (i.e., 87% of all principals), had to undergo the standardised exam, and were competitively selected based on their results.

Using administrative data and an independently conducted survey, we study the effects of this policy on the student outcomes in the national school-leaving exam, the Baccalaureate. Since the policy took place in two waves (one in 2016, one in 2017), the traditional identification strategy is to use a two-way fixed effects difference-in-difference (TWFE DiD) model to assess the change in student outcomes in schools with and without an exam selected principal pre-/post-policy. This comparison of student outcomes between schools with an exam selected principal, and schools with an appointed principal forms our first treatment (T1). Since recent literature (Baker, Larcker and Wang, 2022) has highlighted the biases within the TWFE estimator, we use the estimator developed by Callaway and Sant'Anna (2021) which estimates group-time-specific treatment effects with the correct control groups.

This approach has the advantage that we can use all or a large part of the population of Romanian schools, since we know all schools which have had an exam selected principal. The interpretation of this treatment is limited by the fact that the policy has de facto changed only around 30% of principals. Therefore, using a sample for which we have additional information on principals, we also study the impact of a change in manager on student test scores. Specifically, we estimate the difference in student outcomes between schools with principals who retained their position (henceforth legacy managers) and those who replaced an existing principal as a result of the policy (new managers; this is our second treatment - T2).

Initially we estimate the effect of T1, to understand whether the policy overall had any impact on student outcomes (by comparing schools with a principal selected by the policy with those schools who did not have a principal selected by the policy). We find overall zero average treatment effects of competitively selected principals in a large sample of Romanian schools.[3] This is likely to be due to the fact that a large share (~70%) of schools with exam selected principals retain their pre-policy manager.

We then estimate the effect of T2 by comparing student outcomes for new manager schools, compared with those with legacy managers post-policy. We find that new managers have a positive and significant impact on exam pass rates in the longer-term (3 years after the policy); as expected, these effects are very small and insignificant in the short-term (1 year after the policy). Importantly, we find that the average treatment effect doubles in magnitude when comparing treatment effects from 1 year after the policy to those 3 years after the policy. The ATTs of new managers in 2019, the last year in our data, are 0.07 of a SD for the final Baccalaureate score, and approximately 8 percentage points higher passing rates compared to students in schools with legacy managers in 2012-2016. These findings are in line with literature which suggests that it takes time for managers to have an impact on student outcomes (Coelli and Green, 2011).

[3] We discuss the definition of our working sample in section 4.1.

We also examine the policy impact across the distribution of school performance. We divide schools into three different percentile groups based on the schools' pre-policy average Baccalaureate scores: poor performance schools (<25th percentile), midrange performance schools (25th-75th percentile), and top performance schools (>75th percentile). For new managers selected in 2017, we find a significant positive impact on student outcomes in mid-performing (25th-75th percentile) schools. Specifically mid-performing schools with a 2017 new manager see exam pass rates which are 11.3% greater than the control group mean (a 5.1 percentage point increase). We next turn to the potential mechanisms through which this change may be brought about.

Unlike in many countries, Romanian principals have almost no independence in staff recruitment; hiring and dismissal decisions are made centrally at the county level.[4] Therefore, our results suggest that a change in school managers through merit-based selection can have an impact on student outcomes, even in education systems with little school autonomy. Even slight improvements in outcomes driven by new mangers in low- to mid-performing schools can contribute to reducing education inequality. To explore the mechanisms through which new principals improve attainment in poor performing schools, we provide supplementary survey data we collected in 2017 after the second round of the competitive selection. The survey data reveals suggestive evidence that new managers are more prosocial, indicating that they feel more motivated to fulfil their public mission. Despite this, we also find evidence that legacy managers report higher levels of trust between teachers in their school. This result is not unexpected as at the time of the survey, new managers had been in post for a relatively short period of time; they had not yet had chance to enact positive change.

One area of potential autonomy for Romanian principals is their ability to strategically decide whether to allow marginal students to sit the Baccalaureate exam. Therefore, we turn our attention to Baccalaureate enrolment rates; to do so we use school admission rates, utilising secondary data provided by Munteanu (2021).[5] We find that enrolment rates are significantly lower in schools with a new manager for both the 2018 and 2019 Baccalaureate exam sessions. During each year, there are two sittings for the Baccalaureate exam, though only outcomes from the first sitting are included in school performance targets. Our results suggests that new managers are restricting permission for students to sit the initial Baccalaureate exam and deferring them to sit the second sitting of the exam later in the year. Whilst this provides students with more time to study for the exam, it also serves to mechanically enhance the Baccalaureate performance measures. This finding flags the necessity for more school autonomy in order to enable managers to make effective and sustainable change.

In order to verify the validity of our results, we undertake several robustness checks. To rule out the concern that new managers select into schools which were on differential pre-policy trends in student outcomes, we verify the parallel trend assumption. We run alternative estimations changing the control group (from never to not-yet treated). We also estimate OLS TWFE estimates and based on these, run placebo policy tests which indicate that our results are indeed a consequence of the meritocratic selection policy. Finally, we make corrections to our clustered standard errors for the OLS estimations.

[4] For example, Bohlmark *et al.* (2016) show that in Sweden, principals can recruit and dismiss teachers and have control over the hiring process.

[5] We thank from Diana Coman for centralising Romanian public education data in her repository, which was the original source for Munteanu (2021) (www.ossasepia.com).

Our paper contributes to at least two strands of literature. Firstly, we present complementary evidence to studies linking principal performance to student outcomes. Several studies show a positive correlation between principal performance and student outcomes (e.g. Meyer *et al.*, 2020; Agasisti *et al.*, 2018; Masci *et al.*, 2018). Only a few studies have analysed this relationship through a causal lens, using quasi-experimental designs or tightly controlled econometric models (Coelli and Green, 2011; Dhuey and Smith, 2014; Bloom *et al.*, 2015; Di Liberto *et al.*, 2015; Bohlmark *et al.*, 2015). Generally, principals can influence outcomes by improving management practises (Di Liberto *et al.*, 2015), and higher management quality is shown to be positively associated with outcomes (Bloom *et al.*, 2015). Coelli and Green (2011), demonstrate that principals influence outcomes for students who are already dedicated to improvement; specifically, principals who switch between schools typically have a greater impact in higher performing schools than in poorer performing schools, since students in higher performing schools are more dedicated to improvement. Importantly, the evidence suggests that the principal effects takes a few years to show, which the authors attribute to the time needed for a principal to change the ethos and management practises of the school. Dhuey and Smith (2014) show that principals can improve outcomes in math and reading exams; similar to Coelli and Green (2011), they show that the effect is strongest in higher performing schools.

One of the most important channels managers in many education systems have to influence student outcomes is the hiring and firing of teachers (Jacob *et al.*, 2010). In addition, principals must be able to utilise the skills of their teachers (Agasisti *et al.*, 2018). Meyer *et al.* (2020) highlight the impact which principals can have on the collaboration of teachers when the collective efficacy of teachers is in line with the beliefs of the principal. However, the influence of principals on outcomes depends greatly on the level of autonomy which they have over teacher recruitment and other school decisions. In this vein, our study contributes complementary findings that there are limited short-term gains from meritocratic principal recruitment in a setting where principals have little decision power over key education inputs. In our context, for example, they are unable to influence the hiring and firing of teachers, which we discuss further in section 3.2. The lack of principal autonomy in our setting is a key reason for the short-run null average treatment effect in our study. The positive impact we find on student test scores with a lag in schools with new managers is most likely related to a change in light-touch interventions that newly instated managers are able to introduce: restricting permission for students to sit in the main Baccalaureate exam series, or introducing a more prosocial environment with the school.

Secondly, we contribute to an emerging literature on meritocratic staff selection in public education. Since principals are shown to influence student outcomes, selecting high quality principals is paramount. Meritocratic recruitment processes inherently favour the selection of managers with the highest ability. Ruiz-Tagle (2019) show that meritocratic selection of principals in Chile was correlated with improved student outcomes 6 years after the introduction of the policy. Whilst there were no changes in management practises within these schools, student attendance improved by around 1% per year. Hsiao *et al.* (2012) examine a similar policy in Taiwanese secondary schools, but do not link the policy to student outcomes. Using interview techniques they find differences in management styles; competitively selected principals are more willing to transform management practises and encouraging democratic participation in management than traditionally appointed principals. These papers are based on qualitative or correlation analysis, however we provide a significant contribution by using a robust econometric identification strategy to get close to a causal link between meritocratic principal selection and student outcomes.

Significantly more literature exists in the meritocratic selection of teachers. Jacob *et al.* (2018) use data from Washington DC schools to show that performance of prospective teachers in screening measures, such as written assessments and sample lessons, is highly predictive of job performance. Similarly, Bruno and Strunk (2019) demonstrate that performance in specific screening measures is predictive of contributions to student achievement in the same areas of study. Despite this, Jacob *et al.* (2018) suggest that meritocratic selection of teachers is significantly under-utilised in the hiring decisions of public schools. Estrada (2019) consider differences between discretion based and rule based appointments of public school teachers in Mexico. The findings of Estrada (2019) are particularly strong, since they undertake analysis in a very large education system (of around 1 million teachers), and the estimation strategy controls for several selection-based issues. Estrada (2019) demonstrate that teacher hired based on the discretion of the appointing manager perform considerably worse on the job than those appointed using a rule-based system. As previously discussed, whilst literature highlights the importance of meritocratic selection of teachers on outcomes, there is a gap in the literature surrounding the meritocratic selection of their managers (principals); our study is, to our knowledge, the first to empirically study meritocratic selection at this level.

Outside of the education sector, meritocratic selection is shown to improve performance (Dahis *et al.*, 2020), increase legitimacy (Inter-American Development Bank, 2016), increase the quality of new hires (Hoffman *et al.*, 2018), and reduce discrimination (Tan, 2008). Charron *et al.* (2017) argue that meritocratic selection shifts the motivation of public sector workers away from political criteria and towards professional criteria (see also Yeboah-Assiamah *et al.*, 2014). Dahis *et al.* (2020) show that even among state judges in Brazil, meritocratic selection is helpful in selecting the most competent candidates. Similarly, Hoffman *et al.* (2018) demonstrate meritocratic selection in the private sector improves the overall quality of new hires.

There are considerable difficulties when testing the causal impact of meritocratic selection of principals on student outcomes. The self-selection of principals into specific types of schools, such as inexperienced managers in poor performing schools (Loeb *et al.,* 2018; Branch *et al.*, 2012) threatens the identification of treatment effects. To counter these selection issues, Bohlmark *et al.* (2016) utilise a principal switching system in Sweden, which allows the inclusion of manager and school fixed effects as they change between schools to identify the effect of principals on outcomes. Whilst we cannot utilise a principal switching strategy, since we do not observe the school placement of principals pre-policy, by using the new Callaway and Sant'Anna (2021) estimator, we obtain difference-in-difference estimates that are clean of the bias of two-way fixed effect estimates. In so doing, our main contribution is to bring causal evidence of the impact of school principals on student outcomes in the short- and medium term, using a unique policy and a rigorous identification strategy. We reinforce this evidence with alternative OLS difference-in-difference estimations including school fixed effects and county-specific trends, similar to Estrada (2019).

The paper is structured as follows. Background information, and detail on the policy are outlined in section 2. We present our data sources, working sample and survey in section 3. Section 4 provides detail on the empirical strategy, whilst section 5 presents the results and mechanisms. Section 6 concludes the paper.

1.2 Background

1.2.1 Meritocratic Selection of Managers

Prior to 2016, principal appointments and dismissals were decentralised to the county inspectorates and were often made in a haphazard manner, lacking clear competency criteria and transparency. The typical recruitment process included a CV and operational plan assessment. Contracts were typically three years, but contracts would often be extended on the inspector's order, and on an annual basis. Anecdotal evidence suggests that often appointments and extensions would be granted based on political connections rather than candidate ability. The introduction of meritocratic competition in 2016 sought to rectify this.

The year 2016 was a time of numerous attempted reforms to the Romanian public administration. Following mass protests as a result of the deaths of 64 young people in a club fire in November 2015, which was linked to corruption, the former prime-minister Ponta was forced to resign. As a result, a new interim government, entirely comprised of politically independent technocrats, held power from November 2015 till December 2016. In an attempt to remove the existing political influence within school management, this government introduced the nationwide meritocratic selection process for all public school principals.

In August 2016, the Romanian Ministry of Education announced that the new competition would screen both cognitive and managerial skills; the process applied to both principal and deputy-principal positions in all pre-university institutions, and effectively meant the dismissal of those currently holding the position. Any principal currently in position was therefore forced to run in the competition to continue on as principal of that school; other principals were also able to apply for the post, giving no guarantee that the existing principal would remain. As part of the policy, four year fixed terms were introduced for meritocratically selected principals.

The exam was comprised of three main components: a written test; a CV screening; an interview. The written test was held at a national level, and focused on assessing both cognitive skills (a logical reasoning test, receiving a 66% weight in the overall test score) and knowledge of the newest management literature (33% weight) of applicants.[6] The CV screening and interview were judged by a county panel with the oversight of nationally appointed inspectors; these components focused on the motivation and management competencies of applicants, reflected in their one year operational plan. This process was transparent, and samples of all work were held and monitored by officials. Each component of the exam was awarded a maximum score of 50 points. In order to pass the exam, candidates had to score at least 35 points in each component and therefore a minimum score of 105 points overall – the highest written test score would break any tie. Principals were able submit multiple applications tailored for their preferred schools, however they could sit the test only once. Eligibility criteria to enter the competition included: having the relevant degree; have a "very good" professional, managerial, and moral record for the past 4 years; have had no disciplinary issues for the past 3 years.[7]

[6] Since the written test is a purely objective form of assessment, the introduction of such a test (in the absence of leakage from other sources) purely increases the probability of a principal being meritocratically selected.

[7] One might be concerned that meritocratic selection may still lead to biased appointments. Stravakou (2019) argues that the interview portions of meritocratic selection policies are still likely to lead to subjective assessments. Despite this, we argue that the policy which we study in Romania is a discrete change from the previous system of appointments at the discretion of regional education regulators (often based on obscure criteria or clientelist relationships) and towards a more merit-based form of selection.

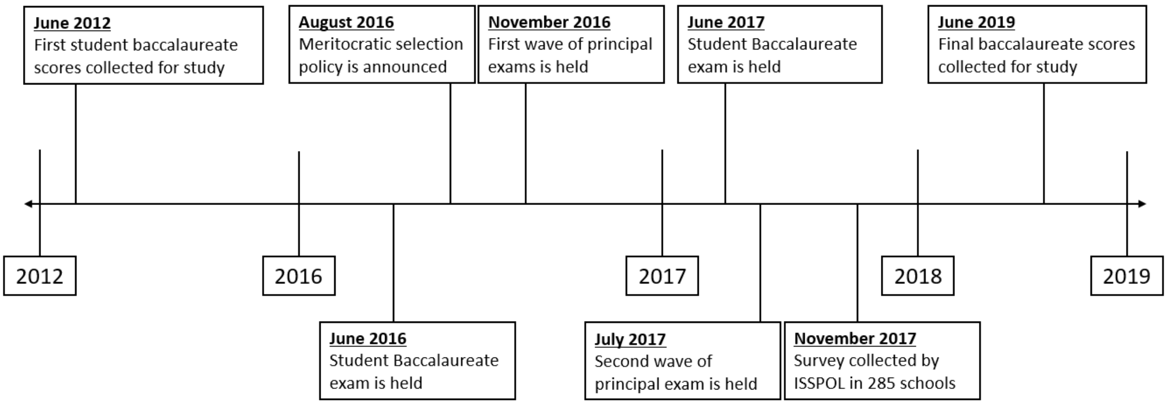
The meritocratic selection process was announced on 31st August 2016, and principals began to sit the exam as soon as October 2016 (applications were open between September 13th and October 2nd). The very short time period between announcement and enforcement left little time for prospective principals to gain an in-depth understanding of the selection process or to game the system. In addition, in this limited period, exam entrants were given considerable novel literature to learn, upon which parts of the written test were based. As a result, many schools saw no candidates apply for the position in 2016, whilst others saw candidates apply but fail the meritocratic selection process; the rest had a principal meritocratically selected in 2016 (around 75% of all schools).[8]

A resulting second wave of meritocratic selection occurred in July 2017, meaning that only schools which had not successfully appointed a principal through selection in 2016 were reopened for candidates to apply. In some instances, principals who failed the exam in 2016 were able to apply to the same school and be selected by exam in 2017. The exam process in 2017 was very similar to that in 2016; however, the written test covered knowledge of education legislation and managerial competencies based on different literature and excluded the logical reasoning test in 2017. Principals selected in 2017, took office in September 2017. Following the selection process in 2017, schools which either had no applicants or candidates selected by exam were appointed a principal by the county school inspector. This staggered intervention allows us to study the effects of competitive principals across time and different schools, and compare effects of the two meritocratic selection waves.

The timing of the exam selection of principals has significant impact in the way we must analyse the policy, a graphical representation of this timeline is shown in Figure 1.1. Since the selection process took place late in 2016, principals were placed in schools during the 2016/2017 academic year and were unable to influence student test scores for exams which took place in the summer of 2016 (2015/2016 academic year). Similarly, the meritocratic selection in 2017 took place after the 2017 Baccalaureate exam. This is important for our analysis as effects of the policy must be considered on the following years student outcomes (2016 selected principals first impact on 2017 outcomes; 2017 selected principals first impact on 2018 outcomes). The timeline (Fig. 1) visually demonstrates the event horizon of the policy implementation, including when our survey was collected (details in the next section).

[8] For the remaining 25% of schools, the technocratic government ruled that neither the incumbent nor the unsuccessful candidate could hold the position of interim principal; this meant that another person must be appointed. This stipulation was overruled by the new Social-Democratic government (appointed in December 2016), and was changed to allow incumbent principals to be reinstated until the meritocratic selection process could run again in 2017.

Figure 1.1. Timeline of Baccalaureate exam and competitive selection of managers


June 2012
First student baccalaureate scores collected for study
August 2016
Meritocratic selection policy is announced
November 2016
First wave of principal exams is held
June 2017
Student Baccalaureate exam is held
June 2019
Final baccalaureate scores collected for study
2012
2016
2017
2018
2019
June 2016
Student Baccalaureate exam is held
July 2017
Second wave of principal exam is held
November 2017
Survey collected by ISSPOL in 285 schools

1.2.2 The Role of Principals in Student Performance

Romanian school principals have very little autonomy in resource allocation and curriculum design. Romania's level of school autonomy is below the OECD average (OECD, 2014). The responsibilities of principals in Romanian schools are regulated through an Education Ministry's act,[9] and are limited to organising the entire educational process and delivering national education objectives. They are accountable for the school's performance, end-of-year evaluation and quality assurance processes. Principals also decide the annual budget and procurement processes, subject to approval from the school executive committee.

In terms of hiring decisions, principals have limited autonomy. They can propose new posts or submit vacancies to the county inspectorate, but the latter decide the final placement of tenured teachers in schools; however the principal's consultation with their school's executive committee regarding the new hires is often symbolic. Principals have some degree of freedom in recruiting substitute teachers (see ROFUIP 2020, Art. 21, paragraph 3). Also under the principal's remit are the training, integration and motivation of staff. Other responsibilities linked to the teaching and learning activities include: drafting internal regulations; assigning form teachers, school and extracurricular project coordinators; forming working groups and coordinating teaching and learning organisation such as timetabling; facilitating the professional development of staff. In terms of staff motivation and sanctions, principals have little autonomy. Salary scales and bonuses in state schools are regulated nationally, while contracts are protected by worker's rights. In extraordinary circumstances, the principals may propose to terminate staff contracts, but this must be approved by the inspectorate. However, principals contribute their statement in teachers' end of year evaluations, which may influence the inspectorate's allocation of merit bonuses. In terms of the routine monitoring of teacher's activities, principals are mandated to observe the teacher attendance records on a daily basis and make recommendations.

In sum, principals of Romanian state schools can, in the long run, influence education quality by maintaining a good school resource base and improving the school performance and reputation; in the short run, principals may improve staff discipline and attendance keeping, use non-pecuniary rewards to motivate staff and lead by example.

[9] The Ministry's act for the organisation and functioning of the pre-university education establishments - "Regulamentul-cadru de organizare si functionare a unitatilor de invatamant preuniversitar" (ROFUIP 2020).

1.2.3 The Baccalaureate Exam

At the start of high school, Romanian students are sorted into subject specific tracks based on previous performance measures (this process is computerised), these are: Theoretical – including humanities and sciences; Technological – including technical training, and natural resource/environment focuses; Vocational – including arts, military and sports. Regardless of their track, Romanian high school students are required, at the end of their 12th grade, to undertake a nationwide, standardized test: The Baccalaureate. This exam process is especially important as strong weighting is applied on the grade obtained in the Baccalaureate during university and labour market considerations. The Baccalaureate exam takes place in June-July (first sit) and August (reassessment for students who fail or miss the first sit) each year, and includes oral and written tests; whilst some exams differ between tracks, all students (regardless of track) take identical exam papers for Romanian language and literature. This allows for a direct comparison between students across tracks. The Baccalaureate scores were problematic before 2011, because the grades were inflated through cheating and corruption; since 2012 tougher punishments have been introduced and the exams have been monitored with CCTV devices. This means that Baccalaureate test scores have become a reliable measure of student performance (for an account of the impact of the anti-corruption campaign and further details about the Baccalaureate exam, see Borcan *et al.*, 2017). Therefore, we analyse the impact of the meritocratic selection of principals on student outcomes in the Baccalaureate exam from 2012-2019.

1.3 Data

1.3.1 Data Sources

We examine the impact of meritocratic selection on student test scores and explore the management mechanisms through which principals could enact change. To do so we use several sources to compile our outcomes, controls and to generate several exploratory variables. These sources are:

1) Principal selection data from each Romanian county and school from the Ministry of Education's administrative records of the meritocratic selection of managers in 2016 and 2017.[10] This data includes information about which candidates successfully passed the exam, and which were selected for the position.[11] From this data, we know which exam wave the principal was selected in, or whether there was no successfully appointed principal. Throughout the paper, we define a principal who was selected as part of the meritocratic selection policy as a "competitively selected principal" and any principals who were appointed without undertaking the meritocratic selection process as a "non-competitively selected principal". Whilst in principle we could use the entire universe of schools (1,637 schools), we restrict attention to a sample of schools for which we have more information on principal characteristics (see the working sample below).

2) Our working (reduced) sample consists of schools with data collected from county inspectorates; this data includes information on the year in which managers started their post. We use this data to divide the sample into managers who took their post as a result of the meritocratic selection policy (new managers), and those who remained managing the same school pre- and post-policy (legacy managers). Data was collected from 16 (out of 42) counties, who answered our Freedom of Information Act requests, and covers all schools within these counties. Our working sample is representative of the full sample (above) in terms of student outcomes both pre- and post- policy (Appendix Table 1.1). Within the working sample, 71% of schools (71.5% of students) had a legacy manager selected whilst 29% of schools (28.5% of students) had a new manager selected (Appendix Table 1.2).

3) Student outcome data, from by the Ministry of Education, which includes all students who enrolled in the Baccalaureate exam from 2012 to 2019.[12] In total, this data contains the following for over 1 million students in the full sample and 431,940 students in our working sample: the student's school and track; the exam subjects; the breakdown of exam scores; the final outcome of the overall Baccalaureate exam (including whether they passed); the rank of the student on a county and country wide level.

4) Survey data we collected independently from a randomly selected sample of 303 high school principals (around 20% of all high schools in Romania), carried out by the

[10] Whilst the meritocratic selection took place in all 41 Romanian counties, administrative data is only publicly available for 39 of 41 counties, and therefore two counties are excluded from analysis as we are unable to determine any principal appointments through meritocratic selection.

[11] This data also includes a full breakdown of exam test scores for each applicant (in some cases only final mark is available); whilst we do not use this data we do account for principal and school selection issues (see section 4).

[12] Only exam results from 2012 onwards are used as in 2012 all Baccalaureate exams in Romania became monitored by camera in an attempt to curb cheating, see Borcan *et al.* (2017).

Romanian "Institute for Social and Political Studies" (hereafter ISSPOL).[13] We identify competitively and non-competitively selected principals and new or legacy principals (also cross-checked with county inspectorate data used to form our reduced working sample). Importantly, the survey includes useful information about the principals' characteristics (demographics and motivation for public service) as well as their managerial practises, which we use to examine mechanisms linking the policy to student outcomes.

1.3.2 Summary Statistics

Our reduced sample includes 510 schools (and 431,940 students) across 16 counties of Romania, over the 2012-2019 period. In the Appendix Table 1.1 we show that there are strong similarities between the average student outcomes in the full sample and the reduced sample. In the full sample, the average overall Baccalaureate score between 2012-2016 was 6.118, whilst in the reduced sample it was 6.176. The similarities also hold for post-policy student outcomes (2017-2019), where the full sample average was 6.832 whilst the reduced sample average was 6.846, as well as for other student outcome metrics examined. Of the schools contained within the reduced sample, 90.78% (463) had an exam selected principal whilst 9.22% had a non-competitively selected principal before/after the policy (Appendix Table 1.2). Of the 90.78% schools with an exam selected principal, 71% had a legacy manager whilst 29% had a new manager. Significantly more new managers were appointed in the 2017 policy wave (48.4% of schools had a new manager in this wave) than in the 2016 wave (24% of schools had a new manager in this wave).

Within the reduced sample, schools with competitively selected versus non-competitively selected principals are comparable pre-policy in terms of the overall Baccalaureate score (student's average overall Baccalaureate score was 6.160 and 6.178 respectively; see Table 1.1), but schools with competitively selected managers have slightly better Romanian scores and pass rates. Conditional on having a competitively selected principal, schools with a legacy manager had significantly lower pre- and post-policy student outcomes than schools with a new manager (pre-policy student average was 6.137 for legacy compared with 6.280 for new managers, and post-policy outcomes were 6.787 compared with 7.045; see Table 1.1).

[13] The Romanian name is "Institutul de Studii Sociale si Politice". The Institute's website link (fully translatable) is available at: <www.isspol.ro>.

Table 1.1 – Student Outcome Variables by (Panel A) No Exam compared with Exam, and (Panel B) Legacy Managers compared with New Managers; Reduced Sample

Panel A	**(1)** **No Exam**	**(2)** **Exam**		
	Mean (sd)	Mean (sd)	Difference (1)-(2)	P Value
	(1)	(2)		
2012-2016 Overall Baccalaureate Score	6.160 (2.290)	6.178 (2.257)	-0.018	0.259
2017-2019 Overall Baccalaureate Score	6.786 (2.130)	6.851 (2.160)	-0.065	0.004***
2012-2016 Romanian Score	5.913 (2.285)	6.072 (2.376)	-0.159	0.000***
2017-2019 Romanian score	6.519 (2.141)	6.729 (2.282)	-0.21	0.000***
2012-2016 Pass rate	0.539 (0.497)	0.558 (0.498)	-0.022	0.000***
2017-2019 Pass rate	0.639 (0.473)	0.662 (0.480)	-0.023	0.000***
Panel B	**(1)** **Legacy Manager**	**(2)** **New Manager**		
	Mean (sd)	Mean (sd)	Difference (1)-(2)	P Value
	(1)	(2)		
2012-2016 Overall Baccalaureate Score	6.137 (2.318)	6.280 (2.278)	-0.143	0.000***
2017-2019 Overall Baccalaureate Score	6.767 (2.073)	7.045 (2.148)	-0.278	0.000***
2012-2016 Romanian Score	6.033 (2.319)	6.174 (2.270)	-0.141	0.000***
2017-2019 Romanian score	6.638 (2.096)	6.941 (2.153)	-0.303	0.000***
2012-2016 Pass rate	0.550 (0.494)	0.577 (0.497)	-0.027	0.000***
2017-2019 Pass rate	0.646 (0.460)	0.697 (0.478)	-0.051	0.000***

Notes: The table displays means and differences for all three student outcome metrics considered, along with a p value for a completed t-test, for schools in our reduced (working) sample. Panel A displays results for competitively selected (exam) and non-competitively selected (no exam), both pre- (2012-2016) and post- (2017-2019) policy. Panel B displays results for new managers and legacy managers, again pre- and post- policy. Standard errors at displayed in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Whilst it is smaller than the reduced sample, the survey sample contains information from a greater number of counties across Romania (303 schools from 39 counties).[14] In Appendix Table 1.2 we show that 277 (90.5%) of surveyed schools have exam selected managers, whilst 29 (9.5%)

[14] The survey sample may not constitute a fully representative sample because there is a selection of schools who decided to respond to the survey questionnaire. For this reason, as well as the sample size, our main results are based on the reduced sample. However, the survey sample gives us extra information on the managers' characteristics and practise, which are useful for examining the mechanisms behind the results.

did not. Of the exam selected principals, 25.6% began as a new manager in the school and the remainder were legacy managers. As in our reduced sample, considerably more new managers were selected during the 2017 policy wave in surveyed schools than in the 2016 wave (43% vs 13%; Appendix Table 1.2). Outcomes both pre- and post-policy are slightly lower in our survey sample than in the full and reduced samples, but are still very similar (Appendix Table 1.1).

These descriptive statistics make it clear that there was systematic selection into competition and change of management. Schools with a competitively selected principal in the 2016 wave were better on average, and new managers tend to be matched with higher performing schools. This has important implications for the identification of the competition impact on schools and student performance, as discussed in the following sections.

1.3.3 The Survey

Our survey was held by telephone interview, designed by us and conducted by ISSPOL – a private agency specialised in social research. The survey was carried out in November 2017, after both waves of the competition were over and when all principals were in office. The main advantages of telephone interviews are that they are less prone to social desirability bias and we could reach principals from a sample spanning almost the entire country. Interviews were between 15 and 20 minutes, and were held one-to-one with principals in Romanian high schools. The questions were adapted from the Perry (1996) survey on motivation for civil service and from the management practise survey of Bloom and Van Reenen (2007). Our survey has four independent sections: General Questions; Motivation for the Role and Management Activity; Evolution of Management and School Performance; Meritocratic Selection Process (see the full survey in the appendix).

In the general questions section we collected information such as the name, age, gender and general statistics about the running of the school. We use the first year in post to identify new and legacy managers. The motivation for the role and management activity module of the survey enquires about the beliefs of the principal regarding their suitability for the role, including ethical concerns, service to the public, their leadership style and management of working relationships. Questions in this portion of the survey are seven-dimension Likert-scales as in Perry (1996), with equal positive and negative worded items and a neutral term. In the evolution of management and school performance part of the survey, principals were asked to rate the performance of management practises compared to pre-policy years on a 3-point scale from "better" to "worse". These questions provide us with suggestive evidence of evolution in management practises which may have influenced changes in student outcomes. The final section of the survey is comprised of questions about meritocratic selection. Here the focus is on the principal's motivation to take part in the meritocratic selection policy (if they did take part), and their perceptions of the policy.

1.4 Empirical Strategy

1.4.1 The Model

We employ a Difference-in-Difference (DiD) strategy that exploits the gradual introduction of the policy in order to estimate the impact of the compulsory meritocratic selection of principals on student outcomes. We estimate the effects of two treatments. First, we compare schools with a competitively versus non-competitively selected principal; we are able to implement a staggered DiD due to the gradual filling of manager posts: a large share of schools received a competitively selected principal in 2016, but the rest deferred the competitive selection by exam until 2017 or had no valid candidates in 2016. The comparison between schools with policy selected principal and those without a policy selected principal forms our first treatment (T1).[15] Second, for schools with competitively filled posts, we estimate the impact of having a new manager (as opposed to a legacy manager) on student outcomes. Legacy (i.e. incumbent) principals remain in post upon passing the competitive selection process. Hence, having a new manager as a result of the competitive selection is our second treatment (T2).

The difference-in-difference two-way fixed effects (TWFE) estimator, which is commonly used to estimate of the average treatment effect on the treated (ATT) is based on an OLS regression:

$$Y_{isct} = \alpha + \beta T_{sct} + \varphi_t + \theta_s + \theta_c \cdot t + \gamma X_{icst} + \varepsilon_{icst} \qquad (1)$$

Here, Y_{isct} is the exam outcome for student i, in school s, situated in county c in year t. T is either T1 or T2, two indicators which are 1 when the units are treated (some begin to be treated in 2016, others in 2017), and 0 before they are treated. θ_s and θ_c are school and county fixed effects (which can also be interacted with a linear trend) and X_{icst} are several student-level pre-treatment covariates: chosen track (theoretical and technologic, where vocational is the omitted category) and, where we have the information, full-time student.

The unbiasedness of the TWFE estimator β (commonly used in such settings to capture the average treatment effect on outcomes) has been challenged by Goodman-Bacon (2021), who showed that it is in fact the "weighted average of all possible two-group/two-period DiD estimators in the data". Importantly, the TWFE estimator is comprised of several 2x2 DiD comparisons, some of which involve comparisons between groups treated at different points in time ("timing only" 2x2s). In our context, we have the gradual introduction of principals selected based on merit (and therefore also a gradual introduction of new principals as opposed to continuing ones), with two treated groups: the schools treated in the 2016 principal selection and those treated in the 2017 selection, while a small share of schools remained untreated through the entire period. This means that the TWFE estimator is an average of the following 2x2 comparisons: between the 2016 treated group vs never-treated, between the 2017-group and never-treated, but also between 2016 and 2017 treated groups (one where the 2017 group *before it is treated* serves as control for the 2016 group, and another where the 2016 group *after it is treated* serves as control group for the 2017 group).

The last "timing-only" component is particularly problematic, because when treatment effects are different over time between early and late treated groups, this component may capture

[15] Schools may not have had a policy selected principal either because there were no candidates who applied to the post, or no candidates who passed the exam.

these dynamic effects. Since OLS applies variance weighting on each 2x2 component to estimate the sample TWFE DiD, the weights on all comparisons will be positive, and this means the problematic 2x2 estimate may bias the average treatment effect (and may even change the sign of the average ATT). The specific bias is sample-dependent, with components for which there is a higher treatment variation receiving a higher weight (thus changing the length of the panel and implicitly the variation in treatment for certain treated groups can alter the estimates). Baker, Larcker and Wang (2022) highlight the fact that the problems associated with dynamic treatment effects and the 2x2 component from comparing the late-treated to the early treated group (as control) ca be mitigated when never-treated groups account for a substantial portion of the sample. In our case, the never-treated group for T1 (schools with principals confirmed in post without a merit-based competition) represent around 13% of the sample, and the never-treated group for T2 (schools with legacy managers) represent 71.06% of the sample (See Appendix Table 1.2). Thus, we expect a reduced bias in estimating the TWFE impact of T2 on outcomes.

In section 3.2 we showed that the early treated groups are systematically different from the later treated groups, and schools with new managers are different to those with legacy managers prior the selection. We may also expect that the treatment effects are heterogeneous across groups, and also that the effects change over time, as suggested in the literature on managers' impact on student performance, which may take time to be felt. For these reasons, the naïve OLS TWFE in model (1) may be biased even if parallel trends are satisfied. Therefore, we need a reliable estimator which can exclude the problematic 2x2 comparisons from the estimation of the overall ATT.

Several recent papers propose alternative estimators that correct the static and dynamic TWFE estimator, using different methodologies, depending on the context and treatment adoption (see Baker, Larcker and Wang (2022) for an overview of the main methods developed by Callaway and Sant'Anna, 2021, Sun and Abraham, 2021, de Chaisemartin, D'Haultføeuille, 2020). We apply the estimator developed by Callaway and Sant'Anna, 2021, which essentially estimates group-time-specific treatment effects through simple 2x2 comparisons with clean control groups (either never-treated, or not-yet-treated units). Each 2x2 comparison is a valid ATT for that specific group and time period, so long as there is no anticipation and there are unconditional parallel trends. The overall ATT for the sample is obtained by aggregating the group-timing 2x2s. The advantages with the Callaway and Sant'Anna estimator are: i) it is particularly suitable for our context where units once treated remain treated for the entire period; ii) it works with repeated cross-sections, as is the case in our data; and iii) it aggregates 2x2s both for each group and for each post-treatment period, such that we can infer the heterogeneous and dynamic treatment effects. This is important because it allows us to capture the changes in the impact of merit-selected and new managers on student outcomes.

To summarise, for our main analysis, we use the estimator from Callaway and Sant'Anna (2021), based on the doubly-robust estimator from Sant'Anna and Zhao (2020), and cluster bootstrap standard errors, clustered at county level. The main outcomes of interest Y_{isct} are: the overall student score of the Baccalaureate exam (observed between 2012-2019), whether the student passed the exam and the scores in a standardised written Romanian exam. We estimate this model on our working sample of 510 schools in the main specifications.

One disadvantage with this approach is that the standard errors are less efficient than the OLS standard errors (Wooldridge, 2021). For comparison, we also report the OLS regression TWFE estimates from model (1) in the appendix. We run the Goodman-Bacon (2021) decomposition (see appendix table 1.3), which shows that our TWFE estimator is in large part based on the outcome differences between the treated schools and the never-treated schools (90% weight).[16] This indicates that the bias in the classic DiD estimator is likely to be small, despite the problematic 2x2 comparisons between late and early treated groups.

We are also interested in whether the estimates differ along the distribution of pre-treatment school performance. For this heterogeneity analysis, we present the Calloway and Sant'Anna (2021) estimates, and for simplicity we also estimate variants of model (1) for different percentile groups (bottom, middle, and top) in the school average exam score distribution in 2012-2015 (see section 5.2).

Our setting is susceptible to selection bias, owing to the possibility that principals of different abilities selected the wave of the competition or the schools for which to compete. The concern is that we may have seen better performance in the schools with exam-selected managers or new managers, even in the absence of the policy. Table 1.1 indicates that pre-policy student performance is indeed higher in the exam-selected treatment group, and in the new managers (compared to legacy) group. The difference-in-difference estimates automatically account for the pre-policy gap in school performance. In the next section we examine the outcome trends before the policy, to rule out differential trajectories of the treatment groups before the competition was introduced. In terms of selection on unobserved characteristics, we also report OLS TWFE estimates controlling for school fixed effects and country trends.

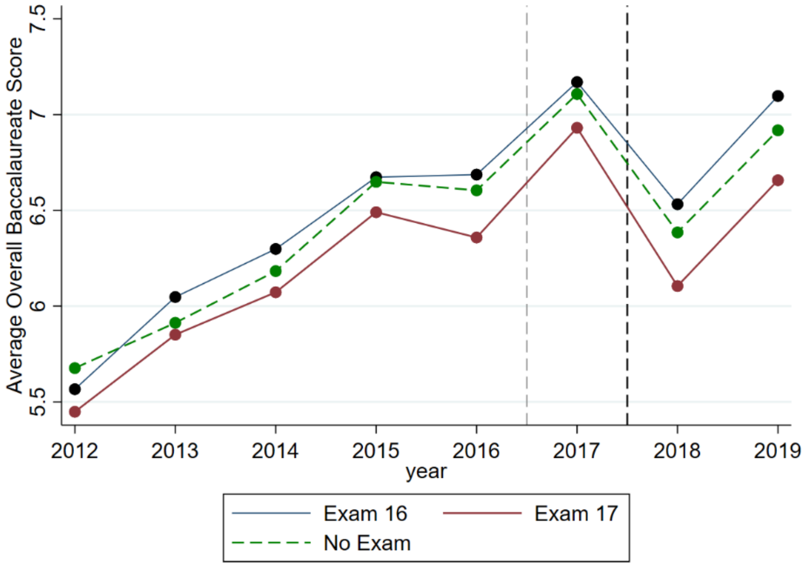
[16] Due to the requirement to have a strongly balanced panel, we retain only the schools that have Baccalaureate data for all eight years of the study time period, and we conduct the analysis on school-level aggregate scores and characteristics, weighted by the number of students in 2019.

1.5 Results

1.5.1 The Effect of a Principal Being Competitively Selected on Student Outcomes

First, we examine graphically the evolution of student outcomes over the study period in schools with a principal that was meritocratically selected in either wave of the policy, compared to schools where the principal was appointed (Figure 1.2). We see a very slight trend difference in overall Baccalaureate exam scores post-2016 based on whether the principal sat the exam in 2016 or 2017, however schools with a competitively selected principal in 2016 have higher student outcomes across the whole study period. This signals that there was selection of motivated principals into higher performing schools; despite this, Figure 1.2 also displays fairly parallel trends between the two groups pre-policy.

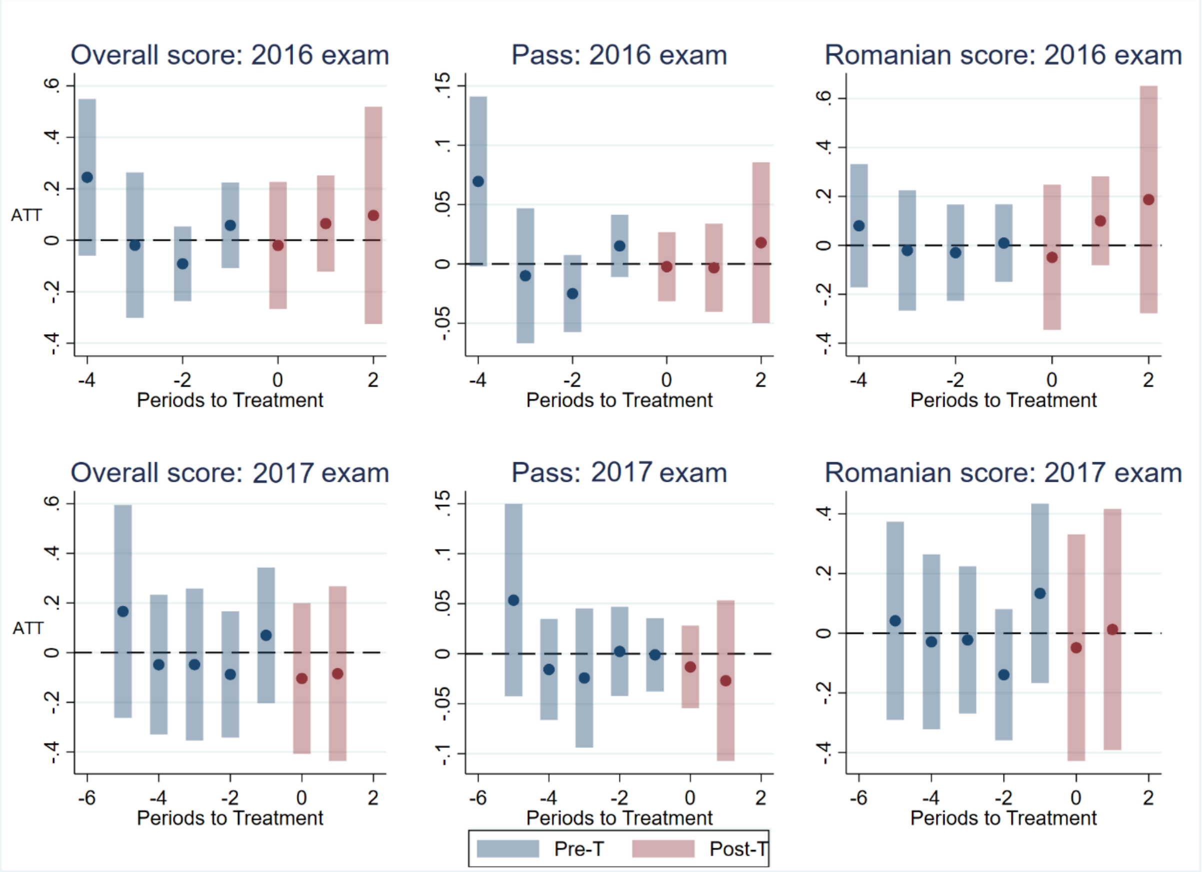
Figure 1.2 – Exam 2016 vs Exam 2017 (Reduced Sample)



Notes: The figure shows average overall Baccalaureate scores for our reduced (working) sample between 2012-2019. We show the average scores for schools with principals selected in the 2016 exam (in blue), and schools with principals selected in the 2017 exam (in red). The light dashed vertical line demarcates the timing of the 2016 exam, whilst the dark dashed vertical line demarcates the timing of the 2017 exam.

We also examine graphically the group-period ATTs estimates obtained using Callaway and Sant'Anna (2021) estimators, without conditioning for pre-treatment covariates, for each outcome and for each group (2016 and 2017 competition, respectively). In Figure 1.3, the top row figures display the ATTs for the 2016 competition group, while the bottom row figures display ATTs for the 2017 group. All pre-treatment ATTs are insignificantly different from zero. Most of the pre-treatment ATTs are close to zero or negative. The post-treatment ATTs for the 2016 competition group are close to zero in the first period after treatment, and become positive and larger in the next periods. The post-treatment ATTs for the 2017 competition group are close to zero or slightly negative. None of these post-treatment 2x2 DiD estimates are significantly different from zero.

Figure 1.3: DID plots (T1=principal exam) – dynamic heterogeneous effects by groups



Notes: The figure shows group-period ATTs from Callaway and Sant'Anna (2021) estimates for the effect of having an exam selected principal on student outcomes. We show the overall Baccalaureate scores, pass rates, and Romanian test scores for schools with principals selected in 2016 (top figures), and those selected in 2017 (bottom figures) separately. Note that these results do not condition for pre-treatment covariates. Pre-treatment ATTs are shown in blue, whilst post-treatment ATTs are shown in red.

We confirm these results in Table 1.2 in which we present the Calloway and Sant'Anna (2021) overall sample average ATT estimate (in column 1), the average ATT estimates for the 2016 treatment group and the 2017 treatment group (columns 2 and 3) and the average ATT estimates for each post-treatment period 2017-2019 (columns 4-6). In Panel A we display the estimates for the overall Baccalaureate score, Panel B displays estimates for the probability of passing the exam, and Panel C displays the estimates for the written Romanian exam. The overall sample ATTs are insignificant, and all group and period ATTs are insignificant. The period ATTs become larger the more time has passed since the competitive selection of a principal for the school, but never exceed 0.03 of a SD for the overall Baccalaureate score and pass, or 0.07 for the Romanian exam score. Thus the effects are precisely estimated, albeit fairly small. The formal tests of the unconditional parallel trend assumption confirm that we have parallel trends in terms of two outcomes: the probability of passing the exam and the Romanian exam score. We can reject the null hypothesis of parallel trends for the overall Baccalaureate score; however, this is mainly driven by the outcomes in the first period – the year 2012; all subsequent periods have ATTs insignificantly different from zero, as Figure 1.3 shows.

Table 1.2 – Exam vs No Exam; ATT and dynamic effects

	Sample	Group 2016	Group 2017	ATT 2017	ATT 2018	ATT 2019
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Overall Baccalaureate Exam Score						
ATT (T = Principal Exam)	0.029	0.046	-0.095	-0.020	0.035	0.066
	(0.116)	(0.123)	(0.144)	(0.126)	(0.092)	(0.199)
P-Value	[0.803]	[0.710]	[0.510]	[0.874]	[0.703]	[0.740]
Observations	428,061					
Chi2 H_0: parallel pre-trends	6.542					
P-Value H_0: parallel pre-trends	[0.000***]					
Panel B: Pass rate						
ATT (T = Principal Exam)	0.001	0.004	-0.020	-0.002	-0.005	0.010
	(0.018)	(0.018)	(0.028)	(0.015)	(0.018)	(0.034)
P-Value	[0.965]	[0.836]	[0.479]	[0.876]	[0.780]	[0.764]
Observations	431,940					
Chi2 H_0: parallel pre-trends	11.962					
P-Value H_0: parallel pre-trends	[0.215]					
Panel C: Romanian written exam score						
ATT (T = Principal Exam)	0.064	0.076	-0.020	-0.049	0.073	0.157
	(0.128)	(0.135)	(0.178)	(0.151)	(0.092)	(0.223)
P-Value	[0.615]	[0.573]	[0.908]	[0.746]	[0.423]	[0.480]
Observations	429,674					
Chi2 H_0: parallel pre-trends	65.450					
P-Value H_0: parallel pre-trends	[0.685]					

Notes: The table displays ATT estimates from difference-in-difference specifications of the effect of T1 (the meritocratic selection policy) on student outcomes, using the double-robust inverse probability weighting estimator from Calloway and Sant'Anna (2021). Panel A displays results for the overall Baccalaureate Score, Panel B for the passing rate, and Panel C for the standard Romanian Written Exam scores. We present estimates for the entire working sample in column (1), ATTs by treatment groups in columns (2) and (3), and ATTs by period in columns (4)-(6). County-clustered standard errors in parentheses for 19 clusters. *** p<0.01, ** p<0.05, * p<0.1

In Appendix Table 1.3, we show the Goodman-Bacon (2021) decomposition for treatment timing groups in our estimation specification. The Goodman-Bacon (2021) decomposition provides a full breakdown of the weighting of the four comparisons made between groups when using a TWFE DiD estimator. We denote the weights on the comparison of: our early treated group (2016) to our untreated group (no exam) as "W never vs 2016"; our late treated group (2017) to our untreated group as "W never vs 2017"; our early treated group to our late treated group pre-policy as "W treatment timings"; and our late treated group to our early treated group post-policy as "W within". As shown in Appendix Table 1.3 (column 1), only 2.7-3.4% of the main ATT components come from the problematic comparison. We expect the bias in the OLS TWFE estimate to be small, owing to the fact that its main ATT components that make up most of the effect (around 90%) are coming from comparing the treated groups with the never treated group, or with the not yet treated group. Therefore, in Appendix Table 1.4 we present the OLS results from model (1) on T1, whether a principal was competitively selected, for our full and working samples.[17] In columns (1) and (3) we present results with year fixed effects (FE) with 2016 as a reference, school FE and our student-level controls (track dummies and full-time student dummy); in column (2) and (4) we present results with the same FE and include county specific trends. Consistent with the results in Table 1.2, we do not find any significant difference in any of the three outcomes in our full sample, and no significant differences in passing probability and Romanian exam score in the working sample. We only find a small effect on the overall Baccalaureate score (around 0.01 of one SD of the no exam pre-policy mean) that is significant at 10 percent significance level in the working sample.[18]

Overall, we estimate a null average treatment effect of the policy of assigning a school a competitively selected principal on student outcomes up to three years after the policy, given the small magnitude of the coefficients and the narrow confidence intervals.

1.5.2 The Effect of a New Competitively Selected Principal on Student Outcomes

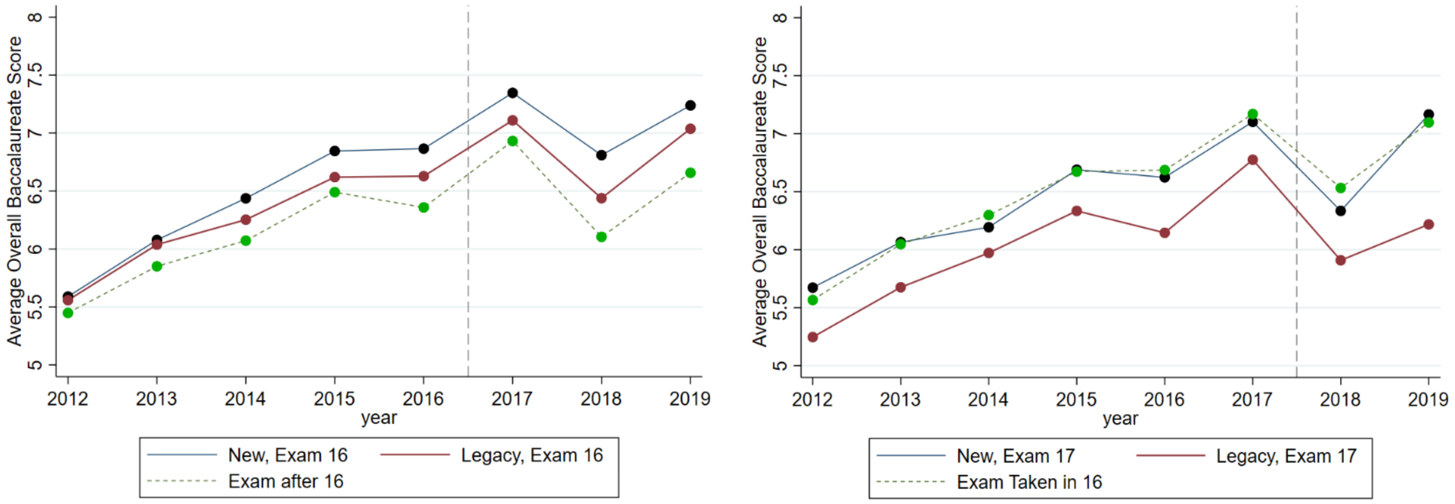
The second treatment we examine is having a new manager, compared to having a legacy manager conditional on the manager being confirmed in post through merit-based selection. It is more likely that new managers change school policies or processes than legacy managers.[19] However, the proportion of managers who remain in their post following the exam is larger than the proportion of new managers, which means that the overall effect of the exam may conceal the difference in impacts between existing and new hires. Figures 1.4.a and 1.4.b display the overall Baccalaureate score trends in schools of new and legacy managers who were competitively selected in 2016 and 2017 respectively; the proportion of new managers in the 2016 exam session was 24%, compared with 48.4% in 2017 (Appendix Table 1.2). Both figures show little difference in pre-policy trends of student outcomes for treatment and control schools. Figure 1.5 shows graphs of the Calloway and Sant'Anna (2021) estimated ATTs for each group and period, suggesting that most pre-policy ATTs are insignificant (parallel trends), except for the third period pre-treatment ATT for the overall Baccalaureate score, which is insignificant, but positive and a bit larger than the rest of the pre-treatment ATTs.

[17] For T1 (schools with competitively selected vs non-competitively appointed managers), we are able to compare DiD estimates in the full and reduced samples, since we know the treatment status for all schools. For T2 (schools with new managers vs legacy managers), we only have the treatment status data for the reduced sample of counties (see section 4). We use the T1 effects comparison as supporting evidence for the representativity of the reduced sample.

[18] We display estimates of T1 on our survey sample of schools in appendix table 1.5. We find no significant differences between competitively selected and appointed principals in any student outcome metric.

[19] For more information on the changes to school policies which Romanian principals can make, see section (2.1).

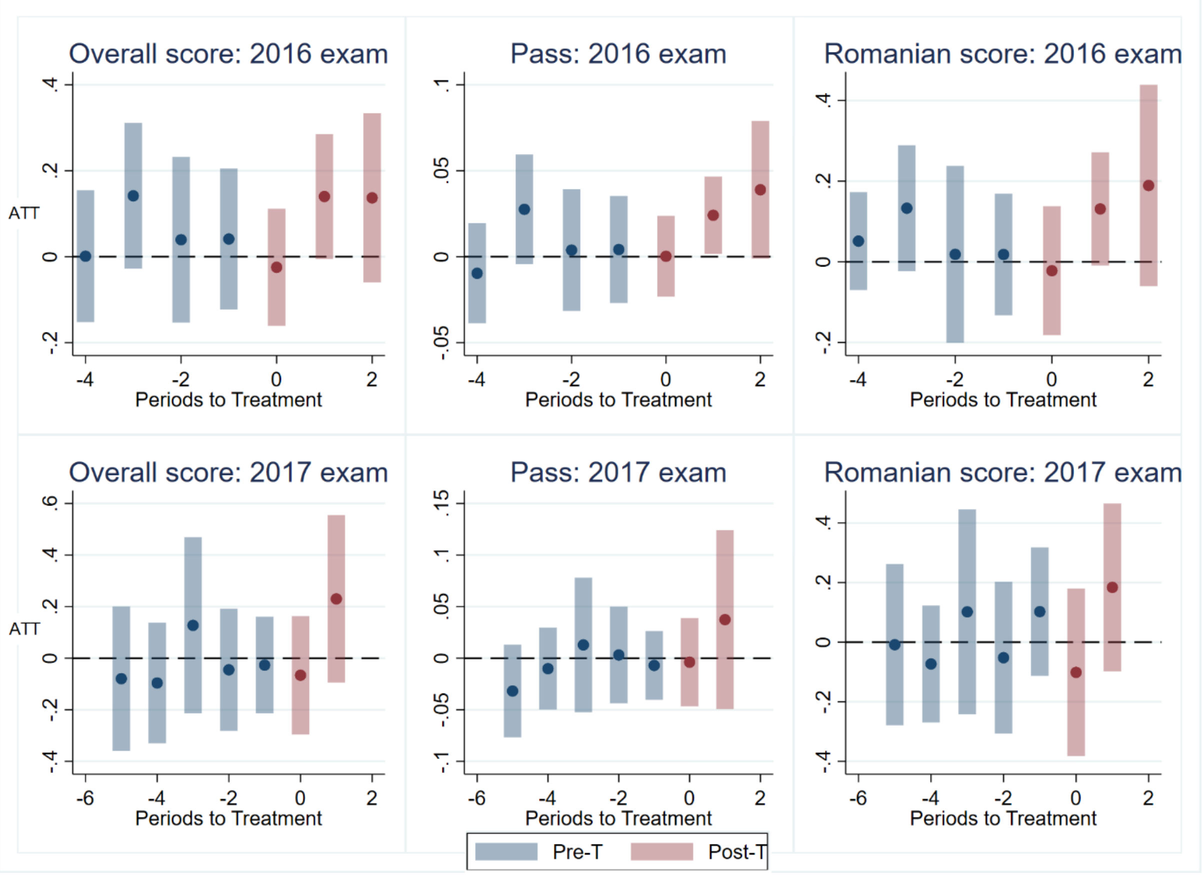
Figure 1.4.a – Exam 2016 New vs Exam 2016 Legacy vs Exam 2017 (all); Figure 1.4.b – Exam 2017 New vs Exam 2017 Legacy vs Exam 2016 (all)



Notes: The figure shows average overall Baccalaureate scores for our reduced (working) sample between 2012-2019. Figure 1.4.a shows average scores for schools with principals selected in the 2016 exam; these are displayed for new managers (in blue) and legacy managers (in red). In addition, we show average outcomes for principals selected in the 2017 exam as a reference (in green). Figure 1.4.b shows average scores for schools with principals selected in the 2017 exam; these are displayed for new managers (in blue) and legacy managers (in red). In addition, we show average outcomes for principals selected in the 2016 exam as a reference (in green). The dashed vertical lines demarcate the timing of the relevant exam.

The post-treatment ATTs are insignificant in the first period post-treatment, but become positive and larger in the second and third periods, and some are significant (for example, for pass probability, for the 2016 group).

Figure 1.5: DID plots (T2=new principal) – dynamic heterogeneous effects by groups



Notes: The figure shows group-period ATTs from Callaway and Sant'Anna (2021) estimates for the effect of having a new principal (compared to a legacy principal) on student outcomes. We show the overall Baccalaureate scores, pass rates, and Romanian test scores for schools with principals selected in 2016 (top figures), and those selected in 2017 (bottom figures) separately. Note that these results do not condition for pre-treatment covariates. Pre-treatment ATTs are shown in blue, whilst post-treatment ATTs are shown in red.

In Table 1.3, following the same format as in Table 1.2, we present the Calloway and Sant'Anna (2021) overall sample average DiD ATT estimate for T2 (new compared to legacy manager schools, in column 1) and again the average ATT estimates for the 2016 treatment group and the 2017 treatment group (columns 2 and 3) and the average ATT estimates for each post-treatment period 2017-2019 (columns 4-6). Note that the sample is now restricted to include only schools which had a principal selected through the competition. The overall sample ATTs for the Baccalaureate score and Romanian exam are insignificant, but the overall ATT for the pass probability is 2 percentage points and statistically significant at 10% significance level. The results differ slightly for the 2016 and 2017 groups, particularly in the Romanian exam score, where the 2016 group ATT is three times larger, albeit still insignificant. The overall sample ATT of having a new manager suggests an increase in overall Baccalaureate score by 0.035 of a SD (1.3% increase) on the legacy manager schools in the pre-policy period. The effect on the pass probability is larger, amounting to 0.04 of a SD (3.6% increase on the legacy schools mean pass rate).

One interesting result is that we see dynamic effects. The period ATTs of having a new manager become larger the more time has passed since the competitive selection of a principal for the school, and are statistically significant in the last period at 10% significance level for the Baccalaureate exam score and for the Romanian score, and significant at 5% for the pass probability. The last period ATTs are double the magnitude of the overall sample ATTs (an increase by 0.07 of a standard deviation increase for the overall Baccalaureate score, pass rates that are around 8 percent larger than in the control group, and an increase by 0.08 of a SD for Romanian exam scores). This indicates that new managers take some time to produce effects in terms of student outcomes, which is in line with the literature (e.g., Coelli and Green, 2011).

Table 1.3 – New vs Legacy Manager; Reduced Sample; ATT and dynamic effects

	Panel A: Overall Baccalaureate Score					
	Sample	Group 2016	Group 2017	ATT 2017	ATT 2018	ATT 2019
	(1)	(2)	(3)	(4)	(5)	(6)
ATT (T = New Manager)	0.081	0.083	0.070	-0.025	0.081	0.162
	(0.061)	(0.068)	(0.095)	(0.069)	(0.070)	(0.093)
P-Value	[0.186]	[0.220]	[0.467]	[0.724]	[0.247]	[0.081*]
Observations	394,409					
Chi2 H_0: parallel pre-trends	52.868					
P-Value H_0: parallel pre-trends	(0.000***)					
	Panel B: Pass					
ATT (T = New Manager)	0.020	0.021	0.015	0.000	0.016	0.039
	(0.011)	(0.012)	(0.027)	(0.012)	(0.011)	(0.019)
P-Value	[0.085*]	[0.088*]	[0.566]	[0.982]	[0.151]	[0.047**]
Observations	398,013					
Chi2 H_0: parallel pre-trends	14.365					
P-Value H_0: parallel pre-trends	(0.110)					
	Panel C: Romanian written score					
ATT (T = New Manager)	0.084	0.097	0.030	-0.022	0.064	0.188
	(0.067)	(0.079)	(0.100)	(0.082)	(0.069)	(0.100)
P-Value	[0.210]	[0.215]	[0.765]	[0.787]	[0.355]	[0.062*]
Observations	395,925					
Chi2 H_0: parallel pre-trends	9.1576					
P-Value H_0: parallel pre-trends	(0.421)					

Notes: The table displays ATT estimates from difference-in-difference specifications of the effect of T2 (the new manager policy) on student outcomes, using the double-robust inverse probability weighting estimator from Calloway and Sant'Anna (2021). Panel A displays results for the overall Baccalaureate Score, Panel B for the passing rate, and Panel C for the standard Romanian Written Exam scores. We present estimates for the entire working sample in column (1), ATTs by treatment groups in columns (2) and (3), and ATTs by period in columns (4)-(6). All specifications display the chi2 and p-values from the tests for the unconditional parallel trends assumption. County-clustered standard errors in parentheses for 19 clusters. *** p<0.01, ** p<0.05, * p<0.1

Parallel pre-treatment trends tests confirm the assumption holds for the probability of passing the exam and the Romanian exam score, but not for the overall Baccalaureate score. Nevertheless, we see consistently significant and higher treatment effects for all outcomes in the last period. In Appendix Table 1.6, we present estimates of our OLS DiD specification using T2, whether a principal is a new or legacy manager. Overall, we find similar or slightly smaller magnitudes to the overall ATT estimates in Table 1.3, none of which are statistically significant.

Having found null contemporaneous and a small, delayed effects, we explore the potential channels by which competitive managers may have made a difference on student outcomes. We examine the differences between new and legacy managers in the next section.

1.5.3 Mechanisms

We examine the differences between management styles and motivations of new and legacy managers in our dataset of surveyed schools. Of the 277 surveyed schools which had an exam selected principal, 206 schools had a legacy manager whilst 71 had a new manager. One caveat of our survey is the absence of one particular reference category: former principals who were replaced by a new principal but who were not competitively selected for another post. Thus, we do not have the ideal counterfactual to inform the mechanisms behind the treatment effect of new managers. However, we are able to compare new principals to competitively selected legacy principals, who likely share some similar traits by virtue of being instated through the policy and managing similar schools. We analyse the difference in survey responses between competitive new and legacy managers in our surveyed schools (Table 1.4).

First, we consider differences in responses to questions related to the motivation for the management role (adapted from Perry's 1996 motivation for civil service survey). These range from beliefs about the importance of ethical behaviour, self-righteousness, and own interpersonal skill, measured on a 7-point Likert scale (e.g. "The ethical behaviour of school directors is as important to me as their competencies"; see the complete questionnaire in the Appendix). On average, new managers are more supportive of peers (more willing to "fight for the rights of others"; $p<0.1$), are typically less rude to others ($p<0.05$), and see financial success as less important ($p<0.05$). These traits might suggest that new managers are more service-minded and more likely to foster an environment of support and cooperation between teachers and management; in particular, their lesser concern with financial success suggests they are more prosocial and more motivated to fulfil their duties.

Table 1.4 – All Survey Responses for Surveyed Schools; P-Value for T-Test between New and Legacy Managers.

Surveyed Schools with Exam	Legacy Manager		New Manager		
	Mean	sd	Mean	sd	P Value
Values and Beliefs					
Ethics are Important	6.828	0.531	6.900	0.302	0.164
Fights for Rights of Others	6.197	1.211	6.414	0.807	0.093*
Civic Issues are Moral Obligations	6.527	0.940	6.657	0.720	0.232
Care for Welfare of Strangers	3.813	2.074	3.943	2.139	0.659
Financial Success is Important	2.616	1.752	2.171	1.569	0.050**
Effective Problem Solver	6.030	0.949	6.143	0.889	0.368
Sometimes Rude to Others	2.739	2.028	2.129	1.769	0.018**
Likes to Listen to and Help Others	6.443	0.796	6.400	0.969	0.737
Easy to Work with Others	6.631	0.800	6.443	1.072	0.183
Working Relationships					
Teacher Trust	5.581	1.013	5.257	1.259	0.054*
Director Trust	5.980	0.802	5.871	0.992	0.409
Share Common Values	6.054	0.961	5.971	1.063	0.566
Trust Enough to Delegate	6.192	0.877	6.186	0.952	0.961
Men Make Better Managers	1.872	1.443	1.529	0.944	0.025**
Management Techniques Compared to Pre-Policy					
Adapt Learning to Student Needs	2.726	0.469	2.594	0.577	0.089*
Tracking Teacher Performance	2.594	0.502	2.623	0.571	0.707
Tracking Objectives	2.764	0.459	2.754	0.526	0.889
Teacher Absence	2.335	0.654	2.265	0.704	0.471
Hiring Skilled Staff	2.470	0.557	2.368	0.621	0.229
Use of Staff Sanctions	2.097	0.551	2.118	0.441	0.756
Motivate Staff	2.502	0.575	2.638	0.593	0.102
Mobilization and Initiative Spirit	2.631	0.523	2.735	0.507	0.146
Manage Authority	2.577	0.505	2.559	0.583	0.818
Student Absence Rate	2.269	0.719	2.294	0.648	0.786
Student Dropout Rate	2.343	0.655	2.235	0.649	0.240
Graduation Rate of 12th Grade	2.624	0.553	2.559	0.608	0.438
Registration Rate for Baccalaureate	2.361	0.671	2.456	0.679	0.322
Observations	203		74		

Notes: The table displays mean survey responses for a full set of survey questions, along with a p value for the difference in means test, for our survey sample. Panel A displays results for the set of survey questions related to manager beliefs about morality and their own qualities in relation to work and co-workers. Panel B displays results for the set of questions related to working relationships. Panel C displays results for the set of questions related to how management techniques have changed compared to pre-policy. In column 1 we show responses for new managers (with standard deviations in column 2), and in column 3 we show responses for legacy managers (with standard deviations in column 4). *** p<0.01, ** p<0.05, * p<0.01

Despite these findings, we also find suggestive evidence that legacy managers are more likely to believe that teachers in their school trust each other (*p<0.1*). In addition, we find some differences between new and legacy managers which may be imprecisely estimated due to a small sample size: legacy managers are more likely to believe that teachers trust management, share common values with staff and trust teachers enough to delegate authority more often than it is the case in schools with new managers. Note that one of our main findings highlights the time it takes for new managers to have a strong impact on outcomes. Therefore, it is unsurprising that new managers perceive trust between teachers and management to be weaker than legacy managers do, as productive relationships within the organisation may take time to build.

We also analyse the differences in responses to questions about the (perceived or observed) change in management practises and school performance indicators since taking on the role. These questions ask whether the principal believes that specific metrics have declined, stayed the same or improved since 2017-2018 compared to previous years (on a scale 1-3, where 1 is declined and 3 is improved). These metrics are: adapting learning to student needs, tracking teacher performance, monitoring objectives, teacher absence, hiring skilled staff, use of sanctions for staff, motivating staff, mobilizing staff, the level of authority over staff, absence amongst students, student dropout rate, graduation rate, baccalaureate enrolment rate. Table 1.4 shows that new managers report significantly less improvement in adapting learning to student needs ($p<0.1$), but that there are no other significant differences between new and legacy managers. Many similarities between new and legacy managers' schools in terms of management practises (especially on motivating staff) are unsurprising, given that school managers have little autonomy in decision making; most key decisions for education quality are made at the county inspectorate or ministry level, including: remuneration, hiring and firing of teachers and principals; school budgets.

We can go one step further and examine whether the objective Baccalaureate enrolment rates have changed differentially in schools with new compared to legacy managers. Managers can decide strategically whether to pass marginal students and allow them to graduate and sit the exam (e.g., only better students sit the exam). Schools are ranked by passing rates in the first sit, and improvements in the passing rates are monitored on an annual basis. Preventing students predicted to fail the exam from taking the first sit (leaving them the option of sitting in the August reassessment) improves passing rates and test scores, and therefore the objective measure of school performance. To test for this, we need data on the Baccalaureate enrolment rate, i.e. the share of students who sit the Baccalaureate of the students who are enrolled in the 12th grade in each school. For our measure of the number of students enrolled in the 12th grade, we use high school admissions data from Munteanu (2021).[20] The data contains the number of admitted students per each track, but excludes students who were in vocational tracks (e.g. theology, pedagogy etc.). As a result, for some schools which do not offer a vocation track we have the total number of students admitted to the high school four years prior the Baccalaureate. However, for schools with vocational tracks, we only have the number for the subset of the students admitted to technical and theoretical tracks. We create the Baccalaureate enrolment rate as the ratio of the number of theoretical or technical track students who sat the Baccalaureate (corrected by the number of returning students) to the number of students admitted to that particular school and track four year prior. Thus, we deal with the missing data by excluding vocational track students from the sample. The disadvantage of restricting the observations in this way is that we lose all vocational schools (typically lower performance schools). We re-estimate the results for the sample of technical and theoretical schools.

The results in Table 1.5 show that the new manager DiD estimates on the Baccalaureate enrolment rate from the restricted sample are negative and significant, both in terms of the overall ATT, implying an overall reduction in Baccalaureate enrolment rates by 8.4 percentage points

[20] Obtained from Diana Coman's repository of Romanian public education data (www.ossasepia.com).

(*p=0.02*), and in terms of the 2018 and 2019 outcomes (the drop is 13.6 percentage points the 2018). The results suggest that one margin which new managers can and do control is the permission for students to enrol in the Baccalaureate exam. Managers have the authority to decide that students whose predicted grades are below the passing threshold defer to the second sit of the exam which takes place two months later, instead of sitting in the main exam in June. This has two effects: students have more time to prepare for the exam and achieve higher grades, and school Baccalaureate performance measures mechanically improve.

Table 1.5 – New vs Legacy Manager; Reduced Sample; ATT and dynamic effects on Baccalaureate Enrolment rates and initial cohort size

	Baccalaureate Enrolment rate					
	Sample	Group 2016	Group 2017	ATT 2017	ATT 2018	ATT 2019
Panel A: Baccalaureate Enrolment	(1)	(2)	(3)	(4)	(5)	(6)
ATT (T=New manager)	-0.084	-0.069	-0.135	-0.010	-0.136	-0.082
	(0.036)	(0.042)	(0.053)	(0.023)	(0.071)	(0.029)
P-value	[0.021**]	[0.101]	[0.011**]	[0.662]	[0.055*]	[0.004***]
Observations	2,848					
Chi2 H_0: parallel pre-trends	5.022					
P-Value H_0: parallel pre-trends	[0.832]					
Panel B: Cohort size						
Admission ATT (T=New manager)	3.876	4.083	3.165	0.979	5.655	4.142
	(3.605)	(3.984)	(9.855)	(3.872)	(4.563)	(5.480)
P-value	[0.282]	[0.306]	[0.748]	[0.800]	[0.215]	[0.450]
Observations						
Chi2 H_0: parallel pre-trends	18.42					
P-Value H_0: parallel pre-trends	[0.030**]					
Controls	No	No	No	No	No	No

Notes: The table displays Calloway and Sant'Anna (2021) estimates of the ATT of new managers on Baccalaureate Enrolment rates and the initial cohort size (number of students admitted four years prior to the Baccalaureate exam), including a test for unconditional parallel trends.

Thus, we provide evidence of two potential channels by which managers may influence the school performance in the Baccalaureate exam. First, survey measures suggest that new managers are more prosocial and more motivated to fulfil their public mission. The main limitation of the survey is its susceptibility to social desirability bias, which is a concern if it is displayed differently by new and legacy managers (however, in that case, we would expect new managers to score higher on more measures than the select few that we see). Despite this and the small sample, the results suggest that selecting new managers can bring renewed energy in improving student achievement. Second, because managers have little autonomy in changing fundamental inputs in education, such as recruiting and incentivising teachers, they use the only available instruments to adjust school performance measures. We show evidence that schools with new managers have proportionally fewer students enrolling in the Baccalaureate exam, starting with the second cohort after the policy. This may be a significant driver of the modest improvement in test scores over time. Nevertheless, it is plausible that along with a selection of better students in sitting the exam, there was also a real improvement in students' performance.

1.5.4 Heterogeneity Analysis in School's Pre-Policy Performance

In order to analyse the differential impact of new and legacy managers depending on past performance of schools, we split our working sample into three groups on the full sample distribution of school average final exam score over the 2012-2016 period: below the 25th percentile (<25th); between the 25th and 75th percentile (25th-75th); above the 75th percentile (>75th) in school average exam scores. We divide the distribution in this way in order to capture the bottom and top performing schools in distinct groups from the midrange.

In columns (1-3) of Table 1.6 we present the Calloway and Sant'Anna (2021) average DiD ATT estimates for T2 (new compared to legacy manager schools) for schools with pre-policy performance under the 25th percentile. In columns (4-6) we present results for the middle part of the distribution, and in columns (7-9) we present ATT estimates for schools with pre-policy performance above the 75th percentile. We report results assuming unconditional parallel trends, We also report the ATT estimates for the 2016 treatment group and the 2017 treatment group (columns 2-3/5-6/8-9 respectively). As with the results in Table 1.3, the sample is now restricted to include only schools which had a principal selected through the competition.

We find negative but insignificant overall ATT's of new managers on student outcomes for schools in the group below the 25th percentile. The new manager effects are positive for the 2017 policy wave for this group, modest in magnitude for the overall exam score and large for the Romanian exam score (the estimate for the Romanian exam score is 0.31, or 7.6% increase compared to mean score in the legacy managers' schools before 2017, and *p=0.157*). Within the 25th-75th percentile group, the overall ATT estimates are positive, modest in magnitude and insignificant. In this group, we find significant positive ATTs for the 2017 policy wave, for the Baccalaureate exam score (estimate 0.168, equivalent to a 3.2% increase on the control group mean score before 2017; *p=0.067*) and for pass rates (estimate 0.051, an 11.3% increase compared to the pass rate of 0.451 in legacy managers' schools before 2017; *p=0.029*);[21] By contrast, we do not find a significant impact of new managers in schools belonging to the group above the 75th percentile. Indeed, many of these estimates are close to 0.

[21] For a complete set of reference group student average outcomes, pre- and post-policy, please see Appendix Table 1.1.

Table 1.6 – New vs Legacy Managers by school pre-policy performance; ATT

	>25th percentile			25th-75th percentile			>75th percentile		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Sample	2016	2017	Sample	2016	2017	Sample	2016	2017
Panel A: Overall Baccalaureate Score									
ATT	-0.076	-0.114	0.052	0.075	0.050	0.168	0.026	0.029	0.013
	(0.073)	(0.089)	(0.164)	(0.082)	(0.088)	(0.092)	(0.057)	(0.062)	(0.105)
P-Value	[0.301]	[0.197]	[0.750]	[0.363]	[0.569]	[0.067*]	[0.650]	[0.643]	[0.901]
Observations	46,953			190,848			156,339		
Chi2 H_0: parallel pre-trends	244.762			25.718			10.731		
P-Value H_0: parallel pre-trends	[0.000***]			[0.002***]			[0.295]		
Panel B: Pass Rate									
ATT	-0.024	-0.024	-0.025	0.022	0.014	0.051	0.004	0.006	-0.002
	(0.014)	(0.018)	(0.037)	(0.019)	(0.019)	(0.023)	(0.007)	(0.009)	(0.019)
P-Value	[0.093*]	[0.177]	[0.491]	[0.238]	[0.471]	[0.029**]	[0.529]	[0.491]	[0.908]
Observations	47,969			193,099			156,661		
Chi2 H_0: parallel pre-trends	47.683			8.162			9.605		
P-Value H_0: parallel pre-trends	[0.000***]			[0.518]			[0.383]		
Panel C: Written Romanian Score									
ATT	-0.025	-0.128	0.310	0.038	0.045	0.014	0.042	0.046	0.024
	(0.114)	(0.111)	(0.219)	(0.105)	(0.104)	(0.185)	(0.070)	(0.072)	(0.152)
P-Value	[0.829]	[0.250]	[0.157]	[0.715]	[0.667]	[0.940]	[0.548]	[0.523]	[0.873]
Observations	47,265			191,878			156,504		
Chi2 H_0: parallel pre-trends	94.556			33.421			8.036		
P-Value H_0: parallel pre-trends	[0.000***]			[0.000***]			[0.531]		

Notes: The table displays ATT estimates from difference-in-difference specifications of the effect of T2 (the new manager policy) on student outcomes by pre-policy performance groups, using the double-robust inverse probability weighting estimator from Calloway and Sant'Anna (2021). Panel A displays results for the overall Baccalaureate Score, Panel B for the passing rate, and Panel C for the standard Romanian Written Exam scores. The sub-samples in terms of the 2012-2016 average exam score are displayed as follows: the low-performing schools (below the 25th percentile) in columns (1)-(3), middle-performing schools (25th-75th percentile) in columns (4)-(6) present, and top-performing (above the 75th percentile) in columns (7)-(9). We present estimates for the entire working sub-sample in columns (1), (4) and (7), ATTs by treatment groups in columns (2) – (3), (5) – (6), and (7)-(8). All specifications display the chi2 and p-values from the tests for the unconditional parallel trends assumption. County-clustered standard errors in parentheses for 19 clusters. *** p<0.01, ** p<0.05, * p<0.1

The caveat in reading these results is that when splitting the sample in this way, we cannot guarantee parallel pre-policy trends in each group. For the low performing schools, which sit below the 25th percentile, as well as for two outcomes for schools in the 25th -75th percentile, we could reject the null hypothesis of parallel pre-trends (both unconditional and conditional). For this reason, the estimates should be read with caution. We can reliably conclude that there were no effects of new managers on top performing schools. However, there is some evidence that middle-performing schools in the 2017 policy wave benefitted significantly from having new managers, at least in terms of pass rates (for which the parallel trend assumption holds), and possibly in terms of the other outcomes. The results for the low-performing schools are inconclusive, and cannot be reconciled with the OLS TWFE estimates in Appendix Table 1.7, which display positive and significant estimates for this low performing group.

Given that we have found an overall positive treatment effect emerges two years after the policy, these results make it plausible to suspect that the policy shifted the left side of the performance distribution up, and that the policy has helped to reduce the performance gap between students in low-to-middle and top schools. We also examine the impact of the overall policy (comparing competitively and non-competitively selected principals) on students in different regions of the school performance distribution, and we display the results in Appendix Table 1.8. We find no impact of the overall policy on any student outcome in any percentile group, which suggests that the manager exam policy only worked insofar as it enabled a change in managers in some of the lower performing schools.

1.5.5 Robustness Checks

In order to verify that our main results are econometrically robust we undertake a series of robustness checks including using the not yet treated as an alternative control group, and running the traditional TWFE OLS models with placebo policies.

We estimate the Calloway and Sant'Anna ATT, for the overall sample, and by timing groups and dynamic effects using the not yet treated units as a control group. This means that units never treated and units who were only treated in 2017 become the control group for units treated in 2016. For the units treated in 2017, the never treated units remain the control group. The results displayed in appendix Tables A9 and A10 (for T1 – exam and T2- new managers) are very similar to the main results in Tables 2 and 3.

Earlier we showed that the TWFE OLS models show results broadly consistent with the Calloway and Sant' Anna ATT. The advantage of OLS is that we can control for time-varying and invariant characteristics of schools and students. One additional check to understand whether the OLS estimates reflect a trend that predated the policy is a placebo test where we artificially instate the policy one year earlier (2015 replacing 2016, and 2016 replacing 2017 selection wave). We would expect to see insignificant coefficients of the placebo policy. Appendix Table 1.10 shows no significant result from our placebo tests in any student outcome metric in the overall sample. There is one significant estimate of the placebo policy on overall Baccalaureate scores for schools which received a new manager, relative to those who retained legacy managers in 2016, but the effect is very small and significant only at 10%. This is further supporting evidence that our main results are not confounded by diverging trends started just before the policy, which partly alleviates the concern that new managers selected schools based on their prior performance trajectory.

Finally, for the OLS TWFE estimator, we turn our attention to the standard errors. Our main estimation includes 19 county clusters, which is lower than a minimum standard number of clusters required to estimate unbiased standard errors. A small number of clusters with a high degree of inter-cluster correlation, as is likely to be the case within Romanian counties, can lead to an underestimation of standard errors and falsely rejecting a true zero effect (Cameron *et al.* 2008). We therefore run alternative regressions (examining T2 – the impact of new managers) with the wild bootstrap correction for the standard errors. We report these in Appendix Tables A12. The wild bootstrap correction of the standard errors does not compute new coefficients, however Table 1.12 displays new t- and p-values. We find no significant results when estimating bootstrap standard errors; this may be expected as our earlier results suggest that new managers only have an impact in the long-term, but our bootstrap standard errors are calculated based on a TWFE estimate which does not include a breakdown of ATTs (as in Callaway and Sant'Anna, 2021).

1.6 Conclusions

This paper aims to understand whether meritocratic selection of public school managers can improve student outcomes. We test whether (1) having a competitively selected manager, and (2) having a new manager selected based on merit, improves a range of student outcome metrics: overall Baccalaureate scores; pass rates; and written Romanian scores.

Specifically, we study the effect of the Romanian meritocratic selection policy, which was introduced in two waves (2016 and 2017 independently), on outcomes from 2012-2019. We utilise administrative data to analyse effects of principals who became new managers in the school, compared to those who remain manager of the same school pre- and post-policy. Further, we examine survey responses of principals and additional outcomes to understand the possible mechanisms through which change is enacted.

At first glance, we do not find a significant impact of the overall policy; that is to say, having an exam selected principal does not inherently improve student outcomes in the short run. However, we do find some evidence that those competitive managers who were new in post had a positive impact on student outcomes. In line with the literature, our results suggest that it takes time for new managers to have an impact on outcomes since the ATT continues to increase in the years post-policy. In addition, we find evidence that new managers who were exam selected in 2017 have a positive impact on outcomes in low to mid- performing schools (25th-75th percentiles). In contrast, new managers seem to have little impact on top- (over 75th percentile) performing schools; note that improvement in these schools may be limited due to them having consistently high performing students.

Whilst autonomy is limited within the Romanian education system, potential mechanisms through which our results are brought about are explored using survey responses. Evidence suggests that new managers are more prosocial individuals and are more motivated to fulfil their public mission through their work. However, legacy managers are more likely to believe that teachers within their school trust each other. We believe that these results are not unexpected since, both in the literature and in our study, new managers are shown to take time to enact positive change. Since the survey was held shortly after the final wave of the policy, changes by new managers were unlikely to have yet been introduced or taken effect.

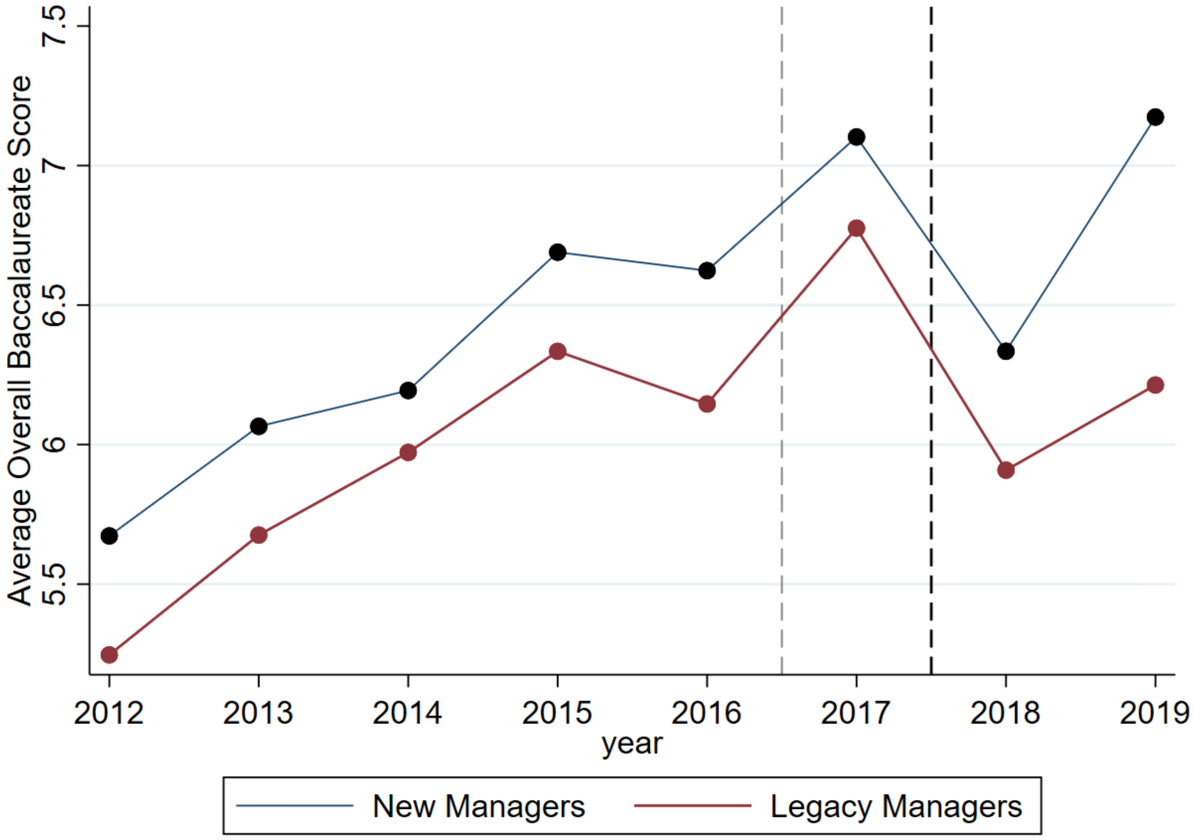
An important lesson from our analysis of the mechanisms is that meritocratic selection of public sector managers may contribute to increasing the performance of students, but little school autonomy also means managers might use shortcuts to mechanically enhance performance indicators. One of the few margins which managers can influence in our context is the rate at which students are enrolled into the Baccalaureate exam. Our results suggest that this is one mechanism through which new managers enact change. By withholding low performing students from taking the exam, managers may deliver higher passing rates for their schools.

The results for the low- and mid-performing schools suggest that competitively selected new managers can influence the outcomes of students at the bottom and middle of the distribution, which is promising in terms of short-term solutions for reducing education inequality. Future work may include a longer time-horizon of student outcomes after the enactment of the meritocratic selection policy, which was not possible in this study due to the systematic changes to the education process following the COVID-19 pandemic.

On balance, the meritocratic selection process allowed more motivated candidates to replace underperforming ones and deploy their skills to bring about medium- and possibly long-term gains in student performance. Despite this, there were relatively fewer new managers in low-performing schools compared to the rest. Therefore, providing incentives for new managers to select into low-performing schools alongside the meritocratic selection of managers is a promising avenue to improving student outcomes and bridging the performance gap between high and poor performing schools. While we do not test it in this paper, the interaction between meritocratic management selection and increased school autonomy may increase competition across schools and lead to more sustained gains in student achievement, a topic which we leave for future research. More broadly, our paper provides robust evidence from a sudden and widespread policy change that merit-based selection of managers in the public sector can generate improvements in public service delivery.

1.7 Appendix

Appendix Figure 1.1 – New vs Legacy Manager; <25th Percentile; Reduced Sample



Notes: The figure shows average overall Baccalaureate scores in schools <25th percentile for our reduced (working) sample between 2012-2019. We show the average scores for schools with new managers (in blue), and schools with legacy managers (in red). The light dashed vertical line demarcates the timing of the 2016 exam, whilst the dark dashed vertical line demarcates the timing of the 2017 exam.

Appendix Table 1.1 – Student Outcomes Pre- and Post-Policy for the Full Sample, Reduced Sample and Survey Sample

	Full Sample	**Reduced Sample**	**Survey Sample**
	Mean (sd)	Mean (sd)	Mean (sd)
	(1)	(3)	(4)
2012-2016 Overall Baccalaureate Score	6.118	6.176	5.990
	(2.348)	(2.288)	(2.283)
2017-2019 Overall Baccalaureate Score	6.832	6.846	6.621
	(2.153)	(2.132)	(2.165)
2012-2016 Romanian Score	6.082	6.060	5.963
	(2.331)	(2.293)	(2.278)
2017-2019 Romanian score	6.739	6.713	6.540
	(2.137)	(2.153)	(2.135)
2012-2016 Pass rate	0.548	0.556	0.519
	(0.498)	(0.497)	(0.500)
2017-2019 Pass rate	0.657	0.660	0.618
	(0.474)	(0.474)	(0.486)
Observations	1,260,671	431,940	248,899

Notes: The table displays mean student outcomes, taken for all student outcome data, for overall Baccalaureate score, written Romanian score and pass rates in our: full sample; reduced (working) sample; and survey sample. Mean outcomes are split into pre- (2012-2016) and post- (2017-2019) student outcomes with standard deviations displayed in parentheses.

Appendix Table 1.2 – Proportion of Schools with New and Legacy Managers (Conditional on Competition)/Exam Selected Managers vs No Exam Selected Managers

	Legacy		New		Exam		No Exam	
	Number	Proportion	Number	Proportion	Number	Proportion	Number	Proportion
Full Sample								
Schools	-	-	-	-	1,425	87.02%	213	12.98%
Students	-	-	-	-	1,098,079	87.08%	162,572	12.92%
Reduced Sample								
Schools	329	71.06%	134	28.94%	463	90.78%	47	9.22%
Students	284,673	71.52%	113,340	28.48%	398,013	92.15%	33,927	7.85%
Schools (2016)	281	75.95%	89	24.05%	370	79.91%	-	-
Students (2016)	245,913	75.23%	80,979	24.77%	326,892	82.13%	-	-
Schools (2017)	48	51.61%	45	48.39%	93	20.09%	-	-
Students (2017)	38,760	54.50%	32,361	45.50%	71,121	17.87%	-	-
Survey Sample								
Schools	206	74.37%	71	25.63%	277	90.52%	29	9.48%
Students	161,855	75.65%	52,111	24.35%	213,966	85.96%	34,933	14.04%
Schools (2016)	182	86.67%	28	13.33%	210	75.81%	-	-
Students (2016)	148,832	85.66%	24,909	14.34%	173,741	81.20%	-	-
Schools (2017)	24	35.82%	43	64.18%	67	24.19%	-	-
Students (2017)	13,023	32.38%	27,202	67.62%	40,225	18.80%	-	-

Notes: The table displays raw number and percentages of legacy and new managers in the overall reduced (working) sample and survey sample; these are further broken down into groups who took part in the 2016 wave and 2017 wave of the competition. In addition, we provide the raw number and percentage of competitively selected (exam) and non-competitively selected (no exam) principals for the full sample, reduced (working) sample, and survey sample; these are again presented separately for the two waves of the policy (2016/2017). Note that in separate waves of the policy, we do not present non-competitively (no exam) selected principals, since due to the definition of principals selected as part of the policy wave, all principals were competitively selected.

Appendix Table 1.3. Bacon-Goodman Diff-in-Diff estimate and decomposition of effects. Strongly balanced panel of schools).

	(1) Exam Policy	(2) New Manager Policy
Panel A: Overall Baccalaureate Score		
DD Exam	0.026	
	(0.048)	
DD New manager		0.029
		(0.035)
Observations	2,936	2,664
Panel B: Pass		
DD Exam	0.005	
	(0.012)	
DD New manager		0.006
		(0.007)
Observations	2,936	2,664
Panel C: Romanian Exam Score		
DD Exam	-0.012	
	(0.073)	
DD New Manager		0.011
		(0.034)
Observations	2,936	2,664
W timing groups	0.431	0.040
W never vs 2016	0.472	0.720
W never vs 2017	0.070	0.205
W within	0.027	0.034
School FE	Yes	Yes
Year FE	Yes	Yes
County-specific trends	Yes	Yes
Controls	Yes	Yes

Note: The table displays the Goodman-Bacon (2021) decomposition of the effects into effects from pairwise conditional mean differences (between the early and late implementing groups, and between the implementers and never-implementers). We use a restricted sample of schools which form a strongly balanced panel (i.e., retaining only schools which have data for the entire time period 2012-2019), which is a requirement for implementing the Goodman-Bacon decomposition. Column (1) presents estimates of T1 (exam) and column 2 estimates of T2 (new manager). All estimations include school and year fixed effects, country specific linear trends and controls (the share of theoretic and technical track students), and regressions are weighted by the number of students in 2019.

Appendix Table 1.4. Full Sample and Reduced Sample

	Full Sample		Reduced Sample	
Overall Baccalaureate Score	(1)	(2)	(3)	(4)
Exam * Policy	0.044	-0.019	0.070	0.077*
	(0.040)	(0.039)	(0.041)	(0.040)
Constant	5.027***	5.081***	4.989***	4.953***
	(0.102)	(0.118)	(0.085)	(0.083)
Observations	1,247,350	1,247,350	428,061	428,061
R-squared	0.505	0.507	0.481	0.484
Pass Rate				
Exam * Policy	0.005	-0.006	0.013	0.014
	(0.008)	(0.008)	(0.008)	(0.009)
Constant	0.395***	0.399***	0.362***	0.354***
	(0.019)	(0.021)	(0.023)	(0.023)
Observations	1,260,651	1,260,651	431,940	431,940
R-squared	0.389	0.391	0.369	0.371
Written Romanian Score				
Exam * Policy	0.067	0.000	0.033	0.027
	(0.045)	(0.034)	(0.069)	(0.058)
Constant	4.552***	4.576***	4.487***	4.364***
	(0.082)	(0.101)	(0.093)	(0.087)
Observations	1,252,704	1,252,704	429,674	429,674
R-squared	0.421	0.423	0.415	0.417
Year FE	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
County Specific Trends	No	Yes	No	Yes

Notes: The table displays OLS estimates from difference-in-difference specifications of the effect of T1 (the meritocratic selection policy) on student outcomes. Panel A displays results for the overall Baccalaureate Score, Panel B for the passing rate, and Panel C for the standard Romanian Written Exam scores. We present estimates for the full sample (all Romanian students) in columns 1 and 2, as well as our reduced (working) sample in columns 3 and 4. All specifications include year and school fixed effects, controls (theoretical and technical track dummies, and full-time student dummy) and county specific trends are included in columns 2 and 4. County-clustered standard errors in parentheses for 19 clusters. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 1.5 – Exam vs No Exam; Survey Sample

	Survey Sample	
Overall Baccalaureate Score	(1)	(2)
Exam * Policy	0.040	0.022
	(0.048)	(0.060)
Constant	5.101***	5.116***
	(0.179)	(0.211)
Observations	246,178	246,178
R-squared	0.471	0.475
Pass Rate		
Exam * Policy	0.001	-0.006
	(0.009)	(0.012)
Constant	0.387***	0.376***
	(0.035)	(0.040)
Observations	248,897	248,897
R-squared	0.363	0.366
Written Romanian Score		
Exam * Policy	0.028	-0.015
	(0.050)	(0.060)
Constant	4.672***	4.588***
	(0.161)	(0.222)
Observations	247,375	247,375
R-squared	0.395	0.397
Year FE	Yes	Yes
School FE	Yes	Yes
Controls	Yes	Yes
County Specific Trends	No	Yes

Notes: The table displays OLS estimates from difference-in-difference specifications of the effect of T1 (the meritocratic selection policy) on student outcomes. Panel A displays results for the overall Baccalaureate Score, Panel B for the passing rate, and Panel C for the standard Romanian Written Exam scores. We present the estimates for overall survey sample in columns 1 and 2. Both specifications include year and school fixed effects and columns 2, includes county-specific trends. County-clustered standard errors in parentheses for 19 clusters. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 1.6 – New vs Legacy Manager; Reduced Sample; Exam Selected Principals for 2016 and 2017

	Reduced Sample		Exam 2016		Exam 2017	
Overall Baccalaureate Score	(1)	(2)	(3)	(4)	(5)	(6)
New Manager * Policy	0.058	0.053	0.074	0.062	0.070	0.142
	(0.043)	(0.036)	(0.051)	(0.043)	(0.109)	(0.104)
Constant	4.914***	4.916***	4.903***	4.874***	4.939***	6.509***
	(0.114)	(0.114)	(0.133)	(0.132)	(0.294)	(0.335)
Observations	394,409	394,409	324,102	324,102	70,307	70,307
R-squared	0.483	0.486	0.484	0.487	0.471	0.477
Pass Rate						
New Manager * Policy	0.010	0.010	0.011	0.010	0.015	0.027
	(0.007)	(0.006)	(0.009)	(0.009)	(0.018)	(0.018)
Constant	0.354***	0.345***	0.345***	0.331***	0.388***	0.546***
	(0.027)	(0.026)	(0.032)	(0.031)	(0.059)	(0.065)
Observations	398,013	398,013	326,892	326,892	71,121	71,121
R-squared	0.372	0.375	0.376	0.379	0.350	0.355
Written Romanian Score						
New Manager * Policy	0.047	0.042	0.054	0.053	0.137	0.151
	(0.036)	(0.029)	(0.044)	(0.033)	(0.102)	(0.093)
Constant	4.436***	4.344***	4.454***	4.344***	4.361***	5.531***
	(0.115)	(0.113)	(0.131)	(0.129)	(0.305)	(0.359)
Observations	395,925	395,925	325,330	325,330	70,595	70,595
R-squared	0.413	0.415	0.409	0.411	0.422	0.426
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
County Specific Trends	No	Yes	No	Yes	No	Yes

Notes: The table displays OLS estimates from difference-in-difference specifications of the effect of T2 (the impact of new managers) on student outcomes. Panel A displays results for the overall Baccalaureate Score, Panel B for the passing rate, and Panel C for the standard Romanian Written Exam scores. We present the estimates for the entire reduced (working) sample (columns 1 and 2), as well as separately for the two waves of the policy (columns 3-4 for the 2016 selection and columns 5-6 for the 2017 selection). All specifications include year and school fixed effects and columns 2, 4 and 6 include county-specific trends. County-clustered standard errors in parentheses for 19 clusters. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 1.7 – New vs Legacy Principals; Reduced Sample; Percentiles Calculated from School Average Final Score 2012-2016

	< 25th Percentile		25th – 75th Percentile		> 75th Percentile	
Overall Baccalaureate Score	(1)	(2)	(3)	(4)	(5)	(6)
New Manager *	0.188	0.263*	0.094	0.044	-0.000	-0.024
Policy	(0.121)	(0.134)	(0.068)	(0.067)	(0.054)	(0.047)
Constant	3.783***	4.380***	4.131***	4.160***	4.936***	4.820***
	(0.337)	(0.374)	(0.171)	(0.176)	(0.287)	(0.294)
Observations	46,953	46,953	190,848	190,848	156,339	156,339
R-squared	0.143	0.151	0.210	0.217	0.187	0.192
Pass Rate						
New Manager *	0.024*	0.039**	0.018	0.010	0.003	0.001
Policy	(0.012)	(0.014)	(0.012)	(0.012)	(0.010)	(0.009)
Constant	0.202***	0.204***	0.181***	0.156***	0.247***	0.250***
	(0.052)	(0.055)	(0.040)	(0.040)	(0.054)	(0.055)
Observations	47,969	47,969	193,099	193,099	156,661	156,661
R-squared	0.060	0.063	0.151	0.156	0.087	0.090
Written Romanian Score						
New Manager *	0.253**	0.238*	0.057	0.049	-0.006	-0.061*
Policy	(0.114)	(0.130)	(0.077)	(0.077)	(0.062)	(0.031)
Constant	3.317***	3.582***	3.811***	3.678***	4.026***	3.894***
	(0.237)	(0.259)	(0.143)	(0.145)	(0.428)	(0.431)
Observations	47,265	47,265	191,878	191,878	156,504	156,504
R-squared	0.160	0.164	0.197	0.200	0.201	0.207
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
County Specific Trends	No	Yes	No	Yes	No	Yes

Notes: The table displays OLS estimates from difference-in-difference specifications of the effect of T2 (the impact of new managers) on student outcomes in our reduced (working) sample. Panel A displays results for the overall Baccalaureate Score, Panel B for the passing rate, and Panel C for the standard Romanian Written Exam scores. We present the estimates for schools in the <25th percentile (columns 1 and 2), 25th-75th percentile (columns 3 and 4) and >75th percentile (columns 5 and 6); percentile groups are calculated based on pre-policy (before 2016) overall Baccalaureate scores. All specifications include year and school fixed effects and columns 2, 4 and 6 include county-specific trends. County-clustered standard errors in parentheses for 19 clusters. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 1.8 – Exam vs No Exam; Reduced Sample; Percentiles Calculated from School Average Final Score 2012-2016

	< 25th Percentile		25th – 75th Percentile		> 75th Percentile	
Overall Baccalaureate Score	(1)	(2)	(3)	(4)	(5)	(6)
Exam * Policy	0.166	0.141	-0.001	-0.022	-0.016	0.016
	(0.181)	(0.156)	(0.077)	(0.073)	(0.077)	(0.079)
Constant	3.776***	4.407***	4.132***	4.170***	4.936***	4.815***
	(0.339)	(0.367)	(0.172)	(0.172)	(0.285)	(0.290)
Observations	46,953	46,953	190,848	190,848	156,339	156,339
R-squared	0.143	0.151	0.210	0.217	0.187	0.192
Pass Rate						
Exam * Policy	0.017	0.013	-0.001	-0.001	0.004	0.007
	(0.023)	(0.022)	(0.022)	(0.022)	(0.016)	(0.016)
Constant	0.201***	0.208***	0.181***	0.158***	0.247***	0.250***
	(0.053)	(0.055)	(0.040)	(0.039)	(0.053)	(0.054)
Observations	47,969	47,969	193,099	193,099	156,661	156,661
R-squared	0.060	0.063	0.151	0.156	0.087	0.090
Written Romanian Score						
Exam * Policy	0.055	0.039	-0.116	-0.121	-0.057	-0.020
	(0.182)	(0.152)	(0.136)	(0.126)	(0.114)	(0.105)
Constant	3.311***	3.611***	3.811***	3.692***	4.025***	3.884***
	(0.231)	(0.249)	(0.143)	(0.144)	(0.426)	(0.428)
Observations	47,265	47,265	191,878	191,878	156,504	156,504
R-squared	0.159	0.164	0.197	0.200	0.201	0.207
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
County Specific Trends	No	Yes	No	Yes	No	Yes

Notes: The table displays OLS estimates from difference-in-difference specifications of the effect of T1 (the meritocratic selection policy) on student outcomes in our reduced (working) sample. Panel A displays results for the overall Baccalaureate Score, Panel B for the passing rate, and Panel C for the standard Romanian Written Exam scores. We present the estimates for schools in the <25th percentile (columns 1 and 2), 25th-75th percentile (columns 3 and 4) and >75th percentile (columns 5 and 6); percentile groups are calculated based on pre-policy (before 2016) overall Baccalaureate scores. All specifications include year and school fixed effects and columns 2, 4 and 6 include county-specific trends. County-clustered standard errors in parentheses for 19 clusters. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 1.9 – Exam vs No Exam; ATT and dynamic effects; Robustness control group not yet treated

	Sample	Group 2016	Group 2017	ATT 2017	ATT 2018	ATT 2019
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Overall Baccalaureate Exam Score					
ATT (T=principal exam)	0.014	0.029	-0.095	-0.068	0.035	0.066
	(0.098)	(0.101)	(0.144)	(0.056)	(0.092)	(0.199)
P-Value	[0.884]	[0.771]	[0.510]	[0.225]	[0.703]	[0.740]
Observations	428,061					
Chi2 H_0: parallel pre-trends	71.354					
P-Value H_0: parallel pre-trends	[0.000***]					
	Panel B: Pass rate					
ATT (T=principal exam)	0.001	0.004	-0.020	-0.002	-0.005	0.010
	(0.016)	(0.015)	(0.028)	(0.009)	(0.018)	(0.034)
P-Value	[0.949]	[0.795]	[0.479]	[0.860]	[0.780]	[0.764]
Observations	431,940					
Chi2 H_0: parallel pre-trends	13.127					
P-Value H_0: parallel pre-trends	[0.157]					
	Panel C: Romanian written exam score					
ATT (T=principal exam)	0.037	0.045	-0.020	-0.140	0.073	0.157
	(0.115)	(0.117)	(0.178)	(0.094)	(0.092)	(0.223)
P-Value	[0.748]	[0.700]	[0.908]	[0.134]	[0.423]	[0.480]
Observations	429,674					
Chi2 H_0: parallel pre-trends	6.638					
P-Value H_0: parallel pre-trends	[0.675]					

Notes: The table displays ATT estimates from difference-in-difference specifications of the effect of T1 (the meritocratic selection policy) on student outcomes, using the double-robust inverse probability weighting estimator from Calloway and Sant'Anna (2021). Panel A displays results for the overall Baccalaureate Score, Panel B for the passing rate, and Panel C for the standard Romanian Written Exam scores. We present estimates for the entire working sample in column (1), ATTs by treatment groups in columns (2) and (3), and ATTs by period in columns (4)-(6). All specifications display the chi2 and p-values from the tests for the conditional parallel trends assumption. County-clustered standard errors in parentheses for 19 clusters. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 1.10 – New vs Legacy Manager; Placebo Year Robustness Test (Policy = 2015/2016); Reduced Sample; Exam Selected Principals for 2016 and 2017

	Reduced Sample		Exam 2016		Exam 2017	
Overall Baccalaureate Score	(1)	(2)	(3)	(4)	(5)	(6)
New Manager *	0.016	0.013	0.006	0.003	0.103	0.097
Policy	(0.034)	(0.034)	(0.037)	(0.037)	(0.117)	(0.112)
New Manager *	0.053	0.051	0.083*	0.072*	-0.039	0.058
Policy t-1	(0.039)	(0.035)	(0.046)	(0.039)	(0.100)	(0.105)
Constant	4.903***	4.903***	4.882***	4.852***	4.939***	6.500***
	(0.117)	(0.118)	(0.136)	(0.136)	(0.294)	(0.338)
Observations	394,409	394,409	324,102	324,102	70,307	70,307
R-squared	0.483	0.486	0.484	0.487	0.471	0.477
Pass Rate						
New Manager *	0.008	0.008	0.004	0.004	0.023	0.019
Policy	(0.006)	(0.006)	(0.007)	(0.007)	(0.024)	(0.023)
New Manager *	0.003	0.004	0.009	0.007	-0.009	0.010
Policy t-1	(0.009)	(0.009)	(0.008)	(0.008)	(0.025)	(0.026)
Constant	0.353***	0.344***	0.343***	0.328***	0.388***	0.544***
	(0.027)	(0.027)	(0.032)	(0.032)	(0.059)	(0.066)
Observations	398,013	398,013	326,892	326,892	71,121	71,121
R-squared	0.372	0.375	0.376	0.379	0.350	0.355
Written Romanian Score						
New Manager *	-0.000	0.000	0.030	0.032	0.108	0.096
Policy	(0.040)	(0.037)	(0.045)	(0.039)	(0.124)	(0.121)
New Manager *	0.060	0.054	0.029	0.026	0.034	0.069
Policy t-1	(0.036)	(0.034)	(0.032)	(0.032)	(0.101)	(0.106)
Constant	4.424***	4.330***	4.447***	4.336***	4.361***	5.520***
	(0.118)	(0.116)	(0.135)	(0.133)	(0.305)	(0.361)
Observations	395,925	395,925	325,330	325,330	70,595	70,595
R-squared	0.413	0.415	0.409	0.411	0.422	0.426
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
County Specific Trends	No	Yes	No	Yes	No	Yes

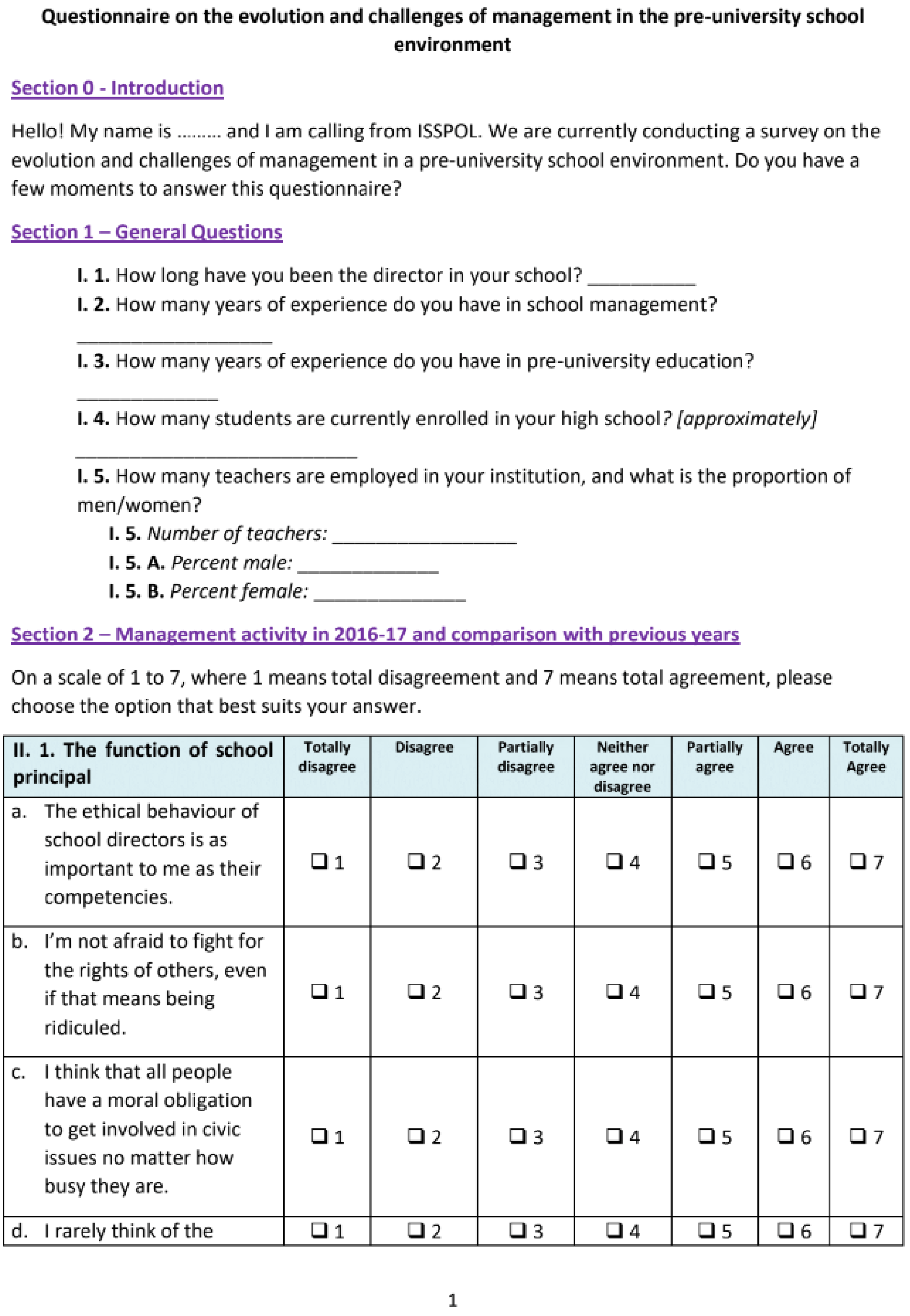
Notes: The table displays OLS estimates from difference-in-difference specifications of the effect of T2 (the impact of new managers) on student outcomes. Placebo treatment variables are included, enacting the policy one year before it took place. Panel A displays results for the overall Baccalaureate Score, Panel B for the passing rate, and Panel C for the standard Romanian Written Exam scores. We present the estimates for reduced (working) sample (columns 1 and 2), 2016 exam selected principals (columns 3 and 4) and 2017 exam selected principals (columns 5 and 6). All specifications include year and school fixed effects and columns 2, 4 and 6 include county-specific trends. County-clustered standard errors in parentheses for 19 clusters. *** p<0.01, ** p<0.05, * p<0.1

Appendix Table 1.11 – New vs Legacy Principals; Wild Bootstrap; Reduced Sample; Percentiles Calculated from School Average Final Score 2012-2015

	Reduced Sample		Exam 2016		Exam 2017	
Overall Baccalaureate Score	(1)	(2)	(3)	(4)	(5)	(6)
New Manager *	0.058	0.053	0.074	0.062	0.070	0.142
Policy	[1.356]	[1.484]	[1.462]	[1.443]	[0.639]	[1.366]
	(0.201)	(0.157)	(0.171)	(0.168)	(0.600)	(0.217)
Constant	4.914***	4.916***	4.903***	4.874***	4.939***	6.509***
	(0.114)	(0.114)	(0.133)	(0.132)	(0.294)	(0.335)
Observations	394,409	394,409	324,102	324,102	70,307	70,307
R-squared	0.483	0.486	0.484	0.487	0.471	0.477
Pass Rate						
New Manager *	0.010	0.010	0.011	0.010	0.015	0.027
Policy	[1.418]	[1.707]	[1.238]	[1.115]	[0.836]	[1.494]
	(0.178)	(0.110)	(0.236)	(0.309)	(0.456)	(0.173)
Constant	0.354***	0.345***	0.345***	0.331***	0.388***	0.546***
	(0.027)	(0.026)	(0.032)	(0.031)	(0.059)	(0.065)
Observations	398,013	398,013	326,892	326,892	71,121	71,121
R-squared	0.372	0.375	0.376	0.379	0.350	0.355
Written Romanian Score						
New Manager *	0.047	0.042	0.054	0.053	0.137	0.151
Policy	[1.309]	[1.464]	[1.214]	[1.596]	[1.336]	[1.625]
	(0.220)	(0.166)	(0.224)	(0.140)	(0.216)	(0.114)
Constant	4.436***	4.344***	4.454***	4.344***	4.361***	5.531***
	(0.115)	(0.113)	(0.131)	(0.129)	(0.305)	(0.359)
Observations	395,925	395,925	325,330	325,330	70,595	70,595
R-squared	0.413	0.415	0.409	0.411	0.422	0.426
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
County Specific Trends	No	Yes	No	Yes	No	Yes

Notes: The table displays OLS estimates from difference-in-difference specifications of the effect of T2 (the impact of new managers) on student outcomes in our reduced (working) sample. Panel A displays results for the overall Baccalaureate Score, Panel B for the passing rate, and Panel C for the standard Romanian Written Exam scores. We present the estimates for schools in the <25th percentile (columns 1 and 2), 25th-75th percentile (columns 3 and 4) and >75th percentile (columns 5 and 6); percentile groups are calculated based on pre-policy (before 2016) overall Baccalaureate scores. All specifications include year and school fixed effects and columns 2, 4 and 6 include county-specific trends. Wild bootstrap t-value is shown in square parentheses and wild bootstrap p-value is shown in round parentheses. *** p<0.01, ** p<0.05, * p<0.1

Full Survey Instrument conducted in November 2017 in collaboration with ISSPOL



Questionnaire on the evolution and challenges of management in the pre-university school environment

Section 0 - Introduction

Hello! My name is ……… and I am calling from ISSPOL. We are currently conducting a survey on the evolution and challenges of management in a pre-university school environment. Do you have a few moments to answer this questionnaire?

Section 1 – General Questions

I. 1. How long have you been the director in your school? ___________

I. 2. How many years of experience do you have in school management? ____________________

I. 3. How many years of experience do you have in pre-university education? ______________

I. 4. How many students are currently enrolled in your high school*? [approximately]* ____________________________

I. 5. How many teachers are employed in your institution, and what is the proportion of men/women?

I. 5. *Number of teachers:* __________________

I. 5. A. *Percent male:* ______________

I. 5. B. *Percent female:* _______________

Section 2 – Management activity in 2016-17 and comparison with previous years

On a scale of 1 to 7, where 1 means total disagreement and 7 means total agreement, please choose the option that best suits your answer.

II. 1. The function of school principal	Totally disagree	Disagree	Partially disagree	Neither agree nor disagree	Partially agree	Agree	Totally Agree
a. The ethical behaviour of school directors is as important to me as their competencies.	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	❑ 7
b. I'm not afraid to fight for the rights of others, even if that means being ridiculed.	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	❑ 7
c. I think that all people have a moral obligation to get involved in civic issues no matter how busy they are.	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	❑ 7
d. I rarely think of the	❑ 1	❑ 2	❑ 3	❑ 4	❑ 5	❑ 6	❑ 7

1

welfare of people that I do not know personally.							
e. It is more important for me to succeed financially than to do good deeds.	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
f. I think that I am an effective problem solver.	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
g. I think that I am a person who can sometimes be rude to others.	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
h. I think that I am a person who listens to and helps others.	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
i. It is easy for me to work with the Deputy Director and other colleagues.	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
II. 2. Working Relationships	**Totally disagree**	**Disagree**	**Partially disagree**	**Neither agree nor disagree**	**Partially agree**	**Agree**	**Totally Agree**
a. Teachers in my high school trust each other.	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
b. Teachers in my high school trust the decisions of the directors.	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
c. If I think about how the teachers in my high school solve challenges, I believe that we have a common set of values.	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
d. I trust colleagues enough to delegate assignments.	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
e. I think that a man is more capable than a woman at running a high school.	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7

II. 3. How do you feel that the following areas of your high school performance have evolved in 2016-17 when compared to pervious years.

	Better	Worse	The Same	N/A
a. Adapting learning activities to student needs.	□ 1	□ 2	□ 3	□ 4
b. Tracking and evaluating the performance of teachers and school activities.	□ 1	□ 2	□ 3	□ 4
c. Establishing, communicating and tracking school objectives related to student outcomes.	□ 1	□ 2	□ 3	□ 4
d. Absence among teachers.	□ 1	□ 2	□ 3	□ 4
e. Hiring and retaining high-skilled staff members.	□ 1	□ 2	□ 3	□ 4
f. Use of staff sanctions (e.g. reprisals, weak evaluations, administrative sanctions, layoffs).	□ 1	□ 2	□ 3	□ 4
g. Motivating your staff (e.g.) through praise, good ratings, bonuses, promotions).	□ 1	□ 2	□ 3	□ 4
h. Mobilization and initivative spirit	□ 1	□ 2	□ 3	□ 4
i. Managing authority (ease of implementation, respect, staff discipline over decisions made by executives).	□ 1	□ 2	□ 3	□ 4
j. Absence among students.	□ 1	□ 2	□ 3	□ 4
k. School dropout rate.	□ 1	□ 2	□ 3	□ 4
l. Gradutation rate of the 12th grade.	□ 1	□ 2	□ 3	□ 4
m. Registration rate for the baccalaureate.	□ 1	□ 2	□ 3	□ 4

II. 3_j_1. What is the absentee rate in your institution? (in the school year 2016-17)
____________________%

II. 3_k_1. What is the dropout rate in your institution? (in the school year 2016-17)
____________________%

II. 3_m_1. What is the baccalaureate enrollment rate in your institution? (in the school year 2016-17)

_____________________%

II. 4. Does your high school have any initiatives in place to improve the school performance of students from disadvantages backgrounds, or high school drop-outs? (e.g. remedial hours, after school hot meal)

YES in 2016/17. Examples ... Pass to II.4.1	□ 1
Yes before 2016	□ 2
No	□ 3

II. 5. (Original: If your high school has applied and has been approved funds in the ROSE project, funded by the World Bank, what amount have you received?)

(in link) Did your high school apply to the ROSE program, financed by the World Bank, and have funds approved?

Did not apply	□ 1
Was not approved	□ 2
Yes. Go to II.5.1	□ 3

II. 5. 1. How much funding did you receive through the ROSE project? ________________________

ROSE is a grant project for eligible high schools (those with low baccalaureate promotion rates and many underprivileged students) aimed at increasing graduate exam outcomes and discouraging school dropouts. The funds are around 70,000-150,000 Euros for each beneficiary school. More details can be found at: https://www.edu.ro/271-de-licee-beneficiaz%C4%83-de-granturi-%C3%AEn-cadrul-proiectului-privind-%C3%AEnv%C4%83%C8%9B%C4%83m%C3%A2ntul-secundarromania

Section 3:

Running the manager contest – Tick all the options that apply to you.

III. 1. Have you run in the competition to become director?

A. Yes, in 2016	□ 1
B. Yes, in 2017	□ 2
C. No, I was seconded to the interest of education	□ 3

III. 2. If you answered 1.A or 1.B: How did you decide to join the competition?

A. Own initiative	□ 1
B. At the request of my colleagues	□ 2
C. To prevent someone from another school gaining the post in my high school	□ 3
D. On recommendation of the inspectors/other people in the education system	□ 4
E. Other reasons (please provide short description)____________________	□ 5

III. 3. If you answered 1.B or 1.C: Why did you decide not to register for the competition in 2016?

A. Too little time to prepare	□ 1
B. The literature was completely new	□ 2
C. I was advised not to sign up	□ 3
I felt that I would be disfavoured/discriminated against during the evaluation	□ 4
E. Other reasons (please provide short description) _______________	□ 5

III. 4. Consider the session running October-November 2016. Winning this competition for the post of director in Romania was based upon:

Based on merit (theoretical knowledge or capabilities)	□ 1
Partially based on merit, partially on political knowledge/connections	□ 2
Based on political knowledge/connections	□ 3

III. 5. Consider the session running July-August 2017. Winning this competition for the post of director in Romania was based upon:

Based on merit (theoretical knowledge or capabilities)	□ 1
Partially based on merit, partially on political knowledge/connections	□ 2
Based on political knowledge/connections	□ 3

III. 6. Do you know people who retired or did not participate at all (in the 2016 or 2017 competitions) because they felt that they would be disfavoured/discriminated against?

Yes	□ 1
No	□ 2

III. 7. Select the one that best suits your experience:

	Totally disagree	Disagree	Partially disagree	Neither agree nor disagree	Partially agree	Agree	Totally Agree
a. The competition tested the skills expected from a school manager.	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
b. The competition attracted candidates with high motivation and abilities and increased the quality of management in schools.	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7
c. A candidate with previous experience as a manager, or worked in that school, has a competitive advantage	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7

over other candidates.							
d. It would have been necessary/helpful to take part in a training program in order to prepare for this exam.	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7

III. 8. Section 3, Question 1, if YES: After winning the post by contest/appointment, what has changed in your high school? (You can refer to any aspect, from efficiency in administration, to ease of money management, to attracting funds and funding, to the attitude of teachers or parents, to student behaviour and school performance.)

____

III. 9. What would you like to change at the administrative level in your high school over the next four years?

____

Demografice:

D1. Name: ____________

D2. Age:_____________

D3. Gender:______________

D4. School/High School: ______________

D5. Town: ________________

D6. County:____________________

D7. Email address: _________________

D8. Telephone Number: _______________

Operator Code: _________________

Chapter 2

When the Going Gets Tough: The Role of Confidence in Complex Real Effort Tasks

2.1 Introduction

Real-effort tasks are frequently used in experimental economics (see Roth, 1987 and Charness *et al.*, 2018 for a useful review). Real-effort tasks include a working task which requires subjects to exert a form of costly effort; as opposed to their chosen effort counterparts, where subjects do not directly exert effort. Typically, real-effort tasks used in the literature are simple tasks, which have a number of benefits. As an example, these tasks may feature answering arithmetic questions (Niederle and Vesterlund, 2007) or counting the number of 0's in a grid (Abeler *et al.*, 2011). When using a simple real-effort task subjects require shorter periods of time to become familiar with the task (Benndorf *et al.*, 2019) as they are simpler to both explain and understand. However, simple real-effort tasks rarely replicate the types of tasks which subjects face in the real world. Moore and Healy (2008) provide evidence that subjects who face relatively more complex trivia questions, have stronger beliefs about their performance. As a result, the beliefs a subject has about their own performance are likely to differ when they face tasks of different complexities; simple tasks may bias these beliefs about performance. Despite this, little is known about the impact of increasing the complexity of the standard real-effort tasks used in experimental economic literature. In this domain, increases in task complexity occur as subjects are presented with more information (Regier *et al.*, 2014), and are subsequently required to complete more steps in order to achieve the answer.

Mitigating biases in subject confidence when facing real-effort tasks is important, since they may impact behaviours which are frequently studied in experimental economics. One domain where this is particularly important, is in studies which require subjects to pick between incentive schemes. For example, Niederle and Vesterlund (2007) ask subjects to pick between a competitive tournament and non-competitive piece rate, whilst Banuri and Keefer (2016) face subjects with a choice of piece rate or flat wage incentive schemes. Here, the decision between incentive schemes often depends on subject confidence. Importantly, the confidence a subject has about their performance in a task is inherently linked to beliefs about performance, since overconfident subjects believe they will perform better than they do in reality.[22] Therefore, there may be unstudied and important interactions between task complexity and confidence which impact the decisions a subject makes, and cause them to make sub-optimal choices based on inaccurate beliefs.

Our paper addresses this gap in the literature by testing the accuracy of predictions when subjects are faced with tasks of different complexities. We undertake an online lab experiment, in which we face subjects with real-effort tasks of different complexities. Before each round of the task, we ask subjects to predict their upcoming performance. We use this prediction to calculate a confidence variable which reflects the distance between a subjects' prediction and actual performance (i.e. confidence = prediction – performance). Our study examines the impact of task complexity on this confidence measure, and consists of a baseline in which subjects face a simple addition task similar to that in Nierdele and Vesterlund (2007); in our experiment subjects must add together three 2-digit numbers on screen. Our remaining treatments implement real-effort tasks which are more difficult or more complex. We define a complex task as one in which a subject has to solve more individual components to get to the final answer; for example first having to clean a dataset before analysing it would be seen as more complex than being provided with a clean dataset to analyse in our setting. Note therefore that tasks can be more difficult, without

[22] Underconfident subjects make lower predictions than their true performance.

necessarily being more complex, if they still involve the same number of components even where those components are more taxing on the subject.

In our first treatment, subjects are tasked with a simple subtraction task in which they must subtract three numbers from each other; since there are no additional components to this task, it represents an increase in difficulty but not in complexity. In our second treatment, we increase task complexity by introducing a grid search component where subjects must first identify the numbers which they then add together. Our third treatment is similar to our second, but here subjects must first identify numbers from a grid, before subtracting the subsequent numbers from each other. Initially, we implement complexity and fix the length of each round, however this makes similar levels of output more difficult to accomplish in complex tasks. As a result, the variation in output differs significantly between our simple and complex treatments. To address this issue, we implement complex treatments in which we increase the length of the rounds (6 minutes instead of 3 minutes).

Our results indicate that subjects who face complex tasks are significantly better at predicting their performance. These results persist in our longer complex treatments (those which last for 6 minutes), when the number of questions attempted is similar between subjects who face the simple and complex task. We provide further evidence that the impact remains significant as subjects progress through more rounds of the task. However, since subjects in our simple task baseline become better at predicting their performance in later rounds, the size of the effect is reduced as time passes. Therefore, we provide evidence that subjects are typically better at predicting performance when faced with complex tasks than when faced with simple tasks.

The findings of this study have important implications for future research within experimental economics. In particular, we shed light on how the choice of real-effort task may independently impact the predictions which subjects make about their performance in the task. This may have ramifications for research which requires subjects to choose between payment incentive schemes, where subject confidence plays an important role in decision making. Since subjects are better at predicting their performance in complex tasks, they may make more accurate decisions when faced with different incentive scheme choices. Future research may shed light on how further increasing the complexity of real-effort tasks impacts the accuracy of subject predictions about performance.

2.2 Literature Review

This study connects with two strands of intertwined literature. Initially, we draw links to past experimental literature which uses real effort tasks to better understand costly effort provision. Secondly, this paper discusses research on the interactions between confidence and effort provision.

2.2.1 Real Effort Tasks

Behavioural and experimental economic experiments typically utilise two methodologies when examining effort; one can employ either stated effort or real effort. Stated effort experiments face subjects with a decision which has pre-determined effort functions mapping choices to outcomes (Charness *et al.*, 2018); in addition, subjects are provided with a full set of information about the benefits and costs of the decision (Erkal *et al.*, 2018). In contrast, real effort experiments require subjects to take part in a working task, meaning that they must exert effort in a real sense (Carpenter and Huet-Vaughn, 2019). Since payments are often linked to output, by exerting more effort subjects are able to increase their performative outcome measure, in turn leading to a higher payment.

A bulk of experimental economics literature utilises real effort tasks, which take many forms. Real effort tasks have included physical tasks such as cracking walnuts (Fahr and Irlenbusch, 2000) and folding paper and stuffing envelopes (Konow, 2000). However more recently, due to the move toward computer based experiments, real effort tasks have included: solving mazes (Gneezy *et al.*, 2003); counting zeros in grids of various sizes (Abeler *et al.*, 2011); and adding two-digit numbers together (Niederle and Vesterlund, 2007).

Real effort tasks are shown to have multiple benefits when compared with chosen effort tasks (Gill and Prowse, 2012; Gill and Prowse, 2019; Charness *et al.*, 2018). Firstly, experiments which include a real effort component represent greater external validity (Bruggen and Strobel, 2007). That is, effort in these tasks more closely represents motivations for exerting effort which subjects face in the real world (Gill and Prowse, 2019). For example, Rosaz and Villeval (2012) introduce a task in which subjects must count the occurrence of words in a paragraph; this task carries over real-world features including concentration during a repetitive task.

In addition, real effort tasks allow experimenters to examine the concept of ability. In a real effort task, the ability of an individual to perform some task may have a strong impact on their performance throughout the experiment. Whereas in a stated effort task, ability of a subject has little to no impact on the choices they make since outcomes are common knowledge and require a simple decision. As a result, studies must control for the natural ability of subjects at the task they are undertaking; for example Kajackaite (2015) compare the output of subjects in the main part of their experiment and in the trial period (or learning stage).

Despite this, real effort tasks may also carry negative effects independent of treatment effects. For example, learning in real-effort tasks may allow subjects to become more efficient at undertaking the task and as a result may interfere with treatment effects (Benndorf *et al.*, 2019). This is of particular importance in specific online experiment platforms, such as MTurk, which have been shown to have non-naïve subjects (Peer *et al.*, 2017); these subjects are likely to have been exposed to similar real effort tasks in previous studies.

Furthermore, it is argued that real effort tasks result in a loss of experimental control when compared to chosen effort tasks (Dutcher *et al.*, 2015). Since the cost of effort varies between subjects, and is simultaneously unknown to the experimenter, quantitative analysis is significantly more difficult when using real effort tasks (Falk and Fehr, 2003). In turn, this loss of control is argued to prevent precise quantitative predictions (Falk and Fehr, 2003). Despite this, real effort tasks allow experimenters to study particular areas of research which were previously out of reach, including: fairness; reciprocity; confidence; and loss aversion (Falk and Fehr, 2003). For example, by using a real effort task Gneezy *et al.* (2003) are able to study performance increases as a result of increasingly competitive environments.

Typically, the real effort tasks used in a lab setting are very simple tasks, allowing experimenters to measure the cost of effort more accurately. Gill and Prowse (2019) highlight the ability of a simple real effort task, such as their "slider task", to provide a graded measure of the cost of effort in relatively few rounds of the task; this allows experimenters to overcome many of the drawbacks which are associated with using real-effort tasks. Simple real effort tasks are also beneficial since they allow subjects to learn the task quickly, preventing the performance of participants improving in later periods as subjects become more efficient at the task (Benndorf *et al.*, 2019).

However, utilising such a simple real effort task directly counteracts many of the benefits which real effort tasks induce since they are far removed from the real-world tasks they are meant to replicate. Since external validity is an important benefit of real effort tasks (Kessler and Vesterlund, 2015), it is vital that external validity is retained when undertaking real effort experiments. Despite this, simple effort tasks reduce external validity since they are comprised of a single component; in contrast, real world subjects face tasks with multiple components. Therefore, the choice of real effort task used in an experiment is of vital importance if one aims to retain external validity of results; specifically multi-component real effort tasks which more closely replicate real world tasks and conditions provide greater ecological validity. This drawback of simple effort tasks is understudied and is the core topic of this paper.

Similar to Lezzi *et al.* (2015) we face subjects with four different types of effort task, however in contrast to this paper we specify that our real effort tasks become increasingly more complex. In this way, our paper is similar to that of Moore and Healy (2008). All of the real-effort tasks which we utilise are based on the number-addition task (Niederle and Vesterlund, 2007), however we include complexity in the form of a grid search in addition to the number-addition task. We class this task as more complex because subjects have to first find the numbers from the grids before adding them together, constituting a task with multiple components in line with our earlier definition of complexity. Our task therefore generates a multiple-component form of the simple number-addition task introduced in Niederle and Vesterlund (2007).

2.2.2 Confidence and Effort

When a subject undertakes a real effort task, a key component of their decision to exert effort relates to their confidence about their ability to perform in the task. Evidence suggests that effort provision is increasing in the beliefs that subjects hold about their own ability at the task (Chen and Schildberg-Horisch, 2019). In a real effort task, the payment which a subject receives often depends on their output. Where payment depends on output, returns to effort are proportional to ability; therefore when a subject is overconfident, their perceived returns to effort are greater since they believe that their ability is inflated (Barron and Gravert, 2021). As a result, overconfident agents are likely to exert greater amounts of effort since they give greater weight to success-contingent payments (de la Rosa, 2011) such as those which feature in real effort tasks.

Whilst Khunen and Tymula (2012) document that people who expect to learn about their rank performance work harder, how a subject responds to feedback about their performance which is not in line with their confidence level is still disputed. On one hand, subjects' overconfidence about performance is also shown to persist even when they are shown feedback which counters their beliefs (Grossman and Owens, 2012). Conversely, some literature suggests that confidence may increase when a subject receives positive feedback (i.e. that which tells them that they are better than their belief) and fall when a subject receives negative feedback (i.e. telling them that they are worse than their belief) (Murad and Starmer, 2021). Similarly, Coutts (2019) suggests that positive feedback is given a stronger weighting than negative feedback. These recent findings might suggest that subjects who receive positive feedback in turn gain a higher level of confidence and therefore exert greater amounts of effort when performing in a real effort task.

Recent evidence has suggested that when eliciting beliefs using a binarized scoring rule, the level of information provided to subjects impacts their incentive to tell the truth (Danz *et al.*, 2022). As more information on the incentives on the outcomes of binary scoring rule lotteries is revealed to subjects, deviations from truth telling occur more frequently (Danz *et al.*, 2022). Whilst this research casts concern over the use of a binary scoring rule for belief elicitation, the scoring rule that we use in our experiment is simple and only the necessary information is provided to subjects.

A core component of our research is the complexity of the real effort task which subjects face. Therefore, a key piece of literature related to our study is Moore and Healy (2008). Importantly Moore and Healy (2008) split overconfidence into two concepts: overplacement and overestimation. Overplacement is taken as a subject who believes that their percentile placement in relative feedback is greater than their true placement (Larrick *et al.*, 2007). Overestimation on the other hand relates only to a subjects' own performance, and is taken as the difference between expected and actual performance (Feld *et al.*, 2017).

Moore and Healy (2008) face subjects with increasingly difficult trivia quizzes, and ask about both their self-confidence and relative confidence related to their performance in the trivia quiz they have undertaken. When faced with more difficult trivia quizzes, subjects tend to overestimate their own performance but underplace themselves; in contrast, when facing easier quizzes, subjects underestimate but overplace (Moore and Healy, 2008).

Our paper adds further evidence to the findings of Moore and Healy (2008). We include forms of real effort task which are more common in experimental economic literature; thus we provide weight to the idea that the complexity of a real effort task is impactful in the same way as the difficulty of a trivia quiz. Our findings, therefore, are more generalisable to economic literature which utilise real effort tasks than those of Moore and Healy (2008). In addition, we provide evidence of the impact of complexity in real effort tasks on confidence in an online subject pool. This is an important feature, since online subject pools are more frequently used since the COVID-19 pandemic.

2.3 Experimental Design

We conduct an online lab experiment using a real effort number adding task. The experiment uses a between-subjects design and varies two main components of the task: task difficulty, and task complexity. Our primary outcome of interest in subject confidence (defined as the difference between subject beliefs and subject performance). We vary the simple number adding task by increasing the difficulty of the task (subtracting numbers), or the complexity of the task (adding a grid search component), or both, yielding a 2-by-2 design. We then add two additional treatments, where we extend the time given per round for the more complex tasks (to account for the greater amount of time needed to perform the complex task), yielding 6 treatments in total.

Across all treatments, subjects are given detailed instructions about the effort task they will face and are then given 30 seconds to practise the task with no financial incentives. This is done so that subjects can work out the best strategies with which to tackle the task. Next, subjects are asked to perform the task again for 30 seconds, but with a piece rate payment based on performance (£0.05 per correct response), which constitutes our measure of ability. Following this, subjects take part in three rounds of the effort task. For our first four treatments, each round lasts for 3 minutes, and for our final two treatments, each round lasts for 6 minutes.[23] Within each round, subjects are paid using a piece rate to incentivise performance (£0.05 per correct response). In addition, we also incentivise predictions: before each round of the task, we ask subjects to predict their performance in the upcoming round, and inform them that if their prediction is correct, they earn an additional small bonus payment (£0.02). We set this payoff to be less than the piece rate so that subjects always have an incentive to maximize their performance regardless of whether they achieve their prediction.

Between each round of the task, subjects are provided feedback about the number of correct answers they scored in the previous round. Subjects are also informed about whether their prediction, made just before the previous round, was correct. Finally, subjects are asked to predict their performance in the next round of the task.

In our baseline, subjects face a simple addition task similar to that in Niederle and Vesterlund (2007); subjects must complete as many three number sums as possible within the time limit. Our first treatment increases the difficulty of the real-effort task; subjects are faced with a three number subtraction problem, based on evidence that subtraction is more difficult than addition (Campbell, 2008). Our second treatment represents an increase in complexity from the baseline, as subjects face a grid-search task (finding the largest number in a grid) before being able to complete the addition task (adding the searched numbers together). In our third treatment, subjects face both difficulty and complexity: they must complete a grid search to identify the numbers, and then face a subtraction problem based on the searched numbers.

Note that finding the correct answer when taking part in the complex addition and subtraction tasks takes more time since there are more components to the task. Hence, the maximum potential output of subjects in a three-minute round is considerably lower in our

[23] Practise rounds are intentionally, and considerably, shorter than the full rounds of the task (30 seconds). By keeping practise rounds short, subjects learn about the task without getting information on exactly what their performance would be in the main rounds of the experiment. This is an important consideration as subjects need to report beliefs about their performance based on limited knowledge about the task.

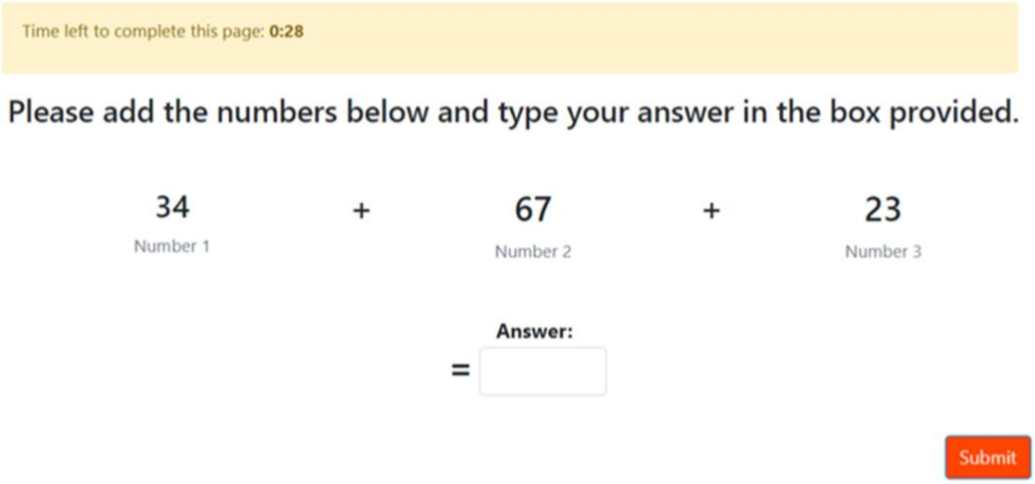
complex task treatments than in our simple task treatments. This results in a greater variance in the predictions made by subjects in the baseline addition and subtraction tasks compared with the complex tasks. With complex tasks, both predictions and performance is closer to zero. This feature can potentially make predictions more accurate mechanically, as the choice set gets restricted. To combat this, we implement an additional two treatments which are identical to our complex addition and complex subtraction treatments, but have longer rounds (six minutes instead of three).

At the end of the experiment, subjects completed a short survey which measured perceptions of the experiment; confidence (using a generalised confidence measure); intrinsic motivation (using a short form of the intrinsic motivation inventory - Ryan, 1982); and attentiveness (using the cognitive reflection test - Frederick, 2005), in addition to standard demographic information.

2.3.1 Real-Effort Task

Our experiment utilizes a simple real-effort task (adapted from Niederle and Vesterlund, 2007), which we then modify to manipulate difficulty and complexity. In the baseline (simple addition) task (displayed in Figure 2.1), subjects were shown three 2-digit numbers on a screen and were tasked with adding the three numbers together.[24] The simple subtraction version of this task is very similar, but requires subjects to subtract the second and third number from the first number displayed on the screen. Both these tasks are similar in that they require the subject to undertake a relatively simple component (arithmetic computation).

Figure 2.1: Simple addition task screenshot



Time left to complete this page: **0:28**

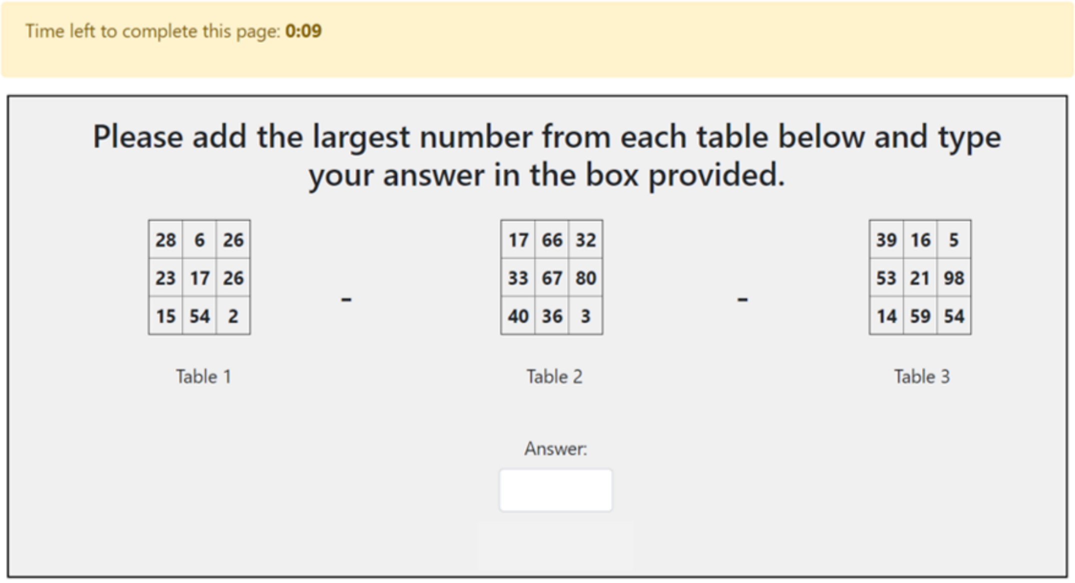
Please add the numbers below and type your answer in the box provided.

34 + 67 + 23

The complex addition and subtraction tasks add a further component to the simple tasks described above: a grid search. In the complex addition task subjects were shown three, 3x3 grids (each containing 9 numbers between 0-99) and were asked to find the largest unique number in each and add them together. Our complex subtraction tasks requires subjects to subtract the largest number from the second and third grids from the largest number in the first grid. An example of our complex subtraction task is shown in Figure 2.2.

[24] We modify the task used by Niederle and Vesterlund (2007), which originally features five 2-digit numbers on the screen. By using fewer numbers in each question, we enable subjects who are good at the task to complete more questions; this results in an easier task, and a wider distribution of output.

Figure 2.2: Complex subtraction task screenshot



Time left to complete this page: **0:09**

Please add the largest number from each table below and type your answer in the box provided.

28	6	26
23	17	26
15	54	2

Table 1

-

17	66	32
33	67	80
40	36	3

Table 2

-

39	16	5
53	21	98
14	59	54

Table 3

Answer:

In each treatment, subjects had to submit a numerical answer in order to move to the next question. Note that subjects were not required to submit a correct answer in order to continue. During the task, subjects were not told whether their answer was correct or incorrect; the only piece of information displayed on screen was the time which was remaining in that round. All numbers used within our real-effort tasks were randomly selected for each subject and for each question, from numbers between 0 – 99. Since all numbers were randomly selected, each subject faced a similar level of difficulty to other subjects who were assigned to the same task. At the start of the round, the timer started to count down (either 3 or 6 minutes). Once the timer had run out, subjects were automatically moved to the next instruction screen.

We measure the performance of each subject as their output in a round of the effort task. Subjects were incentivised by a piece rate, but were also told that they should *"try as hard as possible"* and should answer as many questions as possible within the time limit.

2.3.2 Experimental Procedure

We ran sessions on the Prolific.co platform. Initial sessions were run on the 16th of April 2021, with a second round of data collection on the 19th-20th of July 2021. Our sessions were run asynchronously. Subjects took part in our experiment in isolation of one another; there is no interaction between participants and their performance does not impact anyone else. Subjects were able to join the experiment via Prolific.co between 9am – 5pm on each day and were randomly assigned to treatments.

In our first rounds of sessions (16th April), subjects were assigned to one of the following treatments: simple addition; simple subtraction; short complex addition; short complex subtraction. For each of these treatments we collected data from 20 subjects, for a total of 80 subjects. A second round of sessions (19th-20th July) also assigned subjects into the above treatments plus two additional treatments: long complex addition and long complex subtraction. During this round, we collected data from 20 additional subjects for our simple and short complex treatments alongside data from 40 subjects in each of our long complex treatments. Hence, we have a total of 240 subjects, 40 in each treatment. In the next section, we present demographic information about our sample.

2.3.3 Data

Based on prior power calculations, 40 data points were collected for each treatment: a total of 240 observations. For each subject, demographic information was collected from the Prolific.co platform. This demographic information included: age; sex; student status; and degree status. In addition, we collected information about the attentiveness/cognitive reflection (CRT), intrinsic motivation (IMI), and general confidence as part of our survey. To compile our general confidence measure, we ask subjects a series of general knowledge questions, and subsequently ask them to rate how confident they are about their answer.[25] In the table below, we present the mean and SD of each of our demographic and survey variables, taken as an average for each treatment.

In table 2.1, we see that the average age varies between 27 and 34 years old depending on the treatment; subjects are typically slightly older in our complex addition treatment than in our other treatments. In addition, only 30-40% of subjects in our study are female; despite this, there are no significant differences in gender between our treatments. Almost 50% of subjects in our study are currently students. Note that the minimum age for completing a study on Prolific.co is 18 years old; therefore, those that are students are studying a university level degree. In addition, in most treatments almost 50% of our subjects already hold a degree; subjects in our complex subtraction treatment appear to be less likely to hold an existing degree.

We take the average of correct responses to our CRT survey questions for each subject, and then for all subjects within each treatment. This is represented by our CRT average variable, which shows that subjects typically only answer half of the CRT questions correctly. The lowest average score on our CRT questions was in our simple addition treatment, whilst subjects in our simple subtraction treatment answered the greatest number of questions correctly on average.

Subjects respond to our intrinsic motivation inventory (IMI) questions using a 6 point Likert-scale, where a score of 6 means that they strongly agree with the questions. Table 2.1 indicates that there are no significant differences in the average level of intrinsic motivation among subjects across treatments.

Finally, we turn our attention to the general confidence level of each subject. Subjects are asked how confident they are about their answers to a series of trivia questions, and respond using a 6 point Likert-scale related to each question.[26] We calculate the confidence variable by taking the average of responses from our general confidence measure in the survey. We see from Table 2.1 that subjects are typically underconfident and that there is little difference in the general confidence level between our treatments.

[26] Possible responses included: not confident at all; not very confident; somewhat unconfident; somewhat confident; very confident; certain.

Table 2.1 – Average Responses for Demographic and Survey Variables by Treatment

	Simple Addition	**Simple Subtraction**	**Complex Addition**	**Complex Subtraction**	**Long Complex Addition**	**Long Complex Subtraction**
Age	27.775 (7.416)	29.050 (11.644)	34.150 (14.644)	27.825 (8.756)	25.625 (6.751)	27.200 (7.549)
Sex (1 = Female)	0.400	0.300	0.350	0.375	0.375	0.400
Student (1 = Current Student)	0.400	0.475	0.450	0.475	0.525	0.525
Degree (1 = Has Degree)	0.525	0.500	0.500	0.425	0.500	0.550
CRT Average	0.450 (0.437)	0.558 (0.423)	0.592 (0.374)	0.483 (0.399)	0.500 (0.399)	0.541 (0.425)
IMI Average	4.733 (1.175)	4.983 (1.434)	4.482 (1.101)	4.479 (1.082)	4.700 (1.128)	4.867 (0.935)
General Confidence Average	2.763 (0.906)	2.825 (0.932)	2.838 (0.914)	2.844 (1.075)	2.622 (0.780)	3.050 (0.853)

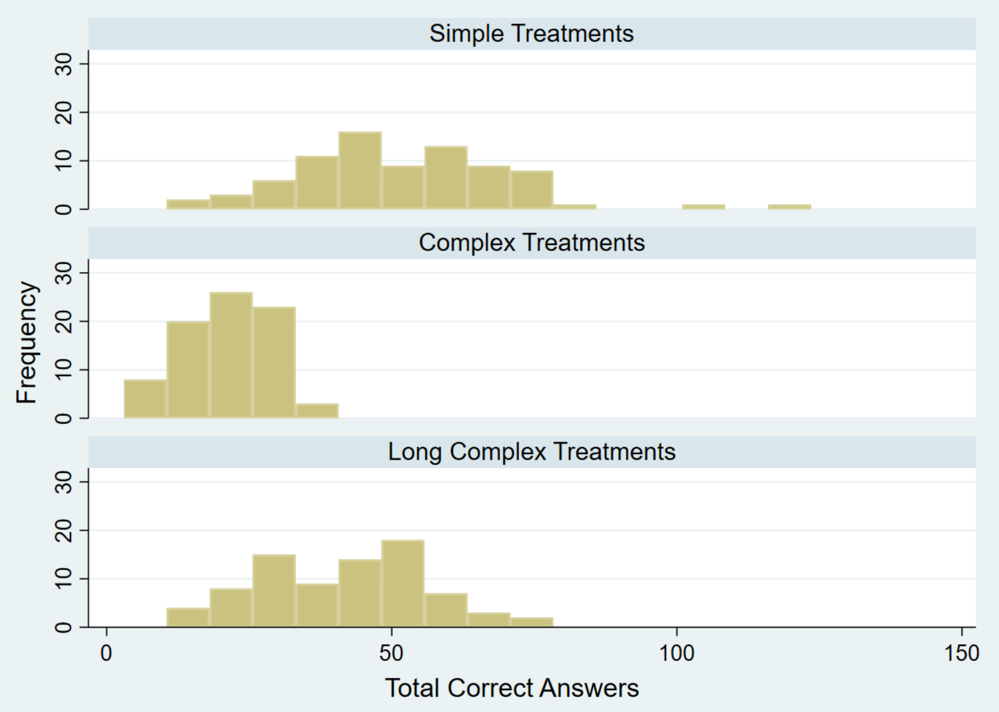
Notes: Demographic responses are shown as the mean of each variable, with the standard deviation shown in parentheses. Sex, student, and degree are binary variables which represent whether the subject is female, currently a student, or has a degree (=1 for each). Therefore, the means of these variables represent the percentage of subjects who match the above descriptions respectively.

2.4 Results

The core outcome measure that we examine is subject confidence about their performance. Confidence is defined as the difference between subject's prediction about their output and their actual output, i.e. confidence = prediction – output. Hence, higher confidence means the subject predicted their output to be higher than their actual output. In terms of confidence, there are no significant differences between the simple and difficult versions of the task (addition and subtraction), and hence we pool these treatments together for the sake of brevity and power. [27] Treatments are pooled into the following three categories: simple tasks (simple addition and simple subtraction); short complex tasks (complex addition and complex subtraction); long complex tasks (long complex addition and long complex subtraction). The appendix presents the full set of results broken out by difficulty for the interested reader.

Figure 2.3 presents the distribution of performance across our three treatment sets. It is clear that subjects in the complex treatment set were able to answer significantly less questions than subjects in our simple treatment set (*p=0.000*), given the same time limit. When the time limit is doubled in the long complex treatment set, the difference in output is still significantly different to the simple treatment set, however the distribution of output is considerably closer (shown in Figure 2.3). Subjects in the simple treatment set had an average output of 51.7, while subjects in the short complex and long complex had an average output of 21.2 and 42.01 respectively, significantly different from each other (*p=0.000*) and significantly lower than the simple treatment set (*p=0.000* for the short and *p=0.000* for the long treatment set).

Figure 2.3: Histogram of the total number of correct answers (output) – all rounds

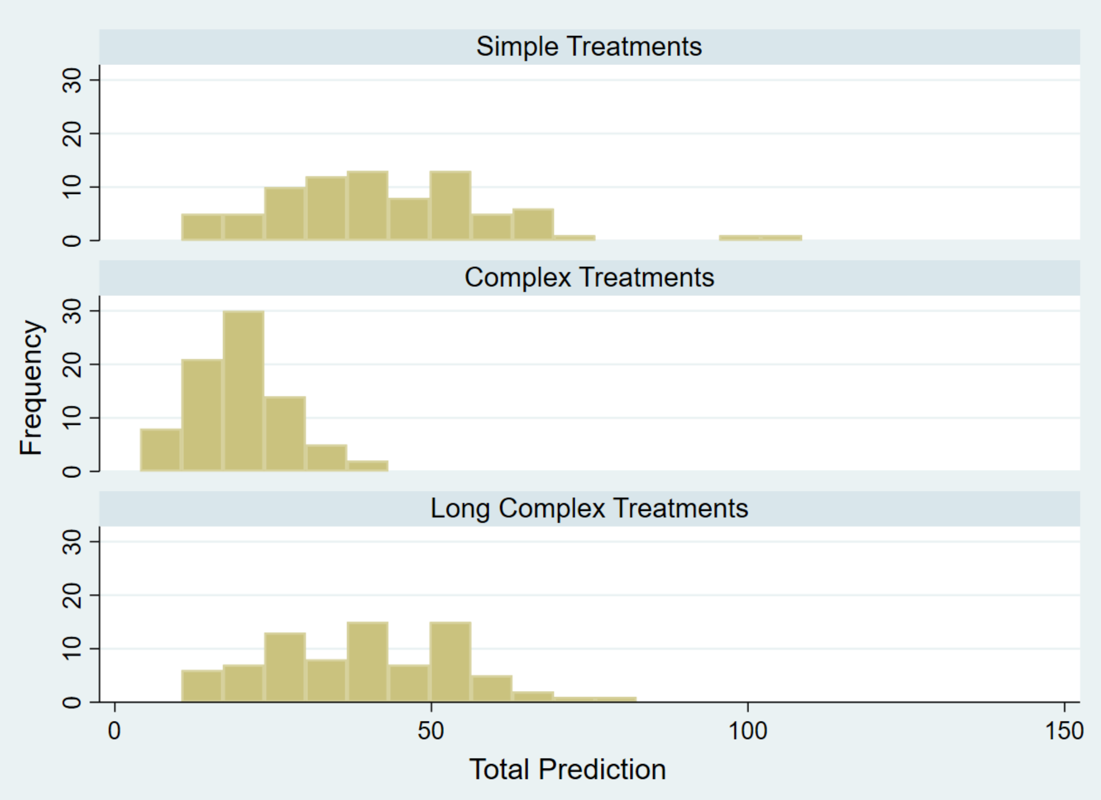


Notes: Treatments are pooled: Simple Treatments (simple addition/subtraction); Complex Treatments (short complex addition/subtraction); Long Complex Treatments (long complex addition/subtraction)

[27] We find no differences between the addition and subtraction treatments for either the simple (*p=0.786*), complex (*p=0.547*), or long complex (*p=0.678*) versions of the tasks, hence we pool them together in the figures. In appendix figures 2.1-2.4, we replicate the figures presented in this section, but broken out by difficulty.

Since our primary outcome variable, confidence, is a composite measure of prediction minus output, we present the distribution of predictions in Figure 2.4. Once again, it is clear that predictions in our complex treatments are significantly lower than in our simple treatments (*p<0.000*). There are no significant differences in predictions between our simple and long complex treatments (*p=0.149*), suggesting that subjects make similar predictions in these treatments.

Figure 2.4: Histogram of the total prediction – all rounds



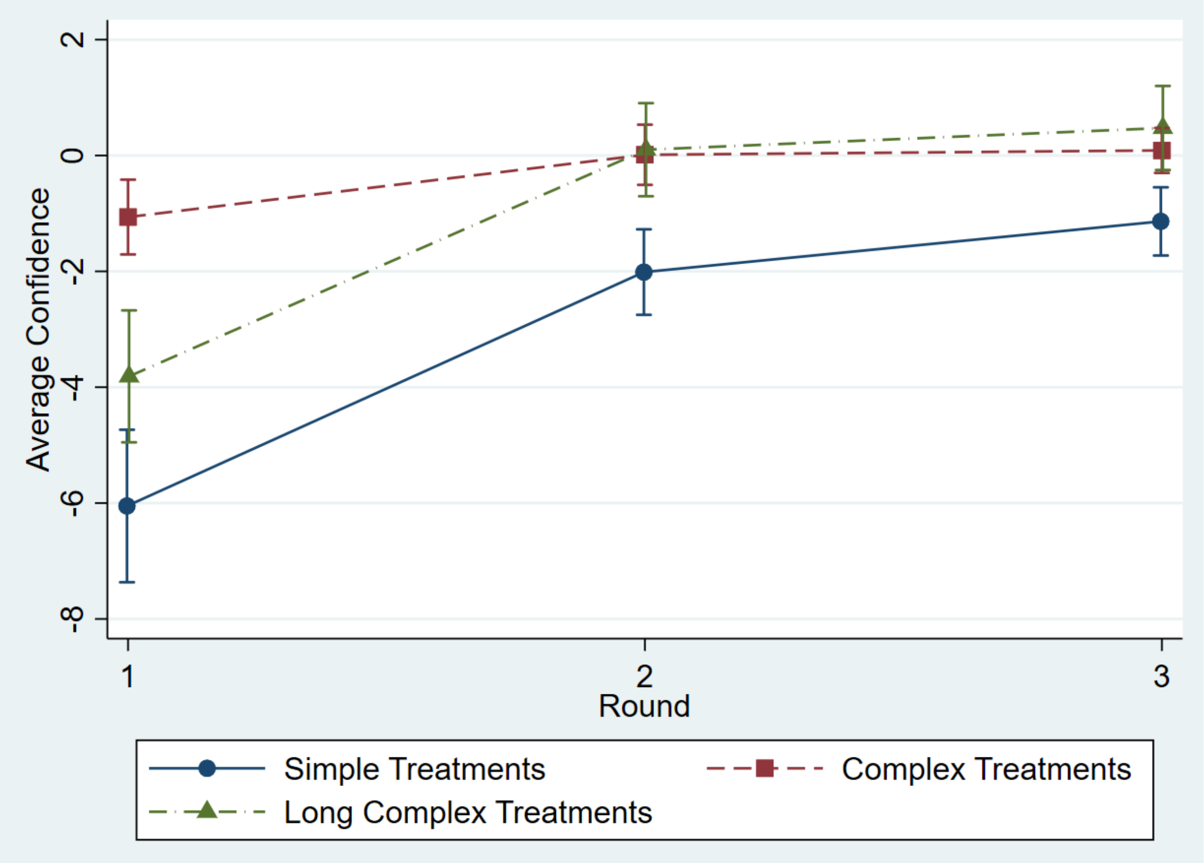
Notes: Treatments are pooled: Simple Treatments (simple addition/subtraction); Complex Treatments (short complex addition/subtraction); Long Complex Treatments (long complex addition/subtraction)

2.4.1 The Evolution of Subject Confidence

After outlining our outcome measure above, we first turn our attention to the evolution of subject confidence over time. We take this approach, since when our subjects make their output prediction before the first round of the task, they only have their performance in the short practise rounds as a guide. These practise rounds are considerably shorter, and only serve to provide the subject an opportunity to practise. As a result, they are unlikely to not provide enough information for subjects to make fully accurate predictions about their performance in the longer rounds. Therefore, we believe that our findings from first round predictions provide an insight into the decisions of subjects who have little information and guidance upon which to make their predictions.

In contrast to this, when our subjects make their prediction before the second round of the task, they have information about their performance in the first round of the same task for the same length of time. Therefore, we might expect predictions to be closer to actual output for the second and third rounds of the task than for the first round. In Figure 2.5 we present a line graph of confidence in each round of the task, by pooled treatment.

Figure 2.5 – Average confidence (prediction – output) in each round for our treatments.



Notes: Total confidence is the sum of confidence (prediction – output) across three rounds of real-effort task. Simple tasks include our simple addition and simple subtraction; Complex tasks includes short complex addition and short complex subtraction ; Long complex tasks includes long complex addition and long complex subtraction. Error bars show the 95% confidence interval.

In Fig. 2.5 (above) we see that our confidence measure is lowest in all treatments in the first round of the task. Since subjects have only faced a short practice round of the task when they face the first full rounds, they are not yet fully informed about their potential to perform over a longer time period. Therefore, subjects make worse predictions in the first round based on their performance in shorter (practise) and less informative past rounds of the task. Since we construct our confidence measure by looking at the difference between predictions and true performance, and since subjects make worse predictions in the first round, this influences the confidence measure downward as subjects are underconfident. Since here subjects have little information about their ability to perform across the full round time, they are unable to make accurate predictions about their performance.

In addition, we see that confidence is closer to zero in all of our complex treatments, across all rounds of the task. However, there are also differences in confidence between our long complex treatments and our shorter complex treatments. Of particular interest is the difference in round 1 confidence. It is likely that subjects in the shorter complex tasks were better at predicting their performance because of the lower magnitude of answers provided, as discussed in earlier sections. However, in round 1 we find that subjects in the long complex treatments have significantly lower confidence about their performance than those in the short complex treatment (two tailed t-test *p=0.000*). Despite this finding, confidence in rounds 2 (*p=0.856*) and 3 (*p=0.350*) is similar in all complex treatments regardless of the length of the task. This highlights how confidence converges on 0 in complex tasks as subjects become more informed, and therefore make better predictions.

In Table 2.2 we present results for our pooled treatments.[28] We find that confidence in our pooled complex treatments (*p=0.000*) and long complex treatments (*p=0.000*) is significantly different from our pooled simple task baseline. In addition, in round 1 the coefficients for complex treatments are over double those for long complex treatments (*p=0.006*). However, in rounds 2 and 3 the magnitude of our coefficients are very similar for both our pooled complex and long complex treatments (*p=0.977* and *p=0.999*). Therefore, we provide evidence that subjects have higher levels of confidence about their performance when faced with complex tasks than when faced with simple tasks; these effects are similar even as the number of questions which a subject can attempt increases in our long complex treatments.

Table 2.2 – The impact of complexity on confidence – by round and pooled treatment

	Round 1		Round 2		Round 3	
Confidence (Predict - Correct)	(1)	(2)	(3)	(4)	(5)	(6)
Complex Treatments	4.987***	4.624***	2.025***	1.998***	1.225***	1.413***
	(0.762)	(0.943)	(0.496)	(0.599)	(0.415)	(0.504)
Long Complex Treatments	2.237***	2.199***	2.112***	1.982***	1.613***	1.413***
	(0.762)	(0.764)	(0.496)	(0.485)	(0.415)	(0.409)
Constant	-6.050***	-5.957**	-2.012***	-1.305	-1.138***	-0.664
	(0.539)	(2.347)	(0.350)	(1.490)	(0.294)	(1.255)

[28] For a breakdown of the results from table 2.4 which include task difficulty please see appendix table 2.2. We see very similar results in appendix table 2.2, but draw attention to two points of interest. Firstly, we see that for the short complex treatments, the magnitude of results in round 1 for subtraction is slightly greater than those for addition. Secondly, we note that the magnitude of coefficients in round 1 for long complex addition is slightly greater than that of long complex subtraction. These results suggest that subjects in round 1 are more confident when facing short complex subtraction tasks (over short complex addition), and when facing long complex addition tasks (over long complex subtraction).

Survey Controls	No	Yes	No	Yes	No	Yes
Observations	240	240	240	240	240	240
R-squared	0.154	0.243	0.089	0.121	0.065	0.110

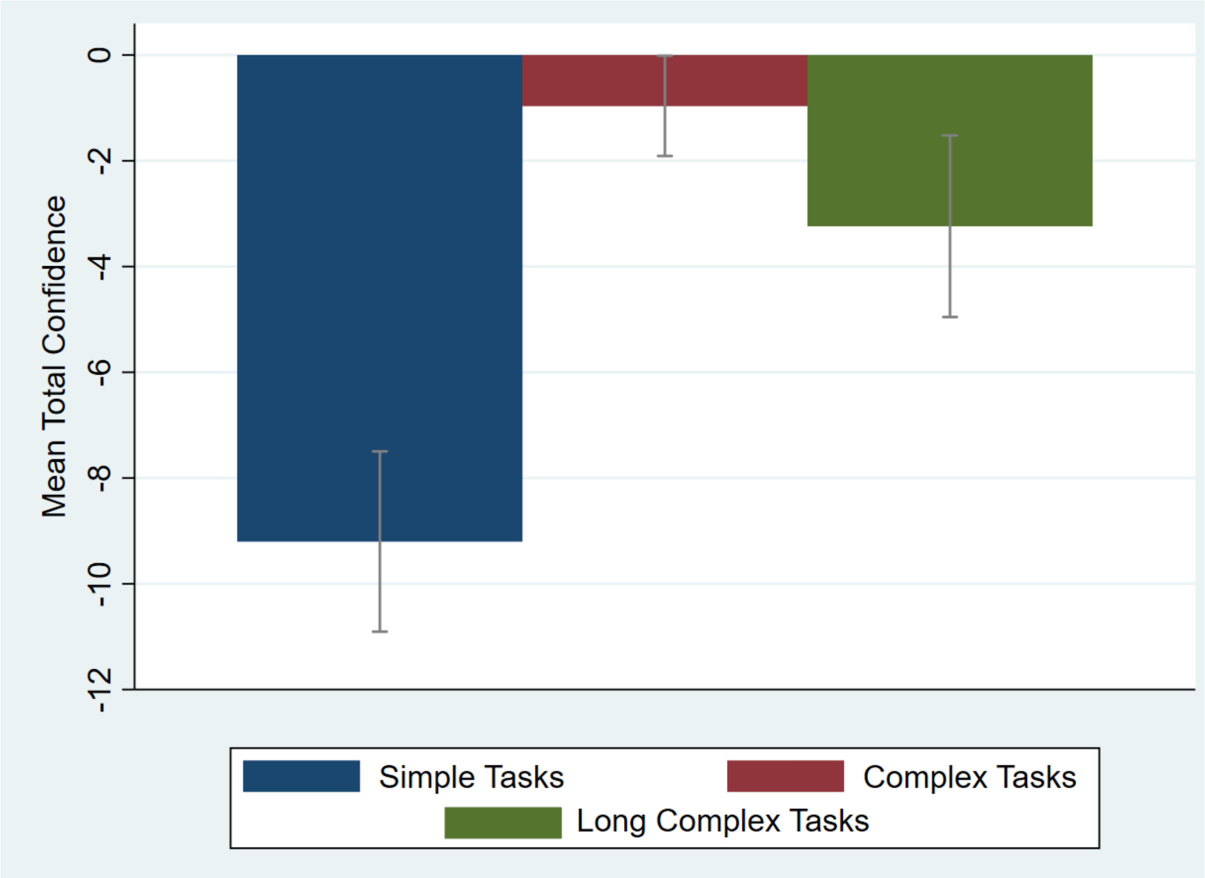
Notes: OLS regression with dependent variable as confidence in each of the 3 rounds of real effort task; the main independent variable is a categorical dummy representing the treatment group. Standard errors in parentheses. (*** p<0.01, ** p<0.05, * p<0.1). Appendix Table 2.3 presents the results including coefficients for all controls.

Table 2.2 shows that subjects in our complex tasks have higher levels of confidence; since the constant is negative, this means that these subjects are less underconfident (i.e. their confidence level is closer to 0). We find similar magnitudes of coefficients for both our short and long complex treatments, providing evidence that subjects are more confident in complex tasks regardless of the length of the task. Our results provide evidence that subjects are better at predicting their performance in complex tasks, and that these differences are greatest when subjects have less information about their true performance (i.e., in round 1).

2.4.2 The Pooled Impact of Complexity

After understanding the effects of complexity on round-by-round confidence, we next turn our attention to the overall effect of each of the pooled treatment sets in our three round setting. We test for differences in the average level of confidence across our pooled treatments. First, we calculate the total average level of confidence for each subject, within each pooled set of treatments. To do so, we deduct the sum of output from the sum of predictions about output for each subject; we then take the average of this confidence for each of our pooled treatments. Figure 2.6 displays the average levels of confidence by treatment: simple tasks (blue); short complex tasks (red); long complex tasks (green). We find that subjects are, on average, under-confident about their performance; this is in line with Snowberg and Yariv (2021) who demonstrate that online subject pools are less confident than university subject pools. Noticeably, we find that subjects are better at predicting their performance (i.e. the gap between confidence and performance is closer to 0) when faced with long complex tasks (*p<0.000*). Confidence is lowest in the simple task treatments. Subjects are more confident in both the complex and long complex set of treatments (*p<0.01* in both cases). In addition, average total confidence for all three treatments is significantly different from 0; for simple tasks (*p=0.000*); short complex tasks (*p=0.049*); and long complex tasks (*p=0.000*). These results suggest that subjects are much better at predicting their performance when faced with complex tasks, relative to when faced with simple tasks.

Figure 2.6: Average total confidence for pooled treatments.



Notes: Total confidence is the sum of confidence (prediction – output) across three rounds of real-effort task. Simple tasks include our simple addition and simple subtraction; Complex tasks includes short complex addition and short complex subtraction ; Long complex tasks includes long complex addition and long complex subtraction. Error bars show the 95% confidence interval.

Next, in Table 2.3 we present regressions that control for ability, attentiveness, and general confidence (the measurement of which is discussed earlier). In addition, we also control for intrinsic motivation (using the intrinsic motivation inventory), calculator use, and a series of demographic variables (including age, gender, and degree status). The table confirms our observations from the figures. Confidence is significantly higher under complex tasks relative to simple tasks. Note that subjects are underconfident on average, and that a positive coefficient indicates that the confidence measure is closer to 0 (and as such is smaller).Using model 2, the difference between prediction and performance in the short complex treatments is 8.035 units smaller than in the simple treatments ($p<0.000$). Furthermore, the long complex treatments are also significantly higher than the baseline ($p<0.000$), and are statistically different from the complex treatments ($p<0.001$).[29] Further, we find that neither of our estimates (plus the constant) for complex and long complex treatments in model 2 below are significant from 0 ($p=0.973$ and $p=0.458$ respectively). In model 2, we also find that the coefficient on the intrinsic motivation

[29] For robustness, we have additionally provided results in Appendix Table 2.4 which show the impact of complexity per individual and round, through the inclusion of dummy variables representing each round (with the first round removed as the baseline) and standard errors clustered at the individual level. We note that in Appendix Table 2.4 our coefficients are considerably smaller. When including additional controls for round and clustering standard errors at the individual level, the difference between prediction and performance in our long complex treatments is still significantly greater (around 1.9 units, $p<0.000$) than in our simple task baseline.

inventory is significant: those that report being motivated by the task as significantly less confident across all treatments (*p<0.000*).[30]

Table 2.3 – The impact of complexity on confidence

Dependent Variable: Total Confidence	(1) Model 1	(2) Model 2
Complex Treatments	8.238***	8.035***
	(1.077)	(1.263)
Long Complex Treatments	5.963***	5.595***
	(1.077)	(1.022)
Ability		0.174
		(0.464)
Calculator Use		-0.851
		(0.953)
Cognitive Reflection Task Average		0.581
		(0.356)
General Confidence		0.742
		(0.484)
Motivation		-1.344***
		(0.370)
Working Status (Working = 1)		0.963
		(0.886)
Student Status (Student = 1)		1.710*
		(0.971)
Degree Status (Degree = 1)		-0.610
		(0.853)
Age (in Years)		0.0577
		(0.0505)
Gender (Female =1)		-0.410
		(0.903)
Constant	-9.200***	-7.927**
	(0.762)	(3.141)
Observations	240	240
R-squared	0.208	0.302

Notes: OLS regression with dependent variable as total (sum of) confidence across 3 rounds of real effort task; the main independent variable is a categorical dummy representing the treatment group. Standard errors in parentheses. (*** p<0.01, ** p<0.05, * p<0.1).

Overall, we find evidence that subject confidence is significantly greater in our complex treatments compared with the simple treatments. Similarly to our round-by-round findings, these results hold when we examine our complex treatments with long rounds, suggesting that the impact of complexity on confidence is not mechanical due to the lower level of performance with complex tasks. Appendix Table 2.1 displays the results for the low and high difficulty versions of the treatments.

[30] We have separately undertaken a subgroup analysis which examines the impact of motivation on confidence in each of our pooled and separate treatments. This analysis is not presented here, as we find no systematic differences in confidence between motivated and unmotivated subjects in any of our treatments.

We present results in Table 2.4 which are in line with our findings regarding the round-by-round difference that subjects who face the more complex versions of tasks are better at predicting their performance. However, we now add to our earlier findings by including results for two sub-groups: subjects who faced addition tasks; subjects who face subtraction tasks. In our setting, when subjects face addition tasks, regardless of the length of the task the impact of complexity on confidence is statistically indistinguishable (*p=0.243*). However, when subjects are faced with subtraction tasks, the version of complexity which includes longer rounds brings about a smaller gain in confidence than the shorter round counterpart (*p<0.1*).

Table 2.4 – The Impact of Complexity on Confidence – Addition Compared to Subtraction

	Addition Tasks		Subtraction Tasks	
Dependent Variable: Total Confidence	Model 1	Model 2	Model 1	Model 2
Complex Treatment	8.775***	8.682***	7.700***	7.636***
	(1.585)	(1.975)	(1.470)	(1.835)
Long Complex Treatment	7.075***	6.442***	4.850***	4.796***
	(1.585)	(1.477)	(1.470)	(1.494)
Ability		0.141		0.423
		(0.735)		(0.651)
Calculator Use		-1.127		-0.194
		(1.367)		(1.408)
Cognitive Reflection Task Average		1.090**		0.0761
		(0.522)		(0.501)
General Confidence		0.941		0.898
		(0.799)		(0.656)
Motivation		-0.550		-2.004***
		(0.564)		(0.527)
Working Status (Working = 1)		0.878		1.377
		(1.319)		(1.266)
Student Status (Student = 1)		0.200		2.929**
		(1.413)		(1.426)
Degree Status (Degree = 1)		-1.367		-0.134
		(1.254)		(1.224)
Age (in Years)		-0.00678		0.0957
		(0.0749)		(0.0817)
Gender (Female =1)		1.004		-1.321
		(1.337)		(1.275)
Constant	-9.525***	-10.69**	-8.875***	-6.861
	(1.121)	(4.473)	(1.039)	(4.978)
Observations	120	119	120	120
R-squared	0.228	0.328	0.194	0.338

Notes: OLS regression with dependent variable as total (sum of) confidence across 3 rounds of real effort task; the main independent variable is a categorical dummy representing the treatment group. In columns 1-2 we restrict the sample to subjects who took part in an addition task, and in columns 3-4 to subjects who took part in a subtraction task. Standard errors in parentheses. (*** p<0.01, ** p<0.05, * p<0.1).

2.5 Conclusion

Real-effort tasks feature extensively in experimental economic literature. These experiments feature a working task which requires subjects to exert costly effort; examples include answering simple arithmetic questions (Niederle and Vesterlund, 2008) and solving mazes (Gneezy *et al.*, 2003). However, real-effort tasks represent a loss of experimenter control as the cost of effort differs between subjects (Dutcher *et al.*, 2015); this is fixed, and common among subjects, when using a chosen-effort task.

Existing literature often relies on simple real-effort tasks, such as the slider task, to reduce variation in the cost of effort. Since these tasks are very simple, they do not closely mimic real-world tasks. As a result, the beliefs a subject has about their performance, and in turn their ability to predict their performance, may be biased. Despite this, little is known about the effects of implementing more complex versions of standard real-effort tasks. Whilst Moore and Healy (2008) document greater levels of overconfidence in complex general knowledge quizzes, their results are not generalisable to the types of real-effort task typically used in experimental settings.

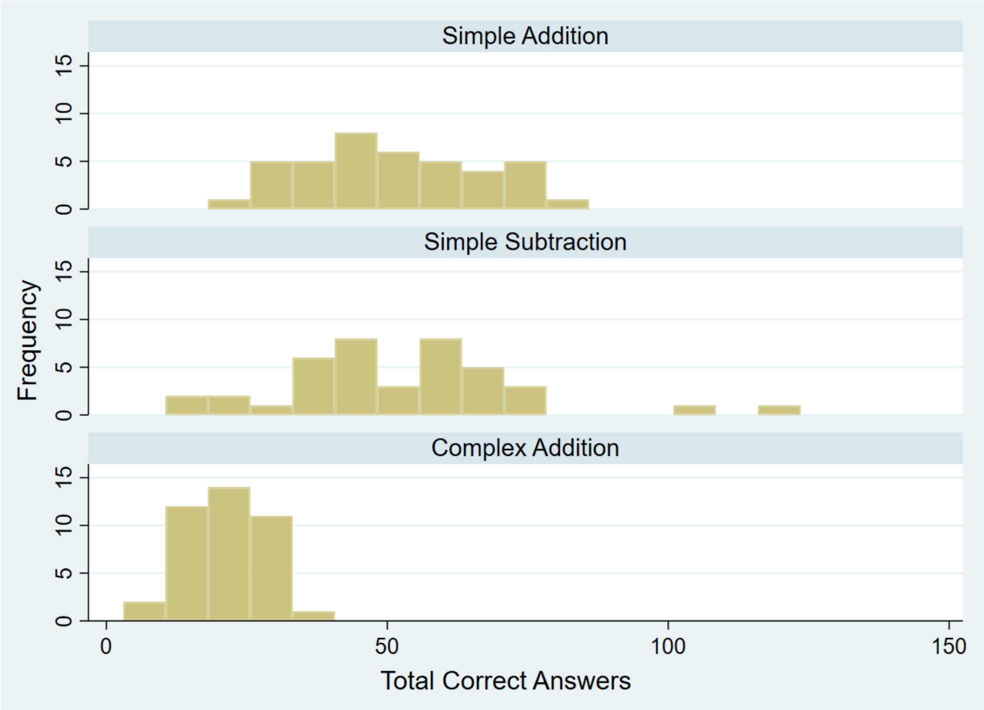
To address this, we implement an online experiment containing six treatments which contain real-effort tasks of different difficulties and complexities. In our baseline, subjects face a simple addition task; in our first treatment we increase the difficulty and introduce a simple subtraction task. In treatments 2 and 3 subjects must first find the largest number in each of three 3x3 grids, before adding them together (treatment 2) or subtracting them from each other (treatment 3). In addition, we implement treatments 5 and 6 which replicate treatments 2 and 3 but double the length of time for each round of the task (from 3 minutes to 6 minutes).

We find that the average subject is underconfident; their beliefs about their performance are lower than their true performance. We find no significant difference in the confidence of subjects when we increase task difficulty, and therefore pool our tasks of similar length and complexity together. However, we find convincing evidence that subjects who face complex tasks are better at predicting their performance, and that our effects remain as subjects complete more rounds of the task. In addition, our results hold when we increase the length of time subjects are given to complete our complex tasks, such that the subjects complete a similar number of correct answers as in our baseline.

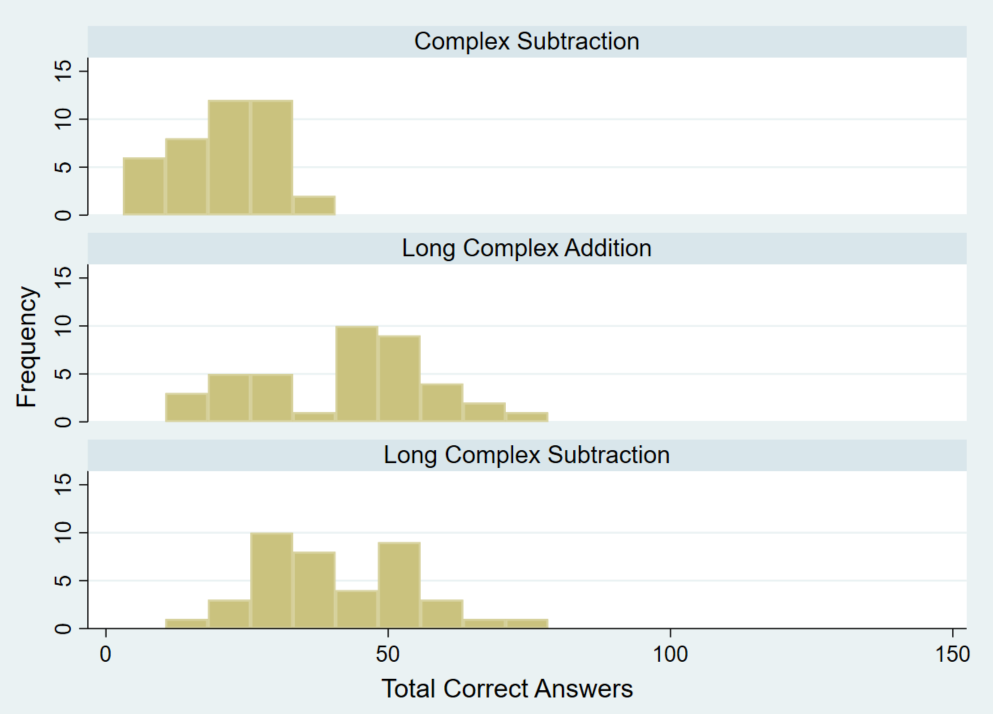
Our results carry implications for future researchers to consider when deciding which type of real-effort task to use in their research. If a simple real-effort task is chosen, then subjects may be significantly less confident than if a more complex task is chosen. In particular, this carries ramifications for studies which require subjects to choose between incentive schemes. Subjects are likely to make sub-optimal choices if they are unable to make accurate predictions about their performance, especially since decisions are based on these predictions. As a result, our findings suggest that the use of more complex real-effort tasks would allow subjects to predict their performance more accurately, and in turn make more informed choices.

2.6 Appendix

Appendix Figures 2.1 and 2.2 – Total Correct Answers by Treatment (Including Difficulty Treatment Breakdown)

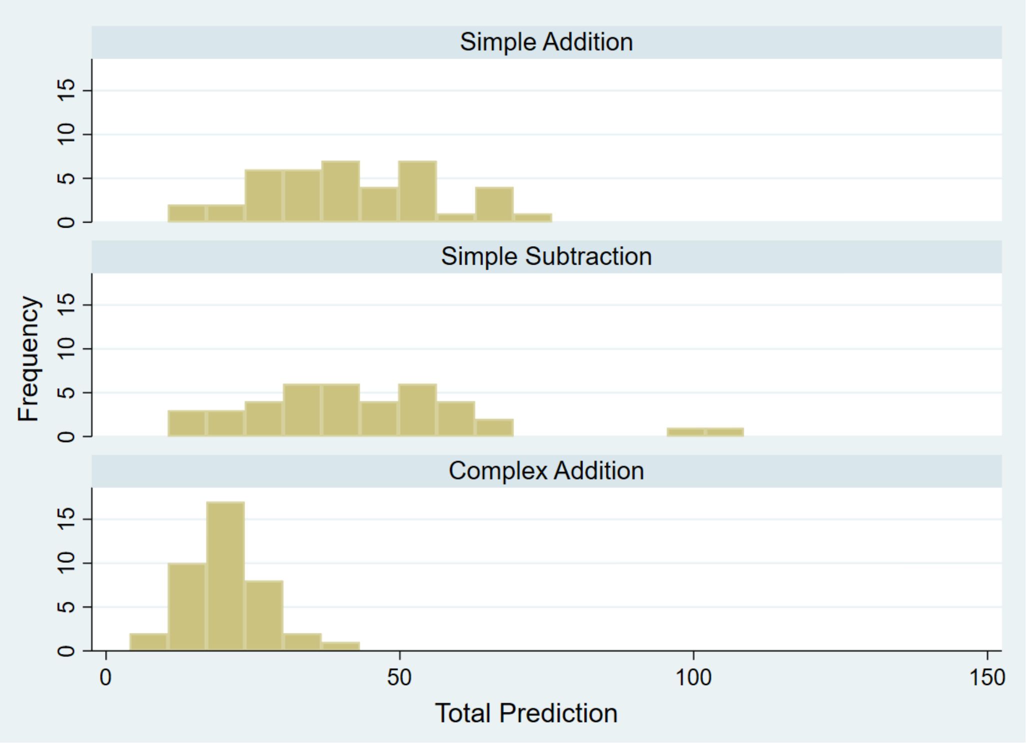


Notes: Frequency of total number of correct answers (sum of correct answers in each round) for our first three treatments.

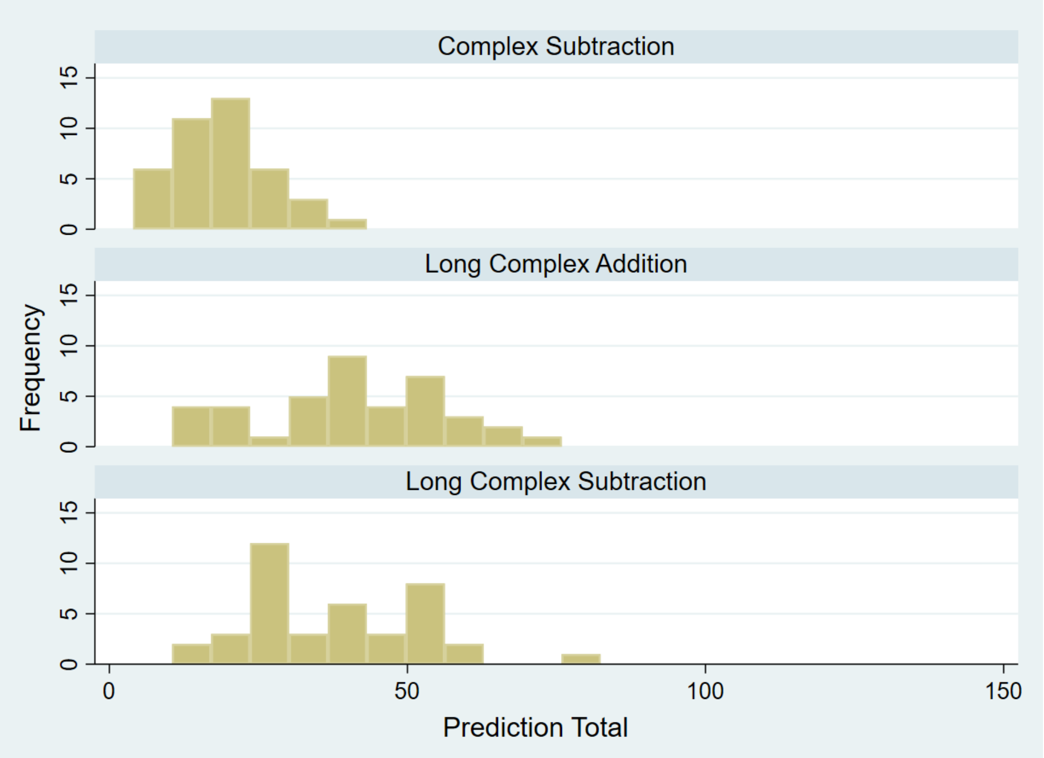


Notes: Frequency of total number of correct answers (sum of correct answers in each round) for our second three treatments.

Appendix Figures 2.3 and 2.4 – Total Prediction by Treatment (Including Difficulty Treatment Breakdown)



Notes: Frequency of total number of total prediction (sum of predictions across all rounds) for our first three treatments.



Notes: Frequency of total number of total prediction (sum of predictions across all rounds) for our first three treatments.

Appendix Table 2.1 – Total Confidence by Full Treatment Breakdown Including Control Coefficients

	(1)	(2)
Dependent Variable: Total Confidence	Model 1	Model 2
Simple Subtraction	0.650	0.580
	(1.529)	(1.416)
Complex Addition	8.775***	8.414***
	(1.529)	(1.635)
Complex Subtraction	8.350***	8.056***
	(1.529)	(1.640)
Long Complex Addition	7.075***	6.400***
	(1.529)	(1.435)
Long Complex Subtraction	5.500***	5.337***
	(1.529)	(1.461)
Ability		0.119
		(0.471)
Calculator Use		-0.852
		(0.961)
Cognitive Reflection Task Average		0.574
		(0.359)
General Confidence		0.795
		(0.493)
Motivation		-1.352***
		(0.374)
Working Status (Working = 1)		1.032
		(0.894)
Student Status (Student = 1)		1.698*
		(0.982)
Degree Status (Degree = 1)		-0.622
		(0.858)
Age (in Years)		0.0556
		(0.0521)
Gender (Female =1)		-0.383
		(0.909)
Constant	-9.525***	-8.157**
	(1.081)	(3.243)
Observations	240	239
R-squared	0.213	0.304

Notes: OLS regression with dependent variable as total (sum of) confidence across 3 rounds of real effort task; the main independent variable is a categorical dummy representing the treatment group. Standard errors in parentheses. (*** p<0.01, ** p<0.05, * p<0.1).

Appendix Table 2.2 – Confidence in Each Round by Full Treatment Breakdown Including Control Coefficients

Dependent Variable:	Round 1		Round 2		Round 3	
Confidence in Each Round	(1)	(2)	(3)	(4)	(5)	(6)
Simple Subtraction	0.600	0.435	0.225	0.267	-0.175	-0.122
	(1.083)	(1.058)	(0.703)	(0.670)	(0.589)	(0.566)
Complex Addition	5.225***	4.608***	2.525***	2.535***	1.025*	1.271*
	(1.083)	(1.222)	(0.703)	(0.774)	(0.589)	(0.654)
Complex Subtraction	5.350***	5.003***	1.750**	1.694**	1.250**	1.359**
	(1.083)	(1.225)	(0.703)	(0.776)	(0.589)	(0.655)
Long Complex Addition	2.725**	2.698**	2.450***	2.152***	1.900***	1.551***
	(1.083)	(1.072)	(0.703)	(0.679)	(0.589)	(0.573)
Long Complex Subtraction	2.350**	2.139*	2.000***	2.055***	1.150*	1.143*
	(1.083)	(1.092)	(0.703)	(0.692)	(0.589)	(0.584)
Ability		-0.018		-0.040		0.177
		(0.352)		(0.223)		(0.188)
Calculator Use		-0.759		-0.637		0.544
		(0.718)		(0.455)		(0.384)
Cognitive Reflection Task Average		0.750***		-0.131		-0.0451
		(0.268)		(0.170)		(0.143)
General Confidence		0.535		0.246		0.0135
		(0.368)		(0.233)		(0.197)
Motivation		-0.800***		-0.204		-0.347**
		(0.279)		(0.177)		(0.149)
Working Status (Working = 1)		0.566		0.0798		0.386
		(0.668)		(0.423)		(0.357)
Student Status (Student = 1)		1.054		0.368		0.276
		(0.734)		(0.465)		(0.392)
Degree Status (Degree = 1)		-0.657		-0.149		0.185
		(0.641)		(0.406)		(0.343)
Age (in Years)		0.0509		0.000		0.005
		(0.039)		(0.025)		(0.021)
Gender (1 = Female)		0.004		-0.571		0.184
		(0.679)		(0.430)		(0.363)
Constant	-6.350***	-6.344***	-2.125***	-1.070	-1.050**	-0.743
	(0.766)	(2.424)	(0.497)	(1.536)	(0.416)	(1.296)
Observations	240	240	240	240	240	240
R-squared	0.155	0.245	0.096	0.128	0.072	0.112

Notes: OLS regression with dependent variable as confidence in each of the 3 rounds of real effort task; the main independent variable is a categorical dummy representing the treatment. Here there is a full breakdown of treatments, including those which represent an increase in difficulty. Standard errors in parentheses. (*** p<0.01, ** p<0.05, * p<0.1).

Appendix Table 2.3 – Confidence in Each Round by Pooled Treatments Including Control Coefficients

Dependent Variable:	Round 1		Round 2		Round 3	
Confidence in Each Round	(1)	(2)	(3)	(4)	(5)	(6)
Complex Treatments	4.987***	4.624***	2.025***	1.998***	1.225***	1.413***
	(0.762)	(0.943)	(0.496)	(0.599)	(0.415)	(0.504)
Long Complex Treatments	2.237***	2.199***	2.112***	1.982***	1.613***	1.413***
	(0.762)	(0.764)	(0.496)	(0.485)	(0.415)	(0.409)
Ability		-0.000		-0.020		0.194
		(0.346)		(0.220)		(0.185)
Calculator Use		-0.743		-0.638		0.530
		(0.712)		(0.452)		(0.381)
Cognitive Reflection Task Average		0.748***		-0.117		-0.0509
		(0.266)		(0.169)		(0.142)
General Confidence		0.503		0.249		-0.0101
		(0.362)		(0.230)		(0.193)
Motivation		-0.804***		-0.187		-0.353**
		(0.277)		(0.176)		(0.148)
Working Status (Working = 1)		0.546		0.0475		0.370
		(0.662)		(0.420)		(0.354)
Student Status (Student = 1)		1.035		0.421		0.254
		(0.725)		(0.461)		(0.388)
Degree Status (Degree = 1)		-0.668		-0.123		0.181
		(0.637)		(0.405)		(0.341)
Age (in Years)		0.048		0.007		0.003
		(0.038)		(0.024)		(0.020)
Gender (1 = Female)		-0.0191		-0.572		0.181
		(0.675)		(0.429)		(0.361)
Constant	-6.050***	-5.957**	-2.012***	-1.305	-1.138***	-0.664
	(0.539)	(2.347)	(0.350)	(1.490)	(0.294)	(1.255)
Observations	240	240	240	240	240	240
R-squared	0.154	0.243	0.089	0.121	0.065	0.110

Notes: OLS regression with dependent variable as confidence in each of the 3 rounds of real effort task; the main independent variable is a categorical dummy representing the treatment group. Standard errors in parentheses. (*** p<0.01, ** p<0.05, * p<0.1).

Appendix Table 2.4 – The impact of complexity on confidence including round as a control and clustering standard errors at the individual level

Dependent Variable: Total Confidence	(1) Model 1	(2) Model 2
Complex Treatments	2.746***	2.678***
	(0..330)	(0.420)
Long Complex Treatments	1.988***	1.865***
	(0.409)	(0.340)
Round 2	3.008***	2.958***
	(0.343)	(0.328)
Round 3	3.450***	3.418***
	(0.382)	(0.328)
Ability		0.058
		(0.154)
Calculator Use		-0.284
		(0.317)
Cognitive Reflection Task Average		0.194
		(0.118)
General Confidence		0.247
		(0.161)
Motivation		-0.448***
		(0.123)
Working Status (Working = 1)		0.321
		(0.294)
Student Status (Student = 1)		0.570*
		(0.323)
Degree Status (Degree = 1)		-0.203
		(0.283)
Age (in Years)		0.019
		(0.017)
Gender (Female =1)		-0.137
		(0.300)
Constant	-5.129***	-4.768***
	(0.429)	(1.061)
Observations	240	240
R-squared	0.212	0.244

Notes: OLS regression with dependent variable as total (sum of) confidence across 3 rounds of real effort task; the main independent variable is a categorical dummy representing the treatment group. Round dummies are included with round 1 removed as the comparison round. Standard errors are clustered at the individual level and are shown in parentheses. (*** p<0.01, ** p<0.05, * p<0.1).

Appendix 2 - Chapter 2 Experimental Protocol

Before Data Collection

- The experiment is hosted on a Heroku server
    - The hosted server should have 1 of each type of agent, and a reset Postgres at the start of data collection
    - The experimenter is responsible for ensuring that the server is hosted correctly at the start of each day of data collection
- Data may only be collected between the times: (9:00am) – (5:00pm), so please ensure that it is currently between these times of day.

During Data Collection

- When data collection has started, subjects will join the experiment via the Prolific.co platform.
- Due to server constraints, no more than 15 spaces should be opened on the Prolific.co platform at any time. Once 15 people have joined the server, the experimenter must wait for them to all finish before opening a further 15 spaces for the next set of subjects to take part in the experiment.
- The experimenter is responsible for maintaining the Prolific.co account during times that the experiment is running. This means that they must check for incoming messages at least every 5 minutes for subjects who are messaging to say that they are having difficulties.
- When a player enters the experiment, the experimenter should take note of their Prolific.co ID, and their start time.
- When a player finishes the experiment, the experimenter should take note of their Prolific.co ID, and their end time.
- If a subject is disconnected from the server, or the server crashes and all subjects are disconnected, then the experimenter must message the subjects in question and provide them with a rejoining link which is specific to their Prolific.co ID. This is provided in the oTree server, but it is important that the experimenter is aware of which ID links to which channel in the server.
- If a subject "times out" on Prolific.co, meaning they have taken too long to finish the experiment, then they must be removed from the server and their data must be dropped from the final dataset.
- Once 15 subjects have finished the task, before allowing a further 15 subjects to enter the server, the experimenter must download the current form of the datasets through the oTree server and save using the format DDMMYY*_numofsubjectscollected_initialofexperimenter* – i.e., 020221_45subjects_jm

After Data Collection

- At the end of each day of data collection, the experimenter is responsible for storing a separate dataset containing the full data which has been collected on that specific day. This should follow the same format as above.
    - A backup of this dataset should also be created in a separate directory.
- The experimenter must cross validate these with the bonus payment variable generated by the oTree code, and store these in a separate excel document. These will be paid once all data is collected.

Chapter 3

Big Fish in Small Ponds: A Lab Experiment on the Impact of Ability Tracking Systems on Effort Provision

3.1 Introduction

In education settings across the world, students are often placed into ability tracked classrooms. Ability tracking is the practice of sorting students into different teaching groups, based on past performance. In some countries, such as Romania, students are placed into schools which study different subjects based on their track; in countries such as the UK, students are placed into different classrooms within the same school and are therefore able to study a full range of disciplines. Fu and Mehta (2018) report that more than 95% of US schools use tracking.[31] Ability tracked classrooms allow teachers to closely target teaching to a specific group of students, who are all of a similar ability level (Duflo *et al.*, 2011). Evidence suggests that tracking systems benefit students of all ability levels (Duflo *et al.*, 2011), however most literature focuses on the impact on high ability students (Imberman *et al.*, 2012; Vardardottir, 2013). One important mechanism through which tracking benefits high ability students is peer effects (McEwan, 2003). High ability students are shown to benefit from studying alongside stronger peers (Vardardottir, 2013). Typically, a student will perform better if they study among higher ability classmates (Ding and Lehrer, 2007). Since peers within the classroom have a greater impact than other school-wide peers (Burke and Sass, 2013), high ability students placed among high ability peers in classrooms are likely to benefit most from tracking. This literature tends to focus on student outcomes (such as test scores) but has spent little attention on one important input: student effort. In this paper, we use a lab experiment to study the impact of ability tracking on effort.

Despite the benefits suggested by Duflo *et al.* (2011) and others, tracking may also have negative impacts on low ability students via effort choices. When students are placed into tracked groups, those placed into a low ability tracks can respond by reducing their effort (Jagacinski and Nicholls, 1990; Carbonaro, 2005). Jagacinski and Nicholls (1990) find that low ability students reduce effort in an attempt to attribute potential failure to low effort, rather than to low ability. In addition, students who are told that they are low ability are less likely to continue studying subjects past the compulsory level (Brown *et al.*, 2008) suggesting that effort becomes more focused on other subjects. These effects may arise from a drop in self-confidence which is shown to be experienced by low-ability students (Francis *et al.*, 2019).

The effort which students exert during their schooling years is often measured by their final exam outcomes. Fu and Mehta (2018) show that high ability student outcomes benefit significantly from tracking systems, whilst Imberman *et al.* (2012) demonstrate that outcomes benefit from studying alongside high ability peers. However, low ability students often find that the quality of their peer group falls under tracking (Epple *et al.*, 2002), and student achievement is shown to worsen when studying alongside low ability peers (Imberman *et al.*, 2012). In sum, the impact of tracking on outcomes is mixed, suggesting that whilst high ability students benefit from tracking (Vardardottir, 2013; Fu and Mehta, 2018) low ability students may be harmed by tracking systems (Carbonaro, 2005), especially when studying alongside low ability peers (Imberman *et al.*, 2012).

In many systems, students are placed into ability tracked groups just once during schooling; in the UK, students are only able to change between tracks once during secondary school. Additionally, ability track placement often determines the maximum grade students can

[31] In addition, Rees *et al.* (1996) show that in 1990, 89.2% of students were reported to be in ability tracked classes
.

achieve in end of school exams.[32] When ability tracked groups are fixed, even if low ability students work hard they remain in the low ability track. This can cause a loss in motivation when students compare themselves to those in the high ability track.

However, even in restrictive systems, students may be able to move between ability tracked groups. For example, UK students are placed into new ability tracked groups when changing school or moving into further education; anecdotal evidence also exists of students moving between ability tracks after improving performance, though no formal process exists within the UK system. We term the movement of a student between ability tracks as 'retracking'. We define 'retracking' as placing students into new ability group compositions at regular intervals, as opposed to the fixed groups which exist in many ability tracked systems. Theoretically, 'retracking' encourages high ability students to exert effort in order to retain their high ability status and provides an incentive for low ability students to exert effort in a bid to move into the higher ability track. Our experiment simulates these two types of tracking systems in the lab in order to identify the effects on effort choices by low and high ability individuals.

Our study is implemented as an online laboratory experiment on the prolific.co platform. Our experiment consists of a mixed ability baseline and two ability tracking treatments. The experiment uses a real effort task to measure effort. In the mixed ability baseline subjects are randomly assigned to either a high- or a low-wage group, reflecting a mixed ability group education system. In the ability tracking treatments, we manipulate the grouping system: In the first ability tracking treatment (referred to as "Tracked Groups"), subjects are placed into either the high or low payoff group based on their initial performance in effort task and remain in these groups for the remainder of the session. In our second ability tracking treatment (referred to as "Retracked Groups"), subjects are placed into new ability tracked groups at the start of each round of the effort task based on their performance in the previous round. In addition to this basic setup, we also measure motivation, ability, cognitive reflection, and general confidence, along with a standard set of demographics.

The lab constitutes a low-cost way to study this research question, and allows us to control other aspects of the educational context so as to focus on the impact on effort, holding all else constant. Furthermore, implementing and testing ability tracking in the real world is costly; placing students into ability tracked groups requires standardised assessment to compare performance and takes up the time of teaching staff. To this end, several studies on ability tracking highlight the potential negative impact that it may have on student outcomes. For example Bolukbas and Gur (2020) highlight the inequality which may arise from tracked schooling, adding further implementation complexities. Our lab setting allows us to abstract away the peer effects mechanism; subjects take part in our experiment asynchronously and are only told about the performance of their competitors. Therefore, we are able to focus on effort as the mechanism which drives differences in responses to our tracking treatments; something which is difficult to control in the real-world.

We find that the second of our two ability tracking treatments (Retracking) increases effort provision over the mixed ability group baseline. The first ability tracking treatment yields nominally

[32] For example, some UK students are placed into ability-based classes at the start of secondary school aged 12 (based on performance in primary school). These same students are again placed into ability-based groups when they begin studying toward their GCSEs at the end of year 9 and remain in these classes for the remainder of their secondary school experience (a further 2 years). We would define this placing of students into new ability-based groups as retracking.

higher output, relative to mixed ability, and is significantly lower than the Retracking treatment, underlining the importance of repeated ability tracking systems. Critically, we find that increases in output are driven by high-ability subjects. Importantly, we do not find evidence that either of our two ability tracking systems have negative effects on low ability subjects: output is not significantly different relative to the mixed ability baseline. We show that these differences are not driven by differences in ability, and we are further able to show that the increase in effort applies to all individuals in the high-ability group. Our results provide robust evidence of the positive impact of ability tracking on effort provision.

3.2 Related Literature

3.2.1 Impact of Tracking

Tracking is shown to improve student outcomes at both the Secondary School (Epple *et al.*, 2002; Duflo *et al.*, 2011; Vardardottir, 2013) and University (Booij *et al.*, 2017) level. Duflo *et al.*, (2011) demonstrate that tracking improves outcomes for both high and low ability students, whilst those in the middle of the distribution improve regardless of their classroom assignment. Fu and Mehta (2018) also highlight the benefit of tracking in schools; if tracking was banned in schools, then the average student outcome would fall. The results in Duflo *et al.* (2011) are in part a result of targeted teaching; teachers are better able to target their teaching when a classroom contains similar ability students. In particular, when there are incentives for teachers with high student performance, outcomes of low ability students are likely to increase (Duflo *et al.*, 2011). These improvements in student outcomes persist past the study period, and outcomes remained significantly higher in tracked schools one year later (Duflo *et al.*, 2011).

In contrast, Vardardottir (2013) finds that increases in student outcomes caused by tracking are exclusive to high achieving students; being assigned to the high ability track causes an increase in both spring exam and end of year exam results by 0.32 and 0.47 SDs respectively (Vardardottir, 2013). Here, the difference between high and low ability tracked students is explained by peer effects, suggesting that students in the high ability group benefit from studying alongside strong peers (Imberman *et al.*, 2012; Vardardottir, 2013).

Whilst literature highlights the impact of peer-effects on outcomes in tracking systems, further mechanisms exist which are more difficult to measure in a real-world setting. Carbonaro (2005) is one of few studies to examine the impact of tracking on effort in the real-world with the understanding that measurement of effort is problematic, and largely rely on anecdotal reports made by teachers to identify effort provision. Despite measurement issues related to effort provision and student outcomes, they find suggestive evidence that tracking positively impacts the effort of high-ability students, but has no impact on effort for low-ability students. Our experiment allows us to abstract away from potential confounding mechanisms through which tracking systems work, such as peer-effects, and provides a reliable measure of effort provision.

The findings of Vardardottir (2013) suggest that tracking can also cause increases in inequality, since outcomes only improve for those in the high ability track. This is significant since those in the low ability tracks are more likely to be from disadvantaged backgrounds.[33] In line with this finding, other studies suggest that tracking increases inequality (Hanushek and Wossmann, 2006; Brunello and Checci, 2007; Betts, 2011; Bolukbas and Gur, 2020). For example, Hanushek and Wossmann (2006) draw attention to the negative impact of tracking on low-ability students, which arises from early years tracking decisions, in a cross-country study. In addition, Imberman *et al.* (2012) show that students from all ability levels are harmed when they are placed among low-ability peers.

Bolukbas and Gur (2020), inspect the difference between high- and low-ability tracked schools in Turkey, and find that the quality of education is significantly lower, and dropout rates are significantly higher, in low-ability schools. Interviewed teaching staff indicate that they feel it necessary to simplify the curriculum in low-ability schools; the simplification of the material is

[33] Evidence suggests that students in disadvantaged primary schools may have a lower academic self-concept (Antecol *et al.*, 2014); tracking may reinforce this low self-concept, as low-ability students are told that they are low-ability.

likely to result in lower student outcomes (Bolukbas and Gur, 2020). This finding highlights a further problem with tracking systems: staff who teach low-ability students may hold prejudice and preconceptions about those they are teaching, leading to lower effort and reduced quality of teaching. In line with this theory, Antecol *et al.* (2014) find that teaching performance in better in higher-achieving classrooms, and features higher quality lesson implementation.

Since tracking systems can negatively impact low-ability students, one potential solution is to place students into mixed ability classrooms through random assignment. In contrast with the findings presented above, Feld and Zolitz (2017) suggest that students in mixed ability classrooms benefit from studying alongside higher quality peers; this is important since low-ability students see the quality of their peer group decline under tracking (Epple, 2002). Despite this, some literature demonstrates how low achieving students can be harmed by studying alongside high ability students, especially when there are substantial interactions between classmates (Carman and Zhang, 2011; Carrell *et al.*, 2013; Yu, 2020). For example, an increase in the average achievement of classroom peers is shown to negatively influence the achievement of low-ability students (Antecol *et al.*, 2014); these results can be explained by a model in which student invidiously compare themselves.

Whilst there is considerable literature studying the impact of tracking systems, the impact of retracking is significantly understudied. In the real-world studying retracking is difficult, since tracking typically only takes place at specific stages of education.[34] In some cases retracking is made impossible by early years tracking decisions which dictate which subjects a student will study in the future. We aim to understand whether tracking systems improve the effort of students, and how effort is impacted by frequent tracking opportunities (retracking). Studying these policies in the real world is risky since it may directly impact student outcomes, and a lack of data availability means that a cross-country analysis is all but impossible. Therefore we propose to study this gap in the literature using a laboratory experiment.

3.2.2 Response to Feedback

Rank-order feedback is a key feature of our experiment, and provides subjects with information about their position within the group. The type of feedback which a student receives is important since it can have interacting effects with effort provision within a tracking system; in the literature, rank order feedback by itself is shown to improve effort provision (Charness *et al.*, 2014; Gill *et al.*, 2019).

Intrinsic motivation is a key factor in the link between rank order feedback and increased effort provision. Inherently, intrinsic motivation provides an incentive for people to complete the task as the end goal itself, rather than for a pecuniary benefit (Fishbach and Woolley, 2022). Charness *et al.* (2014) suggest that, even under a flat wage, intrinsic motivation is reinforced by relative performance feedback and therefore effort provision can increase. Further, evidence that a U-shaped effort response to feedback exists; those at the tail ends of the distribution of effort have the strongest response to feedback. That is to say, those at the top of the distribution aim to retain their high position whilst those at the bottom aim to improve their position (Gill *et al.*, 2019), and as a result both increase effort after being shown rank order feedback.

[34] As before, retracking in the UK occurs when a student enters secondary school and only takes place once more before they finish secondary school. Retracking may also take place as a result of specific decisions made by students, such as switching between public schools or switching between public and private school; however since they decisions are made by specific students it is difficult to reduce selection biases to a sufficient level.

By contrast, in a randomized field experiment, Brade *et al.* (2020) find that feedback on past performance only impacts future performance of those who performed above average. Brade *et al.* (2020) suggest that this is as a result of selective information processing, which leads only those who receive "good news" to update their beliefs after receiving feedback. Therefore, those with top ranks continue to update their beliefs, whilst those with bottom ranks do not. Similarly, Kuhnen and Tymula (2012) draw attention to the fight for dominance which occurs among those with top ranks, which improves productivity among those who already have a high level of output.

Despite these findings, some evidence suggests that feedback may have a negative impact of future effort. For example, Gill and Prowse (2012) show that second movers in a tournament setting exert less effort when first movers exert more effort; this is a discouragement effect from seeing that competition is more difficult (Gill and Prowse, 2012). Gurtler and Harbring (2010) also demonstrate a discouragement effect, which results in low ability individuals exerting less effort when they receive feedback that the ability gap between them and their competitor is large. In addition, feedback about relative performance in a multi-task setting is shown to cause an increase in effort in subjects who outperform others but decrease effort in subjects who underperform compared to others (Hannan *et al.*, 2013).

Finally, individuals are shown to actively avoid feedback when it may negatively harm their self-image. In Mobius *et al.* (2011), around 10% of subjects are willing to bid in order to avoid receiving noiseless feedback about their performance; Eil and Rao (2011) also find evidence that individuals with low expectations about their performance are willing to pay in order to avoid feedback. Whilst our setting does not provide an opportunity to pay to avoid feedback, these findings are important since they highlight the desire of participants to avoid feedback which does not align with their beliefs.

3.3 Experimental Design

3.3.1 Overview

We undertake an online lab experiment to examine the impact of ability tracking on effort, using a between-subjects design. We measure effort using output in a real effort task (described in the next section); subjects take part in 4 rounds of this task during a session, each round lasts for two minutes. Below we detail our baseline treatment (mixed ability groups), before exploring the manipulations to the grouping system which comprise our ability tracking (and retracking) treatments.

In our baseline, to replicate a mixed ability grouping system in the real world, subjects are randomly assigned to either a high wage or a low wage group after they take part in the first round of the task. In the real-world setting, which group a student is placed into during their schooling can impact school-leaving grades and ultimately life-time earnings. To replicate these differences, we incentivise being placed into the high wage group by paying a higher flat-wage for completing a round of the task. Note that subjects are paid a flat wage rather than a piece rate; the use of a piece rate risks crowding out any treatment effect, as subjects would be incentivised to maximise their performance. In our baseline, the existence of high and low wage groupings primarily serves to enable comparison to our tracking and retracking treatments.

In our mixed ability baseline, group placement is not related to subject performance or ability and is randomly assigned. Grouping decisions in both of our ability tracking treatments depend on performance in the real effort task (described below). In both treatments, subjects are placed into either the high or low wage group based on their performance in the first round of the real effort task.[35] If the subject's performance places them into the top performers in round 1, then they are placed into the blue (high-wage) group. However, if their performance places them in the bottom performers in round 1, they are placed into the green (low-wage) group. In the first of our ability tracking treatments (Tracked Groups), subjects remain in these groups for the remainder of the session regardless of their performance in subsequent rounds.

The second ability tracking treatment (Retracked Groups) differs as subjects are able to move between groups based on performance in each previous round. In this treatment, subjects move between groups in each of the subsequent rounds; this means that subjects may both move up (from the low wage group into the high wage group), or move down (from the high wage group into the low wage group) based on their performance.[36] A summary of the baseline and two treatments is shown below in Fig. 3.1.

We believe that the incentive structure used in the tracking and retracking treatments has considerable external validity for application to the school environment despite the differences in timescale over which the payments are received. In our experiment, subjects receive payment for each round of the task that they complete, and subjects with better performance receive higher payments (in the tracking and retracking treatments). In contrast, higher performance at high school does not often result in immediate payoffs; instead, students may earn higher lifelong wages or better career salaries in the future. Despite this, evidence suggests that high school students who are provided with more information on the benefits to lifelong earnings from post-compulsory

[35] Ties were broken at random.

[36] For example, a subject who has low output in round 1 is likely to be placed into the low wage group during round 2. However, if the same subject has a very high output in round 2, they may be moved into the high wage group during round 3.

education are more likely to continue into further education (McGuigan *et al.*, 2016). Furthermore, evidence suggests that students also consider differences in career earnings when deciding whether to take on loans for university (Boatman *et al.*, 2017).

In each of our treatments, subjects are asynchronously matched with 7 other subjects who have previously taken part in the study (this procedure is further outlined in later sections). Since there are two groups, and 8 subjects in total, we place 4 subjects into each group. In our baseline these groups are randomly assigned. In our tracking treatment, the score of the subject is ranked against 7 randomly chosen opponents (who have previously taken part in the study). If the subject is in the top 4 of this ranking they are placed into the blue group, and if they are in the bottom 4 of think ranking they are placed into the green group. The ranking system in our retracking treatment is the same, however this process of ranking the subject against the 7 opponents happens between each of the rounds, and subjects are placed into new blue and green group compositions based on scores in the past round.

Figure 3.1 – Experimental design

Treatment	Description	Number of Subjects
Mixed Ability Groups	Randomly assigned, mixed ability groups which remain fixed throughout.	62
Ability Tracked Groups	Groups based on 1st round performance, which remain fixed throughout.	62
Ability Retracked Groups	Groups based on previous round performance, which may change between rounds.	62

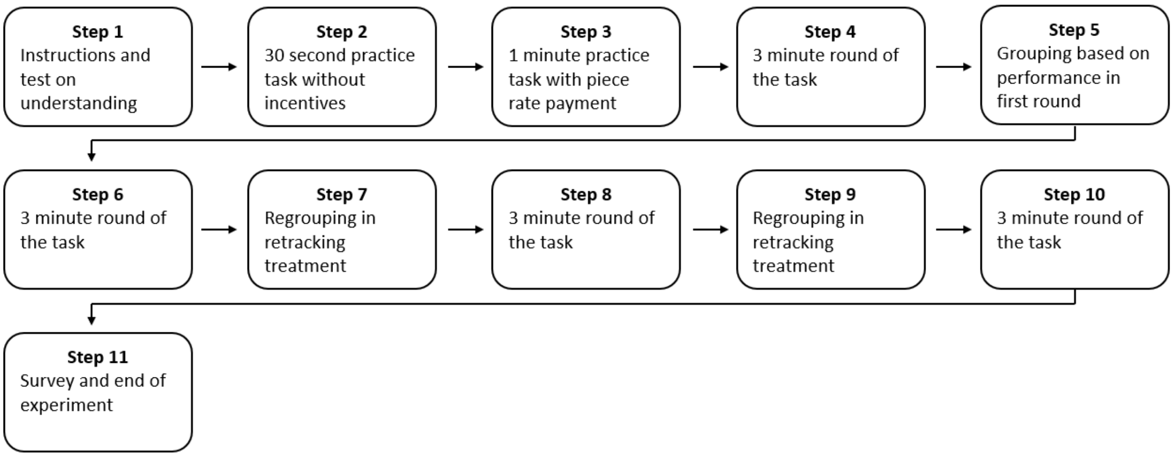
In all treatments, rank-order feedback is provided to subjects in between each round of the task; subjects only receive rank-order feedback about those in the same wage group. For example, those in the high wage group only receive information about their placement compared to others in the high wage group.

An outline of the order of tasks throughout the experiment is shown below (Fig. 3.2). Throughout the experiment, subjects take part in a real effort task (details below). Our experiment begins with an explanation of the task that will follow, during which we ask subjects simple questions about the experiment design to ensure that they understand the instructions. Next, subjects take part in a 30 second round of the task, allowing subjects to practice the task; no bonus payments are earned during this freeform practice. Once subjects understand the task, they take place in a 1 minute practice round of the task. We use the performance of subjects in this practice round as a measure of their ability at the task. In order for us to accurately measure ability during this practice round, we pay a piece rate per correct answer during this paid practice round.[37] Subjects are informed of their performance during this practice round, along with how much they have earned. Then, subjects perform the task for one round (which lasts for 3 minutes). As previously described, based on their performance in this first round of the task, subjects are placed into either the blue (high-wage) or green (low-wage) group. Once these groups have been formed,

[37] By using a piece rate we ensure that extrinsic motivation is high and therefore that subjects exert maximum effort and reveal their true ability.

subjects take part in a further three rounds of the task. In our baseline and tracking treatment, they remain in the same groups throughout these rounds of the task. In our retracking treatment, subjects may swap between groups in between each of these rounds. Between each round of the task, we elicit beliefs about group placement in the following round (based on their performance in the previous round). Once subjects have completed these rounds of the task, they take part in a survey. Our survey metrics include: a general confidence measure (Ortoleva and Snowberg, 2015); a portion of the intrinsic motivation inventory (McAuley et al., 1988); an alternate cognitive reflection task (Thompson and Oppenheimer, 2016); calculator use in each round; and a series of demographics collected from the Prolific.co platform (including gender, age, ethnicity).

Figure 3.2 – Experiment outline



Step 1 Instructions and test on understanding
Step 2 30 second practice task without incentives
Step 3 1 minute practice task with piece rate payment
Step 4 3 minute round of the task
Step 5 Grouping based on performance in first round
Step 6 3 minute round of the task
Step 7 Regrouping in retracking treatment
Step 8 3 minute round of the task
Step 9 Regrouping in retracking treatment
Step 10 3 minute round of the task
Step 11 Survey and end of experiment

3.3.2 Effort Task

Our task is a simple addition task (based on Niederle and Vesterlund, 2007), in which subjects must add three 2-digit numbers together. Our task differs from Niederle and Vesterlund (2007) since we require subjects to add fewer numbers together in each question.[38] We differ in this way in order to ensure a wider distribution of output in each round; our task allows those who excel at the task to complete a greater number of problems since our problems are easier to solve. In addition, each round in our session lasted for 3 minutes compared with the 5 minutes allowed in Niederle and Vesterlund (2007).

In order to move on to the next problem a subject must submit a numeric answer in the box provided on screen (shown below in Fig. 3.3). Subjects are not required to provide the correct answer, and move onto the next three 2-digit number problem on screen regardless of whether their previous answer was correct. Subjects are not informed of how many correct answers they have submitted during the task, nor are they informed of whether an answer is correct or incorrect upon submission. Feedback is only provided between rounds as described above.

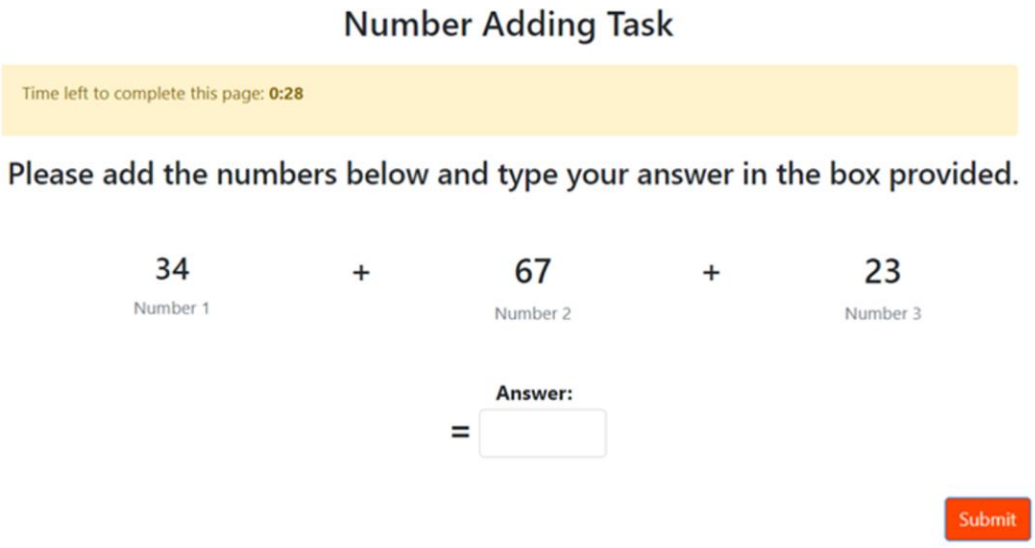
[38] The addition task used in Niederle and Vesterlund (2007) requires subjects to add five 2-digit numbers together before submitting their answer in a box on screen. In contrast, our task requires subjects to add three 2-digit numbers together.

Subjects were informed that they should "*try as hard as possible*", and therefore should complete as many questions (sets of three 2-digit numbers) as possible in each 3 minute round. The performance of each subject is measured as their output in the task. As explained earlier, in our tracking and retracking treatments output in round 1 determines the group which a subject is placed into (either the high- or low-wage group). Further in retracking, output in each round determines the group placement in the subsequent round.

The 2-digit numbers shown to subjects during the task were randomly selected for each question in the task, and independently for each subject. However, numbers were randomly selected between 10-99 and therefore questions are all of similar difficulty. The timer was shown on screen, counting down from 3 minutes. Once the timer had run out, subjects were immediately taken to a belief elicitation screen followed by a feedback screen. On the feedback screen, subjects were only told how many questions they, and the others in their group, answered correctly; subjects were not informed of how many questions they answered incorrectly.

Since our experiment is conducted asynchronously, and using an online platform, we are unable to observe whether subject use a calculator whilst taking part in the task. In order to capture information about calculator use during the task, we ask subjects to truthfully report whether they used a calculator in our survey. Note that subjects have no incentive to lie about calculator use, and there are no punishments or incentives related to answers to survey questions.

Figure 3.3 - Screenshot of the effort task



Number Adding Task

Time left to complete this page: **0:28**

Please add the numbers below and type your answer in the box provided.

3.3.3 Experimental Procedure

Our experimental sessions were run on the Prolific.co platform between 28th October – 12th November 2021. Whilst we understand that experimental data collected from online platforms can have several distinct issues, we are able to gain a more diverse subject pool by undertaking an online experiment. By hosting our sessions on Prolific.co we are not only gain access to a more diverse subject pool (UK citizens between 18-99), but we are also able to collect a greater number of data points.[39]

Subjects took part in the experiment asynchronously, i.e. each subject took part in isolation from other subjects. We provide rank-order feedback to subjects whilst they are taking part in the asynchronous session; since the session is asynchronous, subjects are not ranked against others who are also taking part in a session at the same time, but instead against data collected before their session. In order to generate the rank-order style feedback used in our grouping mechanism and provided to subjects between rounds, a series of pilot studies were run (similar to Buchan *et al.*, 2011). Each of these pilot studies aims to match the incentives of our full study as closely as possible.

In our first pilot, subjects took part in the task for four rounds. They were not compared to any other subjects, took part in the task independently from other subjects, were randomly placed into either the high- or low-wage group by themselves, and received no feedback other than their own performance in each round. By maintaining the high- and low-wage group structure, we created an environment similar to our full experiment. We utilised the data collected from our first pilot in order to construct a performance leader board.

In our second pilot, subjects are informed that their performance will be compared to the performance of past participants; the leader board generated using data from pilot 1 is used as a comparison. To avoid issues of deception, we made clear that the comparison subjects (from pilot 1) faced a similar but not identical environment, and explicitly detailed all differences. During pilot 2, we were able to introduce our treatments since subjects were able to be grouped based on comparisons with subjects from pilot 1.

Those subjects who participated in pilot 2 faced similar conditions to our full experiment, with the exception of the differences between their environment and the environment of those who made up the leader board. During pilot 2 we collected data for the baseline (mixed ability groups) as well as our tracking and retracking treatments. Whilst the data from pilot 2 could be used to generate the leader board for the full study, we go further to minimise the potential impacts of generating leader-board data in this way by conducting a third pilot study.

In pilot 3 subjects face an almost identical environment to subjects in the full study. The leader board which they face is generated using subjects from pilot 2, who faced the same form of feedback, and identical pay structure and grouping decisions. Since they constitute a good comparison, our final study leader board was generated using performance from subjects who took part in pilot 3.

[39] Data collection on Prolific.co can be considerably cheaper than an in-person laboratory. Subjects take part in the session asynchronously, and therefore are not required to wait for subjects before being able to continue through the session; therefore sessions take considerably less time online, costing less.

Pilot study data was collected between 28th October – 4th November 2021. 10 data points were collected during pilot 1, 30 data points were collected during pilot 2 (10 in each treatment), and a further 30 data points were collected during pilot 3 (10 in each treatment). As subjects took part in the experiment asynchronously their performance was compared to a leader board which was generated from 5 randomly selected pilot 3 subjects from the same treatment.

Our full experiment sessions took place between 8th – 12th November. Sessions were run between 9am – 5pm each day, and subjects were randomly assigned to one of the three treatments upon opening the experiment link. Based on prior power calculations, 62 subjects were collected for each of the three treatments: a total of 186 subjects. Demographic characteristics of the subjects follows in the upcoming *data* section.

3.3.4 Data

Based on prior power calculations, 62 data points were collected per treatment, for a total of 186 subjects. Demographic information was collected from Prolific.co, and includes: age; gender; degree holder; student status; and employment status. Table 3.1 (below) includes results of comparisons between these demographics, along with our measure of ability, performance in the cognitive reflection test and intrinsic motivation inventory, and calculator use throughout the experiment. Since subjects were randomly assigned to our control and treatments, we should expect no significant differences in the distribution of demographics.

In Table 3.1 we see that there are sporadic significant differences between variables when comparing specific treatments. Initially we compare our control (randomly assigned groups) to tracked groups in column 1; we see significant differences in working status, age, and gender. Subjects in tracking were more likely to be working (means = 0.645 (control) vs 0.774 (treatment)), were older (33.4 vs 35.4) and also more likely to be male (0.387 vs 0.508) than those in our baseline.

In column 2 we once again compare our control, but this time to retracked groups. Here, we see significant differences in ability (5.274 vs 5.694), IMI (1.944 vs 2.058) and confidence survey responses (0.274 vs 0.194), and age (33.4 vs 35.8). Subjects in retracking have higher ability, are more motivated but have less general confidence, and are older than those in our control.

Finally, we compare tracking to retracking. Here we only see a significant difference in gender (0.508 vs 0.355), where subjects in tracking are significantly more likely to be male than those in retracking. Whilst this may indicate that we do not achieve full random assignment into treatments, we ensure that our results are robust to demographic differences by including the full suite of variables discussed here as control variables in our regressions.

Table 3.1 – Balance Table Including Demographic and Survey Response Variables

	Control	Tracked Groups	Retracked Groups	Control vs Tracked	Control vs Retracked	Tracked vs Retracked
		Group Mean			T-Test P-Value	
	(1)	(2)	(3)	(4)	(5)	(6)
Ability	5.274	5.500	5.694	0.152	0.011**	0.210
	(1.872)	(1.627)	(1.805)			
Average Calculator Use	0.391	0.383	0.464	0.844	0.085*	0.056*
	(0.456)	(0.456)	(0.480)			
CRT Average	0.435	0.478	0.435	0.119	1.000	0.108
	(0.279)	(0.332)	(0.259)			
IMI Average	1.944	1.995	2.058	0.309	0.018**	0.147
	(0.614)	(0.511)	(0.452)			
Confidence	0.274	0.258	0.194	0.685	0.034**	0.086*
	(0.447)	(0.438)	(0.396)			
Working	0.645	0.774	0.726	0.002***	0.053*	0.214
	(0.479)	(0.419)	(0.447)			
Student	0.263	0.288	0.224	0.548	0.331	0.113
	(0.441)	(0.454)	(0.418)			
Degree	0.597	0.565	0.597	0.468	1.000	0.468
	(0.492)	(0.497)	(0.492)			
Age	33.371	35.419	35.839	0.035**	0.019**	0.701
	(10.290)	(11.314)	(12.915)			
Male	0.387	0.508	0.355	0.007***	0.458	0.001***
	(0.488)	(0.501)	(0.479)			
N	248	248	248			

Notes: Means of demographic and control variables (1-3), standard errors in parentheses. Two-tailed t-test p-values with significance for: control compared to tracked (4); control compared to retracked (5); tracked compared to retracked (6). *** p<0.01, ** p<0.05, * p<0.1.

3.4 Results

The experiment is designed to identify the impact of tracking systems on effort. Note that while the literature on evaluating the effects of tracking on student outcomes, such as test scores (Duflo *et al.*, 2011; Lavy *et al.*, 2012), we use the lab to focus on one outcome: student effort. Student outcomes are naturally a function of many different factors including peer effects, confidence, effort, learning, teacher impact, etc. The lab allows us to hone in on one important input, effort. At the same time, we abstract away from other relevant factors inherent to the classroom, which may generate additional effects. In this section, we present the results of our analysis, focusing on the impact of Tracking and Retracking on effort. Subjects undertake effort in 4 rounds of a real effort task. The control condition randomly assigns individuals to a high-wage or a low-wage group after the first round (they keep their assignment for all subsequent rounds). The Tracking treatment assigns individuals to a high-wage, or a low-wage group based on performance in round 1 (they keep their assignment for all subsequent rounds). The Retracking condition assigns individuals to a high- or a low-wage group based on performance in rounds 1, 2, and 3.

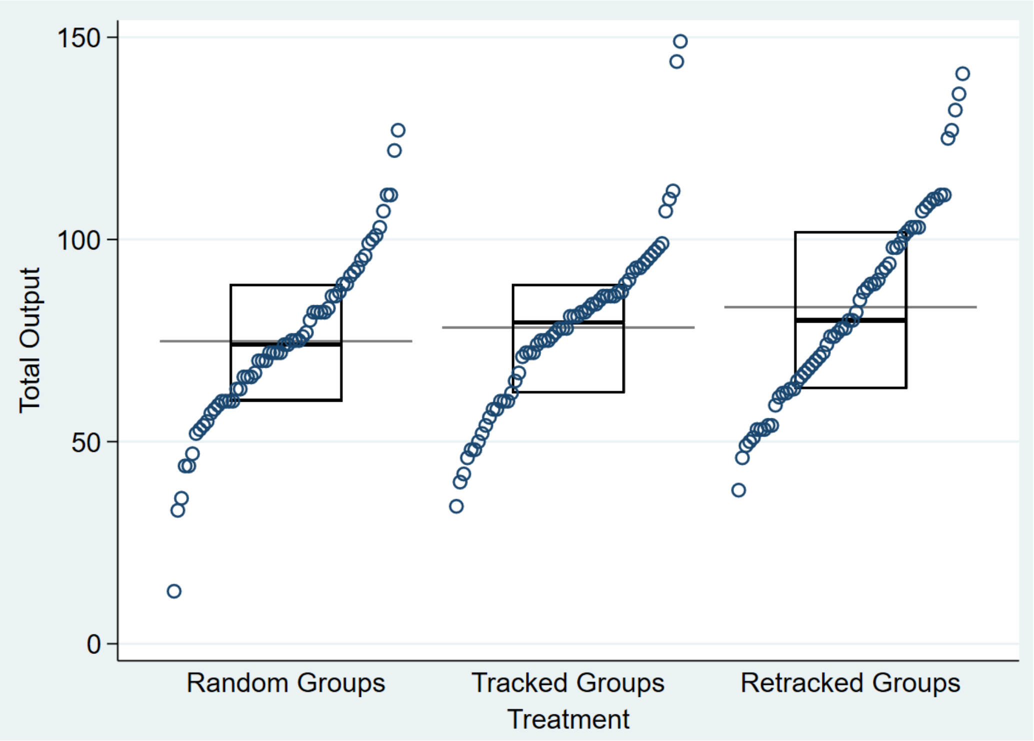
We find that Retracking (in particular) increases effort, and that this increase comes from individuals that are placed in the high wage group. Importantly, we find no evidence of detrimental effects of tracking on individuals placed in the low wage group.[40] In addition, we conduct some exploratory analysis to identify mechanisms. Specifically, we ask whether the increase in effort is driven by subject rankings. Gill et al. (2019) identify a first-place loving effect where those individuals ranked first work harder to maintain their rankings. We find that our results are not driven by any effects of rankings: subjects placed in the high-wage group increase their effort regardless of where they get ranked.

3.4.1 Effects of Tracking on Effort

Our first bit of analysis focuses on the impact of the tracking treatments on output (effort). We begin with measuring total output as the total number of correct answers in the number-adding task across all rounds. We estimate the impact of our two types of tracking systems on output: (1) baseline; (2) Tracked; and (3) Retracked. Figure 4 displays a jittered strip-plot of total output by treatment across all 4 rounds of the task. We find that output is significantly greater in the Retracked treatment (two tailed t-test: $p<0.01$) but not in the Tracked treatment ($p=0.37$), relative to the control. Further to this, we find that output is significantly higher in the Retracked treatment relative to the Tracked treatment ($p<0.05$). This shows that tracking systems affect output, but only when Retracking is available, rather than when students are tracked once.

[40] Note that since our baseline contains mixed ability groups, the effects we detail above may simply be mechanical. That is, since tracking sorts individuals on ability, they would be of higher average ability in tracking relative to the baseline. To rule this out as a potential explanation of our findings, we offer a series of robustness tests to show that the increase in output stems from differences in effort, not ability.

Figure 3.4 - Jittered strip plot of total output by treatment



Notes: Total output is the sum of output in rounds 1-4. Box plots show 25th-75th percentiles and median, thin added line shows mean. Each circle represents the total output of one subject.

One important reason for the differences in output may be due to subject ability. We control for this, as well as other factors, in the regressions below. Model 1 in Table 3.2 estimates the impact of the Tracking and Retracking treatments on total output. Model 2 includes control variables for game specific differences, including ability (as measured by performance in the incentivized practice round), calculator use during the task, attentiveness (as measured by the average score in the cognitive reflection test), and task-based motivation (intrinsic motivation inventory). Finally, in Model 3 we control for our full suite of demographics (gender, age, education, and employment status).

Table 3.2 confirms the findings from the strip plot: total output is significantly greater in the Retracked treatment, relative to the control (mixed ability; $p<0.05$); in our retracking treatment, output was close to 5.7 questions more over the course of the experiment than those in the baseline. Additional results confirm that total output is significantly greater in the Retracked treatment relative to the Tracked treatment ($p<0.1$ under Model 3 – Appendix Table 3.1*)*. Note that in our regression results, there are no significant differences between total output in Tracked systems relative to the control. In addition to the effects of treatment, Table 3.2 further shows significantly greater total output by subjects with higher ability, by those using calculators, those that were more attentive (higher scores on the cognitive reflection test), and are more intrinsically motivated. These results are not surprising, since greater motivation, ability and calculator use all enable a subject to perform better in the real-effort task.

Table 3.2 – The impact of tracking on output

Dependent variable: Total Output	(1) Model 1	(2) Model 2	(3) Model 3
Tracked Groups	3.484	0.593	0.841
	(4.043)	(2.432)	(2.555)
Retracked Groups	8.468**	5.038**	5.739**
	(4.043)	(2.448)	(2.549)
Ability		8.243***	8.131***
		(0.593)	(0.612)
Calculator Use		12.36***	11.05***
(Use = 1)		(2.183)	(2.363)
Cognitive Reflection Task Average		12.49***	10.77***
		(3.589)	(3.818)
Motivation		7.602***	7.013***
		(1.910)	(1.984)
Working Status			0.593
(Working = 1)			(2.482)
Student Status			0.884
(Student = 1)			(2.688)
Has a Degree			2.139
(Degree = 1)			(2.156)
Age			-0.090
(In Years)			(0.103)
Gender			0.412
(Male = 1)			(2.269)
Constant	74.77***	3.558	7.595
	(2.859)	(5.071)	(6.832)
Observations	186	186	173
R-squared	0.024	0.657	0.664

Notes: OLS regression with dependent variable as total (sum of) output across 4 rounds of real effort task; the main independent variable is a categorical dummy representing the treatment group. Standard errors in parentheses. (*** p<0.01, ** p<0.05, * p<0.1).

The table above shows that output is higher for subjects in the Retracked treatment relative to the control, and to the Tracked treatment. Overall, the Retracked treatment increases subject effort (close to 5.7 questions throughout the experiment), though we do not observe a similar effect for the Tracked treatment. However, recall that in the Tracked treatment, subjects remain in the same ability group beyond on the first round, and hence have lower incentives to continue to exert effort beyond the first round. The next subsection focuses on output across rounds.

3.4.2 The Evolution of Output

The next question is how output differs over rounds. The Tracked treatment used output in the first round to assign people to the high and low wage groups for the rest of the treatment. The Retracked treatment used output in each of the first three rounds to assign people to the high and low wage groups for the next round. The baseline condition gives subjects little incentive to exert effort, since effort has no bearing on assignment to the high or low wage groups (and hence, on earnings). We might reasonably expect effort to be higher (relative to the control) in the first round for both the Tracked and Retracked treatments, and then effort to be higher in just the Retracked treatment for rounds 2 and 3. Hence, effort overall may be higher in the Retracking treatment simply because a greater proportion of the rounds are relevant for earnings. To test to see if this is the case, we study output in each round separately.

In Table 3.3, we report OLS regressions using the same specification as in model 3 (all controls) in Table 3.2 above but run separately for each round of the task. Model I reports results for round 1 and so on. The results confirm the findings earlier, across all rounds where effort is incentivized by the Retracking treatment, subject output is significantly higher (*p<0.05*). We see that subjects in the Retracked treatment answered close to 1.5 questions in the first round, 1.7 questions in the second round, and 2.2 questions in the third round, more than subjects in the baseline treatment. We see a significant and positive difference in output for all rounds bar the fourth in our Retracked condition, relative to the baseline.

Table 3.3 – Treatment Effects on Output (by Round)

Dependent variable: Output in Round	(1) Round 1	(2) Round 2	(3) Round 3	(4) Round 4
Tracked Groups	0.658	0.371	0.660	-0.848
	(0.661)	(0.738)	(0.729)	(0.752)
Retracked Groups	1.545**	1.744**	2.202***	0.247
	(0.660)	(0.736)	(0.727)	(0.750)
Ability	2.132***	2.123***	1.852***	2.024***
	(0.158)	(0.177)	(0.174)	(0.180)
Calculator Use	2.343***	2.349***	2.953***	3.408***
	(0.611)	(0.683)	(0.674)	(0.695)
CRT Average	2.170**	2.545**	3.191***	2.869**
	(0.988)	(1.103)	(1.089)	(1.124)
IMI Average	1.563***	1.690***	1.205**	2.556***
	(0.513)	(0.573)	(0.566)	(0.584)
Working Status	0.136	0.506	0.0673	-0.115
	(0.642)	(0.717)	(0.708)	(0.730)
Student Status	-0.372	0.467	0.532	0.257
	(0.696)	(0.777)	(0.767)	(0.791)
Has a Degree	0.249	0.518	0.514	0.857
	(0.558)	(0.623)	(0.615)	(0.635)
Age	-0.029	-0.027	-0.033	-0.002
	(0.027)	(0.030)	(0.029)	(0.030)
Gender	0.053	-0.298	0.303	0.355
	(0.587)	(0.656)	(0.647)	(0.668)
Constant	1.384	1.479	3.991**	0.741
	(1.768)	(1.974)	(1.948)	(2.011)
Observations	173	173	173	173
R-squared	0.649	0.609	0.585	0.606

Notes: OLS regression with dependent variable as output in each of the 4 rounds of real effort task; the main independent variable is a categorical dummy representing the treatment group. Standard errors in parentheses. (*** p<0.01, ** p<0.05, * p<0.1).

We find that Retracking increases effort, but find no similar impact of the Tracked treatment, relative to the baseline. Moreover, effort in the Retracking treatment is significantly higher than in the Tracking treatment across rounds 2 and 3 (*$p<0.1$* and *$p<0.05$* respectively), but not in round 1 (*$p=0.183$*). There are no significant differences in rounds 1 and 4 output between our Tracking and Retracking treatments; this is expected, since incentives are identical between the two conditions in rounds 1 and 4, but differ in rounds 2 and 3. Overall, we conclude that tracking systems increase subject effort, but these effects are contained wholly within the rounds where they are payoff relevant. The next section looks at differences in effort between the high and low ability groups.

3.4.3 Effort Differences by Group Assignment

A natural split in our data is the ability groups which subjects were placed into. During the session, each subject was placed into either the low-wage (green) group or the high-wage (blue) group starting at round 2.[41] In the baseline, both the low- and high-wage groups contain mixed ability subjects. However, in the two tracking treatments the group composition is determined by performance: the top 3 performers (in round 1) are assigned to the high-wage group, while the bottom 3 performers are assigned to the low-wage group.[42]

In Table 3.4 we present OLS regressions for two wage groups separately (sub-samples: high-wage (columns 1-3) and low-wage (columns 4-6) group members).[43] In the high-wage group, we find that effort in rounds 2, 3 and 4 is significantly greater in the Retracked treatment relative to the control; close to 2.5 questions extra in rounds 1 and 3, and 3.8 questions in round 2. Further testing also highlights differences in effort between the Tracked and Retracked treatment for the high-wage group (Appendix Table 3.2). We find no significant differences in effort in the Tracked treatment relative to the control.

Interestingly, effort is nominally, but not significantly, lower in the tracking treatments relative to the control, for the low-wage groups. In rounds where effort is payoff relevant for the Retracking treatment, effort is not significantly different across the three conditions. This suggests that subjects assigned to the low-wage group do not reduce their effort in response to treatment.[44] Hence, the overall positive impact of retracking on output stems from subjects assigned to the high-wage group increasing their effort, but with no significant decrease observed in the low-wage group. Overall, this suggests that one additional benefit of tracking systems (specifically Retracking) is that it improves overall outcomes through an increase in effort from high-ability individuals, with no appreciable decrease in effort from low-ability individuals.

[41] Since subjects are not placed into ability groupings until the start of round 2, we exclude round 1 from our ability sub-group analysis.

[42] There are inherent differences in group composition between the randomly assigned treatment and the Tracking and Retracking treatments. The randomly assigned treatment groups contain mixed ability subjects, therefore the top performing subject could still be randomly placed into the low-wage group. This cannot happen in other treatments since groups are created based on performance. Whilst our estimates provide some control for this by using our measure of ability as a secondary independent variable, we provide further robustness checks on this issue later in the section.

[43] Analysis for round 1 is excluded, since subjects had not yet been placed into groups.

[44] Note that effort is significantly lower in both the Tracked and Retracked treatments, relative to the control, in round 4; this is caused by an increase in round 4 effort in the control, rather than a decrease in effort in either condition (Appendix Figure 2). It is unclear why effort is increasing in the control in round 4 however.

Table 3.4 – Treatment Effects on Effort (by Round and Group)

	(1)	(2)	(3)	(4)	(5)	(6)
		High-Wage Group			Low-Wage Group	
Output in Round	Round 2	Round 3	Round 4	Round 2	Round 3	Round 4
Tracked Groups	0.667	1.580	0.298	-0.914	-1.022	-2.895***
	(1.073)	(1.093)	(1.059)	(0.944)	(0.919)	(0.973)
Retracked Groups	2.423**	3.827***	2.570**	-0.400	-0.480	-2.943***
	(1.079)	(1.126)	(1.109)	(0.957)	(0.893)	(0.924)
Ability	1.942***	1.533***	1.827***	1.914***	1.514***	1.558***
	(0.265)	(0.279)	(0.263)	(0.260)	(0.249)	(0.254)
Constant	5.894*	6.353**	0.540	3.791	6.756***	5.529**
	(2.990)	(3.091)	(2.965)	(2.631)	(2.388)	(2.542)
Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	99	97	94	74	76	79
R-squared	0.574	0.556	0.611	0.598	0.544	0.585

Notes: OLS regression with dependent variable as output rounds 2, 3 and 4 of the real effort task; the main independent variable is a categorical dummy representing the treatment group. High-wage group analysis is shown in columns 1-3, and low-wage group analysis is shown in columns 4-6. Standard errors in parentheses. Controls included are the full suite of controls included in Model 3 (Table 3.2). (*** p<0.01, ** p<0.05, * p<0.1).

3.4.4 Additional Robustness Checks

Note that the differences we highlight above (higher output by individuals in the high-wage group in the Retracking treatment) may simply be mechanical, in that the high-wage group contains high ability individuals in the tracking treatments, and mixed ability subjects in the control. Despite controlling for ability, it may be argued that our findings are the result of different grouping systems. We tackle this issue by, ex-post, ranking subjects from the control and artificially constructing groups based on performance. Specifically, we construct artificial high-wage groups in the control by assigning high ability subjects to the high-wage (blue) group. This means that the high-wage group in the control now contains high-ability individuals (whereas earlier, it contained mixed ability individuals). Columns 1-3 in Table 3.5 present the results of this analysis. We find a significant impact of the Retracking treatment on effort, relative to the control, even when subjects in the control condition are now all high ability.

The table below shows that the results of the effects of Retracking on effort remain the same when redefining the groups in the control condition to rule out the possibility that lower ability individuals assigned to the high-wage group are driving our results. One further issue is that in the control condition, individuals are randomly assigned to high- or low-wage groups, meaning that some high ability individuals get assigned to a low-wage group. In columns 1-3 we resort them based on ability such that high ability individuals are now placed into the high-wage group regardless of their actual wage. Since the low wage could have an effect on output, we next construct high-wage groups based on just those subjects that were both high ability, and assigned to a high-wage group. This reduces the number of observations in the control (and thus, power). In columns 4-6 of Table 3.5, we present the results of this analysis. We find that the effect of Retracking remains positive and significant, indicating that the effect of Retracking on effort is not explained by the way the groups were constructed in the control. Finally, we undertaken an identical exercise for the low-wage group, and confirm our findings of no significant effects of either of the tracking treatments on effort. Hence, the positive effects of tracking on the high-wage group are not offset by a negative effect on the low-wage group, yielding a positive impact of tracking on effort overall. The results for the low-wage group can be found in Appendix Table 3.3.

Table 3.5 – Treatment effects on effort (by round) – High-wage group only

	(1)	(2)	(3)	(4)	(5)	(6)
		Resorted Groups			Resorted Groups and Correct Wage	
Output in Round	Round 2	Round 3	Round 4	Round 2	Round 3	Round 4
Tracked Groups	0.440	1.205	0.306	0.014	1.580	0.298
	(1.016)	(0.960)	(0.929)	(1.168)	(1.093)	(1.059)
Retracked Groups	2.244**	3.430***	2.617***	1.765	3.827***	2.570**
	(1.020)	(0.245)	(0.976)	(1.173)	(1.126)	(1.109)
Ability	1.667***	1.621***	1.783***	1.702***	1.533***	1.827***
	(0.251)	(0.245)	(0.233)	(0.277)	(0.279)	(0.263)
Constant	5.088	5.358*	0.371	7.389**	6.353**	0.540
	(3.348)	(2.838)	(2.723)	(3.641)	(3.091)	(2.965)
Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	97	105	102	89	97	94
R-squared	0.469	0.570	00.619	0.447	0.556	0.611

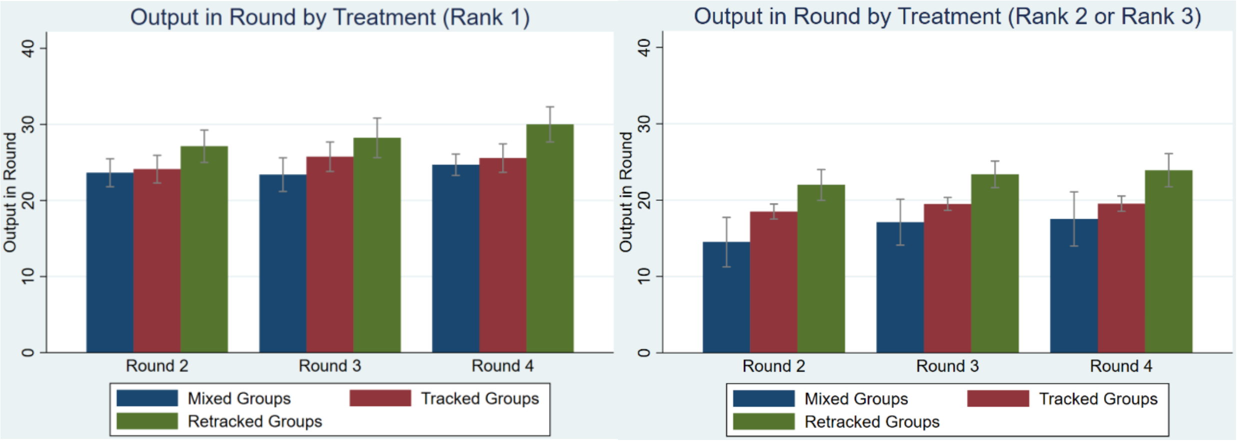
Notes: OLS regression with dependent variable as output rounds 2, 3 and 4 of real effort task, and the main independent variable as a categorical dummy representing the treatment group. Ex-post resorted high-wage group analysis is shown in columns 1-3, and ex-post resorted and restricted (only those also from the high-wage group during the session) high-wage group analysis is shown in columns 4-6. Standard errors in parentheses. Controls include the full suite of controls included in Model 3 (Table 1). (*** p<0.01, ** p<0.05, * p<0.1).

3.4.5 A Note on Mechanisms

We find that effort is significantly greater in the high-wage group under the Retracking treatment. We now turn our attention to a potential mechanism. We explore the presence of a first-place loving effect (Gill *et al.*, 2019) to see if being ranked in first place drives the increase in effort amongst our subjects. We restrict our analysis to the high-wage groups. In Figure 3.5.a we display effort in each round for subjects who were ranked 1st in the high-wage group in the previous round. We show that effort is significantly higher in rounds 2, 3 and 4 for those ranked first in the high-wage group in a Retracking system (*$p<0.05$*;*$p<0.01$*;*$p<0.05$*). compared with the control. We confirm this result using regressions (reported in Appendix Table 3.4 columns 1-3).

As a result of our asynchronous design, estimation of a treatment effect for those ranked 2nd or 3rd in the high-wage is biased due to small sample size.[45] In Figure 3.5.b we show the effort exerted by a pooled sub-sample of subjects who were either ranked 2nd or 3rd in the high-wage group. Here we find significantly greater output for those ranked 2nd or 3rd in the high-wage group under Retracking systems in rounds 2 and 3 (*$p<0.05$; $p<0.05$*), relative to the control (Appendix Table 3.4, columns 4-6). Overall, we find that all subjects in the high-wage group increased their effort in response to the Retracking treatment, not just those ranked in first place.

Figure 3.5.a (left) and Figure 3.5.b (right) – Output in rounds 2, 3 and 4 for those shown rank 1 in previous round (3.5.a) and those shown rank 2 or 3 in the previous round (3.5.b).



Notes: Mixed ability group treatment (blue), ability tracked groups (red), ability retracked groups (green). Error bars show 95% confidence intervals.

[45] Participants were ranked compared to a randomly selected, but previously generated, leaderboard. Therefore, we do not have a balanced panel of subjects across ranks.

3.5 Conclusion

Ability tracking is a prominent feature of education systems around the world, including more than 95% of classrooms in the USA (Fu and Mehta, 2018). Tracked classrooms may allow teachers to target the material to a group of similar students (Duflo *et al.*, 2012), and additionally benefits high-ability students who study among others of high-ability (Imberman *et al.*, 2012; Vardardottir, 2013). Despite this, tracking is shown to increase inequality as low-ability students suffer whilst studying alongside low-ability peers (Hanushek and Wossman, 2006; Imberman *et al.*, 2012). These negative effects are less likely to occur in a mixed ability group system where low-ability students study among higher-ability peers (Feld and Zoltz, 2017). Nevertheless, evidence on the effects of ability tracking on student effort is thin. To study this, and to control for confounding effects, we conduct a lab experiment where we replicate the core aspects of ability tracking: group assignment based on ability. We implement two types of ability tracking, one where subjects are tracked once (Tracked Groups), and another where subjects are tracked multiple times. We find that the effects of ability tracking on effort are positive for high ability individuals, but find no evidence of any negative impacts for low ability individuals. Hence, the overall effects of ability tracking (and specifically, Retracking) are positive. This corresponds to inequality, however, given that high ability subjects fare better when placed with their peers, even when peer effects are not possible. This occurs because individuals that are placed in high ability groups increase their effort levels to maintain their place in the high ability group.

The literature goes a long way in identifying peer effects as the main mechanism through which tracking brings about change. However, we cleanly identify another mechanism which is difficult to measure in the real-world: effort. In an experimental setting, we examine the impact of tracking, and retracking, on effort. Our results show that retracking provides a pareto-improvement in effort provision, over both our mixed ability baseline and our tracking treatment. We go further to provide evidence that the effort improvements in the retracking condition are driven by subjects in the high-wage group. Finally, through an ex-post re-sorting of subjects in our mixed ability baseline, we demonstrate that increases in effort provision are not as a result of the group formation. Our results carry important implications for structuring tracking systems and serve as a useful guide for policymakers.

3.6 Appendix

Appendix Table 3.1 – The Impact of Retracking Systems on Total Effort

Total Output	(1) Model 1	(2) Model 2	(3) Model 3
Retracked Groups	4.984	4.274*	4.847*
	(4.133)	(2.529)	(2.691)
Ability		9.104***	9.065***
		(0.755)	(0.788)
Calculator Use		12.13***	11.40***
		(2.739)	(3.072)
CRT Average		12.54***	11.10**
		(4.347)	(4.660)
IMI Average		7.406***	6.061**
		(2.664)	(2.807)
Working Status			1.187
			(3.392)
Student Status			0.0499
			(3.577)
Has a Degree			0.812
			(2.783)
Age			-0.138
			(0.126)
Gender			1.262
			(2.917)
Constant	78.26***	-0.0775	6.661
	(2.922)	(6.754)	(9.676)
Observations	124	124	116
R-squared	0.012	0.650	0.660

Notes: OLS regression with dependent variable as total output across 4 rounds of real effort task and the main independent variable as a categorical dummy representing the treatment group. Our Tracked condition is taken as the baseline for our Retracked condition. (*** p<0.01, ** p<0.05, * p<0.1).

Appendix Table 3.2 – Effort in Rounds 2, 3 and 4 for the High-Wage (1-3) and Low-Wage (4-6) Group

	(1)	(2)	(3)	(4)	(5)	(6)
	High-Wage Group			Low-Wage Group		
Effort in Round	Round 2	Round 3	Round 4	Round 2	Round 3	Round 4
Retracked Groups	1.855**	2.185**	2.262**	1.033	0.953	0.151
	(0.903)	(0.982)	(0.989)	(1.216)	(1.040)	(1.226)
Ability	1.903***	1.414***	1.825***	1.779***	1.246***	1.706***
	(0.322)	(0.367)	(0.337)	(0.423)	(0.379)	(0.404)
Constant	7.345*	13.53***	-0.363	5.995	6.617*	4.277
	(4.028)	(4.479)	(4.258)	(4.437)	(3.507)	(3.921)
Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	73	71	68	43	45	48
R-squared	0.489	0.424	0.542	0.456	0.396	0.415

Notes: OLS regression with dependent variable as output rounds 2, 3 and 4 of real effort task, and the main independent variable as a categorical dummy representing the treatment group. Our Tracked conditions is taken as the baseline for our Retracked conditions. High-wage group analysis is shown in columns 1-3, and low-wage group analysis is shown in columns 4-6. Standard errors in parentheses. Controls include the full suite of controls included in Model 3 (Table 1). (*** p<0.01, ** p<0.05, * p<0.1).

Appendix Table 3.3 – Output in Rounds 2, 3 and 4 for Ex-Post Resorted Low-Wage Group

	(1)	(2)	(3)	(4)	(5)	(6)
		Resorted Groups			Resorted Groups and Correct Wage	
Output in Round	Round 2	Round 3	Round 4	Round 2	Round 3	Round 4
Tracked Groups	-0.724	-0.493	-2.436**	-0.720	-0.561	-2.846***
	(1.060)	(0.906)	(1.000)	(0.981)	(0.902)	(1.037)
Retracked Groups	0.405	0.631	-1.872*	0.097	0.323	-2.658**
	(1.081)	(0.887)	(0.973)	(1.013)	(0.897)	(1.026)
Ability	1.675***	1.100***	1.460***	1.577***	0.937***	1.410***
	(0.338)	(0.285)	(0.300)	(0.307)	(0.282)	(0.307)
Constant	4.773*	6.031***	5.305**	6.219**	9.460***	7.327**
	(2.697)	(2.184)	(2.404)	(2.841)	(2.403)	(2.767)
Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	76	78	81	66	68	71
R-squared	0.440	0.466	0.493	0.458	0.400	0.467

Notes: OLS regression with dependent variable as output rounds 2, 3 and 4 of real effort task, and the main independent variable as a categorical dummy representing the treatment group. Ex-post resorted low-wage group analysis is shown in columns 1-3, and ex-post resorted and restricted (only those also from the low-wage group during the session) low-wage group analysis is shown in columns 4-6. Standard errors in parentheses. Controls include the full suite of controls included in Model 3 (Table 1). (*** p<0.01, ** p<0.05, * p<0.1).

Appendix Table 3.4 – Effort in Each Round for High-Wage Group Subjects Ranked 1st In Previous Round (1-3) and Ranked 2nd or 3rd In Previous Round (4-6)

	(1)	(2)	(3)		(4)	(5)	(6)
Output in Round	Round 2	Round 3	Round 4	Output in Round	Round 2	Round 3	Round 4
	Ranked 1st In Previous Round				Ranked 2nd/3rd in Previous Round		
Tracked Groups	-0.529	0.990	0.937	Tracked	1.335	1.682	-0.608
	(1.432)	(1.480)	(1.545)	Groups	(1.627)	(1.488)	(1.424)
Retracked Groups	3.750**	4.899***	3.940**	Retracked	3.552**	3.021**	2.612
	(1.669)	(1.559)	(1.615)	Groups	(1.578)	(1.426)	(1.554)
Constant	9.642*	15.19***	4.990	Constant	7.403*	0.122	2.714
	(5.107)	(5.174)	(5.439)		(3.737)	(3.530)	(3.797)
Controls	*Yes*	*Yes*	*Yes*	Controls	*Yes*	*Yes*	*Yes*
Observations	47	53	40	Observations	52	46	54
R-squared	0.585	0.436	0.604	R-squared	0.634	0.747	0.626

Notes: OLS regression with dependent variable as output rounds 2, 3 and 4 of real effort task, and the main independent variable as a categorical dummy representing the treatment group. Our Tracked conditions is taken as the baseline for our Retracked conditions. Subjects ranked 1st in previous round in columns 1-3, and subjects ranked 2nd or 3rd in previous round in columns 4-6. Standard errors in parentheses. Controls include the full suite of controls included in Model 3 (Table 3.1). (*** p<0.01, ** p<0.05, * p<0.1).

Appendix 3 - Chapter 3 Experimental Protocol

Before Data Collection

- The experiment is hosted on a Heroku server
  - The hosted server should have 1 of each type of agent, and a reset Postgres at the start of data collection
  - The experimenter is responsible for ensuring that the server is hosted correctly at the start of each day of data collection
- Data may only be collected between the times: (9:00am) – (5:00pm), so please ensure that it is currently between these times of day.

During Data Collection

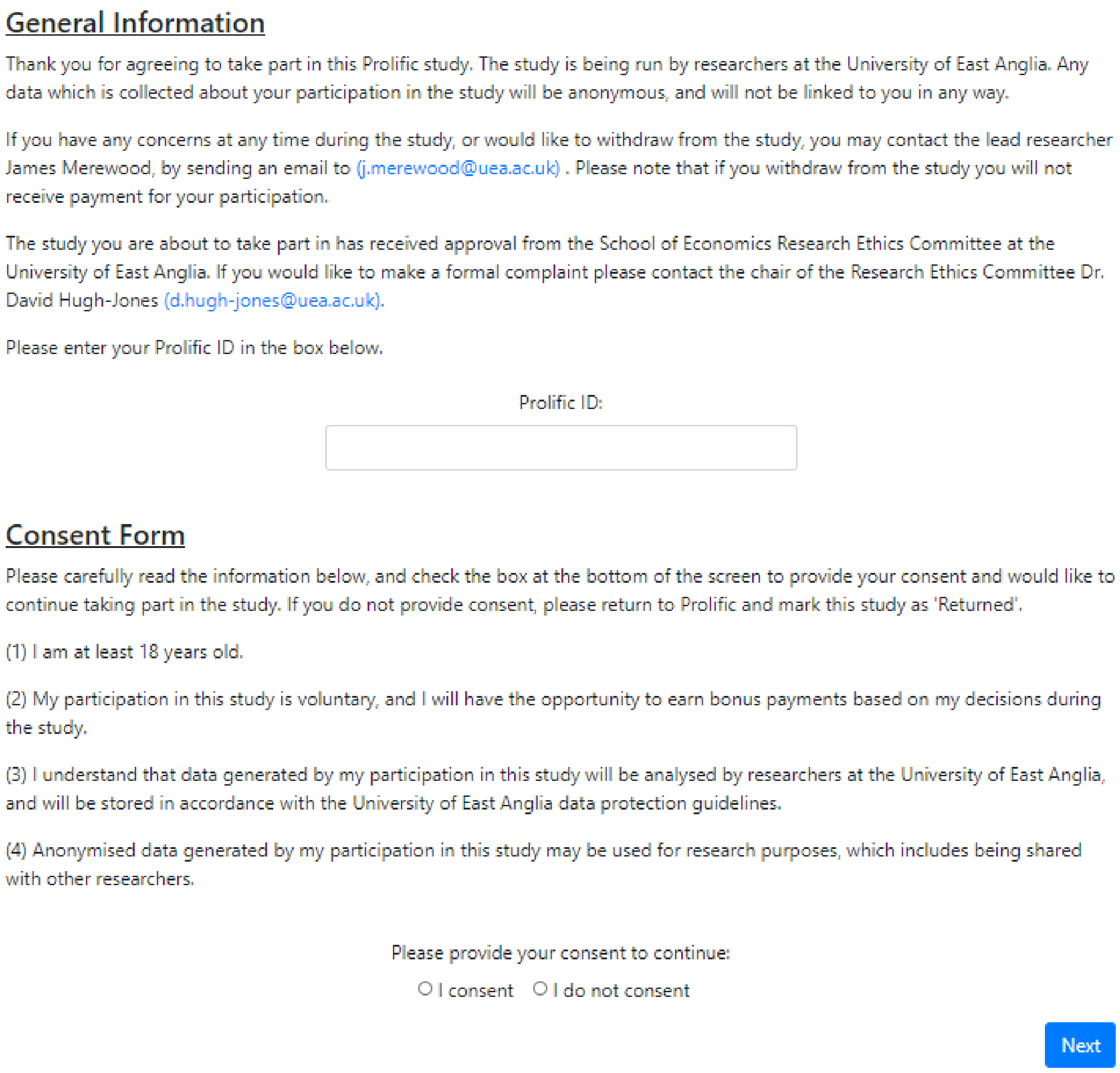
- When data collection has started, subjects will join the experiment via the Prolific.co platform.
- Due to server constraints, no more than 10 spaces should be opened on the Prolific.co platform at any time. Once 10 people have joined the server, the experimenter must wait for them to all finish before opening a further 10 spaces for the next set of subjects to take part in the experiment.
- The experimenter is responsible for maintaining the Prolific.co account during times that the experiment is running. This means that they must check for incoming messages at least every 5 minutes for subjects who are messaging to say that they are having difficulties.
- The time and date for a player entering and leaving the experiment is now coded into the experiment data output, but should still be collected by the experimenter for cross validation.
- If a subject is disconnected from the server, or the server crashes and all subjects are disconnected, then the experimenter must message the subjects in question and provide them with a rejoining link which is specific to their Prolific.co ID. This is provided in the oTree server, but it is important that the experimenter is aware of which ID links to which channel in the server.
- If a subject "times out" on Prolific.co, meaning they have taken too long to finish the experiment, then they must be removed from the server and their data must be dropped from the final dataset.
- Once 15 subjects have finished the task, before allowing a further 15 subjects to enter the server, the experimenter must download the current form of the datasets through the oTree server and save using the format DDMMYY*_numofsubjectscollected_initialofexperimenter* – i.e., 101022_160subjects_jm

After Data Collection

- At the end of each day of data collection, the experimenter is responsible for storing a separate dataset containing the full data which has been collected on that specific day. This should follow the same format as above.
  - A backup of this dataset should also be created in a separate directory.
- The experimenter must cross validate these with the bonus payment variable generated by the oTree code, and store these in a separate excel document. These will be paid once all data is collected.

Annex

Annex 1 – Chapter 2 Instructions for the Experiment



General Information

Thank you for agreeing to take part in this Prolific study. The study is being run by researchers at the University of East Anglia. Any data which is collected about your participation in the study will be anonymous, and will not be linked to you in any way.

If you have any concerns at any time during the study, or would like to withdraw from the study, you may contact the lead researcher James Merewood, by sending an email to (j.merewood@uea.ac.uk) . Please note that if you withdraw from the study you will not receive payment for your participation.

The study you are about to take part in has received approval from the School of Economics Research Ethics Committee at the University of East Anglia. If you would like to make a formal complaint please contact the chair of the Research Ethics Committee Dr. David Hugh-Jones (d.hugh-jones@uea.ac.uk).

Please enter your Prolific ID in the box below.

Prolific ID:

Consent Form

Please carefully read the information below, and check the box at the bottom of the screen to provide your consent and would like to continue taking part in the study. If you do not provide consent, please return to Prolific and mark this study as 'Returned'.

(1) I am at least 18 years old.

(2) My participation in this study is voluntary, and I will have the opportunity to earn bonus payments based on my decisions during the study.

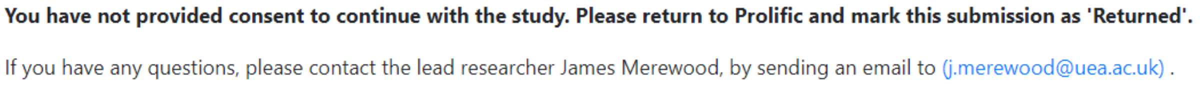
(3) I understand that data generated by my participation in this study will be analysed by researchers at the University of East Anglia, and will be stored in accordance with the University of East Anglia data protection guidelines.

(4) Anonymised data generated by my participation in this study may be used for research purposes, which includes being shared with other researchers.

Please provide your consent to continue:

○ I consent ○ I do not consent

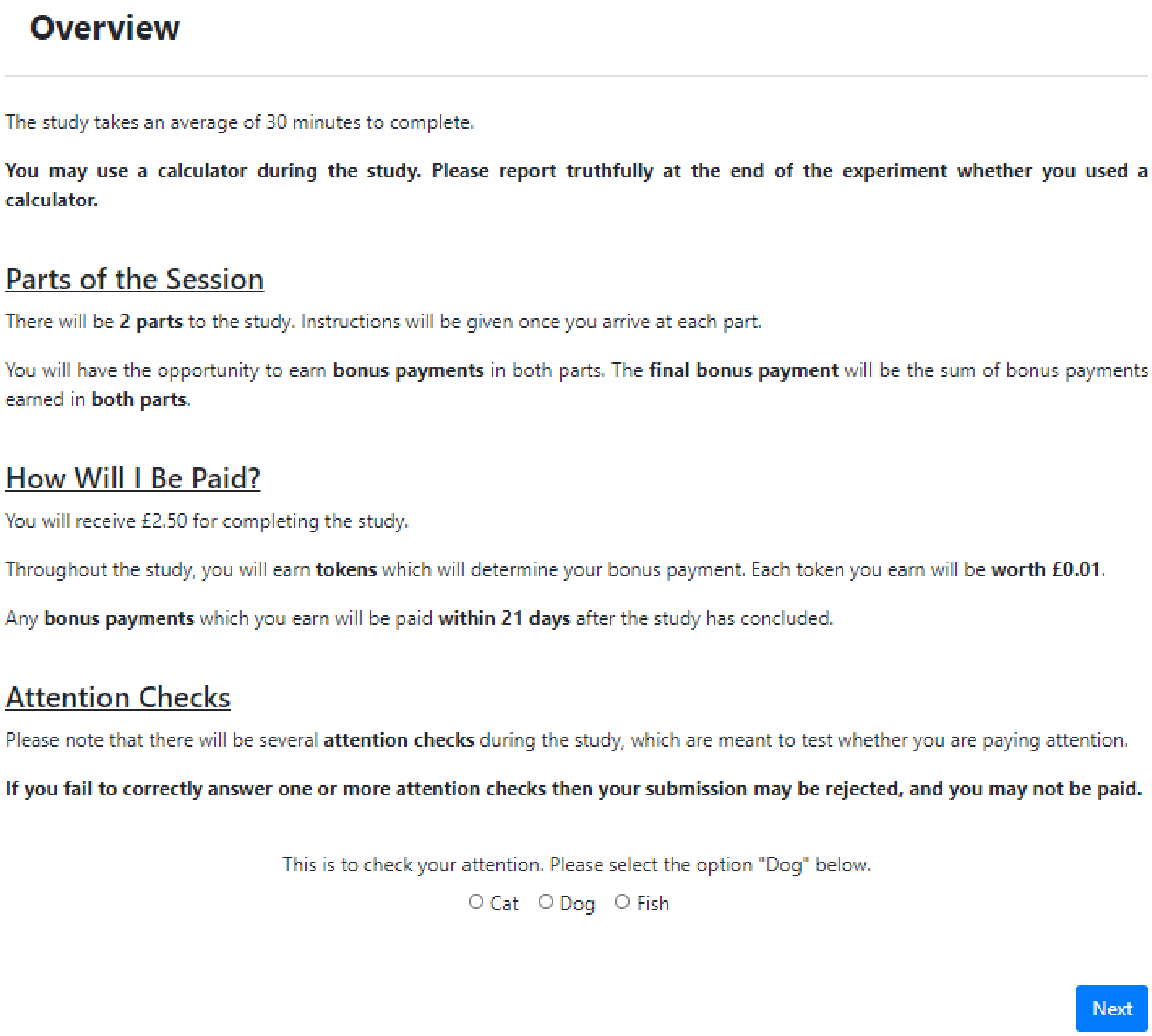
{If the participant does not consent to the study – No back button so that they cannot continue with the study}



You have not provided consent to continue with the study. Please return to Prolific and mark this submission as 'Returned'.

If you have any questions, please contact the lead researcher James Merewood, by sending an email to (j.merewood@uea.ac.uk) .

[Otherwise if they do consent to the study}



Overview

The study takes an average of 30 minutes to complete.

You may use a calculator during the study. Please report truthfully at the end of the experiment whether you used a calculator.

Parts of the Session

There will be **2 parts** to the study. Instructions will be given once you arrive at each part.

You will have the opportunity to earn **bonus payments** in both parts. The **final bonus payment** will be the sum of bonus payments earned in **both parts**.

How Will I Be Paid?

You will receive £2.50 for completing the study.

Throughout the study, you will earn **tokens** which will determine your bonus payment. Each token you earn will be **worth £0.01**.

Any **bonus payments** which you earn will be paid **within 21 days** after the study has concluded.

Attention Checks

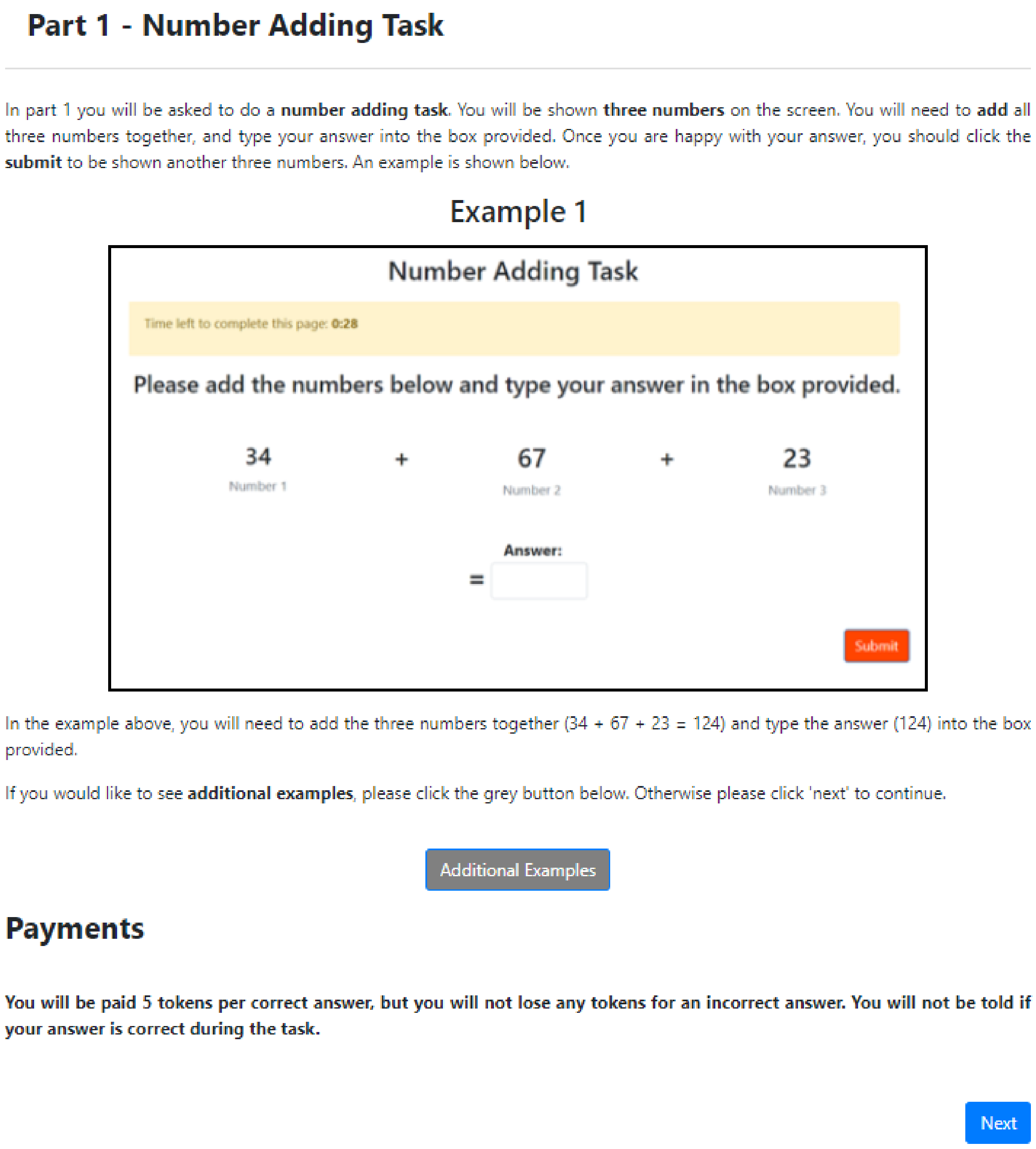
Please note that there will be several **attention checks** during the study, which are meant to test whether you are paying attention.

If you fail to correctly answer one or more attention checks then your submission may be rejected, and you may not be paid.

This is to check your attention. Please select the option "Dog" below.

○ Cat ○ Dog ○ Fish

[TREATMENT 1 – SIMPLE ADDITION]



Part 1 - Number Adding Task

In part 1 you will be asked to do a **number adding task**. You will be shown **three numbers** on the screen. You will need to **add** all three numbers together, and type your answer into the box provided. Once you are happy with your answer, you should click the **submit** to be shown another three numbers. An example is shown below.

Example 1

Number Adding Task

Time left to complete this page: **0:28**

Please add the numbers below and type your answer in the box provided.

34	+	67	+	23
Number 1		Number 2		Number 3

Answer:

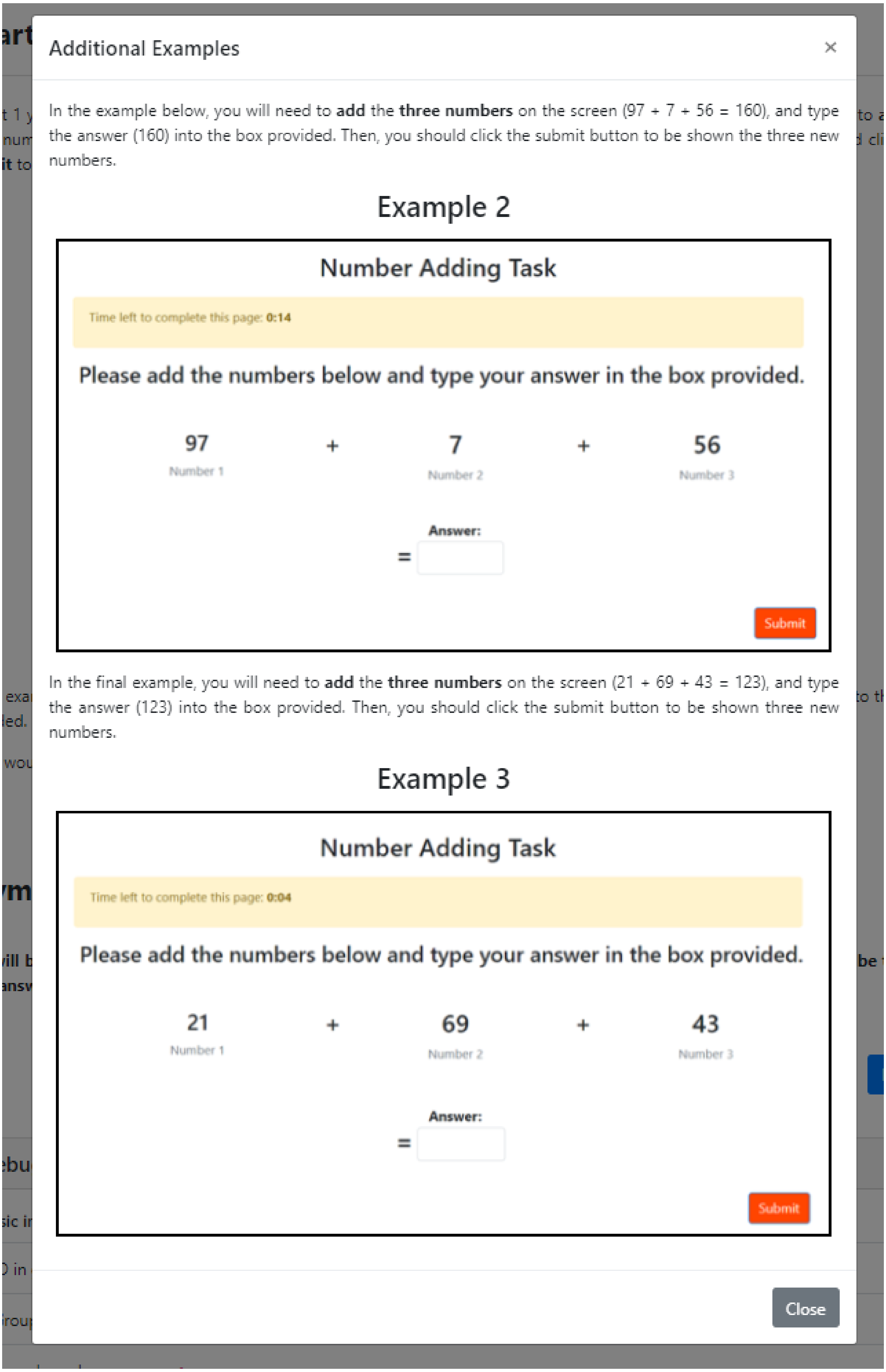
=

In the example above, you will need to add the three numbers together (34 + 67 + 23 = 124) and type the answer (124) into the box provided.

If you would like to see **additional examples**, please click the grey button below. Otherwise please click 'next' to continue.

Payments

You will be paid 5 tokens per correct answer, but you will not lose any tokens for an incorrect answer. You will not be told if your answer is correct during the task.



Additional Examples

In the example below, you will need to **add** the **three numbers** on the screen (97 + 7 + 56 = 160), and type the answer (160) into the box provided. Then, you should click the submit button to be shown the three new numbers.

Example 2

Number Adding Task

Time left to complete this page: **0:14**

Please add the numbers below and type your answer in the box provided.

97	+	7	+	56
Number 1		Number 2		Number 3

Answer:

=

In the final example, you will need to **add** the **three numbers** on the screen (21 + 69 + 43 = 123), and type the answer (123) into the box provided. Then, you should click the submit button to be shown three new numbers.

Example 3

Number Adding Task

Time left to complete this page: **0:04**

Please add the numbers below and type your answer in the box provided.

21	+	69	+	43
Number 1		Number 2		Number 3

Answer:

=

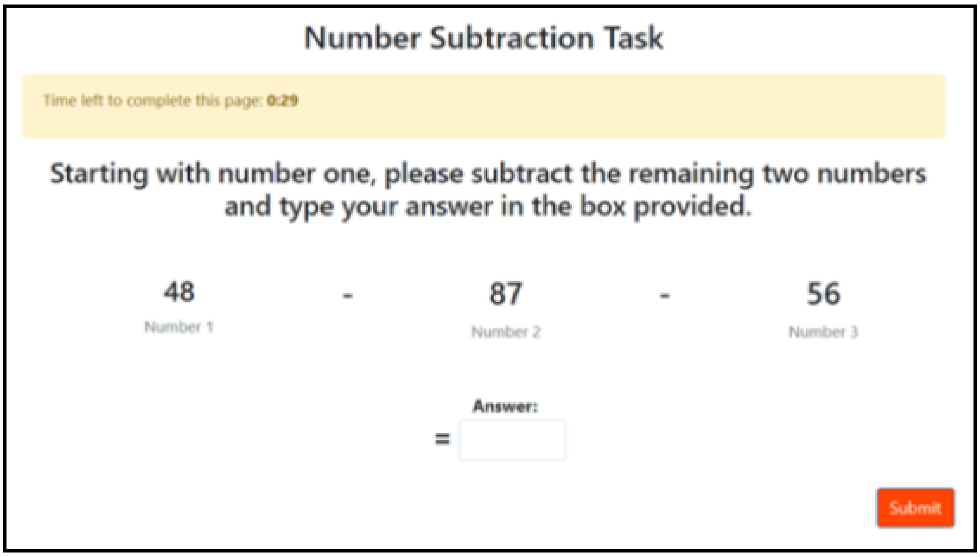
[TREATMENT 2 – SIMPLE SUBTRACTION]

Part 1 - Number Subtraction Task

In part 1 you will be asked to do a **number subtraction task**. You will be shown **three numbers** on the screen. Starting with the first number, you should **subtract** the remaining two numbers, and type your answer into the box provided.

If the answer is a negative number, please enter a minus sign (-) in front of the answer. Once you are happy with your answer, you should click the **submit button** to be shown another three numbers. An example is shown below.

Example 1



Number Subtraction Task

Time left to complete this page: **0:29**

Starting with number one, please subtract the remaining two numbers and type your answer in the box provided.

48 - 87 - 56

To complete the example above, you should start with the first number (48) and subtract the remaining two numbers (48 – 87 – 56 = -95) and type the answer (-95) into the box provided.

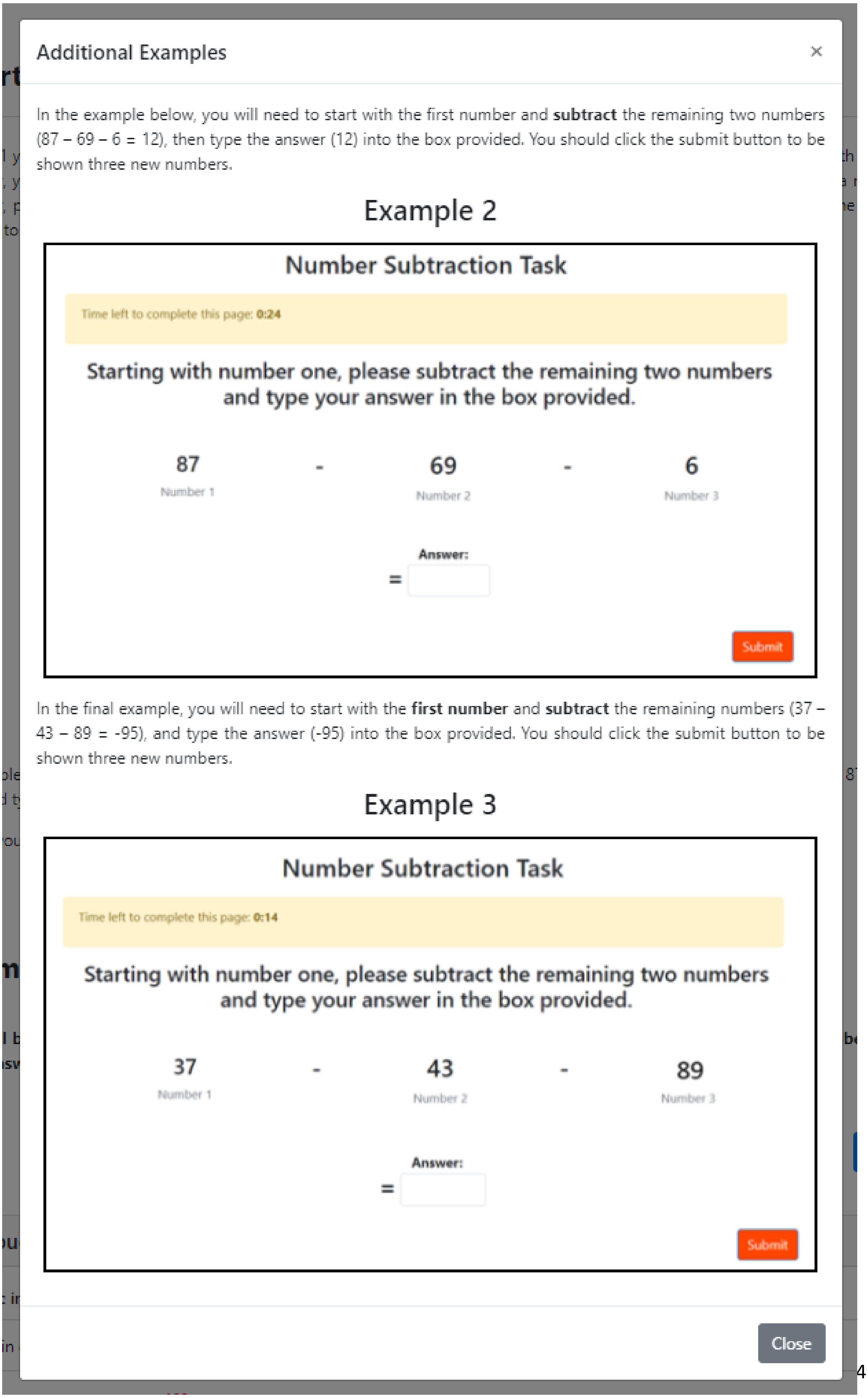
If you would like to see **additional examples**, please click the grey button below. Otherwise please click 'next' to continue.

Additional Examples

Payments

You will be paid 5 tokens per correct answer, but you will not lose any tokens for an incorrect answer. You will not be told if your answer is correct during the task.

Next



Additional Examples

In the example below, you will need to start with the first number and **subtract** the remaining two numbers (87 – 69 – 6 = 12), then type the answer (12) into the box provided. You should click the submit button to be shown three new numbers.

Example 2

Number Subtraction Task

Time left to complete this page: **0:24**

Starting with number one, please subtract the remaining two numbers and type your answer in the box provided.

87	-	69	-	6
Number 1		Number 2		Number 3

Answer:

=

In the final example, you will need to start with the **first number** and **subtract** the remaining numbers (37 – 43 – 89 = -95), and type the answer (-95) into the box provided. You should click the submit button to be shown three new numbers.

Example 3

Number Subtraction Task

Time left to complete this page: **0:14**

Starting with number one, please subtract the remaining two numbers and type your answer in the box provided.

37	-	43	-	89
Number 1		Number 2		Number 3

Answer:

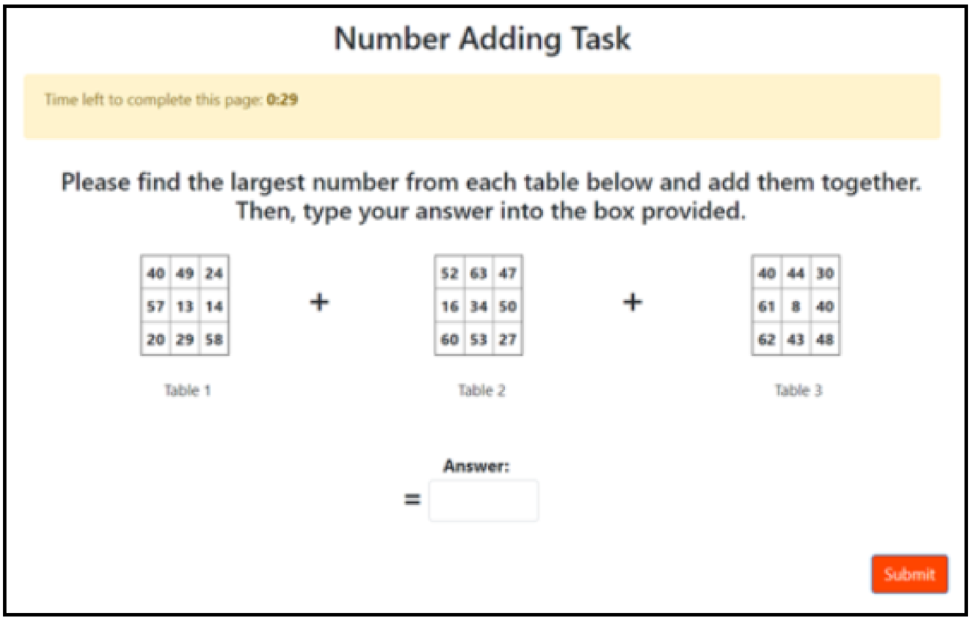
=

[TREATMENT 3 – COMPLEX ADDITION]

Part 1 - Number Adding Task

In part 1 you will be asked to do a **number adding task**. You will be shown **three tables** on the screen. You should **add** the **largest number** from each of the three tables together, and type your answer into the box provided. Then, you should click the **submit button** to be shown another three tables to solve. An example is shown below.

Example 1



Number Adding Task

Time left to complete this page: **0:29**

Please find the largest number from each table below and add them together. Then, type your answer into the box provided.

40	49	24
57	13	14
20	29	58

Table 1

+

52	63	47
16	34	50
60	53	27

Table 2

+

40	44	30
61	8	40
62	43	48

Table 3

Answer:

=

To complete the example above, you will need to find the largest number from each table, which are: 58 from table 1; 63 from table 2; 62 from table 3. You will need to add these together (58 + 63 + 62 = 183) and type the answer (183) into the box provided.

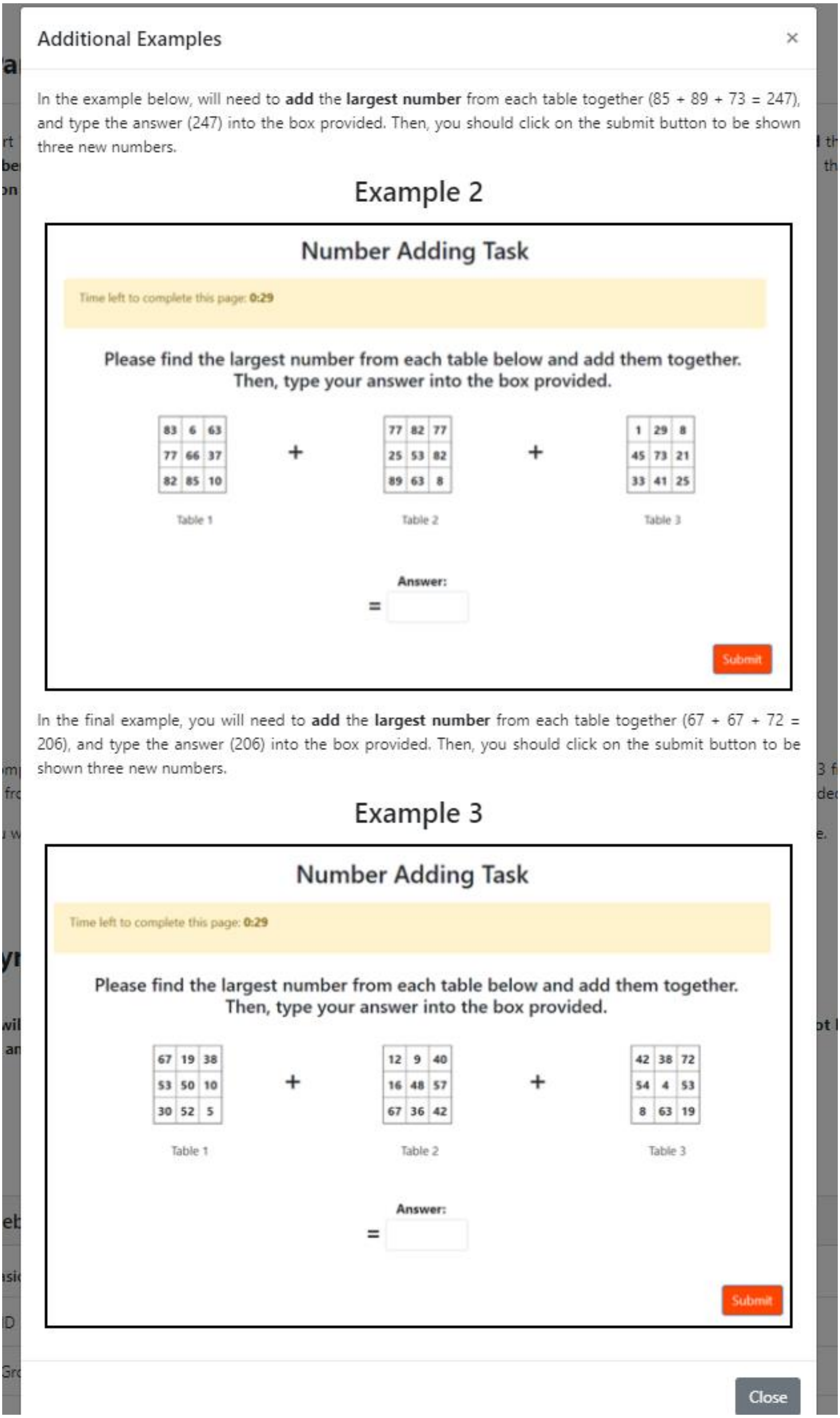
If you would like to see **additional examples**, please click the grey button below. Otherwise please click 'next' to continue.

Additional Examples

Payments

You will be paid 5 tokens per correct answer, but you will not lose any tokens for an incorrect answer. You will not be told if your answer is correct during the task.

Next



Additional Examples ×

In the example below, will need to **add** the **largest number** from each table together (85 + 89 + 73 = 247), and type the answer (247) into the box provided. Then, you should click on the submit button to be shown three new numbers.

Example 2

Number Adding Task

Time left to complete this page: **0:29**

Please find the largest number from each table below and add them together. Then, type your answer into the box provided.

83	6	63
77	66	37
82	85	10

Table 1

+

77	82	77
25	53	82
89	63	8

Table 2

+

1	29	8
45	73	21
33	41	25

Table 3

Answer:

=

In the final example, you will need to **add** the **largest number** from each table together (67 + 67 + 72 = 206), and type the answer (206) into the box provided. Then, you should click on the submit button to be shown three new numbers.

Example 3

Number Adding Task

Time left to complete this page: **0:29**

Please find the largest number from each table below and add them together. Then, type your answer into the box provided.

67	19	38
53	50	10
30	52	5

Table 1

+

12	9	40
16	48	57
67	36	42

Table 2

+

42	38	72
54	4	53
8	63	19

Table 3

Answer:

=

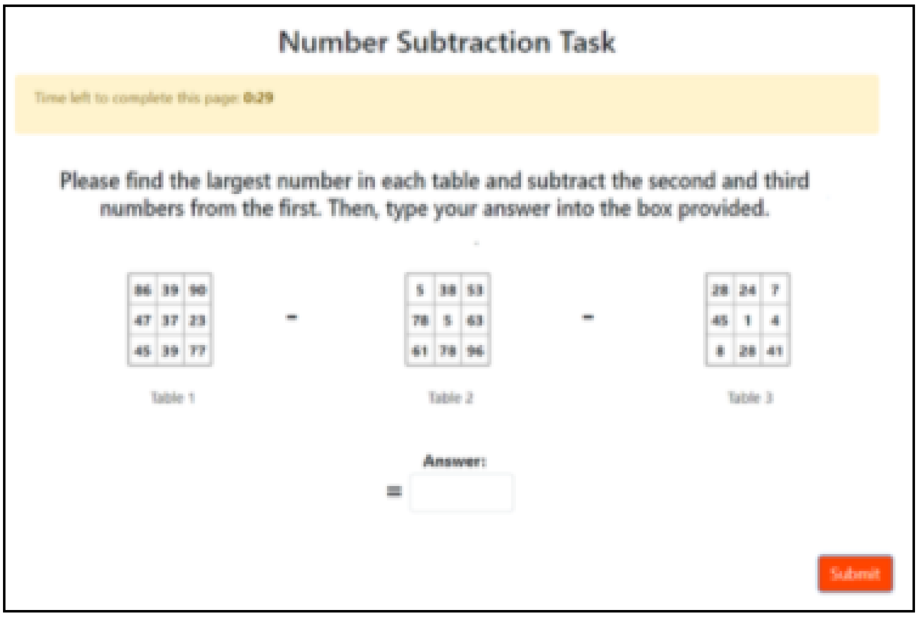
[TREATMENT 4 – COMPLEX SUBTRACTION]

Part 1 - Number Subtraction Task

In part 1 you will be asked to do a **number subtraction task**. You will be shown **three tables** on the screen. Starting with the first table, you should **subtract** the **largest number** from each of the three tables, and type your answer into the box provided.

If the answer is a negative number, you should enter a minus sign (-) in front of the answer. Once you are happy with your answer, you should click the **submit button** to be shown another three tables to solve. An example is shown below.

Example 1



Number Subtraction Task

Time left to complete this page: **0:29**

Please find the largest number in each table and subtract the second and third numbers from the first. Then, type your answer into the box provided.

86	39	90
47	37	23
45	39	77

Table 1

-

5	38	53
78	5	63
61	78	96

Table 2

-

28	24	7
45	1	4
8	28	41

Table 3

Answer:

=

To complete the example above, you will need to find the largest number from each table, which are: 90 from table 1; 96 from table 2; and 45 from table 3. You will need to start with the largest number from the first table and subtract the remaining numbers (90 – 96 – 45 = -51) and type the answer (-51) into the box provided.

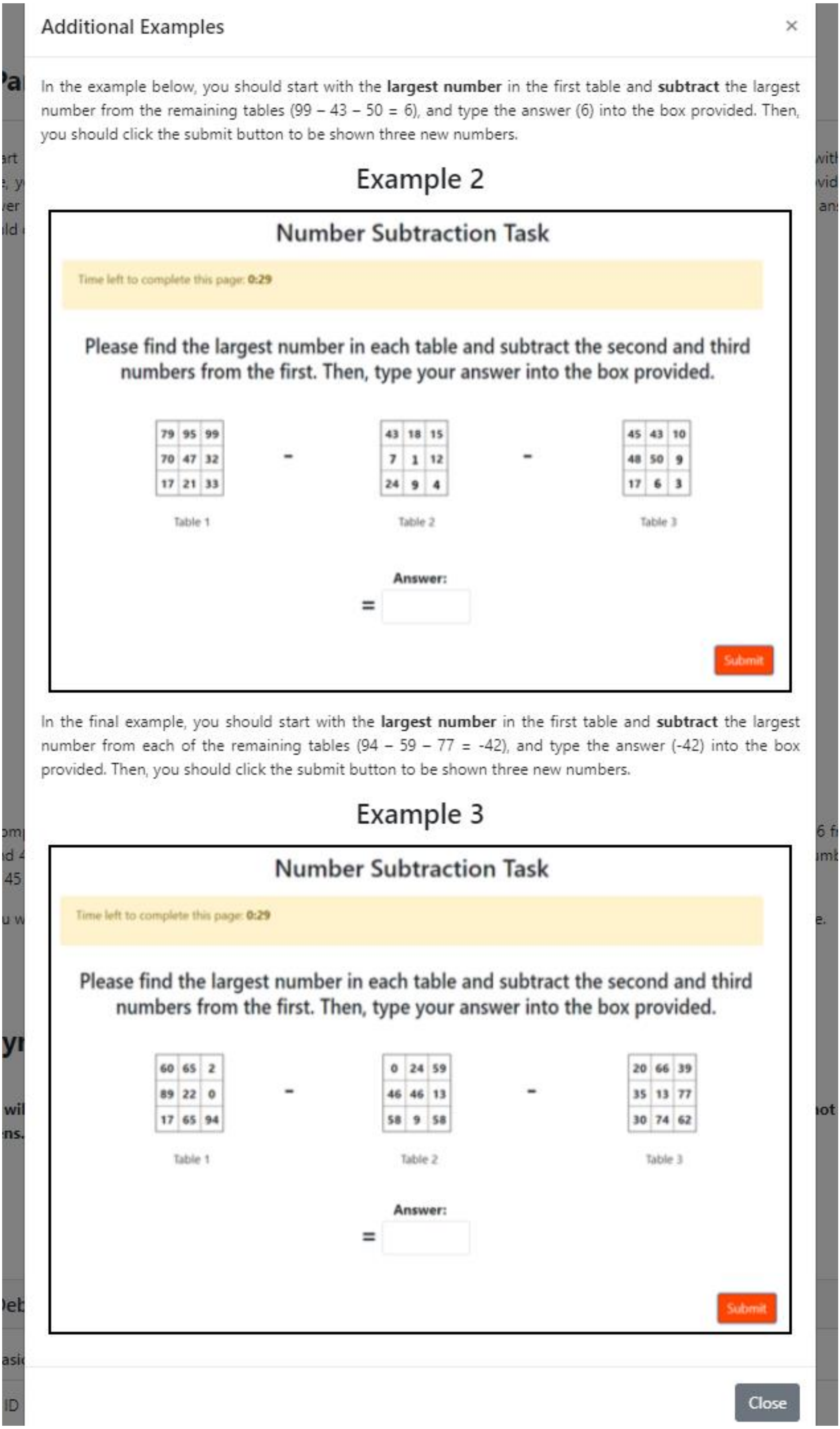
If you would like to see **additional examples**, please click the grey button below. Otherwise please click 'next' to continue.

Additional Examples

Payments

You will be paid 5 tokens per correct answer, but you will not lose any tokens for an incorrect answer. You will not be told if your answer is correct during the task.

Next



Additional Examples ×

In the example below, you should start with the **largest number** in the first table and **subtract** the largest number from the remaining tables (99 – 43 – 50 = 6), and type the answer (6) into the box provided. Then, you should click the submit button to be shown three new numbers.

Example 2

Number Subtraction Task

Time left to complete this page: **0:29**

Please find the largest number in each table and subtract the second and third numbers from the first. Then, type your answer into the box provided.

79	95	99
70	47	32
17	21	33

Table 1

-

43	18	15
7	1	12
24	9	4

Table 2

-

45	43	10
48	50	9
17	6	3

Table 3

Answer:

= □

Submit

In the final example, you should start with the **largest number** in the first table and **subtract** the largest number from each of the remaining tables (94 – 59 – 77 = -42), and type the answer (-42) into the box provided. Then, you should click the submit button to be shown three new numbers.

Example 3

Number Subtraction Task

Time left to complete this page: **0:29**

Please find the largest number in each table and subtract the second and third numbers from the first. Then, type your answer into the box provided.

60	65	2
89	22	0
17	65	94

Table 1

-

0	24	59
46	46	13
58	9	58

Table 2

-

20	66	39
35	13	77
30	74	62

Table 3

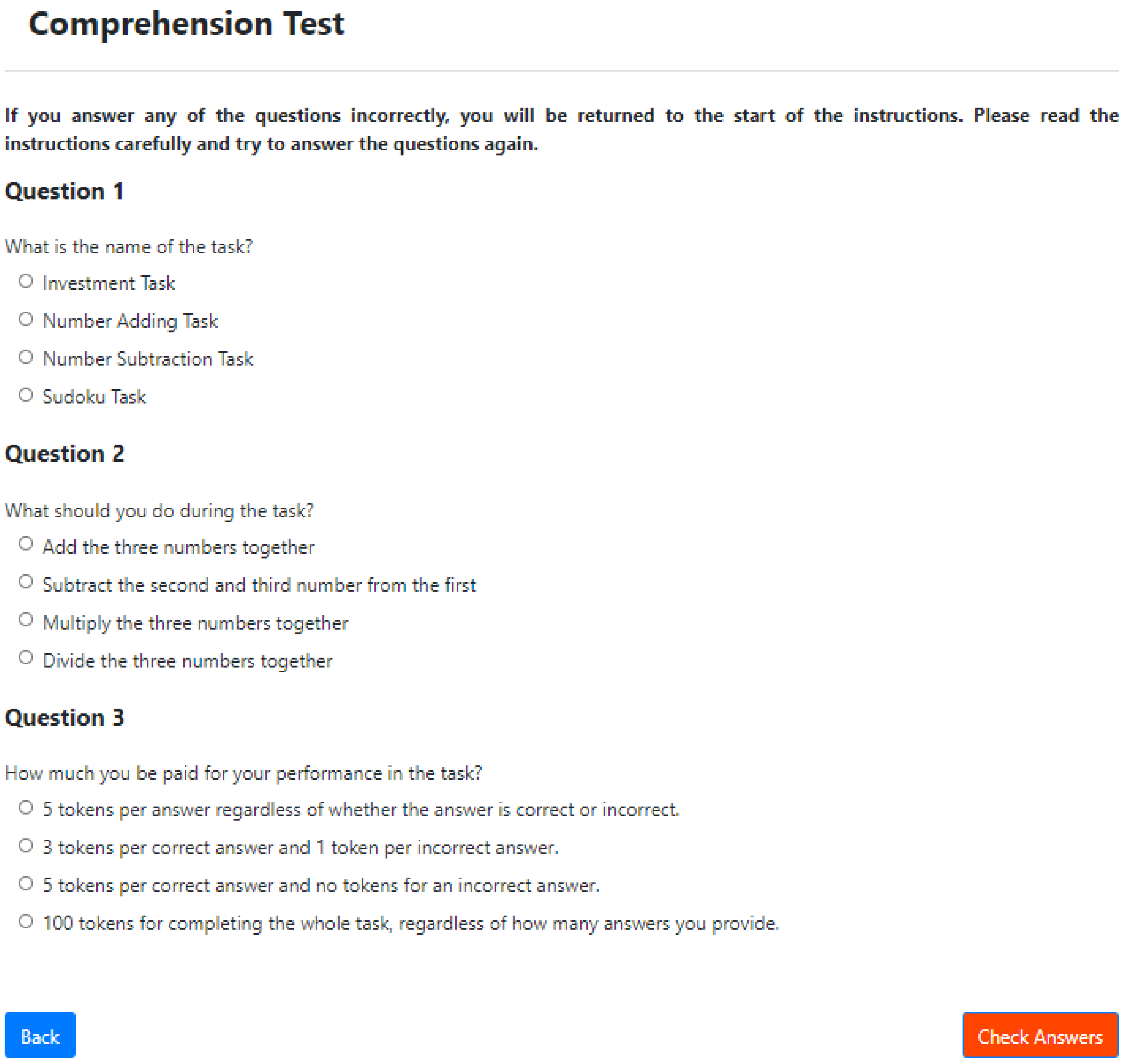
Answer:

= □

Submit

Close

[TREATMENT 1 – SIMPLE ADDITION]



Comprehension Test

If you answer any of the questions incorrectly, you will be returned to the start of the instructions. Please read the instructions carefully and try to answer the questions again.

Question 1

What is the name of the task?

○ Investment Task

○ Number Adding Task

○ Number Subtraction Task

○ Sudoku Task

Question 2

What should you do during the task?

○ Add the three numbers together

○ Subtract the second and third number from the first

○ Multiply the three numbers together

○ Divide the three numbers together

Question 3

How much you be paid for your performance in the task?

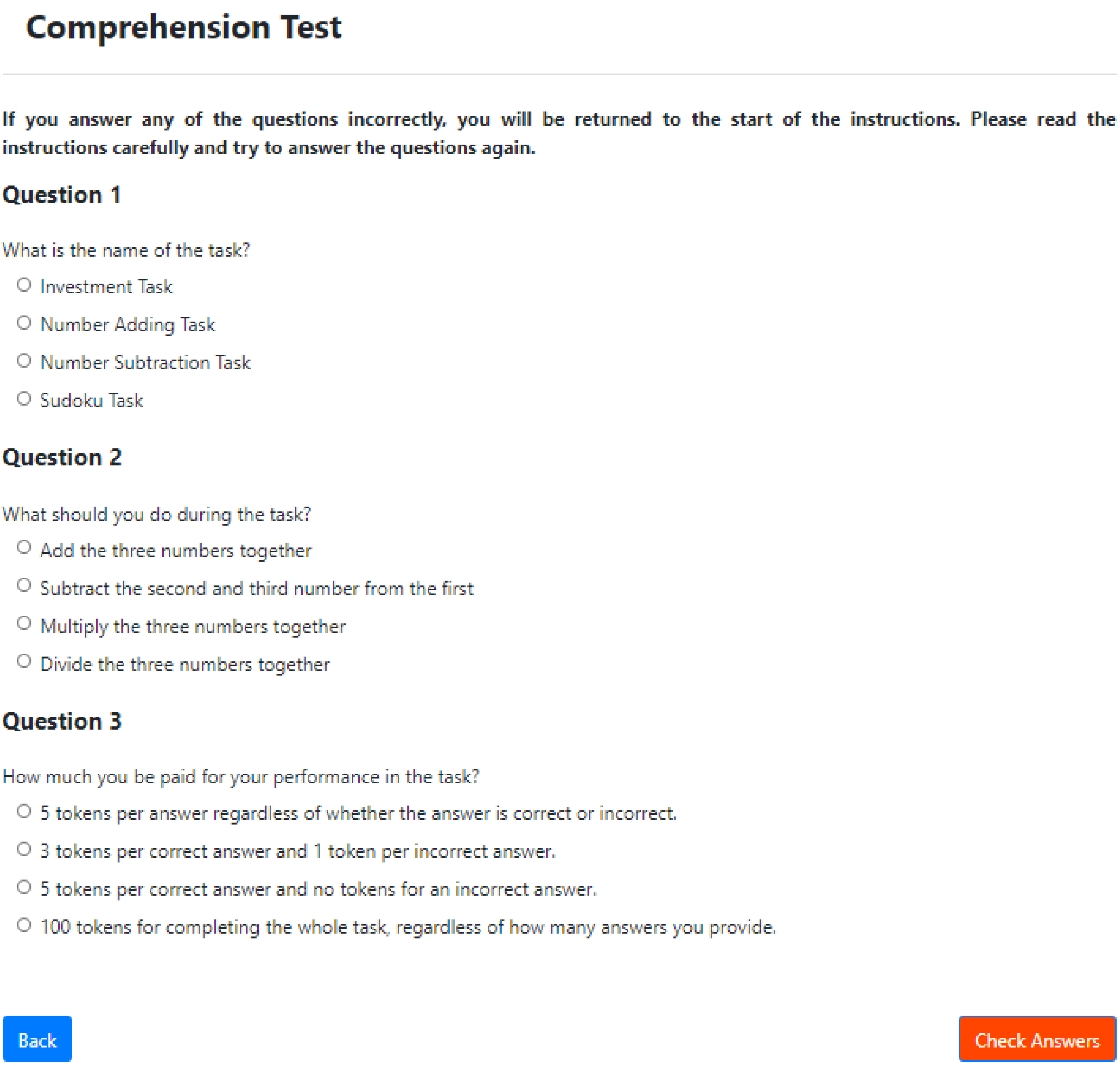
○ 5 tokens per answer regardless of whether the answer is correct or incorrect.

○ 3 tokens per correct answer and 1 token per incorrect answer.

○ 5 tokens per correct answer and no tokens for an incorrect answer.

○ 100 tokens for completing the whole task, regardless of how many answers you provide.

[TREATMENT 2 – SIMPLE SUBTRACTION]



Comprehension Test

If you answer any of the questions incorrectly, you will be returned to the start of the instructions. Please read the instructions carefully and try to answer the questions again.

Question 1

What is the name of the task?

○ Investment Task

○ Number Adding Task

○ Number Subtraction Task

○ Sudoku Task

Question 2

What should you do during the task?

○ Add the three numbers together

○ Subtract the second and third number from the first

○ Multiply the three numbers together

○ Divide the three numbers together

Question 3

How much you be paid for your performance in the task?

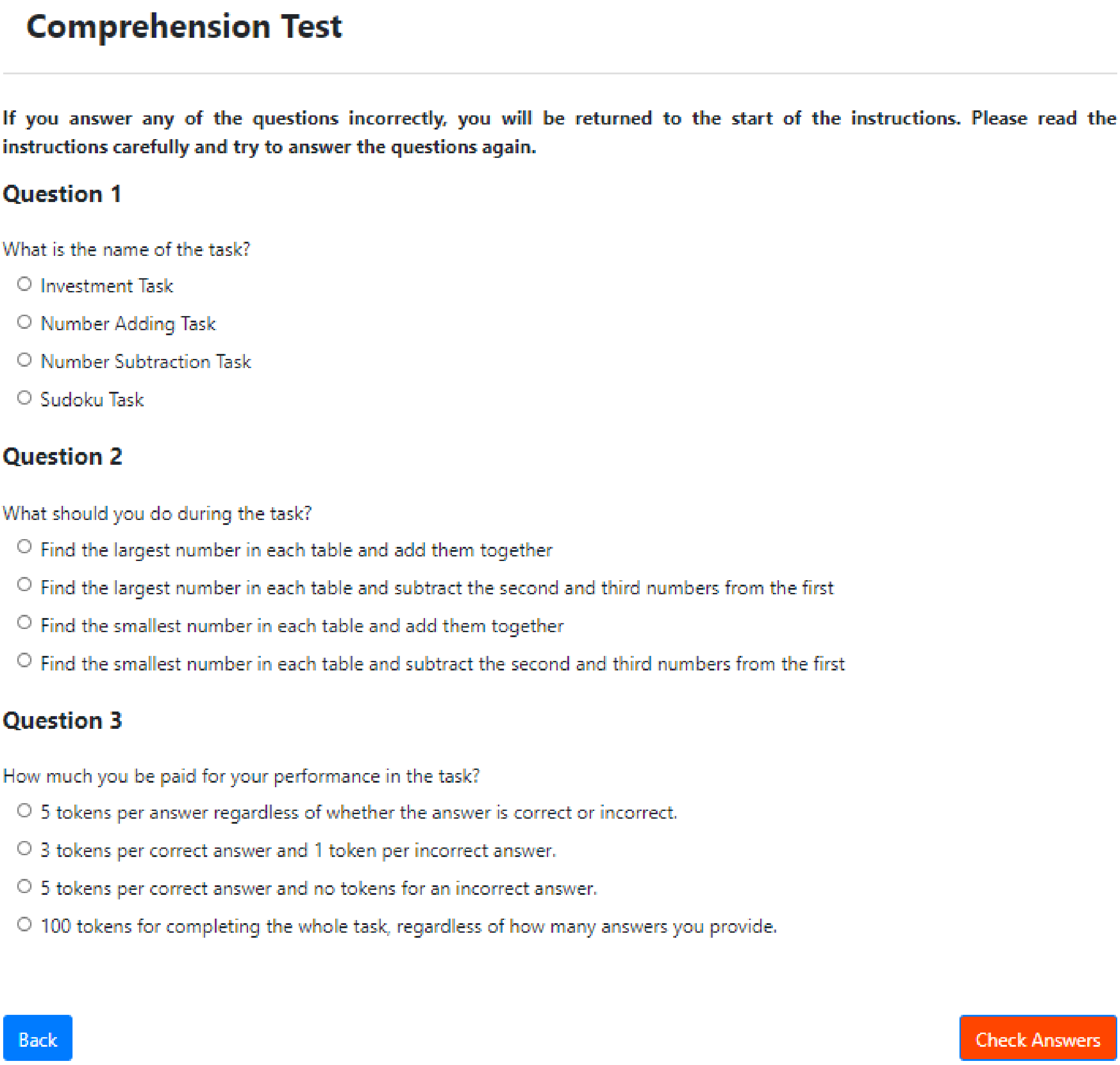
○ 5 tokens per answer regardless of whether the answer is correct or incorrect.

○ 3 tokens per correct answer and 1 token per incorrect answer.

○ 5 tokens per correct answer and no tokens for an incorrect answer.

○ 100 tokens for completing the whole task, regardless of how many answers you provide.

[TREATMENT 3 – COMPLEX ADDITION]



Comprehension Test

If you answer any of the questions incorrectly, you will be returned to the start of the instructions. Please read the instructions carefully and try to answer the questions again.

Question 1

What is the name of the task?

○ Investment Task

○ Number Adding Task

○ Number Subtraction Task

○ Sudoku Task

Question 2

What should you do during the task?

○ Find the largest number in each table and add them together

○ Find the largest number in each table and subtract the second and third numbers from the first

○ Find the smallest number in each table and add them together

○ Find the smallest number in each table and subtract the second and third numbers from the first

Question 3

How much you be paid for your performance in the task?

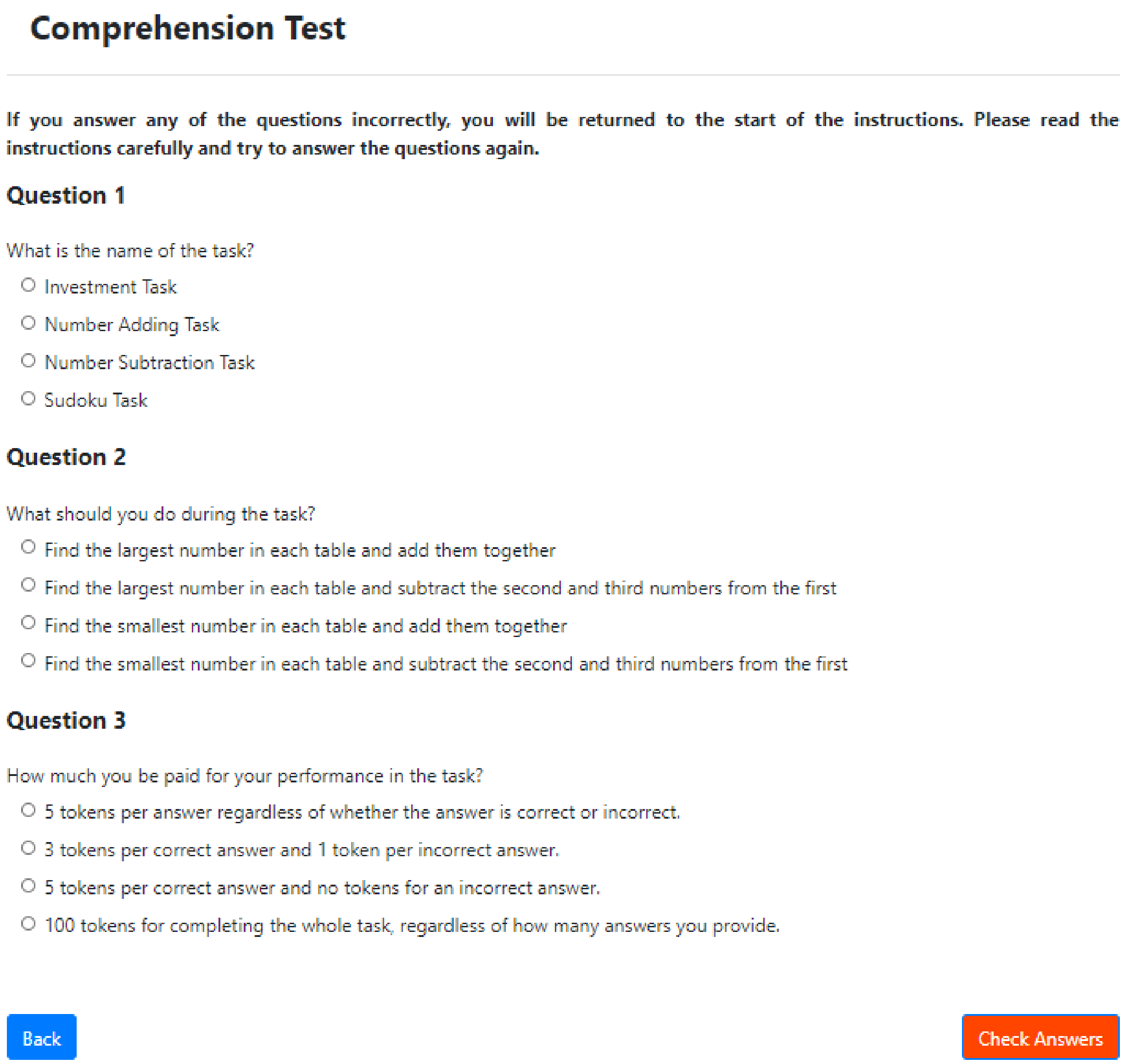
○ 5 tokens per answer regardless of whether the answer is correct or incorrect.

○ 3 tokens per correct answer and 1 token per incorrect answer.

○ 5 tokens per correct answer and no tokens for an incorrect answer.

○ 100 tokens for completing the whole task, regardless of how many answers you provide.

[TREATMENT 4 – COMPLEX SUBTRACTION]



Comprehension Test

If you answer any of the questions incorrectly, you will be returned to the start of the instructions. Please read the instructions carefully and try to answer the questions again.

Question 1

What is the name of the task?

○ Investment Task

○ Number Adding Task

○ Number Subtraction Task

○ Sudoku Task

Question 2

What should you do during the task?

○ Find the largest number in each table and add them together

○ Find the largest number in each table and subtract the second and third numbers from the first

○ Find the smallest number in each table and add them together

○ Find the smallest number in each table and subtract the second and third numbers from the first

Question 3

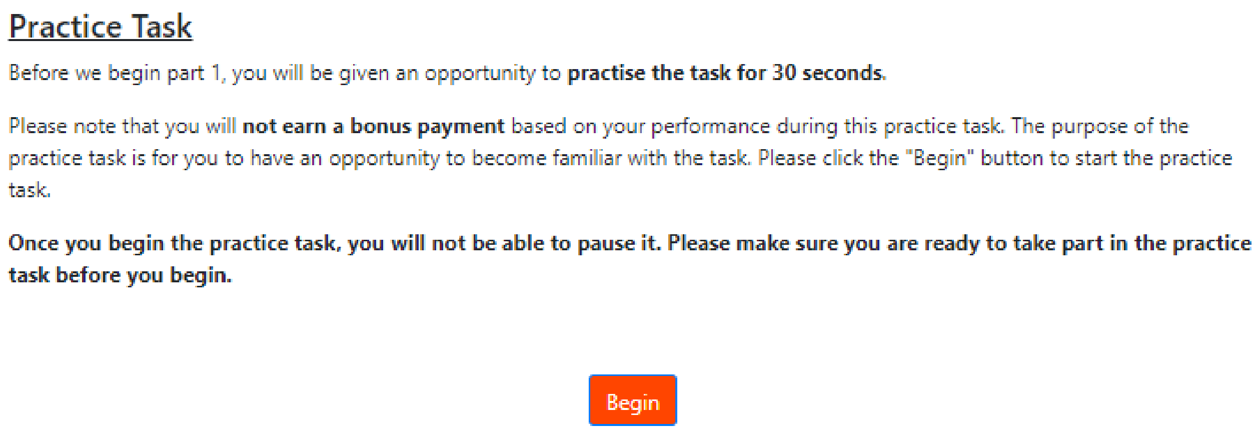
How much you be paid for your performance in the task?

○ 5 tokens per answer regardless of whether the answer is correct or incorrect.

○ 3 tokens per correct answer and 1 token per incorrect answer.

○ 5 tokens per correct answer and no tokens for an incorrect answer.

○ 100 tokens for completing the whole task, regardless of how many answers you provide.

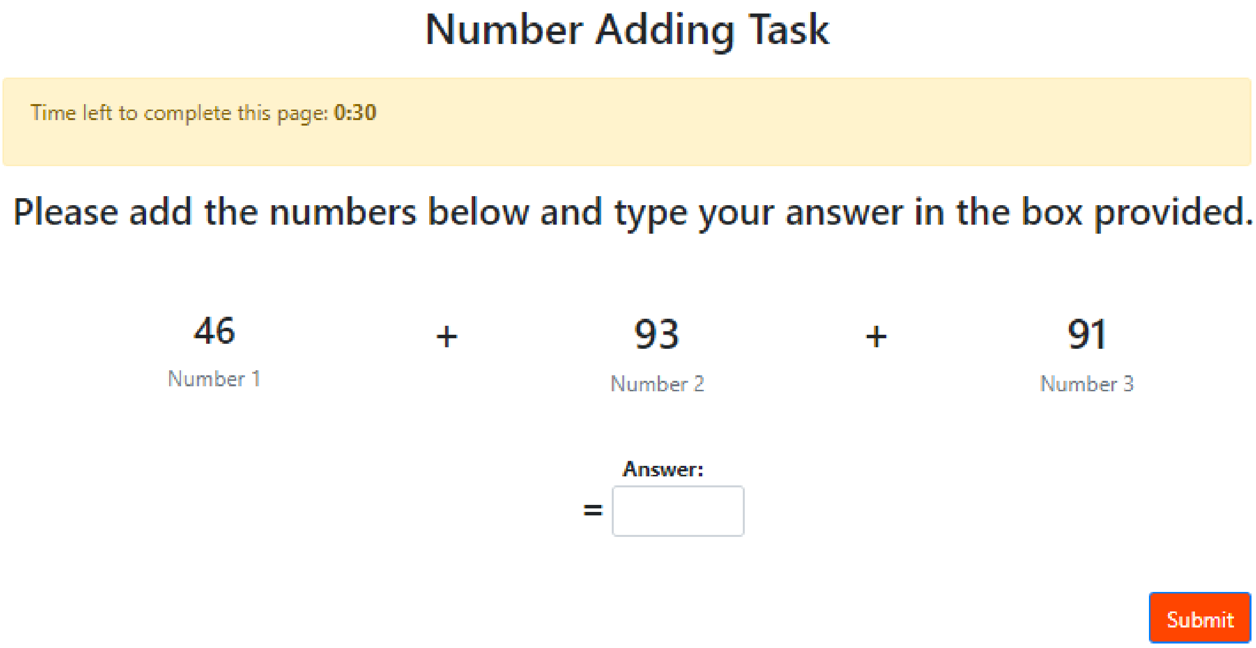


Practice Task

Before we begin part 1, you will be given an opportunity to **practise the task for 30 seconds**.

Please note that you will **not earn a bonus payment** based on your performance during this practice task. The purpose of the practice task is for you to have an opportunity to become familiar with the task. Please click the "Begin" button to start the practice task.

Once you begin the practice task, you will not be able to pause it. Please make sure you are ready to take part in the practice task before you begin.



Number Adding Task

Time left to complete this page: **0:30**

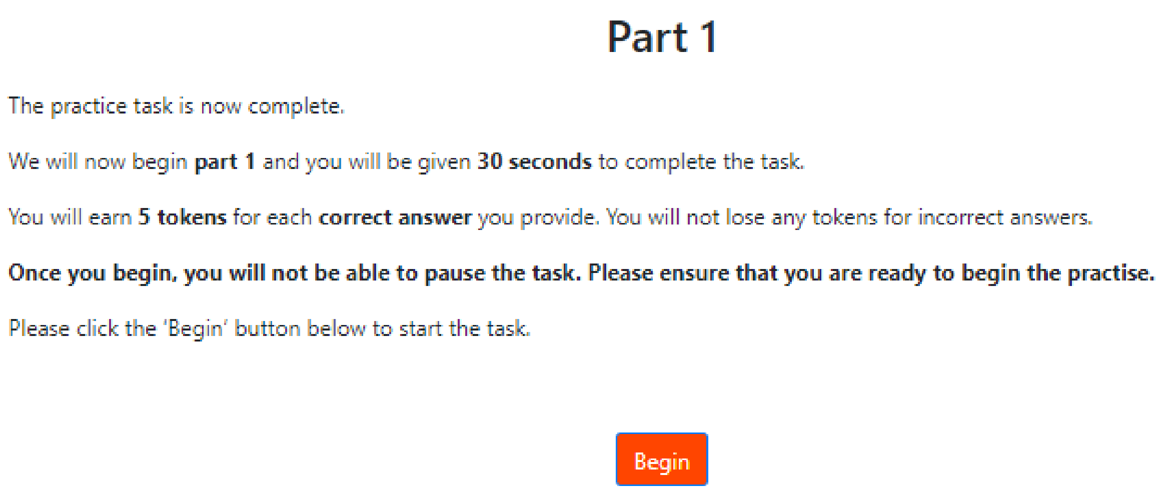
Please add the numbers below and type your answer in the box provided.

46 + 93 + 91

Number 1 Number 2 Number 3

Answer:

=



Part 1

The practice task is now complete.

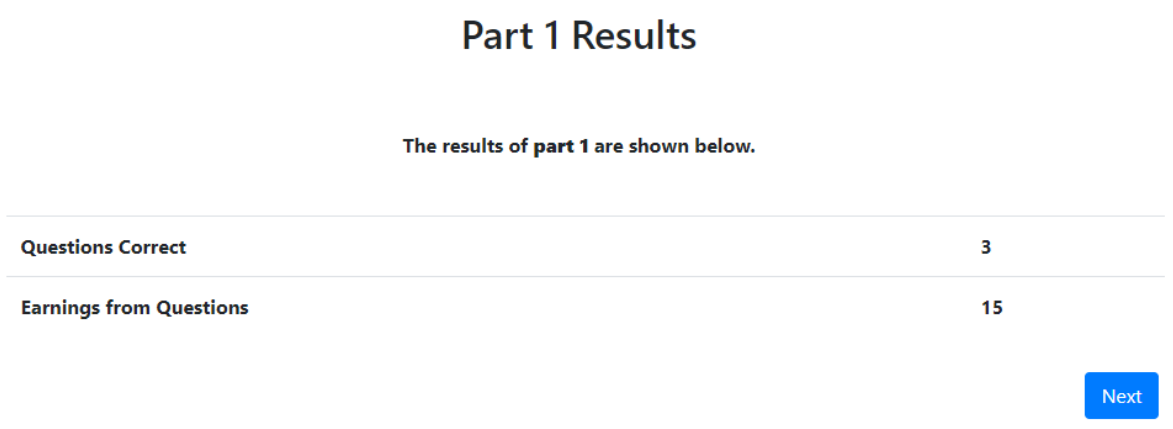
We will now begin **part 1** and you will be given **30 seconds** to complete the task.

You will earn **5 tokens** for each **correct answer** you provide. You will not lose any tokens for incorrect answers.

Once you begin, you will not be able to pause the task. Please ensure that you are ready to begin the practise.

Please click the 'Begin' button below to start the task.

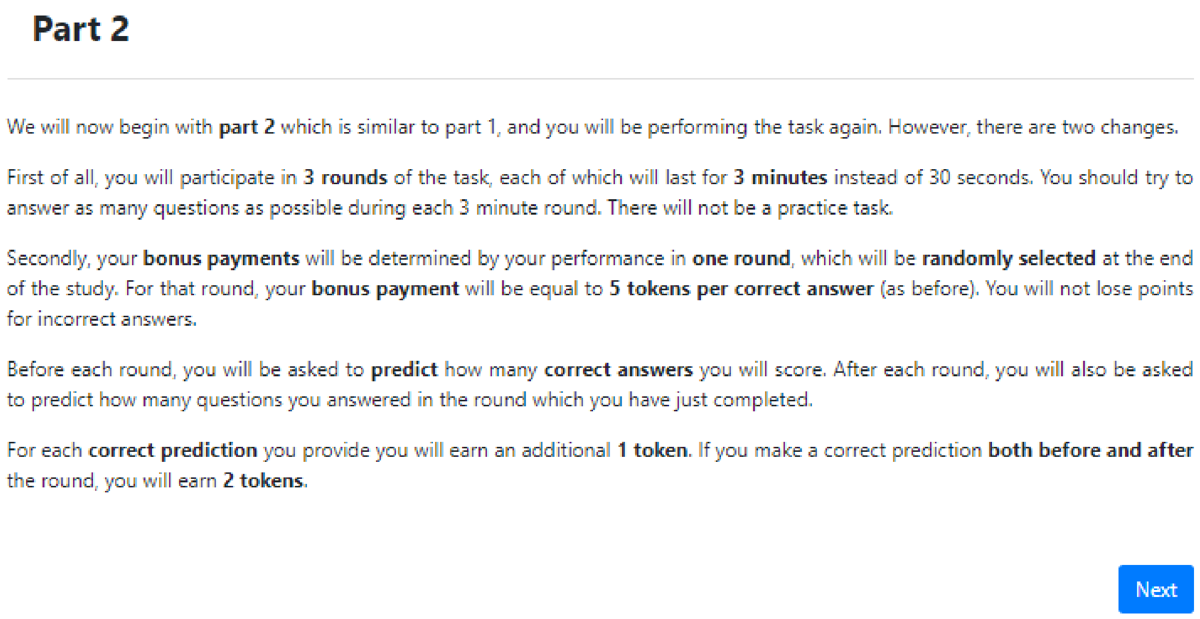
TASK PAGE IS SHOWN AGAIN HERE



Part 1 Results

The results of part 1 are shown below.

Questions Correct	**3**
Earnings from Questions	**15**



Part 2

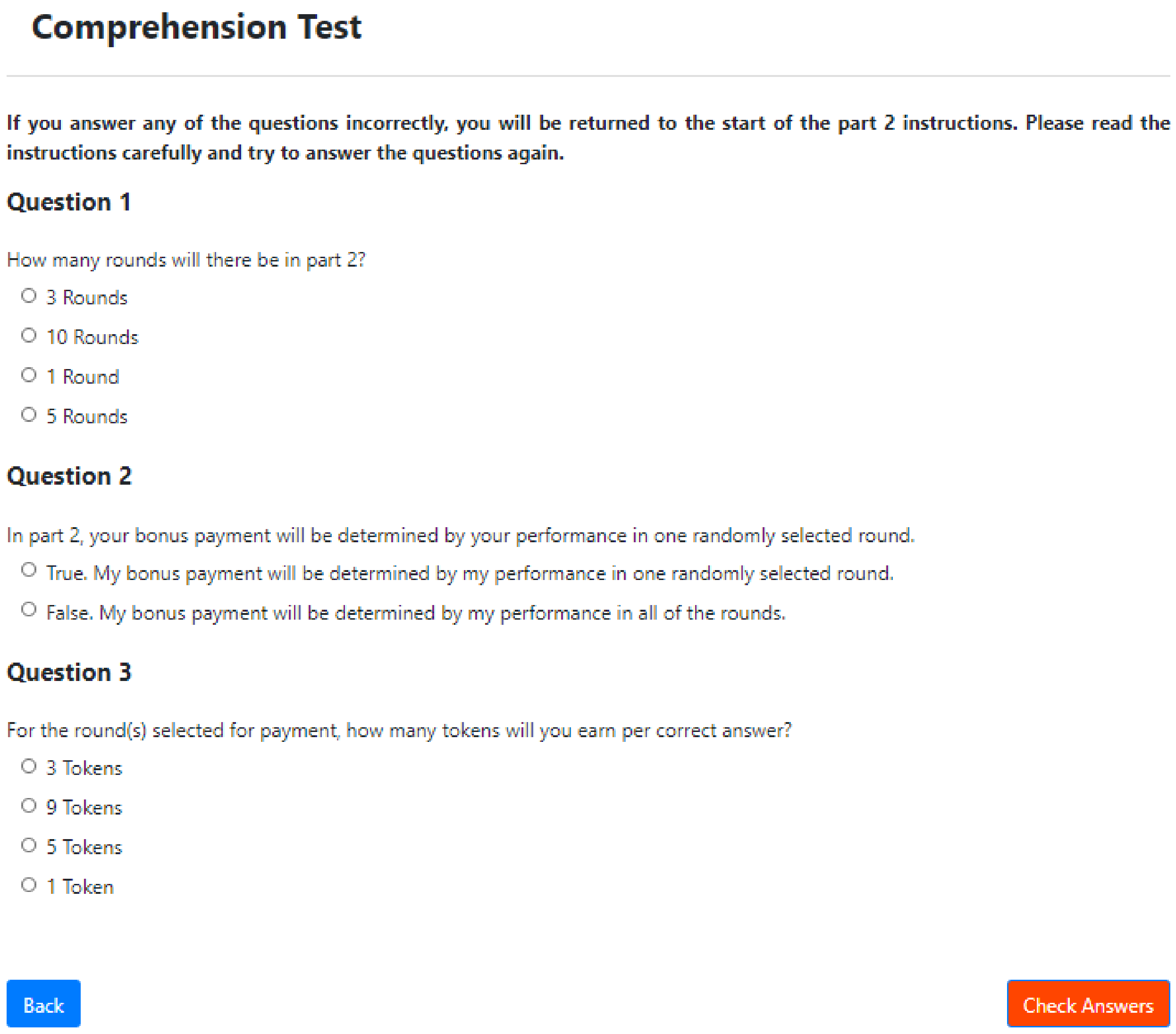
We will now begin with **part 2** which is similar to part 1, and you will be performing the task again. However, there are two changes.

First of all, you will participate in **3 rounds** of the task, each of which will last for **3 minutes** instead of 30 seconds. You should try to answer as many questions as possible during each 3 minute round. There will not be a practice task.

Secondly, your **bonus payments** will be determined by your performance in **one round**, which will be **randomly selected** at the end of the study. For that round, your **bonus payment** will be equal to **5 tokens per correct answer** (as before). You will not lose points for incorrect answers.

Before each round, you will be asked to **predict** how many **correct answers** you will score. After each round, you will also be asked to predict how many questions you answered in the round which you have just completed.

For each **correct prediction** you provide you will earn an additional **1 token**. If you make a correct prediction **both before and after** the round, you will earn **2 tokens**.



Comprehension Test

If you answer any of the questions incorrectly, you will be returned to the start of the part 2 instructions. Please read the instructions carefully and try to answer the questions again.

Question 1

How many rounds will there be in part 2?

○ 3 Rounds

○ 10 Rounds

○ 1 Round

○ 5 Rounds

Question 2

In part 2, your bonus payment will be determined by your performance in one randomly selected round.

○ True. My bonus payment will be determined by my performance in one randomly selected round.

○ False. My bonus payment will be determined by my performance in all of the rounds.

Question 3

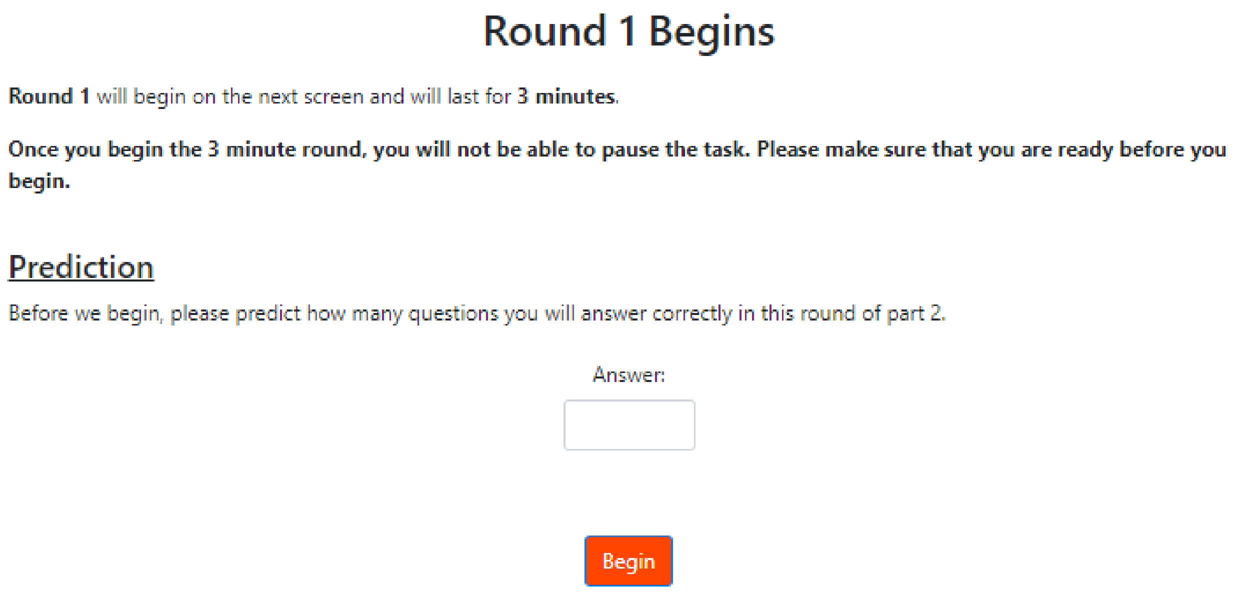
For the round(s) selected for payment, how many tokens will you earn per correct answer?

○ 3 Tokens

○ 9 Tokens

○ 5 Tokens

○ 1 Token



Round 1 Begins

Round 1 will begin on the next screen and will last for **3 minutes**.

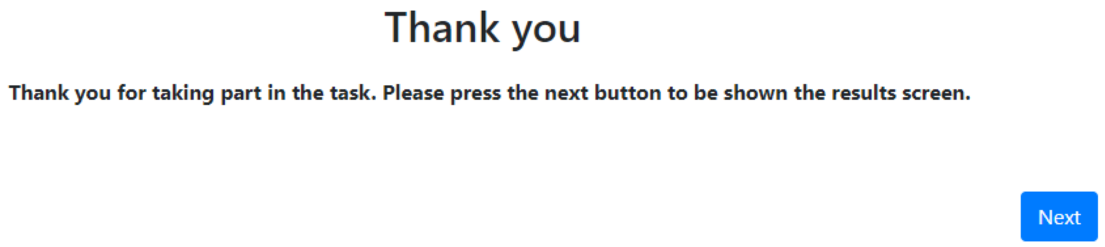
Once you begin the 3 minute round, you will not be able to pause the task. Please make sure that you are ready before you begin.

Prediction

Before we begin, please predict how many questions you will answer correctly in this round of part 2.

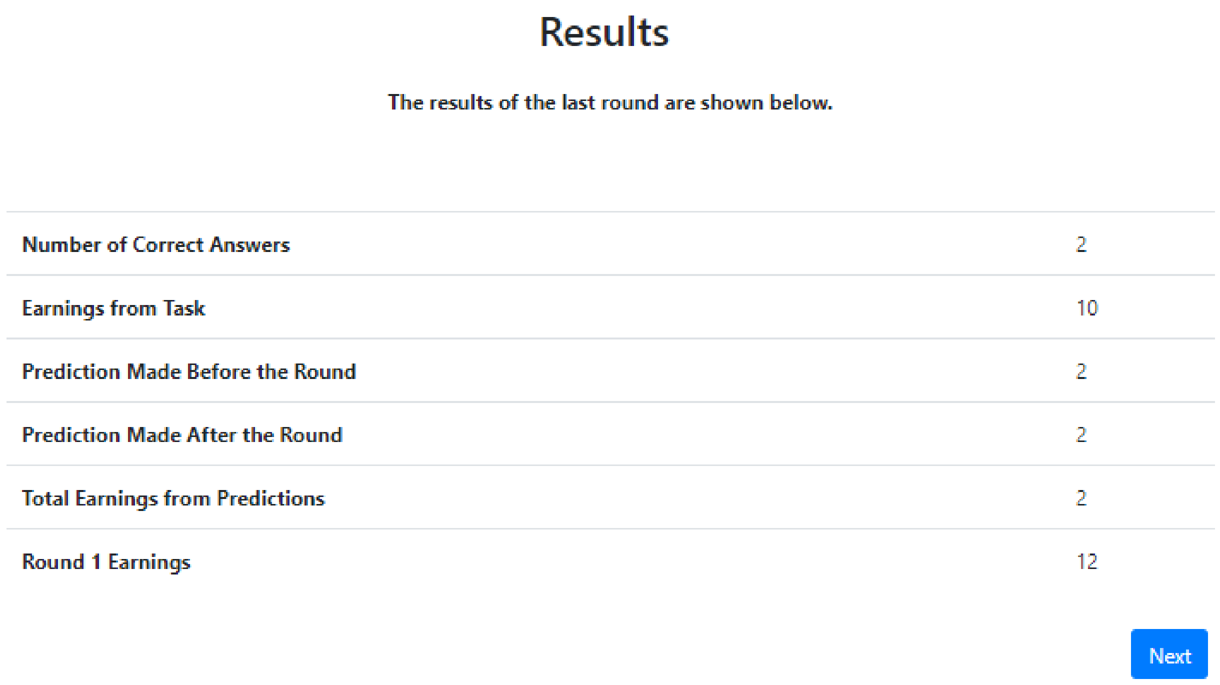
Answer:

TASK PAGE IS SHOWN AGAIN HERE



Thank you

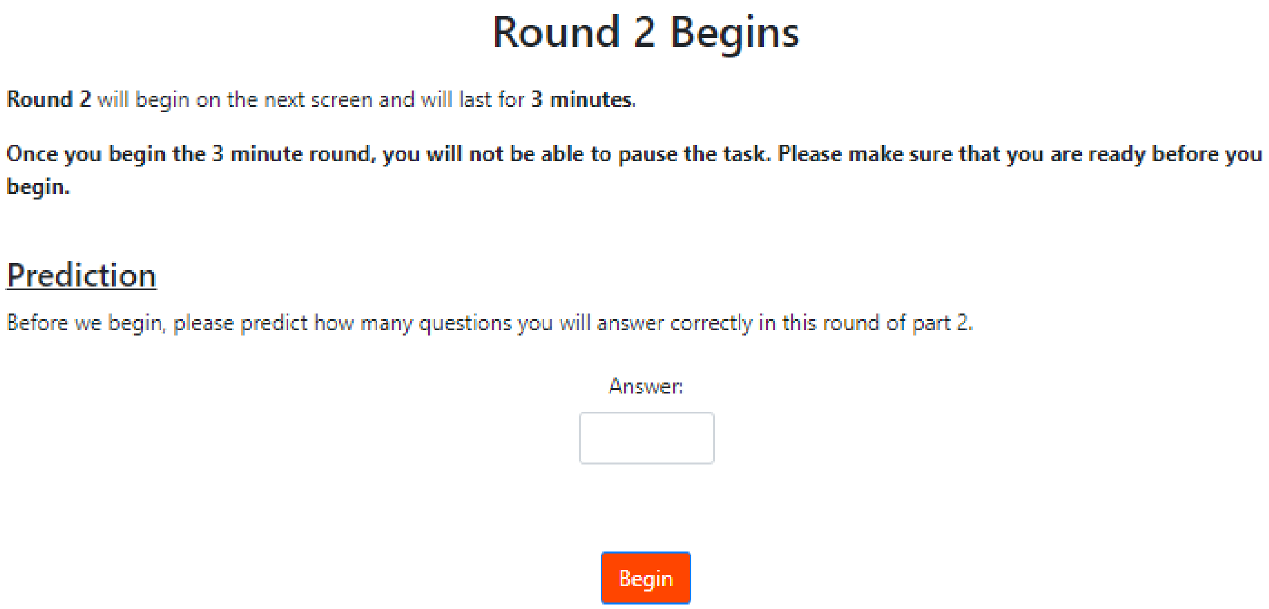
Thank you for taking part in the task. Please press the next button to be shown the results screen.



Results

The results of the last round are shown below.

Number of Correct Answers	2
Earnings from Task	10
Prediction Made Before the Round	2
Prediction Made After the Round	2
Total Earnings from Predictions	2
Round 1 Earnings	12



Round 2 Begins

Round 2 will begin on the next screen and will last for **3 minutes**.

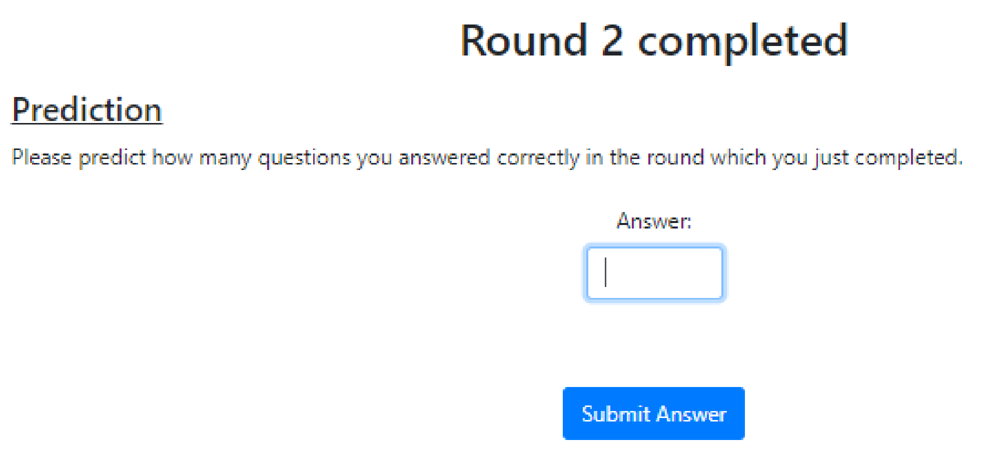
Once you begin the 3 minute round, you will not be able to pause the task. Please make sure that you are ready before you begin.

Prediction

Before we begin, please predict how many questions you will answer correctly in this round of part 2.

Answer:

TASK PAGE IS SHOWN AGAIN HERE

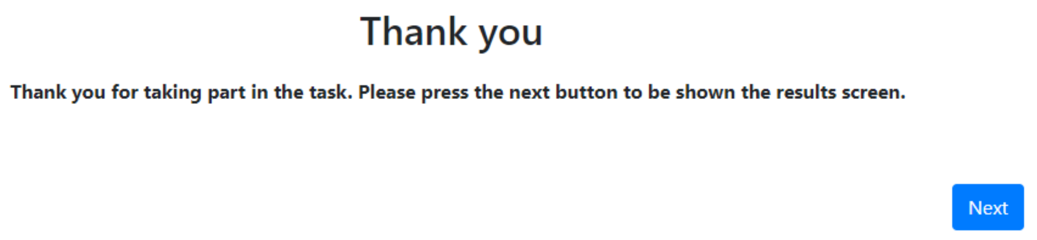


Round 2 completed

Prediction

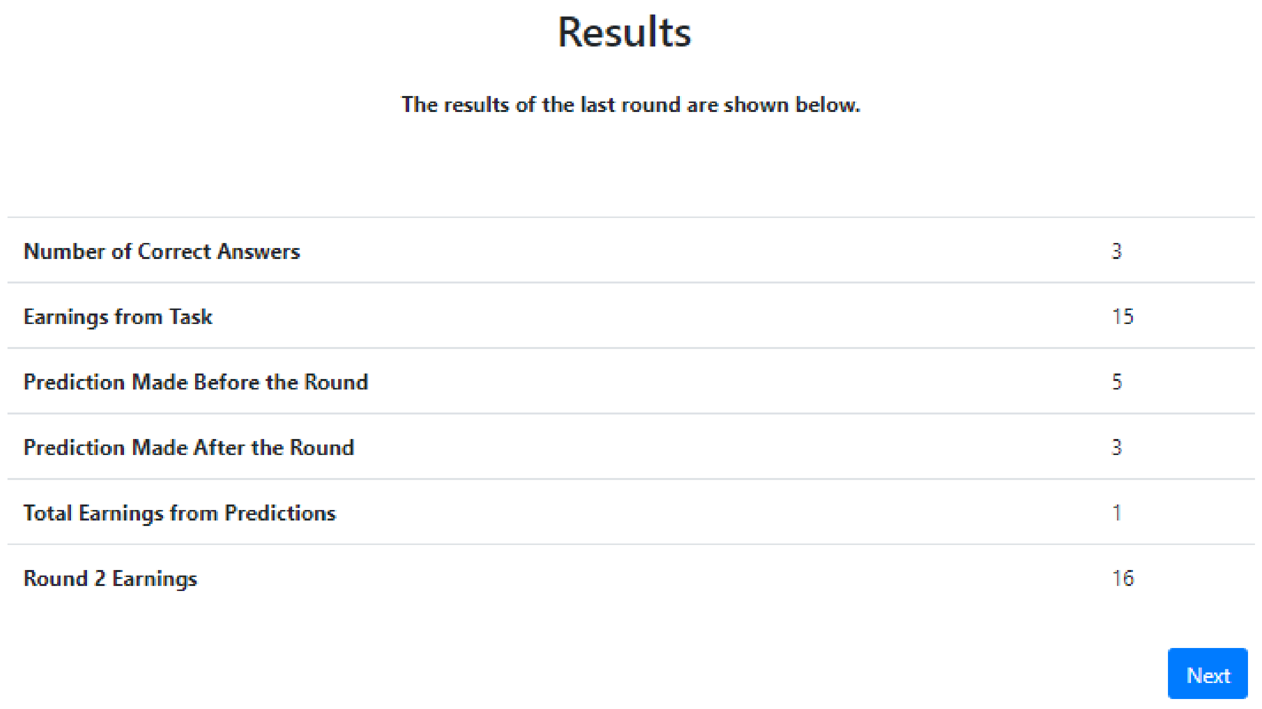
Please predict how many questions you answered correctly in the round which you just completed.

Answer:



Thank you

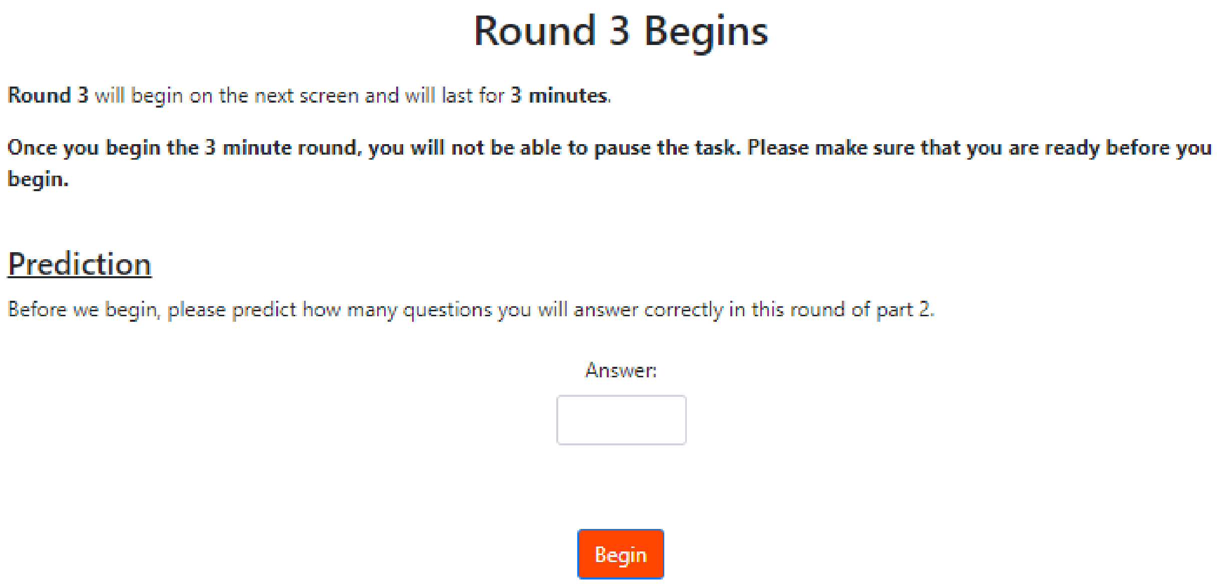
Thank you for taking part in the task. Please press the next button to be shown the results screen.



Results

The results of the last round are shown below.

Number of Correct Answers	3
Earnings from Task	15
Prediction Made Before the Round	5
Prediction Made After the Round	3
Total Earnings from Predictions	1
Round 2 Earnings	16



Round 3 Begins

Round 3 will begin on the next screen and will last for **3 minutes**.

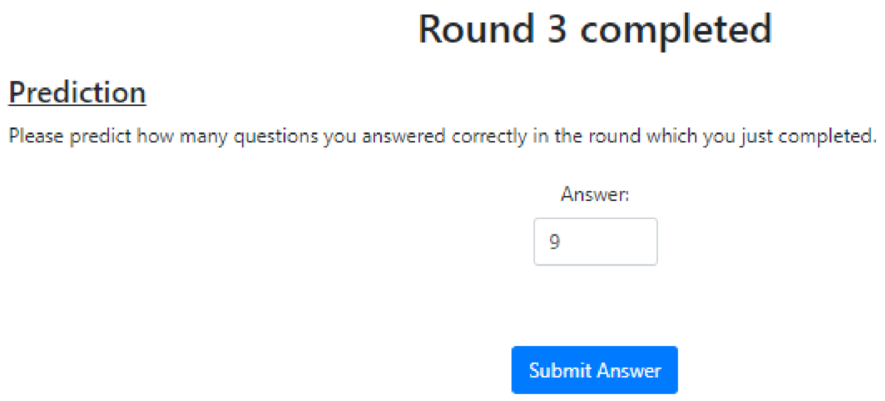
Once you begin the 3 minute round, you will not be able to pause the task. Please make sure that you are ready before you begin.

Prediction

Before we begin, please predict how many questions you will answer correctly in this round of part 2.

Answer:

TASK PAGE IS SHOWN AGAIN HERE

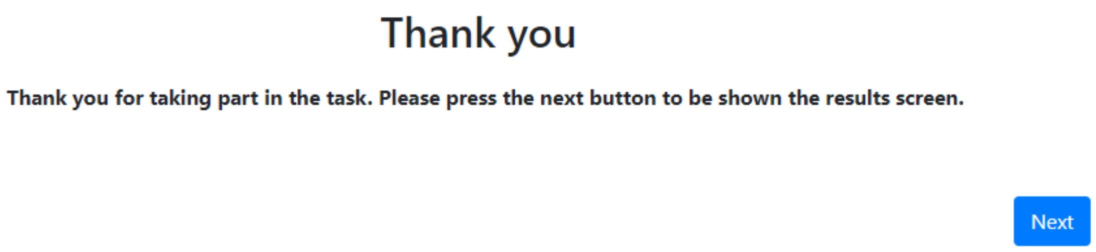


Round 3 completed

Prediction

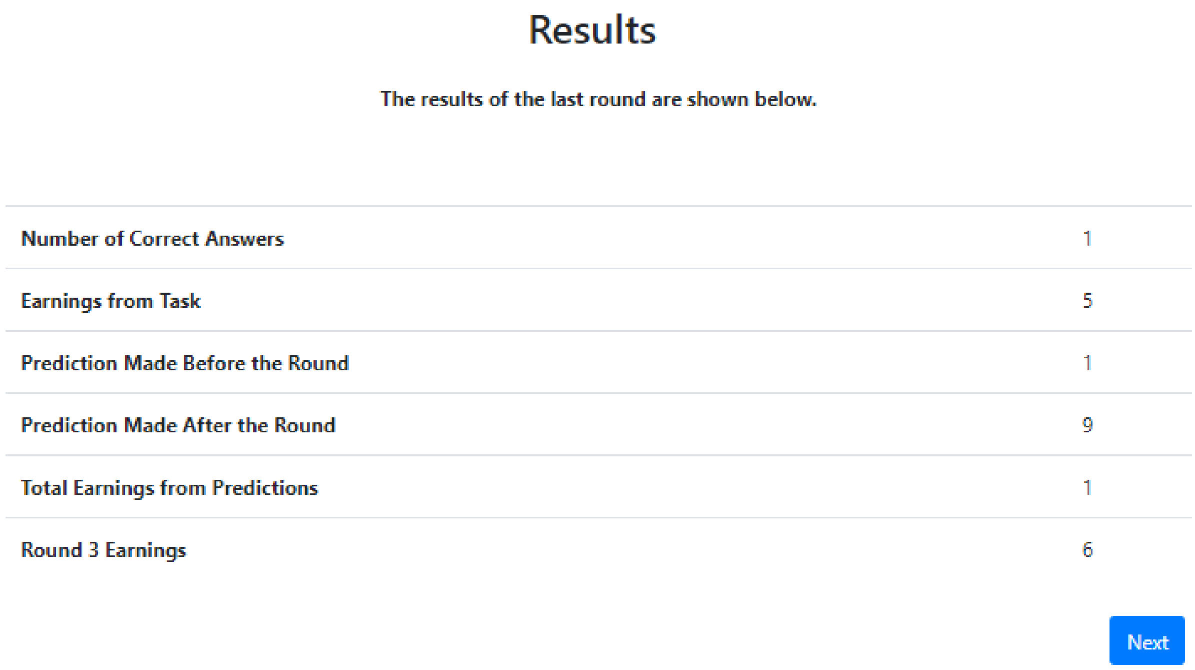
Please predict how many questions you answered correctly in the round which you just completed.

Answer: 9



Thank you

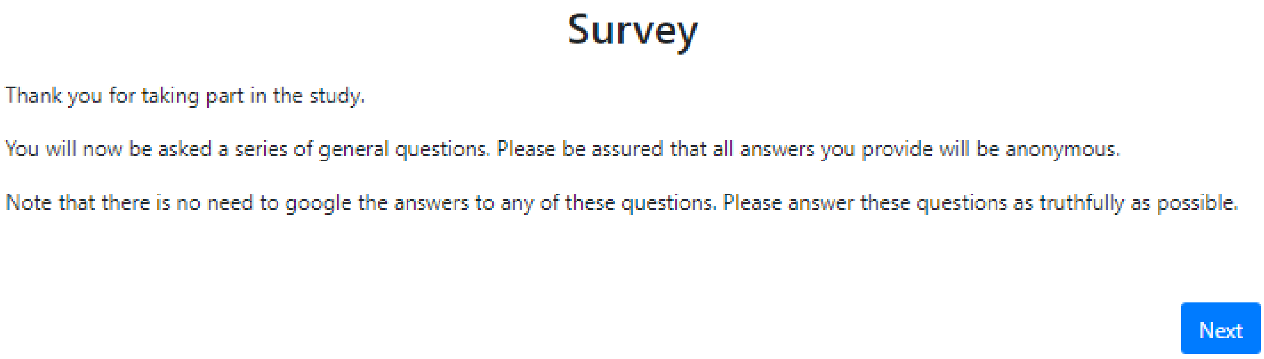
Thank you for taking part in the task. Please press the next button to be shown the results screen.



Results

The results of the last round are shown below.

Number of Correct Answers	1
Earnings from Task	5
Prediction Made Before the Round	1
Prediction Made After the Round	9
Total Earnings from Predictions	1
Round 3 Earnings	6

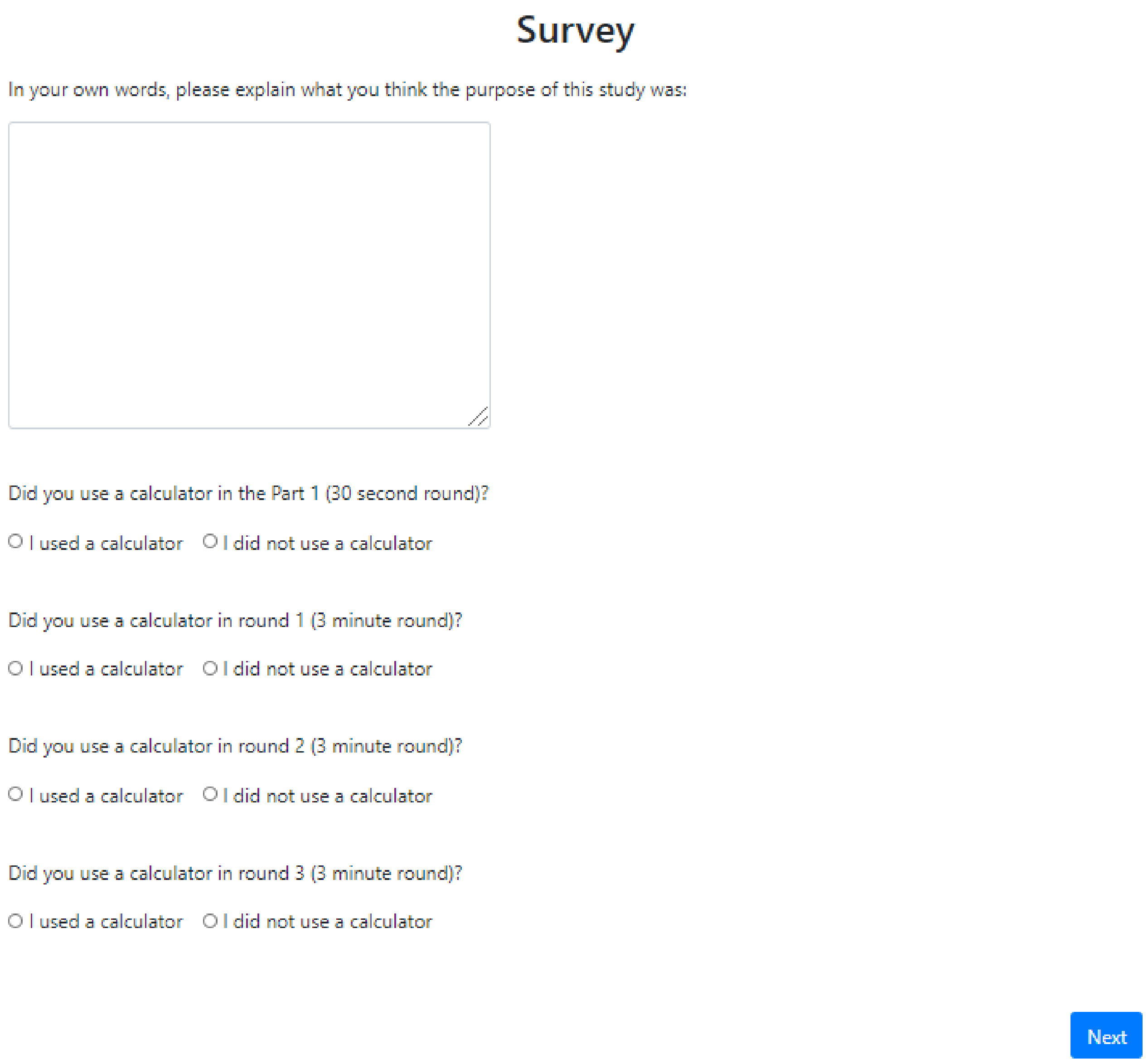


Survey

Thank you for taking part in the study.

You will now be asked a series of general questions. Please be assured that all answers you provide will be anonymous.

Note that there is no need to google the answers to any of these questions. Please answer these questions as truthfully as possible.



Survey

In your own words, please explain what you think the purpose of this study was:

Did you use a calculator in the Part 1 (30 second round)?

○ I used a calculator ○ I did not use a calculator

Did you use a calculator in round 1 (3 minute round)?

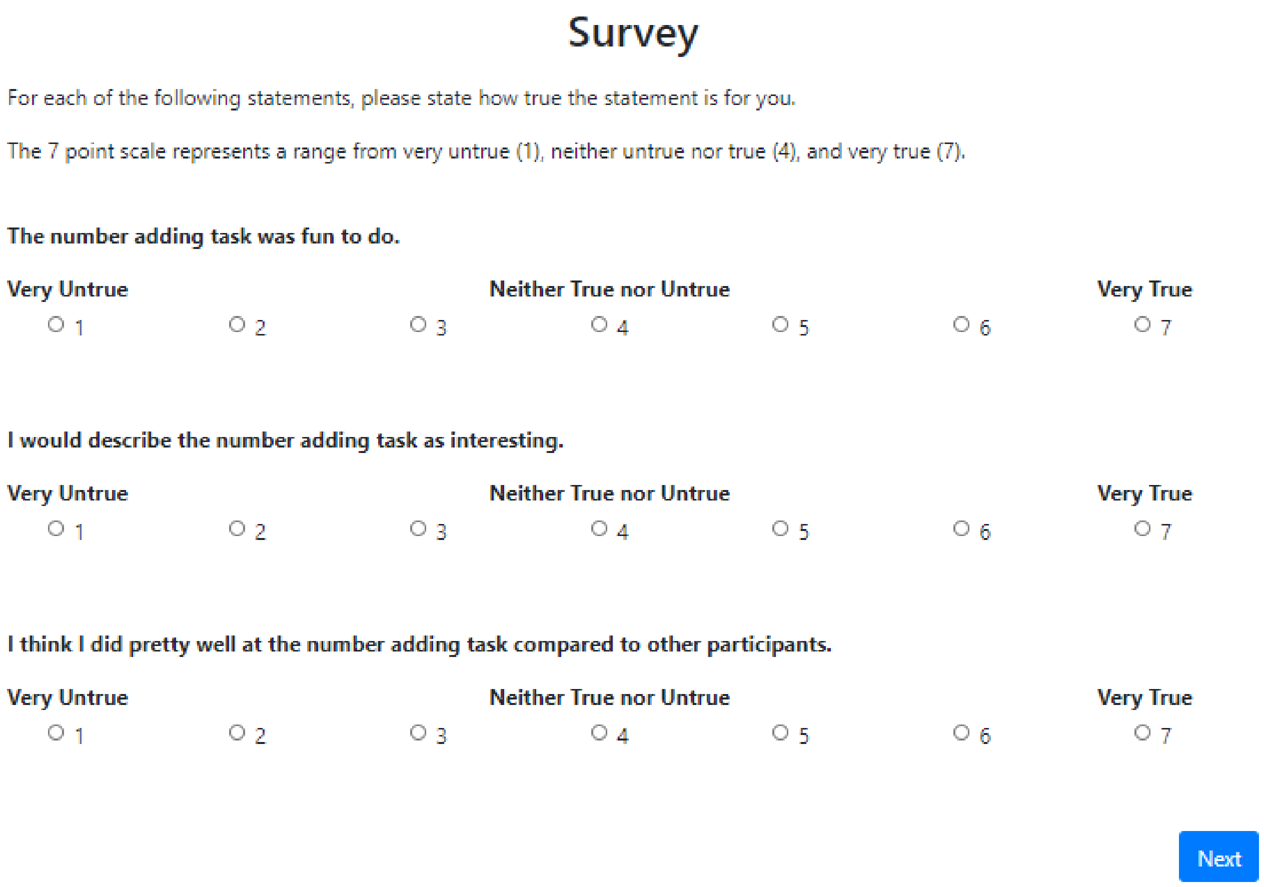
○ I used a calculator ○ I did not use a calculator

Did you use a calculator in round 2 (3 minute round)?

○ I used a calculator ○ I did not use a calculator

Did you use a calculator in round 3 (3 minute round)?

○ I used a calculator ○ I did not use a calculator



Survey

For each of the following statements, please state how true the statement is for you.

The 7 point scale represents a range from very untrue (1), neither untrue nor true (4), and very true (7).

The number adding task was fun to do.

Very Untrue **Neither True nor Untrue** **Very True**

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

I would describe the number adding task as interesting.

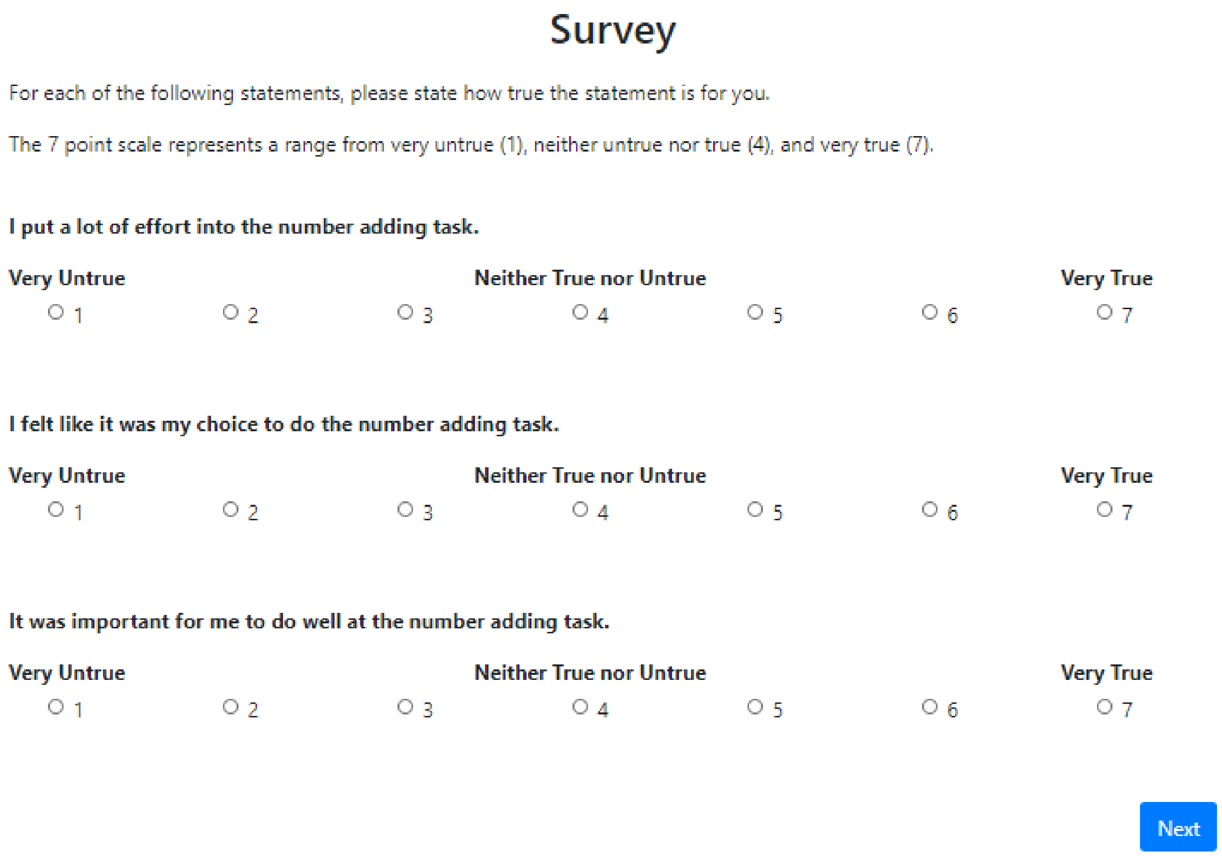
Very Untrue **Neither True nor Untrue** **Very True**

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

I think I did pretty well at the number adding task compared to other participants.

Very Untrue **Neither True nor Untrue** **Very True**

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7



Survey

For each of the following statements, please state how true the statement is for you.

The 7 point scale represents a range from very untrue (1), neither untrue nor true (4), and very true (7).

I put a lot of effort into the number adding task.

Very Untrue **Neither True nor Untrue** **Very True**

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

I felt like it was my choice to do the number adding task.

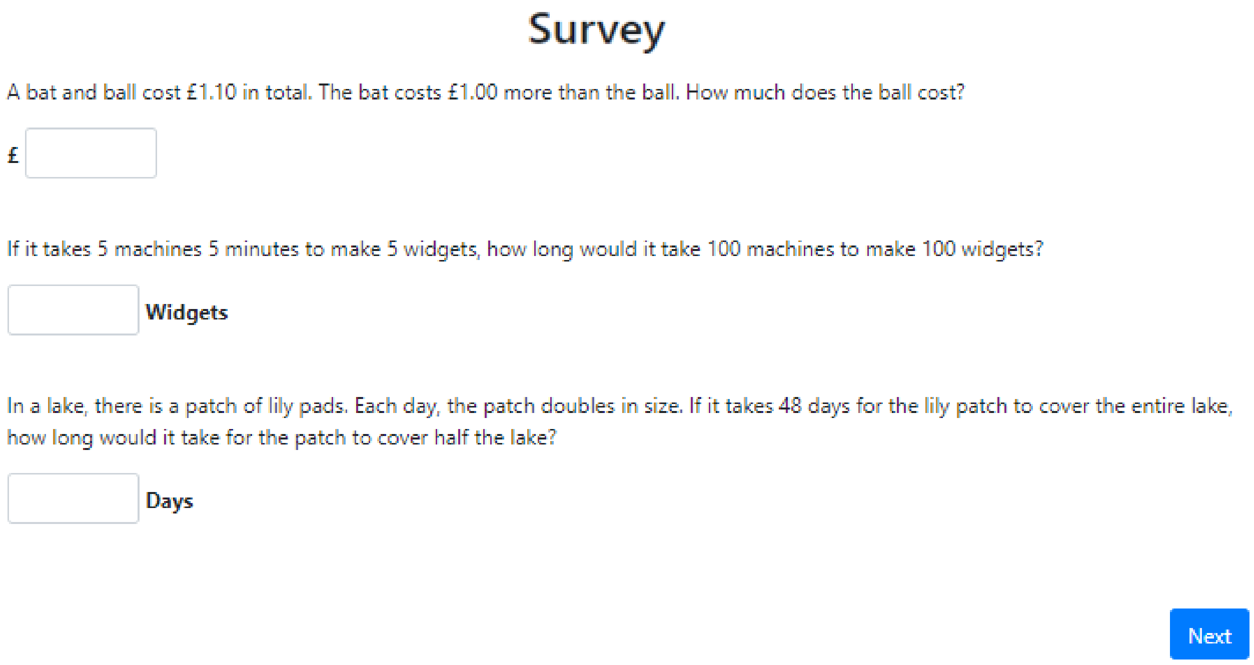
Very Untrue **Neither True nor Untrue** **Very True**

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

It was important for me to do well at the number adding task.

Very Untrue **Neither True nor Untrue** **Very True**

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7



Survey

A bat and ball cost £1.10 in total. The bat costs £1.00 more than the ball. How much does the ball cost?

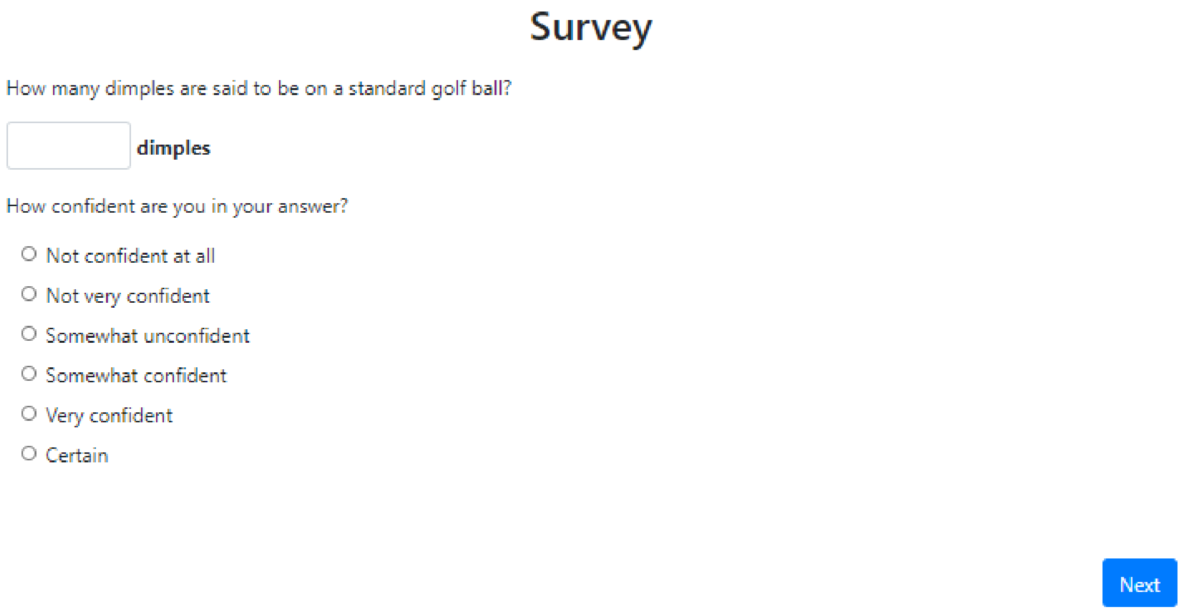
£

If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

Widgets

In a lake, there is a patch of lily pads. Each day, the patch doubles in size. If it takes 48 days for the lily patch to cover the entire lake, how long would it take for the patch to cover half the lake?

Days



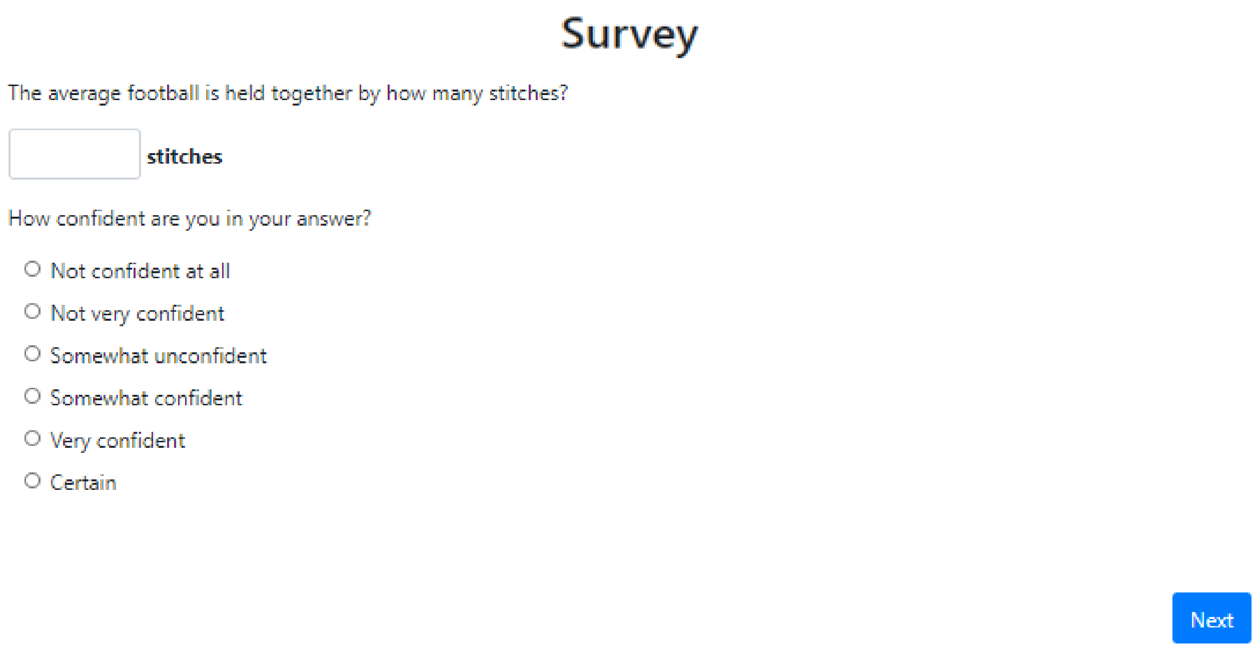
Survey

How many dimples are said to be on a standard golf ball?

dimples

How confident are you in your answer?

○ Not confident at all
○ Not very confident
○ Somewhat unconfident
○ Somewhat confident
○ Very confident
○ Certain



Survey

The average football is held together by how many stitches?

□ **stitches**

How confident are you in your answer?

○ Not confident at all

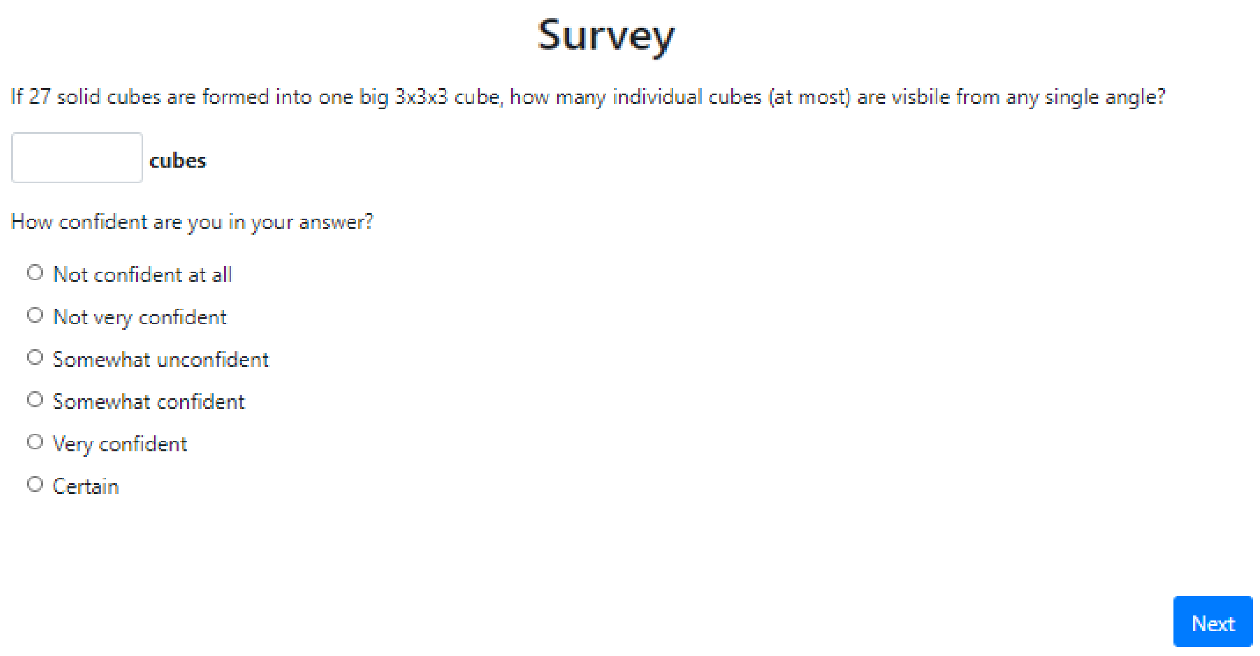
○ Not very confident

○ Somewhat unconfident

○ Somewhat confident

○ Very confident

○ Certain



Survey

If 27 solid cubes are formed into one big 3x3x3 cube, how many individual cubes (at most) are visbile from any single angle?

□ **cubes**

How confident are you in your answer?

○ Not confident at all

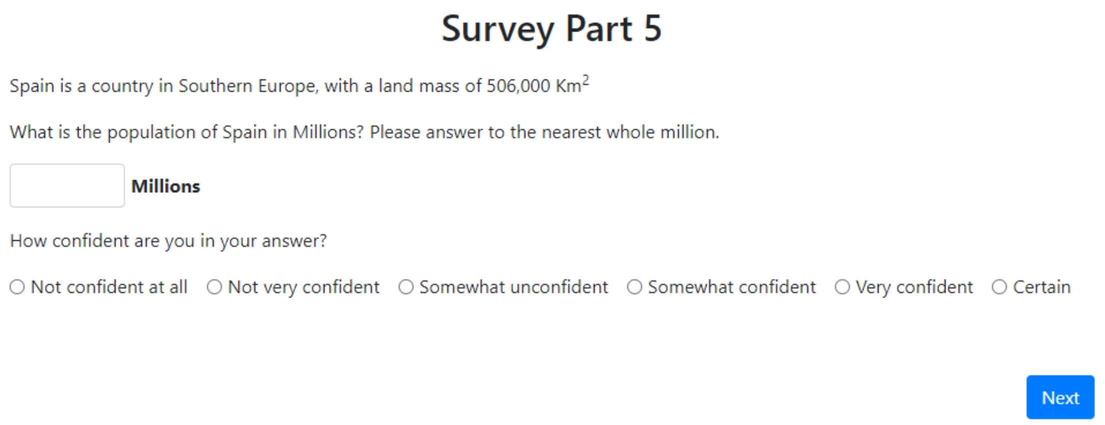
○ Not very confident

○ Somewhat unconfident

○ Somewhat confident

○ Very confident

○ Certain



Survey Part 5

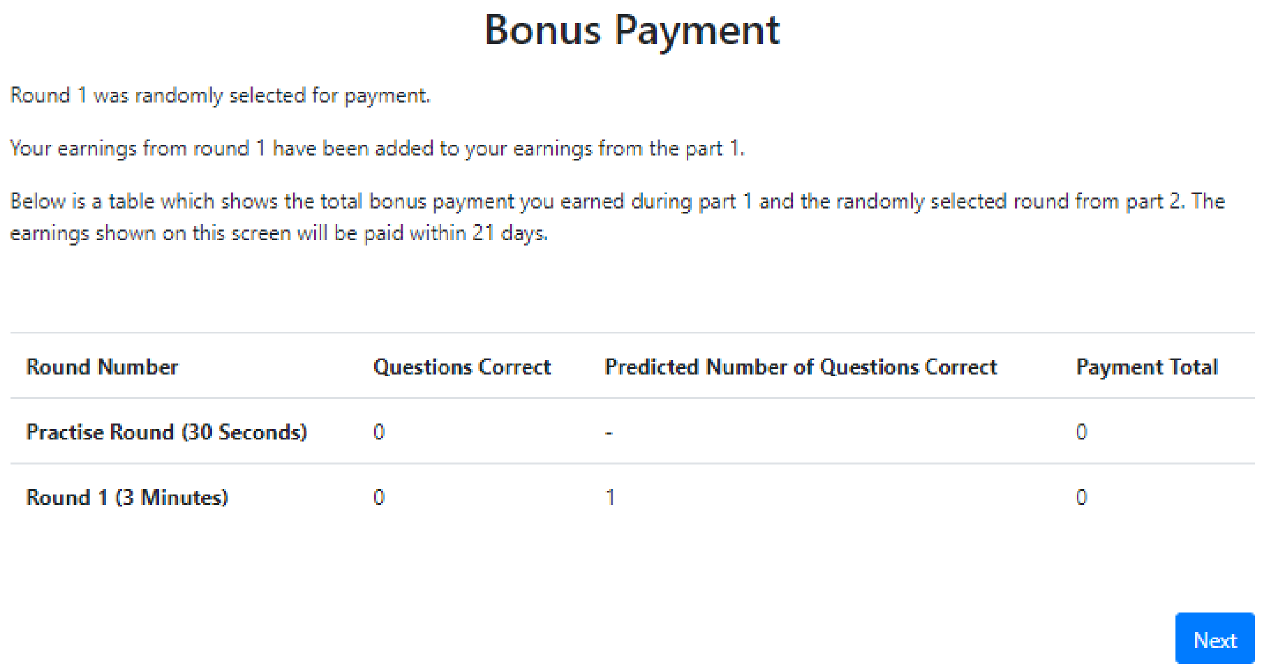
Spain is a country in Southern Europe, with a land mass of 506,000 Km^2

What is the population of Spain in Millions? Please answer to the nearest whole million.

Millions

How confident are you in your answer?

○ Not confident at all ○ Not very confident ○ Somewhat unconfident ○ Somewhat confident ○ Very confident ○ Certain



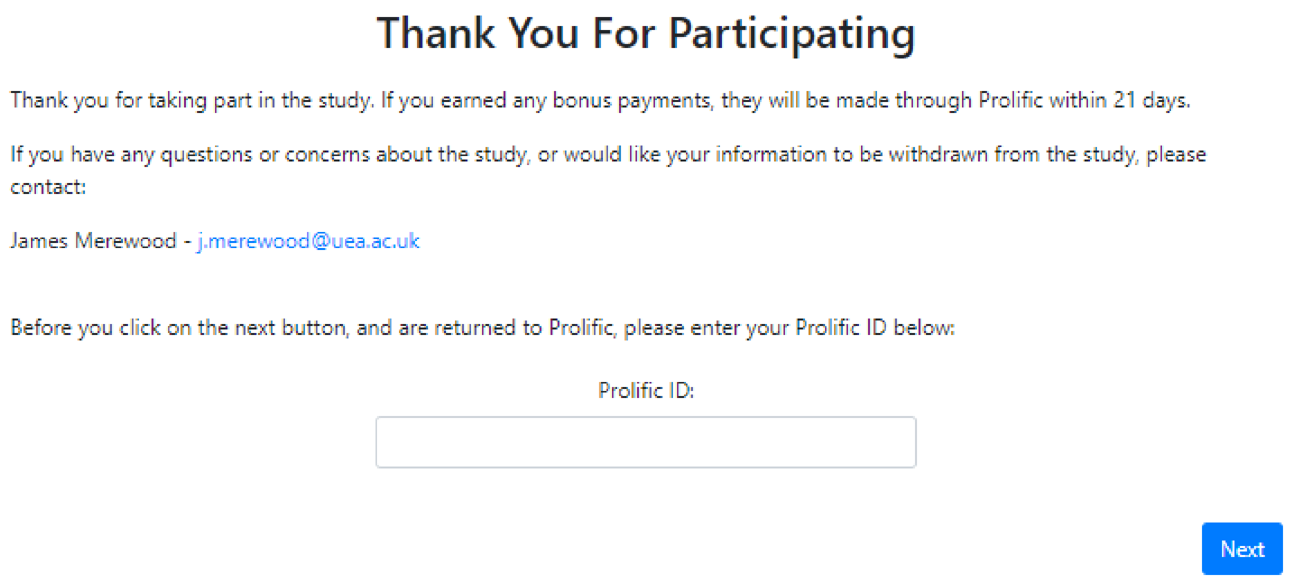
Bonus Payment

Round 1 was randomly selected for payment.

Your earnings from round 1 have been added to your earnings from the part 1.

Below is a table which shows the total bonus payment you earned during part 1 and the randomly selected round from part 2. The earnings shown on this screen will be paid within 21 days.

Round Number	Questions Correct	Predicted Number of Questions Correct	Payment Total
Practise Round (30 Seconds)	0	-	0
Round 1 (3 Minutes)	0	1	0



Thank You For Participating

Thank you for taking part in the study. If you earned any bonus payments, they will be made through Prolific within 21 days.

If you have any questions or concerns about the study, or would like your information to be withdrawn from the study, please contact:

James Merewood - j.merewood@uea.ac.uk

Before you click on the next button, and are returned to Prolific, please enter your Prolific ID below:

Prolific ID:

Annex 2 – Chapter 3 Instructions for the Experiment

General Information

Thank you for agreeing to take part in this Prolific study. The study is being run by researchers at the University of East Anglia. Any data which is collected about your participation in the study will be anonymous, and will not be linked to you in any way.

If you have any concerns at any time during the study, or would like to withdraw from the study, you may contact the lead researcher James Merewood, by sending an email to (j.merewood@uea.ac.uk) . Please note that if you withdraw from the study you will not receive payment for your participation.

The study you are about to take part in has received approval from the School of Economics Research Ethics Committee at the University of East Anglia. If you would like to make a formal complaint please contact the chair of the Research Ethics Committee Dr. David Hugh-Jones (d.hugh-jones@uea.ac.uk).

Please enter your Prolific ID in the box below.

Consent Form

Please carefully read the information below, and check the box at the bottom of the screen to provide your consent and would like to continue taking part in the study. If you do not provide consent, please return to Prolific and mark this study as 'Returned'.

(1) I am at least 18 years old.

(2) My participation in this study is voluntary, and I will have the opportunity to earn bonus payments based on my decisions during the study.

(3) I understand that data generated by my participation in this study will be analysed by researchers at the University of East Anglia, and will be stored in accordance with the University of East Anglia data protection guidelines.

(4) Anonymised data generated by my participation in this study may be used for research purposes, which includes being shared with other researchers.

(5) The researcher will collect my anonymised demographic information which I have previously provided to Prolific, and which will be linked to data generated by my participation in this study.

Overview

The study takes an average of 45 minutes to complete.

You may use a calculator during the study. Please report truthfully at the end of the study whether you used a calculator.

Stages of the Session

There will be a total of 4 parts to this study, which we will refer to as Stages. The instructions for each Stage will be provided once you get to each Stage.

In each Stage you will have the opportunity to earn bonus payments. The final bonus payment will be the sum of the bonus payments you have earned in all Stages.

How Will I Be Paid?

You will receive £3.75 for completing the study.

Throughout the study, you will earn tokens which will determine your bonus payment. Each token you earn will be worth £0.01.

Any bonus payments which you earn will be paid within 21 days after the study has concluded.

Attention Checks

Please note that there will be several attention checks during the study, which are meant to test whether you are paying attention.

If you fail to correctly answer two or more attention checks then your submission may be rejected, and you may not be paid.

Part 1 - Number-adding Task

In the session, you will have the opportunity to earn bonus payments. These payments will be generated by participating in a Number-adding task as described below. As this may be your first time performing such a task, there are detailed instructions provided below, and a chance for you to practice the task to ensure that you will do well.

The task consists of three numbers shown on screen. You will need to add all three numbers together, and type your answer into the box provided. Once you are happy with your answer, you will click on the submit button, and your response will be recorded. Then you will be shown another three numbers, and you will be expected to repeat the task.

An example is shown below.

In the example above, you will need to add the three numbers together (34 + 67 + 23 = 124) and type the answer (124) into the box provided.

If you would like to see **additional examples**, please click the grey button below. Otherwise please click 'next' to continue.

Practice Task

Now you will be given an opportunity to practise the task for 30 seconds.

Please note that you will **not** earn a bonus payment based on your performance during this free-form practice. The purpose of the free-form practice is for you to have an opportunity to become familiar with the task. Please click on the begin button to start free-form practise.

Please note that once you begin free-form practise, you will not be able to pause it. Please make sure that you are ready to proceed.

General Information - Paid Practise

Thank you for completing the free-form practice. Before we begin the Stages, you have the opportunity to practice the task once more for 30 seconds.

However, this time you will be paid 2 tokens per correct answer. It is important to note that you will earn money for correct answers, but will not lose money for incorrect answers. Note further that during the task itself, you will not be told if your answer is correct or incorrect.

On the next screen you will be asked a series of comprehension questions regarding the task. It is important that you answer all the questions correctly before you can proceed.

Comprehension Test

If you answer any of the questions incorrectly, you will be returned to the start of the instructions. Please read the instructions carefully and try to answer the questions again.

Question 1
What is the name of the task?

- ○ Investment Task
- ○ Number-adding Task
- ○ Number Subtraction Task
- ○ Sudoku Task

Question 2
How much you be paid for your performance in the task during the practise stage?

- ○ 5 tokens per answer regardless of whether the answer is correct or incorrect.
- ○ 3 tokens per correct answer and 1 token per incorrect answer.
- ○ 2 tokens per correct answer and no tokens for an incorrect answer.
- ○ 100 tokens for completing the whole task, regardless of how many answers you provide.

Question 3
What should you do during the task?

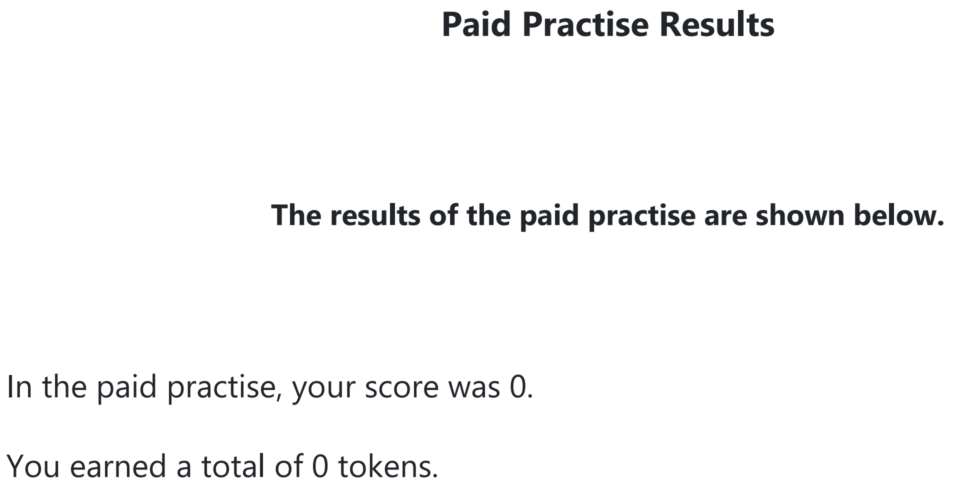
- ○ Add the three numbers together.
- ○ Subtract the second and third number from the first.
- ○ Multiply the three numbers together.
- ○ Divide the three numbers together.

Paid Practise Begins

Thank you for answering all comprehension test answers correctly. We will now begin paid practice. You will be given 30 seconds to complete the practice.

You will earn 2 tokens for each correct answer you provide. You will not lose any tokens for incorrect answers.

Once you begin paid practice, you will not be able to pause it. Please make sure that you are ready.



Paid Practise Results

The results of the paid practise are shown below.

In the paid practise, your score was 0.

You earned a total of 0 tokens.

General Instructions

Thank you for completing the paid practice. Your earnings have been recorded and will be added to the bonus payment.

Before we begin the first stage, there are a number of features specific to the session today that we need to inform you about. Following this, you will be given the instructions for Stage 1. It is important that you read all the instructions as you will be asked to pass comprehension tests before you are allowed to proceed further.

Groups

In the session today, you are one of six individuals participating together. You will be participating along with 5 other people (who have previously completed the study). These other people have completed the same task as you, faced the same pay structure as you, and completed the same number of Stages as you. Their data was collected ahead of time to reduce the amount of time you have to wait for others to complete their decisions. In the same fashion, this also means that the decisions you make may be used in other participants' sessions.

The decisions that these other people have made are important, because they may affect your bonus payment. Similarly, your decisions may affect the bonus payments of others that are yet to participate in the session. How this occurs will be explained shortly, but for now it is important that you remember that you are participating in the sessions with 5 other people (so, 6 participants altogether).

Treatment 1

Here is why these other participants are important. In the first Stages in the session, we will rank you and the other participants in Stage 1 at random. These ranks will be given entirely at random and will not depend on your performance. We will rank you from 1st to 6th at random.

These rankings are important because they **affect the bonus payment** you will earn in all future Stages. Based on the rank, those randomly allocated to 1st – 3rd are placed into a "Blue group" while those randomly allocated to 4th – 6th are placed into a "Yellow group". You will stay in these groups for all remaining Stages. The bonus payments are as follows:

- Blue group members earn 15 tokens from the Stage
- Yellow group members earn 5 tokens from the Stage

More detail will be provided in later Stages, but for now it is important to remember that your randomly allocated rank will impact future bonus payments.

Treatment 2 & 4

Here is why these other participants are important. In the first Stages in the session, your performance in that particular Stage will be compared with the performance of the other participants in that same Stage.

We will compare your performance in Stage 1. Based on your performance, and other participants performance in Stage 1, we will rank you from highest performer (1st) to lowest performer (6th). Note that if two people both have the same level of performance, one will be randomly selected to be ranked higher.

These rankings are important because they **affect the bonus payment** you will earn in all future Stages. Based on the rank, the 3 highest performers are placed into a "Blue group" while the 3 lowest performers are placed into a "Yellow group". You will stay in these groups for all remaining Stages. The bonus payments are as follows:

- Blue group members earn 15 tokens from the Stage
- Yellow group members earn 5 tokens from the Stage

More detail will be provided in later Stages, but for now it is important to remember that your performance in the first Stage can affect the group that you are placed into in later Stages. Therefore, it is in your best interest to try as hard as possible in each Stage.

Treatment 3 & 5

Here is why these other participants are important. In some of the Stages in the session, your performance in that particular Stage will be compared with the performance of the other participants in that same Stage.

For example, we may compare your performance in Stage 1. Based on your performance, and other participants performance in Stage 1, we will rank you from highest performer (1st) to lowest performer (6th). Note that if two people both have the same level of performance, one will be randomly selected to be ranked higher.

These rankings are important because they **affect the bonus payment** you will earn in the next Stage (for this example, in Stage 2). Based on the rank, the 3 highest performers are placed into a "Blue group" while the 3 lowest performers are placed into a "Yellow group." The bonus payments are as follows:

- Blue group members earn 15 tokens from the Stage
- Yellow group members earn 5 tokens from the Stage

More detail will be provided in later Stages, but for now it is important to remember that your performance in a particular Stage can affect the group that you are placed into in later Stages. Therefore, it is in your best interest to try as hard as possible in each Stage.

Feedback

In the session today, you will receive feedback about the number of questions you have answered correctly.

You will also receive feedback about the performance of the other individuals. However, you will not receive feedback about all other individuals, just for the individuals in your group.

For example, if you were in the Blue group, you would receive feedback about the other individuals in the Blue group. Similarly, if you were in the Yellow group, you would receive feedback about the other individuals in the Yellow group.

All feedback which you are shown about other participants in your group will always be accurate.

Treatment 1

Summary

In this session, you will participate in a series of Stages, each containing the number adding task you practiced earlier.

In Stage 1 we will randomly assign a rank to you and the 5 other participants in your session. Based on that randomly assigned rank, the first three ranks will be placed into the Blue group and the last three ranks will be placed into the Yellow group. You will remain in this group for the remaining Stages.

Once you are placed in a group, your earnings for that Stage will be determined by the group. Blue group members earn 15 tokens, while yellow group members earn 5 tokens for the Stage.

Your bonus payment will be the sum of payments in all Stages.

Treatment 2 & 4

Summary

In this session, you will participate in a series of Stages, each containing the number adding task you practiced earlier.

Your performance in Stage 1 will be compared with the 5 other participants in your session. Based on your performance and that of the other participants, the highest three performers will be placed into the Blue group and the lowest three performers into the Yellow group. You will remain in this group for the remaining Stages.

Once you are placed in a group, your earnings for that Stage will be determined by the group. Blue group members earn 15 tokens, while yellow group members earn 5 tokens for the Stage.

Your bonus payment will be the sum of payments in all Stages.

Treatment 3 & 5

Summary

In this session, you will participate in a series of Stages, each containing the number adding task you practiced earlier.

Your performance will be compared with the 5 other participants in your session. Based on your performance and that of the other participants, the highest three performers will be placed into the Blue group and the lowest three performers into the Yellow group.

Once you are placed in a group, your earnings for that Stage will be determined by the group. Blue group members earn 15 tokens, while yellow group members earn 5 tokens for the Stage.

Your bonus payment will be the sum of payments in all Stages.

Comprehension Test

If you answer any of the questions incorrectly, you will be returned to the start of the instructions. Please read the instructions carefully and try to answer the questions again.

Question 1
How many people, including you, are in the session?

- ○ 1 Person
- ● 6 People
- ○ 3 People
- ○ 10 People

Question 2
After we have ranked all the people in your session, which group will the highest 3 performers be placed into?

- ○ Blue group
- ○ Green group

Question 3
After each stage you will be shown feedback about your performance and the performance of others in your group.

- ○ False, I will be shown feedback about everyone in my session.
- ○ True, I will be shown feedback about my performance and those in my group.
- ○ False, I will only be shown feedback about my own performance.

Treatment 1

Stage 1

Thank you for completing the comprehension test. We are now ready to begin Stage 1.

In this Stage, you will be performing the Number-adding task. For this Stage, and for later Stages, the Number-adding task will last for 3 minutes. The rules for the Number-adding task are exactly the same as earlier. Click here for a reminder of the instructions for this task.

Note that you will earn a bonus payment of 5 tokens for this Stage, regardless of your performance in the Number-adding task.

Recall that once you have completed the task, we will randomly assign you to either the Blue or the Yellow group. Once in a group, Blue group members are paid 15 tokens, while Yellow group members are paid 5 tokens.

Please click the button below to take a comprehension test over these instructions.

Treatment 2

Thank you for completing the comprehension test. We are now ready to begin Stage 1.

In this Stage, you will be performing the Number-adding task. For this Stage, and for later Stages, the Number-adding task will last for 3 minutes. The rules for the Number-adding task are exactly the same as earlier. Click here for a reminder of the instructions for this task.

Note that you will earn a bonus payment of 5 tokens for this Stage, regardless of your performance in the Number-adding task. This is done so as to place you in either the Blue or the Yellow group for later stages.

Recall that the first three ranks are placed into the Blue group, while the last three ranks are placed into the Yellow group. Once in a group, Blue group members are paid 15 tokens, while Yellow group members are paid 5 tokens. You will remain in these groups for the remainder of the session.

Please click the button below to take a comprehension test over these instructions.

Treatment 3

Thank you for completing the comprehension test. We are now ready to begin Stage 1.

In this Stage, you will be performing the Number-adding task. For this Stage, and for later Stages, the Number-adding task will last for 3 minutes. The rules for the Number-adding task are exactly the same as earlier. Click here for a reminder of the instructions for this task.

Note that you will earn a bonus payment of 5 tokens for this Stage, regardless of your performance in the Number-adding task. However, your performance can affect your bonus payment for later Stages: your performance in this Stage will be compared with other individuals in this session. This is done so as to place you in either the Blue or the Yellow group for later stages.

Recall that the highest 3 performers are placed into the Blue group, while the lowest 3 performance are placed into the Yellow group. Once in a group, Blue group members are paid 15 tokens, while Yellow group members are paid 5 tokens.

Please click the button below to take a comprehension test over these instructions.

Treatment 4

Thank you for completing the comprehension test. We are now ready to begin Stage 1.

In this Stage, you will be performing the Number-adding task. For this Stage, and for later Stages, the Number-adding task will last for 6 minutes. The rules for the Number-adding task are exactly the same as earlier. Click here for a reminder of the instructions for this task.

Note that you will earn a bonus payment of 5 tokens for this Stage, regardless of your performance in the Number-adding task. This is done so as to place you in either the Blue or the Yellow group for later stages.

Recall that the first three ranks are placed into the Blue group, while the last three ranks are placed into the Yellow group. Once in a group, Blue group members are paid 15 tokens, while Yellow group members are paid 5 tokens. You will remain in these groups for the remainder of the session.

Please click the button below to take a comprehension test over these instructions.

Treatment 5

Thank you for completing the comprehension test. We are now ready to begin Stage 1.

In this Stage, you will be performing the Number-adding task. For this Stage, and for later Stages, the Number-adding task will last for 6 minutes.   The rules for the Number-adding task are exactly the same as earlier. Click here for a reminder of the instructions for this task.

Note that you will earn a bonus payment of 5 tokens for this Stage, regardless of your performance in the Number-adding task. However, your performance can affect your bonus payment for later Stages: your performance in this Stage will be compared with other individuals in this session. This is done so as to place you in either the Blue or the Yellow group for later stages.

Recall that the highest 3 performers are placed into the Blue group, while the lowest 3 performance are placed into the Yellow group. Once in a group, Blue group members are paid 15 tokens, while Yellow group members are paid 5 tokens.

Please click the button below to take a comprehension test over these instructions.

Comprehension Test – Questions differ based on treatment with answers corresponding to the correct instructions from that particular treatment

If you answer any of the questions incorrectly, you will be returned to the start of the instructions. Please read the instructions carefully and try to answer the questions again.

Question 1
How long will stage 1 last for?

- ○ 1 Minute
- ○ 5 Minutes
- ○ 3 Minutes

Question 2
How many tokens will you earn for completing stage 1?

- ○ 5 tokens regardless of your performance.
- ○ 2 tokens per correct answer.
- ○ 10 tokens regardless of your performance

Question 3
Your performance in stage 1 might influence your earnings in future stages of the study.

- ○ True - The group I will be placed into in stage 2 will depend on my performance in stage 1.
- ○ False - I will be randomly ranked and placed into a group, and this will influence the earnings which I can make in future stages.

Treatments 1, 2, and 3

Stage 1 Begins

Stage 1 will begin on the next screen and will last for 3 minutes.

Once you begin the 3 minute stage, you will not be able to pause the task. Please make sure that you are ready before you begin.

Treatments 4 and 5

Stage 1 Begins

Stage 1 will begin on the next screen and will last for 6 minutes.

Once you begin the 6 minute stage, you will not be able to pause the task. Please make sure that you are ready before you begin.

Treatment 1

Stage 1 Complete

Thank you for completing stage 1.

You have earned **5 tokens** for this stage. This will be added to your bonus payment.

You will now be randomly assigned a rank from 1st to 6th and the software will place the first three ranked performers into the Blue group and the last three ranked performers into the Yellow group. You will remain in these groups for the rest of the session.

Please click on the button below to receive your feedback for this stage.

Treatment 2 & 4

Stage 1 Complete

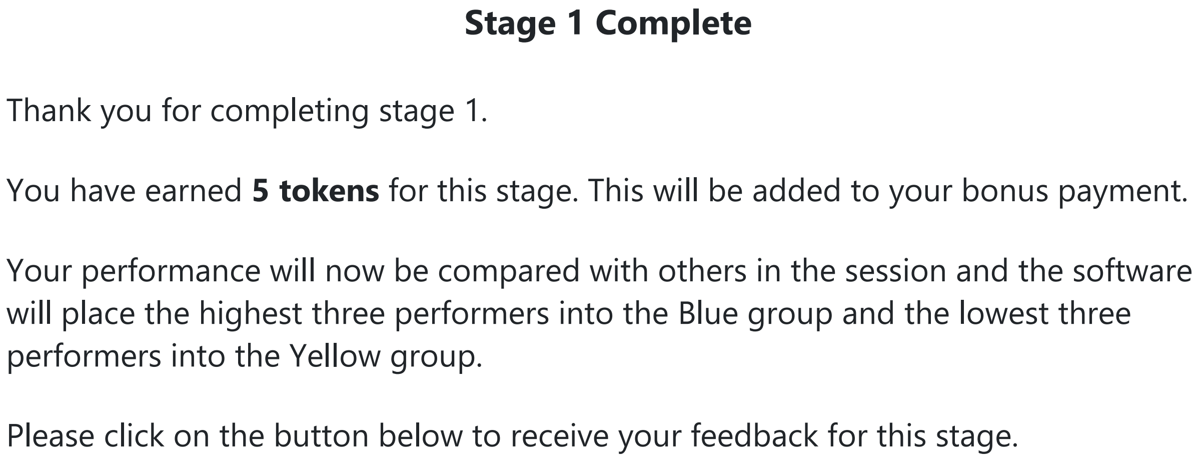
Thank you for completing stage 1.

You have earned **5 tokens** for this stage. This will be added to your bonus payment.

Your performance will now be compared with others in the session and the software will place the highest three performers into the Blue group and the lowest three performers into the Green group. You will remain in these groups for the rest of the session.

Please click on the button below to receive your feedback for this stage.

Treatment 3 & 5



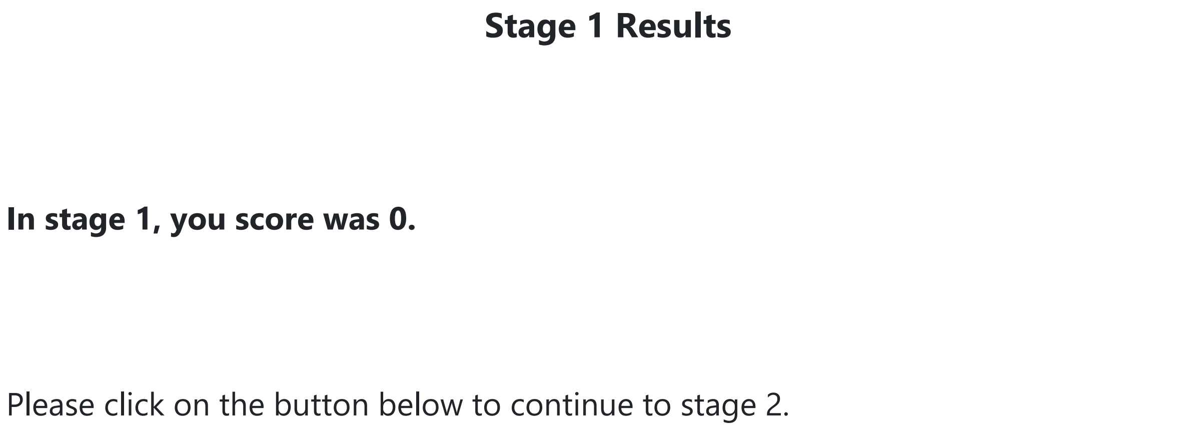
Stage 1 Complete

Thank you for completing stage 1.

You have earned **5 tokens** for this stage. This will be added to your bonus payment.

Your performance will now be compared with others in the session and the software will place the highest three performers into the Blue group and the lowest three performers into the Yellow group.

Please click on the button below to receive your feedback for this stage.



Stage 1 Results

In stage 1, you score was 0.

Please click on the button below to continue to stage 2.

Treatment 1

Stage 2 Instructions

Here are the instructions for Stage 2. In this Stage, you will be performing the Number-Adding Task again as you did before.

For this Stage, you will earn 5 tokens because you are in the Yellow group.

You were randomly assigned a rank from 1st to 6th. You were randomly ranked in the group from 4th – 6th and are therefore in the Yellow group.

You will remain in the Yellow group for the remaining Stages.

After Stage 2 you will be shown feedback about how your performance compares with other participants your group.

Treatments 2 & 4

Here are the instructions for Stage 2. In this Stage, you will be performing the Number-Adding Task again as you did before.

For this Stage, you will earn 5 tokens because you are in the Yellow group.

We ranked you alongside the 5 other participants in the session based on your performance in Stage 1. You were ranked in the 3 lowest participants and are therefore in the Yellow group.

You will remain in the Yellow group for the remaining Stages.

After Stage 2 you will be shown feedback about how your performance compares with other participants your group.

Treatments 3 & 5

Here are the instructions for Stage 2. In this Stage, you will be performing the Number-Adding Task again as you did before.

For this Stage, you will earn 5 tokens because you are in the Yellow group.

We ranked you alongside the 5 other participants in the session based on your performance in Stage 1. You were ranked in the 3 lowest participants and are therefore in the Yellow group.

At the start of the next Stage we will place the highest 3 performing participants into the Blue group and the lowest 3 performing participants into the Yellow group. Therefore it is in your best interest to try as hard as possible.

After Stage 2 you will be shown feedback about how your performance compares with other participants your group.

Comprehension Test

If you answer any of the questions incorrectly, you will be returned to the start of the instructions. Please read the instructions carefully and try to answer the questions again.

Question 1
How many people have been placed into the Blue group?

- ○ 6 people
- ○ 3 people
- ○ 2 people

Question 2
How many tokens will you earn if you have been placed into the Yellow group?

- ○ 2 tokens per correct answer.
- ○ 5 tokens regardless of performance.
- ○ 15 tokens regardless of performance.

Treatments 1, 2 & 3

Stage 2 Begins

Stage 2 will begin on the next screen and will last for 3 minutes.

For this Stage, you will earn 5 tokens because you are in the Yellow group.

Once you begin the 3 minute stage, you will not be able to pause the task. Please make sure that you are ready before you begin.

Treatments 4 & 5

Stage 2 Begins

Stage 2 will begin on the next screen and will last for 6 minutes.

For this Stage, you will earn 5 tokens because you are in the Yellow group.

Once you begin the 3 minute stage, you will not be able to pause the task. Please make sure that you are ready before you begin.

Treatment 1

Before we Continue

Before you are shown the results of this Stage, we would like you to make a quick prediction.

You will remain in the same group for all future Stages. However we are interested in whether you think that you would switch between groups if this were possible.

Imagine that we took the performance of the other 5 participants in your session, and ranked you from highest (1st) to lowest (6th) based on your performance and the performance of the other participants in your session in Stage 2. Now imagine that, based on this ranking, we placed you into a new group, where the highest performing 3 participants were in a new Blue group and the lowest 3 performing were in a new Yellow group.

Please predict which group you think that you would be placed into for Stage 3, if the above scenario were to really happen. If you think you would be placed into the group which would contain the highest 3 participants, please select "Blue" from the following options. If you think you would be placed into the group which would contain the lowest 3 participants, please select "Yellow".

Treatment 2 & 4

Before you are shown the results of this Stage, we would like you to make a quick prediction.

You will remain in the same group for all future Stages. However we are interested in whether you think that you would switch between groups if this were possible.

Imagine that we took the performance of the other 5 participants in your session, and re-ranked you from highest (1st) to lowest (6th) based on your performance in Stage 3. Now imagine that, based on this ranking, we placed you into a new group, where the highest 3 people were in a new Blue group and the lowest 3 were in a new Yellow group.

Please predict which group you think that you would be placed into for Stage 4, if the above scenario were to really happen. If you think you would be placed into the group which would contain the highest 3 participants, please select "Blue" from the following options. If you think you would be placed into the group which would contain the lowest 3 participants, please select "Yellow".

Treatment 3 & 5

Before we continue, please answer the question below.

We will be reranking you and the other participants in the session based on your and their performance in Stage 2. The highest 3 participants will be placed into the 'Blue' group and the lowest 3 participants into the 'Yellow' group.

Which group do you think that you will be placed in, in Stage 3? If you think that you will be in the Blue group, please select Blue. If you think that you will be in the Yellow group, please select Yellow.

Results

You were in the Yellow group, and earned 5 tokens for this stage.

The score of the first ranked person in your group was 11

The score of the second ranked person in your group was 7

Your rank in this round was 3 and your score was 0

Treatment 1

Stage 3 Instructions

Here are the instructions for Stage 3. In this Stage, you will be performing the Number-Adding Task again as you did before.

For this Stage, you will earn 5 tokens because you are in the Yellow group.

You were randomly assigned a rank from 1st to 6th in Stage 1. You were randomly ranked in the group from 4th – 6th and are therefore in the Yellow group.

You will remain in the Yellow group for all remaining Stages.

After Stage 3 you will be shown feedback about how your performance compares with other participants your group.

Treatments 2 & 4

Here are the instructions for Stage 3. In this Stage, you will be performing the Number-Adding Task again as you did before.

For this Stage, you will earn 5 tokens because you are in the Yellow group.

We ranked you alongside the 5 other participants in the session based on your performance in Stage 1. You were ranked in the 3 lowest participants and are therefore in the Yellow group.

You will remain in the Yellow group for all remaining Stages.

After Stage 3 you will be shown feedback about how your performance compares with other participants your group.

Treatments 3 & 5

Here are the instructions for Stage 3. In this Stage, you will be performing the Number-Adding Task again as you did before.

For this Stage, you will earn 5 tokens because you are in the Yellow group.

We ranked you alongside the 5 other participants in the session based on your performance in Stage 1. You were ranked in the 3 lowest participants and are therefore in the Yellow group.

At the start of the next Stage we will place the highest 3 performing participants into the Blue group and the lowest 3 performing participants into the Yellow group. Therefore it is in your best interest to try as hard as possible.

After Stage 3 you will be shown feedback about how your performance compares with other participants your group.

Treatments 1, 2 & 3

Stage 3 Begins

Stage 3 will begin on the next screen and will last for 3 minutes.

For this Stage, you will earn 5 tokens because you are in the Yellow group.

Once you begin the 3 minute stage, you will not be able to pause the task. Please make sure that you are ready before you begin.

Treatments 4 & 5

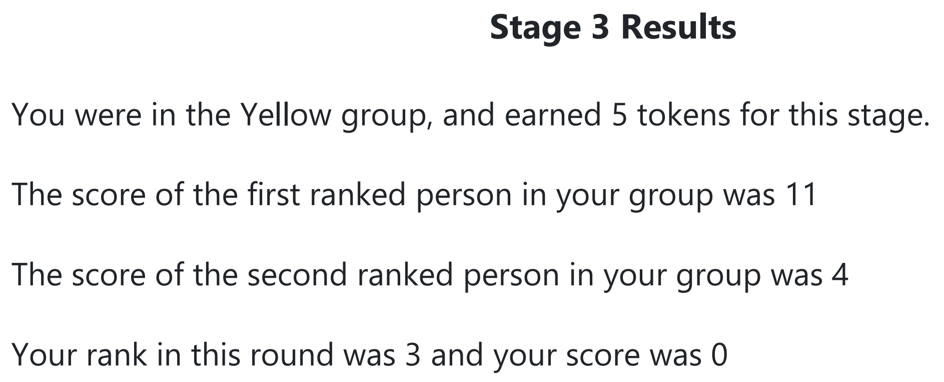
Stage 3 Begins

Stage 3 will begin on the next screen and will last for 6 minutes.

For this Stage, you will earn 5 tokens because you are in the Yellow group.

Once you begin the 3 minute stage, you will not be able to pause the task. Please make sure that you are ready before you begin.

Prediction Screens Show Again As Before Depending on Treatment



Stage 3 Results

You were in the Yellow group, and earned 5 tokens for this stage.

The score of the first ranked person in your group was 11

The score of the second ranked person in your group was 4

Your rank in this round was 3 and your score was 0

Treatment 1

Stage 4 Instructions

Here are the instructions for Stage 4. In this Stage, you will be performing the Number-Adding Task again as you did before. This will be the final round of the task.

For this Stage, you will earn 5 tokens because you are in the Yellow group.

You were randomly assigned a rank from 1st to 6th in Stage 1. You were randomly ranked in the group from 4th – 6th and are therefore in the Yellow group.

After Stage 4 you will be shown feedback about how your performance compares with other participants your group.

Treatment 2 & 4

Here are the instructions for Stage 4. In this Stage, you will be performing the Number-Adding Task again as you did before. This will be the final round of the task.

For this Stage, you will earn 5 tokens because you are in the Yellow group.

We ranked you alongside the 5 other participants in the session based on your performance in Stage 1. You were ranked in the 3 lowest participants and are therefore in the Yellow group.

After Stage 4 you will be shown feedback about how your performance compares with other participants your group.

Treatment 3 & 5

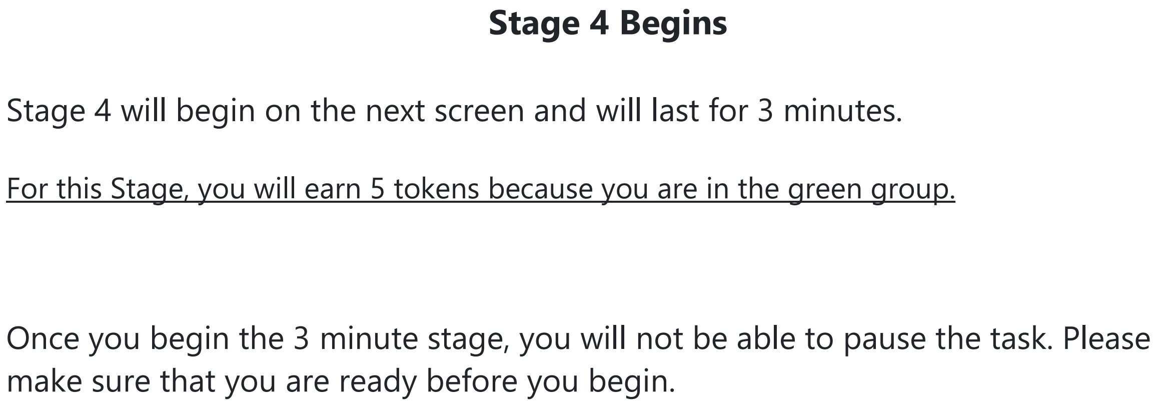
Here are the instructions for Stage 4. In this Stage, you will be performing the Number-Adding Task again as you did before. This will be the final round of the task.

For this Stage, you will earn 5 tokens because you are in the Yellow group.

We ranked you alongside the 5 other participants in the session based on your performance in Stage 1. You were ranked in the 3 lowest participants and are therefore in the Yellow group.

After Stage 4 you will be shown feedback about how your performance compares with other participants your group.

Treatments 1, 2 & 3



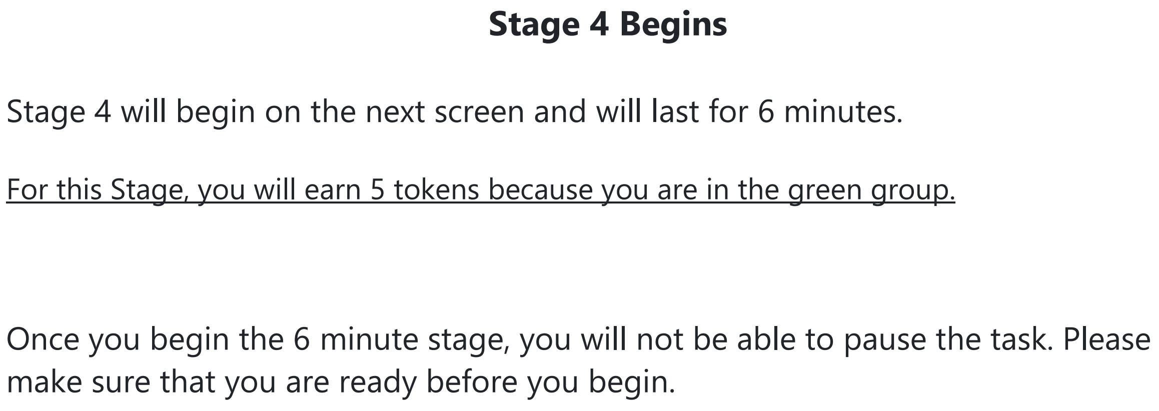
Stage 4 Begins

Stage 4 will begin on the next screen and will last for 3 minutes.

For this Stage, you will earn 5 tokens because you are in the green group.

Once you begin the 3 minute stage, you will not be able to pause the task. Please make sure that you are ready before you begin.

Treatments 4 & 5



Stage 4 Begins

Stage 4 will begin on the next screen and will last for 6 minutes.

For this Stage, you will earn 5 tokens because you are in the green group.

Once you begin the 6 minute stage, you will not be able to pause the task. Please make sure that you are ready before you begin.

Predictions Play as Before Depending on Treatment

Stage 4 Results

You were in the Yellow group, and earned 5 tokens for this stage.

The score of the first ranked person in your group was 9

The score of the second ranked person in your group was 2

Your rank in this round was 3 and your score was 0

Survey

Thank you for taking part in the study.

You will now be asked a series of general questions. Please be assured that all answers you provide will be anonymous.

Note that there is no need to google the answers to any of these questions. Please answer these questions as truthfully as possible.

Survey Plays out As in Chapter 2 Instructions for Experiment (Annex 1)

Bibliography

Abeler, J; Falk, A; Goette, L; Huffman, D. (2011). *Reference Points and Effort Provision*. American Economic Review, Vol. 101, No. 2, Pp 470-492.

Agasisti, T; Bowers, A; Soncin, M. (2018). *School Principals' Leadership Types and Student Achievement in the Italian Context: Empirical Results from a Three-Step Latent Class Analysis*. Educational Management Administration and Leadership, Vol. 47, No. 6, Pp 860-886.

Antecol, H; Eren, O; Ozbeklik, S. (2014). *Peer Effects in Disadvantaged Primary Schools: Evidence from a Randomized Experiment*. Journal of Human Resources, Vol. 51, No. 1, Pp 95-132.

Banuri, S; Keefer, P. (2016). *Pro-Social Motivation, Effort and the Call to Public Service*. European Economic Review, Vol. 83, Pp 139-164.

Barker, A; Larcker, D; Wang, C. Y. (2022). *How Much Should We Trust Staggered Difference-in-Difference Estimates?*. Journal of Financial Economics, Vol. 144, No. 2, Pp 370-395.

Barron, K; Gravert, C. (2021). *Confidence and Career Choices: An Experiment*. The Scandinavian Journal of Economics, doi:10.1111/soje.12444

Benndorf, V; Rau, H; Solch, C. (2019). *Minimising Learning in Repeated Real-Effort Tasks*. Journal of Behavioural and Experimental Finance, Vol. 22, Pp 239-248.

Bertrand, M; Schoar, A. (2003). *Managing with style: The effects of managers on firm policies*. The Quarterly Journal of Economics, Vol. 118, No. 4, Pp 1169-1208.

Betts, J. (2011). *The Economics of Tracking in Education*. In *Handbook of the Economics of Education*. Amsterdam: Elsevier.

Blase, J. (1988). *The politics of favouritism: A qualitative analysis of the teachers' perspective*. Educational Administration Quarterly, Vol. 24, No. 2, Pp 152-177.

Bloom, N; Lemos, R; Sadun, R; Van Reenen, J. (2015). *Does management matter in schools?*. The Economic Journal, Vol. 125, Pp 647-674.

Bloom, N; Van Reenen, J. (2007). *Measuring and explaining management practises across firms and countries*. The Quarterly Journal of Economics, Vol. 122, No. 4, Pp 1351-1408.

Boatman, A; Evans, B; Soliz, A. (2016). *Understanding Loan Aversion in Education: Evidence from High School Seniors, Community College Students, and Adults*. Aera Open, Vol. 3, No. 1.

Bohlmark, A; Gronqvist, E; Vlachos, J. (2016). *The headmaster ritual: The importance of management for school outcomes*. The Scandinavian Journal of Economics, Vol. 118, No. 4, Pp 912-940.

Bohlmark, A; Lindahl, M. (2007). *The impact of school choice on pupil achievement, segregation and costs: Swedish evidence*. IZA Discussion Papers, No. 2786.

Bolukbas, S; Gur, B. S. (2020). *Tracking and Inequality: The Results from Turkey*. International Journal of Educational Development, Vol. 78.

Booij, A; Leuven, E; Oosterbeek, H. (2017). *Ability Peer Effects in University: Evidence from a Randomized Experiment*. The Review of Economic Studies, Vol. 84, No. 2, Pp 547-578.

Borcan, O; Lindahl, M; Mitrut, A. (2017). *Fighting Corruption in Education: What Works and Who Benefits?* American Economic Journal: Economic Policy, Vol. 9, No. 1, Pp 180-209.

Brade, R; Himmler, O; Jackle, R. (2020). *Relative Feedback and Academic Performance – Field Experiment and Replication*. Working Paper, University of Nuerenberg.

Branch, G; Hanushek, E; Rivkin, S. (2012). *Estimating the effect of leaders on public sector productivity: The case of school principals*. NBER Working Paper No. 17803.

Brown, S; Tramayne, S; Hoxha, D; Telander, K; Fan, X; Lent, R. (2008). *Social Cognitive Predictors of College Students' Academic Performance and Persistance: A Meta-Analytic Path Analysis*. Journal of Vocational Behavior, Vol. 72, No. 3, Pp 298-308.

Bruggen, A; Strobel, M. (2007). *Real Effort Versus Chosen Effort in Experiments*. Economics Letters, Vol. 96, No. 2, Pp 232-236.

Brunello, G; Checci, D. (2007). *Does School Tracking Affect Equality of Opportunity? New International Evidence*. Economic Policy, Vol. 22, No. 52, Pp 782-861.

Bruno, P; Strunk, K. (2019). *Making the Cut: The Effectiveness of Teacher Screening and Hiring in the Los Angeles Unified School District*. Educational Evaluation and Policy Analysis, Vol. 41, No. 4, Pp 426-460.

Buchan, N; Brewer, M; Grimalda, G; Wilson, R; Fatas, E; Foddy, M. (2011). *Global Social Identity and Global Cooperation*. Psychological Science, Vol. 22, No. 6, Pp 821-828.

Burke, M; Sass, T. (2013). *Classroom Peer Effects and Student Achievement*. Journal of Labour Economics, Vol. 31, No. 1, Pp 51-82.

Callaway, B; Sant'Anna, P. (2021). *Difference-In-Differences with Multiple Time Periods*. Journal of Econometrics, Vol. 225, No. 2, Pp 200-230.

Cameron, C; Gelbach, J; Miller, D. (2008). *Bootstrap-Based Improvements for Inference with Clustered Standard Errors*. Review of Economics and Statistics, Vol. 90, No. 3, Pp 414-427.

Campbell, J. (2008). *Subtraction by Addition*. Memory and Cognition, Vol. 36, No. 6, Pp 1094-1102.

Carbonaro, W. (2005). *Tracking, Students' Effort, and Academic Achievement*. Sociology of Education, Vol. 78, No. 1, Pp 27-49.

Carman, K; Zhang, L. (2011). *Classroom Peer Effects and Academic Achievement: Evidence from a Chinese Middle School*. China Economic Review, Vol. 23, Pp 223-237.

Carpenter, J; Huet-Vaughn, E. (2019). *Chapter 19: Real-Effort Tasks* in Schram, A; Ule, A (ed.) *Handbook of Research Methods and Applications in Experimental Economics*. Cheltenham: Edward Elgar Publishing Ltd.

Charness, G; Gneezy, U; Henderson, A. (2018). *Experimental Methods: Measuring Effort in Economics Experiments*. Journal of Economic Behavior and Organization, Vol. 149, Pp 74-87.

Charness, G; Masclet, D; Villeval, M. C. (2014). *The Dark Side of Competition for Status*. Management Science, Vol. 60, No. 1, Pp 38-55.

Charron, N; Dahlstrom, C; Fazekas, M; Lapuente, V. (2017). *Careers, connections and corruption risks: Investigating the impact of bureaucratic meritocracy on public procurement processes*. The Journal of Publics, Vol. 79, No. 1, Pp 89-104.

Chen, S; Schildberg-Horisch, H. (2019). *Looking at the Bright Side: The Motivational Value of Confidence.* European Economic Review, Vol. 120: 103302.

Coelli, M; Green, D. (2011). *Leadership effects: School principals and student outcomes*. Economics of Education Review, Vol. 31, No. 1, Pp 92-109.

Colonnelli, E; Teso, E; Prem, M. (2018). *Patronage in the allocation of public sector jobs*. Job Market Paper available at: <https://www.dropbox.com/s/r3qaegqfreorcfb/EdoardoTeso_JMP.pdf?dl=0>

Cortazar, J; Fuenzalida, J; Lafuente, M. (2016). *Merit Based Selection of Public Managers: Better Public Sector Performance*. Inter-American Development Bank.

Coutts, A. (2019). *Good News and Bad News are Still News: Experimental Evidence on Belief Updating*. Economic Letters, Vol. 22, No. 2, Pp 369-395.

Dahis, R; Schiavon, L; Scot, T. (2020). *Selecting Top Bureaucrats: Admission Exams and Performance in Brazil*. Working Paper, SSRN: 3584725.

Danz, D; Vesterlund, L; Wilson, A. J. (2022). *Belief Elicitation and Behavioral Incentive Compatibility*. American Economic Review.

De la Rosa, LE. (2011). *Overconfidence and Moral Hazard*. Games and Economic Behaviour, Vol. 73, No. 2, Pp 429-451.

Dhuey, E; Smith, J. (2014). *How important are school principals in the production of student achievement?* Canadian Journal of Economics, Vol. 47, No. 2, Pp 635-662.

Di Liberto, A; Schivardi, F; Sulis, G. (2015). *Managerial practises and student performance*. Economic Policy, Vol. 30, No. 84, Pp 683-728.

Ding, W; Lehrer, S. (2007). *Do Peers Affect Student Achievement in China's Secondary Schools?* The Review of Economics and Statistics, Vol. 89, No. 2, Pp 300-312.

Dufflo, E; Dupas, P; Kremer, M. (2011). *Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya*. American Economic Review, Vol. 101, No. 5, Pp 1739-1774.

Dutcher, G; Salmon, T; Saral, K. (2015). *Is "Real" Effort More Real?* Available at SSRN: 2701793.

Eil, D; Rao, J. (2011). *The Good News – Bad News Effect: Asymmetric Processing of Objective Information About Yourself.* American Economic Journal: Microeconomics, Vol. 3, No. 2, Pp 114-138.

Epple, D; Newlon, E; Romano, R. (2002). *Ability Tracking, School Competition, and the Distribution of Educational Benefits.* Journal of Public Economics, Vol. 83, No. 1, Pp 1-48.

Erkal, N; Gangadharan, L; Koh, BH. (2018). *Monetary and Non-Monetary Incentives in Real-Effort Tournaments*. European Economic Review, Vol. 101, Pp 528-545.

Estrada, R. (2019). *Rules Versus Discretion in Public Service: Teacher Hiring in Mexico.* Journal of Labor Economics, Vol. 37, No. 2, Pp 545-579.

Evens, P; Rauch, J. (1999). *Bureaucracy and Growth: A Cross-National Analysis of the Effect of Weberian State Structure on Economic Growth.* American Sociological Review, Vol. 64, Pp 748-765.

Fahr, R; Irlenbusch, B. (2000). *Fairness as a Constraint on Trust in Reciprocity: Earned Property Rights in a Reciprocal Exchange Experiment.* Economic Letters, Vol. 66, No. 3, Pp 275-282.

Falk, A; Fehr, E. (2003). *Why Labour Market Experiments?* Labour Economics, Vol. 10, No. 4, Pp 399-406.

Faucette, N; Graham, G. (1986). *The impact of principals on teachers during in-service education: A qualitative analysis.* Journal of Teaching in Physical Education, Vol. 5, No. 2, Pp 79-90.

Feld, J; Sauermann, J; de Grip, A. (2017). *Estimating the Relationship Between Skill and Overconfidence.* Journal of Behavioural and Experimental Economics, Vol. 68, Pp 18-24.

Fishbach, A; Woolley, K. (2022). *The Structure of Intrinsic Motivation.* Annual Review of Organizational Psychology and Organizational Behavior, Vol. 9, Pp 339-363.

Francis, B; Craig, N; Hogden, J; Taylor, B; Tereshchenko, A; Connolly, P; Archer, L. (2019). *The Impact of Tracking by Attainment on Pupil Self-Confidence Over Time: Demonstrating the Accumulative Impact of Self-Fulfilling Prophecy.* British Journal of Sociology of Education, Vol. 41, No. 5, Pp 626-642.

Frederick, S. (2005). *Cognitive Reflection and Decision Making.* Journal of Economic Perspectives, Vol. 19, Pp 25-42.

Fu, C; Mehta, N. (2018). *Ability Tracking, School and Parental Effort, and Student Achievement: A Structural Model and Estimation.* Journal of Labor Economics, Vol. 36, No. 4, Pp 923-979.

Gill, D; Kissova, Z; Lee, J; Prowse, V. (2019). *First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision.* Management Science, Vol. 65, No. 2, Pp 494-507.

Gill, D; Prowse, V. (2012). *A Structural Analysis of Disappointment Aversion in a Real Effort Competition.* American Economic Review, Vol. 102, No. 1, Pp 469-503.

Gill, D; Prowse, V. (2019). *Measuring Costly Effort Using the Slider Task.* Journal of Behavioural and Experimental Finance, Vol. 21, Pp 1-9.

Gneezy, U; Niederle, M; Rustichini, A. (2003). *Performance in Competitive Environments: Gender Differences.* The Quarterly Journal of Economics, Vol. 118, No. 3, Pp 1049-1074.

Goodman-Bacon, A. (2021). *Difference-in-Differences with Variation in Treatment Timing.* Journal of Econometrics, Vol. 225, No. 2, Pp 254-277.

Grossman, Z; Owens, D. (2012). *An Unlucky Feeling: Overconfidence and Noisy Feedback.* Journal of Economic Behavior and Organization, Vol. 84, No. 2, Pp 510-524.

Gurmu, T. (2020). *Primary School Principals in Ethiopia: Selection and Preparation.* Educational Management Administration and Leadership, Vol. 48, No. 4, Pp 651-681.

Gurtler, O; Harbring, C. (2010). *Feedback in Tournaments Under Commitment Problems: Experimental Evidence.* Journal of Economic Management Strategy, Vol. 19, No. 3, Pp 771-810.

Hannan, R; McPhee, G; Newman, A; Tafkov, I. (2013). *The Effect of Relative Performance Information on Performance and Effort Allocation in a Multi-Task Environment.* Account Review, Vol. 88, No. 2, Pp 553-575.

Hanushek, E; Woessmann, L. (2006). *Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence Across Countries.* The Economic Journal, Vol. 106, Pp 63-76.

Hoffman, M; Kahn, L; Li, D. (2018). *Discretion in Hiring.* The Quarterly Journal of Economics, Vol. 133, No. 2, Pp 765-800.

Hsiao, H; Lee, M; Tu, Y. (2012). *The Effects of Reform in Principal Selection on Leadership Behaviour of General and Vocational High School Principals in Taiwan.* Educational Administration Quarterly, Vol. 49, No. 3, Pp 421-450.

Huseyin, A; Mustafa, T. (2008). *Nepotism, favouritism and cronyism: A study of their effects on job stress and job satisfaction in the banking industry of north Cyprus.* Social Behaviour and Personality: An International Journal, Vol. 36, No. 9.

Ido, Erev; Haruvy, E. (2016). *Learning and the Economics of Small Decisions.* In *The Handbook of Experimental Economics, Volume 2.* Princeton: University Press.

Imberman, S; Kugler, A; Sacerdote, B. (2012). *Katarina's Children: Evidence on the Structure of Peer Effects from Hurricane Evacuees.* American Economic Review, Vol. 102, No. 5, Pp 2048-2082.

Jacob, B. (2010). *Do Principals Fire the Worst Teachers?.* Educational Evaluation and Policy Analysis, Vol. 33, No. 4, Pp 403-434.

Jacob, B; Rockoff, J; Taylor, E; Lindy, B; Rosen, R. (2018). *Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools.* Journal of Public Economics, Vol. 116, Pp 81-97.

Jagacinski, C; Nicholls, J. (1990). *Reducing Effort to Protect Perceived Ability: "They'd Do It but I Wouldn't".* Journal of Educational Psychology, Vol. 82, No. 1, Pp 15-21.

Kajackaite, A. (2015). *If I Close My Eyes, Nobody Will Get Hurt: the Effect of Ignorance on Performance in a Real-Effort Experiment.* Journal of Economic Behavior and Organization, Vol. 116, Pp 518-524.

Kessler, J; Vesterlund, L. (2015). *The External Validity of Laboratory Experiments: The Misleading Emphasis on Quantitative Effects.* Handbook of Experimental Economic Methodology, Chapter 18, Pp 392-405.

Konow, J. (2000). *Fair Shares: Accountability and Cognitive Dissonance in Allocative Decisions.* American Economic Review, Vol. 90, No. 4, Pp 1072-1091.

Kuhnen, C; Tymula, A. (2012). *Feedback, Self-Esteem, and Performance in Organizations.* Management Science, Vol. 58, No. 1, Pp 94-113.

Larrick, R; Burson, K; Soll, J. (2007). *Social Comparison and Confidence: When Thinking You're Better Than Average Predicts Overconfidence (and When it Does Not).* Organizational Behavior and Human Decision Processes, Vol. 102, No. 1, Pp 76-94.

Levitt, S. D; List, J. A. (2007). *What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?* Journal of Economic Perspectives, Vol. 21, No. 2, Pp 153-174.

Lezzi, E; Fleming, P; Zizzo, D. (2015). *Does It Matter Which Effort Task You Use? A Comparison of Four Effort Tasks When Agents Compete for a Prize.* Available at SSRN: 2594659.

Loeb, S; Kalogrides, D; Horng, E. (2010). *Principal Preferences and the Uneven Distribution of Principals Across Schools*. Educational Evaluation and Policy Analysis, Vol. 32, No. 2, Pp 205-229.

Masci, C; De Witte, K; Agasisti, T. (2018). *The Influence of School Size, Principal Characteristics, and School Management Practises on Educational Performance: An Efficiency Analysis of Italian Students Attending Middle Schools*. Socio-Economic Planning Sciences, Vol. 61, Pp 52-69.

McAuley, E; Duncan, T; Tammen, V. (1989). *Psychometric Properties of the Intrinsic Motivation Inventory in a Competitive Sport Setting: A Confirmatory Factor Analysis*. Research Quarterly for Exercise and Sport, Vol. 60, No. 1, Pp 48-58.

McEwan, P. (2003). *Peer Effects on Student Achievement: Evidence from Chile*. Economics of Education Review, Vol. 22, No. 2, Pp 131-141.

McGuigan, M; McNally, S; Wyness, G. (2017). *Student Awareness of Costs and Benefits of Educational Decisions: Effects of an Information Campaign*. Journal of Human Capital, Vol. 10, No. 4, Pp 482-519.

Meyer, A; Richter, D; Hartung-Beck, V. (2020). *The Relationship Between Principal Leadership and Teacher Collaboration: Investigating the Mediating Effect of Teachers' Collective Efficacy*. Educational Management Administration and Leadership.

Mobius, M; Nierderle, M; Niehaus, M; Rosenblat, T. (2011). *Managing Self-Confidence: Theory and Experimental Evidence*. National Bureau of Economic Research, Working Paper 17014.

Moore, D; Healy, P. (2008). *The Trouble with Overconfidence*. Psychological Review, Vol. 115, No. 2, Pp 502-517.

Murad, Z; Starmer, C. (2021). *Confidence Snowballing and Relative Performance Feedback*. Journal of Economic Behavior and Organization, Vol. 190, Pp 550-572.

Niederle, M; Vesterlund, L. (2007). *Do Women Shy Away From Competition? Do Men Compete Too Much?* Quarterly Journal of Economics, Vol. 122, No. 3, Pp 1067-1101.

OECD (2014), PISA 2012 Results: What Makes a School Successful? Resources, Policies, and Practises, Volume IV, OECD, Paris.

Ortoleva, P; Snowberg, E. (2015). *Overconfidence in Political Behavior*. American Economic Review, Vol. 105, No. 2, Pp 504-535.

Palmer, B; Mullooly, J. (2015). *Principal Selection and School District Hiring Cultures: Fair or Foul?* Journal of Education and Social Policy, Vol. 2, No. 2, Pp 26-37.

Peer, E; Brandimarte, L; Samat, S; Acquisti, A. (2017). *Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioural Research*, Vol. 70, Pp 153-163.

PEIS. (2007). *The Survey on Higher Education Policies*. Open Society Foundation; Gallup Romania.

Perry, J. (1996). *Measuring public service motivation: An assessment of construct reliability and validity*. Journal of Public Administration Research and Theory, Vol. 6, No. 1, Pp 5-22.

Poocharoen, O; Brillantes, A. (2013). *Meritocracy in Asia Pacific: Status, issues and challenges*. Review of Public Personnel Administration, Vol. 33, No. 2, Pp 140-163.

Rauch, J; Evans, P. (2000). *Bureaucratic structure and bureaucratic performance in less developed countries*. Journal of Public Economics, Vol. 75, No. 1, PP 49-71.

Rees, D; Argys, L. (1996). *Tracking in the United States: Descriptive Statistics from NELS*. Economics of Education Review, Vol. 15, No. 1, Pp 83-89.

Regier, D; Watson, V; Burnett, H; Ungar, W. (2014). *Task Complexity and Response Certainty in Discrete Choice Experiments: An Application to Drug Treatments for Juvenile Idiopathic Arthritis*. Journal of Behavioural and Experimental Economics, Vol. 50, Pp 40-49.

Rosaz, J; Villeval, M. C. (2012). *Lies and Biased Evaluation: A Real-Effort Experiment*. Journal of Economic Behavior and Organization, Vol. 84, No. 2, Pp 537-549.

Roth, A. E. (1987). *Laboratory Experimentation in Economics*. Advances in Economic Theory, Fifth World Congress. Cambridge: University Press.

Ruiz-Tagle, C. (2019). *Selection of School Principals Based on Competitive Processes: Evidence from One Policy in Chile*. Calidad en la Educacion, No. 51, Pp 85-130.

Ryan, R. (1982). *Control and Information in the Intrapersonal Sphere: An Extension of Cognitive Evaluation Theory*. Journal of Personality and Social Psychology, Vol. 43, Pp 450-461.

Scoppa, V. (2009). *Intergenerational transfers of public sector jobs: A shred of evidence on nepotism*. Public Choice, Vol. 141, No. 1-2, Pp 167-188.

Snowberg, E; Yariv, L. (2021). *Testing the Waters: Behavior Across Participant Pools*. American Economic Review, Vol. 111, No. 2, Pp 687-719.

Stravakou, P. (2019). *Selecting School Principals in Greece in the Last Fifteen Years – A Theoretical Approach*. Journal of Advances in Education and Philosophy, Vol. 3, No. 8, Pp 277-282.

Strouse, D. (2004). *A qualitative case study of the impact of principal leadership and school performance awards on eight Maryland schools*. University of Maryland Digital Repository: Teaching, Learning, Policy & Leadership Theses and Dissertations

Sundell, A. (2014). *Are formal civil service examinations the most meritocratic way to recruit civil servants? Not in all countries*. Public Administration, Vol. 92, No. 2.

Tan, K. (2008). *Meritocracy and Elitism in a global city: Ideological shifts in Singapore*. International Political Science Review, Vol. 29, No. 1, Pp 7-27.

Thomson, K; Oppenheimer, D. (2016). *Investigating an Alternate From of the Cognitive Reflection Test*. Judgement and Decision Making, Vol. 11, No. 1, Pp 99-113.

Vardardottir, A. (2013). *Peer Effects and Academic Achievement: A Regression Discontinuity Approach*. Economics of Education Review, Vol. 36, Pp 108-121.

Venkatakrishnan, H; Wiliam, D. (2003). *Tracking and Mixed-Ability Grouping in Secondary School Mathematics Classrooms: A Case Study 1*. British Educational Research Journal, Vol. 29, No. 2, Pp 189-204.

Walker, A; Kwan, P. (2012). *Principal Selection Panels: Strategies, Preferences and Perceptions*. Journal of Educational Administration, Vol. 50, No. 2, Pp 188-205.

Yeboah-Assiamah, E; Asamoah, K; Osei-Kojo, A. (2014). *Corruption Here, Corruption There, Corruption Everywhere: A Framework for Understanding and Addressing Public Sector Corruption in Developing African Democracies*. Journal of Public Administration and Governance, Vol. 4, No. 3.

Yu, H. (2020). *Am I The Big Fish? The Effect of Ordinal Rank on Student Academic Performance in Middle School.* Journal of Economic Behaviour and Organization, Vol. 176, Pp 18-41.