Apascaritei, Paula; Radl, Jonas; Swarr, Madeline

Material incentives moderate gender differences in cognitive effort among children

Learning and Individual Differences

Provided in Cooperation with:
WZB Berlin Social Science Center

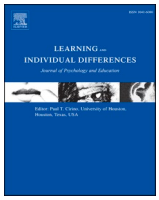
Suggested Citation: Apascaritei, Paula; Radl, Jonas; Swarr, Madeline (2024) : Material incentives moderate gender differences in cognitive effort among children, Learning and Individual Differences, ISSN 1873-3425, Elsevier, Amsterdam, Vol. 114, pp. 1-20, https://doi.org/10.1016/j.lindif.2024.102494

Learning and Individual Differences 114 (2024) 102494

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif



Material incentives moderate gender differences in cognitive effort among children

Paula Apascaritei [a], Jonas Radl [b,c,*], Madeline Swarr [b]

[a] Independent Researcher, Madrid, Spain
[b] Department of Social Sciences, Universidad Carlos III de Madrid, Getafe, Madrid, Spain
[c] WZB Berlin Social Science Center, Berlin, Germany

ARTICLE INFO

Keywords:
Cognitive effort
Gender
Incentives
Children
Laboratory experiments

ABSTRACT

Effort is crucial for academic performance and varies by gender. However, it is not clear at what age nor under what circumstances gender differences in effort arise. Using behavioral measures of executive function from 799 fifth-grade students, we find no gender differences in cognitive effort in the absence of rewards. However, boys exert more effort than girls when materially incentivized. Adding a status incentive on top of material rewards does not further increase the gender gap. According to expectancy-value theory, the degree to which incentives moderate the gender effect may depend on ability. We find that while low-ability girls work as hard as high-ability girls when no incentives are present, low-ability boys tend to disengage from effortful tasks. High-ability girls increase effort more than low-ability girls when material incentives are added, and high-ability boys increase effort more than low-ability boys when status incentives are added.

Educational relevance and implications statement

Effort is essential for young students as it boosts learning and achievement. Educators and parents often use rewards to get students to try harder, under the implicit assumption that they are equally motivating to everyone. Research has shown, however, that there are gender differences in reward preferences, but the differential effect of incentives on effort by gender has yet to be measured accurately. Our study among fifth-grade students confirms that incentives are effective in boosting effort overall, but that boys are more motivated by material rewards than girls. Contrary to the widespread notion that girls are less competitive than boys, status incentives do not significantly add to boys' overall effort edge. Differential motivational effects by gender and ability should be considered carefully when implementing incentivization schemes in schooling contexts.

1. Introduction

Measuring learning achievement poses a significant challenge for educators and researchers: while data on students' grades are readily available, it is not clear to what extent it captures students' effort. Moreover, most largescale grading data comes from low-stakes assessments. Scholars have argued that when the stakes of an exam are not sufficiently motivating, such as with the Trends in International Mathematics and Science Study, or TIMSS, where results hold no external benefits for the test-taker, low performance scores may actually be measuring lower effort rather than lower learning attainment (O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005; Wise & DeMars, 2005). Therefore, what is observed as under-achievement may partially reflect lower levels of test-taking motivation. In order to boost subjects' effort and thus improve the construct validity of test scores, economists have suggested offering external incentives to test-takers, such as performance-based monetary pay or academic awards and certificates (Gneezy et al., 2019; Levitt, List, Neckermann, & Sadoff, 2016). One side effect, however, is the differential impact that these rewards may have on motivation depending on heterogenous preferences, particularly regarding gender.

Differences in motivational orientations by gender have been widely researched using self-reports or inferred via gender gaps in achievement (see Meece et al. (2006) for a historical review). In many Western industrialized countries, girls have shown greater performance on indicators of learning attainment in relation to boys (Buchmann, DiPrete, & McDaniel, 2008; DeAngelo et al., 2011; Voyer & Voyer, 2014). Gender differences in motivation are supported by the consistent finding that

* Corresponding author at: Universidad Carlos III de Madrid, Calle Madrid 135, 28903 Getafe, Spain.
E-mail address: jonas.radl@uc3m.es (J. Radl).

https://doi.org/10.1016/j.lindif.2024.102494
Received 31 July 2023; Received in revised form 17 May 2024; Accepted 18 June 2024
Available online 16 July 2024

boys score significantly lower on self-reported measures of academic effort than girls (DeMars et al., 2013). There is robust evidence in the behavioral economics literature according to which women are less sensitive and perhaps even averse to competition (Niederle & Vesterlund, 2007), while men thrive on it (Gneezy, Niederle, & Rustichini, 2003). Thus, while girls may be more motivated than boys during low-stakes assessments, a shrinking or even reversal of the gender gap favoring boys and men may occur when the stakes of the exams involve competing for accolades or limited admission placements (Ors, Palomino, & Peyrache, 2013).

However, the extent to which motivation directly translates into *effortful behavior*, either by choosing to direct effort towards a specific task or determining how much effort to dedicate to the task, remains unknown. In the current study, we compare performance on a battery of tests of executive function while modifying incentive conditions to test for the possible interaction between gender and incentives on children's cognitive effort and assess how the degree to which incentives moderate the gender effect depends on ability.

1.1. Motivation and effort

Cognitive effort is required to achieve just about anything. In school, for example, it takes effort to learn a new mathematical concept and to demonstrate its understanding by completing one's homework. However, cognitive effort is generally aversive because it draws on limited resources and binds attention (Inzlicht, Shenhav, & Olivola, 2018). Therefore, individuals need motivation to cover the costs that come with engaging in effortful tasks. Expectancy-value theory (EVT) (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000), elaborated by Jacquelynne Eccles and colleagues, posits that achievement and achievement-related decisions are determined at the highest level by an individual's expectations for success as well as the subjective value that achievement provides to the individual. Under this framework, expectancy is affected by factors such as self-perceived ability and efficacy beliefs, with those believing that they will succeed being more willing to exert effort. On the value side, decisions are made about how to expend effort towards a task in function of *intrinsic interest, attainment value, utility value, and cost*.

Intrinsic interest refers to the propensity of pursuing an activity for its own sake, and might reflect inherent intellectual curiosity or need for cognition that is satisfied by performing a particular behavior (Inzlicht, Shenhav, & Olivola, 2018; Ryan & Deci, 2000a). Research has shown that individuals can allocate more cognitive effort to a task while feeling less depleted based on their intrinsic interest (Segal, 2012; Thoman, Smith, & Silvia, 2011). *Attainment value*, on the other hand, is derived by the personal importance of completing or doing well on the task, reflecting internalized motivations such as sense of purpose or alignment with self-concept constructed from an interactive array of socializing forces as well as individual goals and needs. In this way, intrinsic interest and attainment value are closely linked to self-determination theory's (SDT) distinction of different forms of autonomous motivation – those stemming from a task being fun (intrinsic motivation), compatible with one's sense of self (integrated motivation), or of personal importance (identified motivation) (Eccles, 2005; Ryan & Deci, 2000b).

Utility value, in contrast, is what drives behaviors performed as a means to an end, for example, to achieve a reward or avoid a punishment. This is frequently equated to controlled motivation under the SDT framework. EVT suggests that providing attractive incentives will boost performance as it compensates for the cost of effort, especially when intrinsic interest or attainment value are low. Many behavioral studies have focused on understanding how motivation is controlled by looking at how externally-placed incentives, such as prizes or symbolic recognition awarded for a particular behavior, increase performance (see Rios (2021) for a meta-analysis).

Effort investments are not only determined by intrinsic, attainment, and utility value, but also by *cost*, the final factor of EVT's value component. Cost is generally viewed as comprising all the negative aspects of engaging in an effortful task. Costs can too range from the internal, such as fear of failure or performance anxiety, to the external, such as lost opportunities from dedicating time and effort to one task rather than to another.

1.2. Gender differences in motivation and effort

Most studies show that girls report higher levels of autonomous motivation than boys in school (Ratelle, Guay, Vallerand, Larose, & Senécal, 2007; Vecchione, Alessandri, & Marsicano, 2014). Gender stereotypes cultivate differences in academic culture and motivation that may lead to gender gaps in effort (Boutyline, Arseniev-Koehler, & Cornell, 2023; Butler, 2014; Jones & Myhill, 2004; Legewie & DiPrete, 2012). Girls tend to value effort more, both in themselves and others (Hirt & McCrea, 2009). Moreover, girls are on average more self-disciplined than boys (Duckworth & Seligman, 2006), more self-regulating when it comes to performance motivation (Wolters & Benzon, 2013), and are more likely to possess other personality traits, such as higher conscientiousness and openness, that are assumed to indicate a disposition towards effort exertion (Neuenschwander, Cimeli, Röthlisberger, & Roebers, 2013).

Many studies have also investigated whether the degree to which behavior changes in response to certain externally-placed incentives is dependent on gender. When monetary incentives are used, to what extent and for whom they motivate depends on *how* and *for what* performance is rewarded. Masclet, Peterle, & Larribeau, 2015 find that women do better when reward is not contingent on task performance, e.g. in a flat-wage scheme. Evidence on gender differentials under piece-rate reward schemes, where payoff is proportional to performance, show that boys tend to outperform girls, though a statistically significant difference is not consistently found (Buser et al., 2014; Dreber, Von Essen, & Ranehill, 2014; Niederle & Vesterlund, 2010; Sutter, Glätzle-Rützler, Balafoutas, & Czermak, 2016). Analyzing within-individual behavior changes when incentive schemes are experimentally modulated, Levitt, List, Neckermann, & Sadoff, 2016 find that the introduction of low-stake and short-term financial rewards have a greater impact on performance improvement for boys than for girls, which they suggest may be partially due to gender differences in time preferences.

Further evidence on gender gaps that arise from heterogeneities in behavioral responses to incentives has shown that women tend to be less willing to compete than men (Gneezy & Rustichini, 2004; Niederle & Vesterlund, 2007). This gender-graded preference is associated with differential beliefs about the benefits of competing, such as that it enhances performance, builds character, and leads to innovative solutions (Kesebir, Lee, Elliot, & Pillutla, 2019). Moreover, competition brings about honor incentives associated with winning, which may be more attractive to men who exhibit greater affinity with status hierarchy (Beutel & Marini, 1995; Brandts, Gërxhani, & Schram, 2020). In terms of performance under competition, studies report that girls tend to do worse than boys in winner-take-all-style contest settings where only top performers are rewarded (Gneezy, Niederle, & Rustichini, 2003; Horn, Kiss, & Lénárd, 2022; Schram, Brandts, & Gërxhani, 2019; Sutter, Zoller, & Glätzle-Rützler, 2019).

Effort costs may also differ by gender. Research on gender differentials in test non-compliance have shown greater rates of absenteeism among boys for low-stakes exams (Swerdzewski, Harmes, & Finney, 2009). While girls are praised for their hard work as the reason for their success (Heyder & Kessels, 2017), boys may be socially rewarded for their "effortless" talent often associated with inherent ability or intelligence. As a result, boys more often display "work-avoidant" behavior and expend just enough effort to reach the minimum for a passing grade, aligning their behavior with this appreciation of easy, effortless success (Chouinard & Roy, 2008).

1.3. Incentives effects on cognitive effort and the moderating role of ability

Simplifying the EVT framework, achievement is a function of ability, composed of expectancy or competency beliefs, and effort, derived from subjective task value or motivation. Therefore, to isolate the effect of incentives on the effort component requires the ability component to first be neutralized. Those who have greater expectations for success but are less motivated may exhibit similar levels of achievement as someone with lower expectations but more motivation. The type of task is thus crucial. Indeed, the most prominent advantages for girls are seen in subjects such as reading and language, for which they tend to exhibit greater competence (Roivainen, 2011). Consequently, ability and effort are often correlated since tasks that favor one's competencies tend to also favor their interests and support goal-directed behavior. Competency beliefs thus interact with motivational orientations, and evidence suggests that this relationship may depend on gender. Jackson (2003) argues that boys may be more likely to respond to difficult tasks by disengaging from them – a sort of self-preservation or way of preventing others from attributing low performance to low ability. On the other hand, girls are more likely to attribute their own successes to hard work, attenuating the effect of ability beliefs on achievement motivation (Meece et al., 2006). It has also been suggested that ability plays a differential role on gender sensitivity to competition, as evidenced by studies that find that high-ability boys tend to be more self-confident and thus more optimistic about their chances of winning, while high-ability girls suffer from relative under-confidence and tend to shy away from competition (Niederle & Vesterlund, 2011; Tang & Zhao, 2023).

1.4. Measuring effort

Although there is a general understanding that increases in motivation drive increases in effort, effort remains an elusive phenomenon, making it complicated for researchers to understand its direct relationship with motivation. Conventionally, psychologists have measured effort through self or other-reported surveys, which are subject to various types of bias and limitations such as social desirability bias, reference bias, discriminatory bias, and lack of insight or information (Duckworth & Yeager, 2015). Indeed, many studies have found an empirical disconnect between informant-report questionnaires and behavioral measures of effort-related cognitive processes (Duckworth & Kern, 2011). Thus, the use of “real-effort” tasks, where genuine cognitive effort is required to complete a task and thus inferred from task performance (Heckman et al., 2021), is preferred when researchers want to capture trade-offs that can be deemed applicable to real life situations (Dutcher, Salmon, & Saral, 2015). To evoke more accurately the motivation-performance relationship present in work or school settings, neuroscientific research has increasingly implemented behavioral tasks under differing incentive conditions (Buser et al., 2014; Frömer, Lin, Dean Wolf, Inzlicht, & Shenhav, 2021). Such real-effort tasks require executive function (EF), which are top-down mental processes necessary for solving problems and achieving goals (Miyake & Friedman, 2012). EF is considered effortful in that it represents what is not “automatic” about the brain's functioning. There is not full consensus across the literature about its exact subdomains, but most definitions (see Miyake & Friedman, 2012 for examples) stipulate that they primarily involve: (i) information processing and updating; (ii) cognitive flexibility and switching between different activities; (iii) inhibition and control; and (iv) planning and goal prioritization.

Gender-neutrality is crucial in this setting to allow the identification of effort. In experimental studies, similarly, girls have been shown to have an advantage when gender-typed tasks such as rope-skipping were used (Khachatryan, Dreber, Von Essen, & Ranehill, 2015), whereas boys displayed superior performance when studies employed tasks such as solving mazes (Gneezy, Niederle, & Rustichini, 2003). While the intentional use of gender-typed tasks could be useful for recreating certain real-world contexts (Buser et al., 2014; Khachatryan, Dreber, Von Essen, & Ranehill, 2015) to understand how related social self-perceptions affect effort investments (Dreber, Von Essen, & Ranehill, 2014), they are ill-suited to unconfound ability from effort. Tests of executive function are better equipped to understand the independent effect of incentives on effort by gender as they do not considerably favor the abilities and interests of either boys or girls. Grissom & Reyes, 2019 review of gender differences in EF concludes that there are no systematic advantages by gender in any EF component. Thus, performance increases on tests of EF more accurately represent motivational increases and translate directly into the effortful processes needed to achieve higher-order goals.

1.5. The current study

The study draws on a balanced sample of 799 fifth-grade students and uses an experimental setting to understand how effort investments among children are shaped differently by the introduction of external incentives.[1] To overcome notorious difficulties of measuring effort (Apascaritei, Demel & Radl, 2021; Radl & Miller, 2021), we compare performance on three distinct real-effort tasks across three different incentive conditions: an unincentivized condition where no external rewards are given for performance, a monetary incentive condition where points that can later be traded in for toys are awarded in function of performance, and a status incentive condition where the top three performers are additionally awarded a diploma at the end of the session. We additionally account for the confounding effects of key non-cognitive skills (i.e. personality traits). Our measure of cognitive effort draws on three real-effort tasks that target different dimensions of executive function: information processing and updating as assessed by the slider task, cognitive flexibility and switching as assessed by the AX-Continuous Performance Task, or AX-CPT, and inhibition as assessed by a variant of the Simon task.

Given the previous theoretical considerations and existing evidence, we formulate the following hypotheses regarding differential effort responses by gender to varying incentive conditions, while assuming constant ability and intrinsic interest:

H1a. Girls exert more effort than boys in the absence of external rewards.

H2a. Boys increase their effort more than girls do in response to the introduction of performance-based monetary incentives.

H3a. Boys increase their effort more so than girls when competing for an additional status incentive placed on top of performance-based monetary incentives.

Assuming that, in agreement with the EVT hypothesis, high-ability subjects exert more effort than low-ability subjects and that subjects make accurate inferences about their own ability, we hypothesize the following:

H1b. In the absence of incentives, the gap in cognitive effort favoring high- versus low-ability students is greater among boys than among girls.

H2b. When performance-based monetary incentives are introduced, the gap in cognitive effort by ability becomes less dependent on gender, as low-ability boys catch up with high-ability boys.

[1] The larger research project under which this study falls primarily investigates how the incentive-effort relationship varies by parental socioeconomic status (Radl et al., 2024). Further analyses by gender were elaborated in the data exploitation plan for the project given their importance for addressing relevant interdisciplinary debates.

H3b. When an additional status incentive is placed on top of performance-based monetary incentives, the increase in the gap in cognitive effort favoring high- versus low-ability students will be greater for boys than for girls.

2. Materials & methods

2.1. Sample and experimental procedures

Data was collected from a random sample of schools stratified by neighborhood income quartile and school type (public or private) in the urban region of Madrid, Spain. Fifth graders in schools were invited for a one-day visit to a university campus in Madrid. In total, 799 fifth-grade students (419 girls, 380 boys) from 35 classes representing 19 schools visited the campus between October 2019 and March 2022.[2] Each child participated only if they had their parent's written informed consent and signed data protection agreement, in accordance with stipulations of the ethics board and data protection officer at the university. Unfortunately, gender-specific non-participation cannot be determined as data of those who dropped out or did not have parental consent is unavailable.[3]

In each session, all students performed one task in the absence of incentives. Then, all three tasks were performed under a piece-rate scheme where points for correct answers were "cashed in" at the end of the session for prizes selected from a menu of toys with varying associated prices. Finally, one task was performed with the addition of a status incentive, where the three top performers at the end of the rounds were awarded diplomas, on top of the same monetary incentive scheme that was employed in the preceding rounds; thus, the status competition element is the only difference to the monetary incentive condition. The introduction of incentives always happened in the same order and additively to avoid "crowding-out" effects on intrinsic motivation that occur once extrinsic rewards for task performance are removed (Gneezy, Meier, & Rey-Biel, 2011; James, 2005). Tasks, however, were allocated to incentive conditions counterbalancing by experimental session.[4]

The structure of the within-subject design is detailed in Fig. 1. During the experimental sessions, each task-incentive condition pair was performed twice, resulting in a total of 10 observations per student (10 "rounds" per student). Before starting each round, children had a choice between completing the task or playing a computer-based leisure activity with the understanding that they would earn no points if they did the leisure activity. The purpose of such design was to model the cost of effort as an opportunity cost and more accurately represent the realities that children face outside of the laboratory, such as the decision to do homework instead of play video games (Kurzban, Duckworth, Kable, & Myers, 2013). The two leisure games, one where a soccer ball must be kept in the air via mouse clicks and the other where a sliding picture puzzle must be solved on the computer screen, were selected so that they would appeal similarly to both boys and girls.

2.2. The real-effort tasks

Each task engages different components of EF. The slider task primarily covers the information processing and updating subdomain. In this task, the participants are presented with 48 horizontal lines. A dial is presented on each line and the participant must click and drag the dial so that it is exactly at the midpoint, which corresponds to 50 on a scale from 0 to 100. The AX-CPT task tests cognitive flexibility as measured through having to switch between proactive and reactive control. In this task, participants must press a certain button when the letter "X" appears after a probe of the letter "A". When an "A" appears followed by any letter apart from an "X" (reactive condition, or A-Y condition), or when any letter other than "A" is shown as a probe (proactive condition, or B-X condition), the subject must press an alternative button. Finally, the Simon task tests the inhibition subdomain. In the Simon task, participants must press a certain button on the left side of the keyboard when a left-pointing arrow appears on screen and a different button on the right side of the keyboard when a right-pointing arrow appears on screen. Arrows can be randomly shown in a position opposite to their direction (in the incongruent condition), congruent to their direction, or in the middle of the screen (in a neutral condition). The three tasks cover three of the four main aspects of EF. The fourth subdomain, planning and goal prioritization, is not as readily measurable as the other three, and somewhat more distinct in that it does not involve continual concentration and attention. Therefore, this domain is estimated via the decision to complete the real-effort task instead of playing the leisure game. The layouts of each task are illustrated in Fig. 2.[5] All tasks as well as the survey were programmed in OpenSesame (Mathôt, Schreij, & Theeuwes, 2012).

The tasks chosen for the experiment avoid targeting cognitive abilities that have been consistently associated with gender-based advantages, such as mental rotation and verbal abilities (Hirnstein, Coloma Andrews, & Hausmann, 2014). Furthermore, no gender stereotypes were elicited by the experimenters prior to task performance, and there is little basis to believe that the subjects carried internalized gender-based stereotypes or intrinsic interest based on task characteristics considering the non-familiar and mundane nature of the tasks. The lack of statistically significant gender differences in the self-reported measures of task effort engagement and likeability (see Table A1) supports the assumption that these tasks are not gender-typed.

2.3. Measures

2.3.1. Real-effort score

Raw scores were assigned according to the number of correct answers given per two-minute round, with scores recorded as zeros for any rounds for which the task was not performed. Real-effort score is estimated as the raw score per two-minute round, standardized within the distribution of all scores from the same task. Because of the crucial role of expectancy beliefs discussed above, all models controlled for fluid intelligence as a proxy for self-perceived ability.

2.3.2. Real-effort task completion

Real-effort task completion is a binary indicator of whether the subject completed the task or played the leisure activity for each round.

[2] Subjects were classified as either boy or girl according to what gender they indicated on self-reported surveys. 4 subjects did not report their gender. Prospective power analyses were carried out to determine the minimum sample size necessary to detect a small effect (Cohen's $d = 0.1$) with a Type I error probability of $p = .05$. To do so with 80 % power, a sample size of $n = 782$ was determined.

[3] In most schools, all invited fifth-grade classes took part in the study. We have reasons to believe that within-class non-participation was minimal as the median number of participating students from each class is 23, and the median number of students per class enrolled in fifth and sixth grades in Madrid in the 2019–2020 school year was 25 (Ministerio de Educación y Formación Profesional, 2021). A good share of non-participation can be attributed to sickness related absences that are likely unrelated to gender. Furthermore, the sample is relatively representative of the fifth-grade population in Madrid and therefore should be relatively balanced by gender. The average proportion of boys in each class who participated is 47.7 %, while the overall proportion of boys enrolled in obligatory secondary education was 51.3 % (Mañas Antón, 2019).

[4] All sessions took place in the experimental economics laboratory on campus that is equipped with standard desktop computers and cubicles.

[5] Each task was explained to the entire class by one experimenter. Children had the opportunity to ask clarification questions, answered control questions to ensure their correct understanding of tasks and incentive conditions, and performed several practice trials that provided feedback. After performing these steps successfully, it was assumed that all subjects correctly understood tasks and felt capable and ready to carry out the experimental trials.

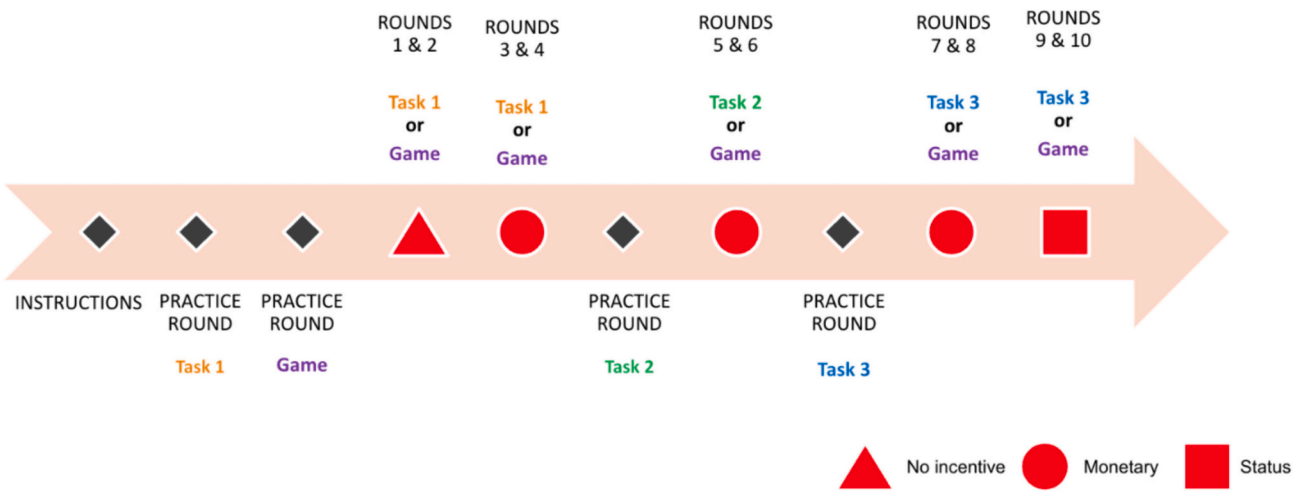


Fig. 1. Experimental set-up.
The diagram shows design and structure of the experiment using the three real-effort tasks – the slider task, AX-CPT task, and the Simon task. Which task came first, second, and third varied across sessions to avoid order effects. Before each round the student could choose to do the task or play the leisure game. Each round lasted for 2 min.

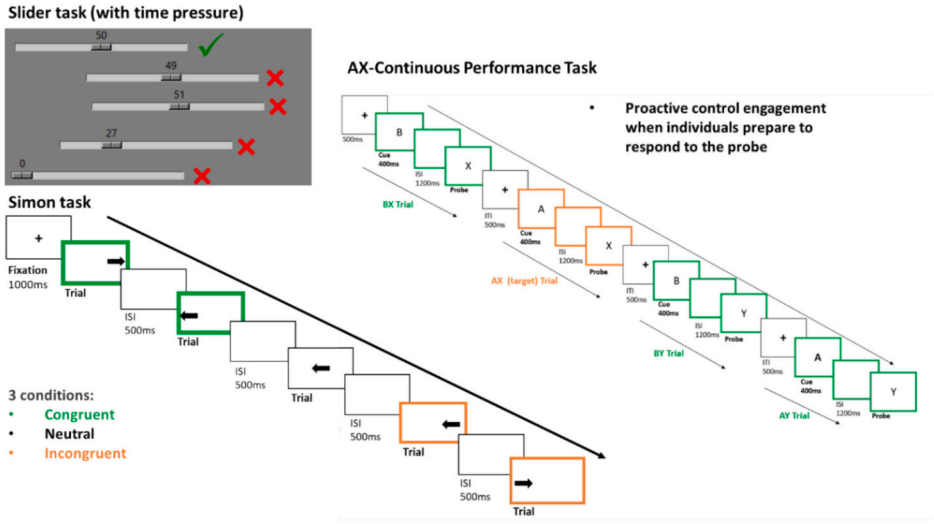


Fig. 2. Illustration of the real-effort tasks.

2.3.3. *Fluid intelligence*

Fluid intelligence is measured via the Raven's progressive matrices test (Raven & Court, 1998). The total number of correct answers given by a subject are standardized within the distribution of all other scores given in the Madrid sample.

2.3.4. *Personality measures*

The selected personality scales measure need for cognition, risk taking, and delay of gratification, as well as the Big Five traits – conscientiousness, agreeableness, openness, neuroticism, and extraversion. These were selected to capture potential heterogeneity in personality traits by gender that may be explaining the relationship between gender and effort. Need for cognition is measured as the sample standardization of the average of the scores given by each subject on a four-item series of questions, with each individual item measured on a 5-point Likert agreement scale (Beißert et al., 2014). Risk taking is measured as done in the Global Preference Survey, calculated as the sample-standardized score given on an 11-point scale from 0 to 10, where 0 indicates that the subject is not willing to take risks and 10 indicates that he or she is very willing to take risks (Falk et al., 2018; Falk, Becker, Dohmen, Huffman, & Sunde, 2023). Delay of gratification is measured as a binary indicator representing whether a subject would prefer to receive a reward immediately or delay its receipt in exchange for a double of the reward (Blossfeld, Von Maurice, & Schneider, 2011). The Big Five traits are measured via the Pictorial Personality Traits Questionnaire for Children, and values are calculated as the sample standardization of the average of the scores given by each subject on a series of items measuring each trait, with each individual item measured on a 5-point Likert agreement scale (Maćkiewicz & Cieciuch, 2016). See Table B1 of Appendix B for a complete description of each personality dimension and their corresponding items/measures.

2.3.5. *Other subject-level explanatory variables*

Age is measured in whole months, considering subjects were born on the first day of the corresponding month. *Mouse use* is coded as a 4-level ordinal variable of how often the subject uses a desktop computer with a mouse. *Computer gaming* is coded as 5-level ordinal variable indicating daily computer, tablet, or mobile use for videogames.

2.4. Modelling and statistical analysis

All regression models in the main analyses are three-level random intercept hierarchical models. The units of analysis are individual task-incentive condition rounds (level 1) grouped at the subject level (level 2) and experimental session level (level 3) to account for random variation in baseline real-effort scores, task completion, and reaction time/error rate between individual students as well as between classes (Figs. 3–7, Tables 1–2, A4–A6). Dependent variables are modeled as linear continuous, including task completion, which takes a value of one if the student opted to do the task over play the leisure game in a specific round, and zero otherwise.[6]

The nested modelling for the principal analyses (Tables 1–2) was done in four steps. Model (1) estimates the effect of gender, incentive condition, and the interaction of gender and incentive condition on real-effort score and task completion.[7] Model (2) enters fluid intelligence as a control to examine potential heterogeneities in the effort-gender relationship by ability. Model (3) includes the three-way interaction between gender, incentive condition, and fluid intelligence to see if effort investment gaps by gender differ depending on ability level. Finally, model (4) includes personality measures to assess potential heterogeneities in personality traits by gender that might explain potential effort-gender correlations. As a prerequisite to performing individual *t*-tests for the significance of between-group comparisons of the interaction terms, a restricted model that constrains the interaction effect of gender and incentive condition to zero was compared with model (1) via a (type III) Likelihood Ratio test, indicating superior fit of the unrestricted model and evidence for a statistically significant interaction term (Table 1: $\chi^2 = 22.39$, $p < .001$; Table 2: $\chi^2 = 6.63$, $p = .04$). An independent test comparing model (4) with a restricted model that constrains the interaction effect of gender, incentive condition, and fluid intelligence to zero was also performed, with evidence supporting a statistically significant interaction term (Table 1: $\chi^2 = 14.41$, $p < .001$; Table 2: $\chi^2 = 11.32$, $p = .003$). The linear mixed-effects models were fit via maximum likelihood estimation using the *lmer* function in the R (v.4.2.2) package *lme4* v.1.1.34 (Bates, Maechler, Bolker, & Walker, 2023). Model fit is evaluated with Akaike's Information Criterion as well as the marginal and condition *r*-squared, which respectively indicate how much of the total variance can be explained by the fixed effects variance and how much can be explained by the fixed and random effects variance combined. Due to the difficulty of counting parameters in multilevel models with crossed random effects, *p*-values are calculated using the Kenward-Roger approximation to get approximate degrees of freedom (Baayen, Davidson, & Bates, 2008; Luke, 2017).

Experimental and survey data were preprocessed in Stata (v.18) (StataCorp., 2023). Final study datasets, statistical analyses, and reporting were produced in R (v.4.2.2) (R Core Team, 2023).

3. Results

3.1. Gender differences in real-effort scores

Fig. 3 displays gender gaps in real-effort scores. There is no statistically significant difference in cognitive effort between boys and girls in the absence of rewards (effect of being a boy $\beta = 0.01$ *SD*, $p = .76$). In the monetary-incentive condition, average effort across all students increases by 1.38 *SD* ($p < .001$), but boys outwork girls on average by 0.17 *SD* ($p < .001$). Similarly, in the status-incentive condition (which also includes piece-rate payoffs), average effort increases by 0.24 *SD* ($p < .001$) as compared to the monetary incentive condition, with boys again outworking girls.[8] However, the gender gap in effort in the monetary incentive condition does not widen significantly with the addition of a status incentive ($\beta = 0.06$ *SD*, $p = .19$).

The marginal effect estimates representing the interaction of gender, incentives, and fluid intelligence are plotted in Fig. 4. When no external incentives are present, an effort gap by fluid intelligence emerges, driven by its differential positive impact among boys, with the effect of one *SD* advantage in fluid intelligence leading to an increase in real-effort score that is 0.10 *SD* ($p = .02$) greater for boys than for girls (see Appendix Fig. A3). The overall gap by fluid intelligence significantly widens when monetary incentives are introduced, driven by its differential positive impact among girls ($\beta_{(\text{gender=girl})\times(\text{incentive=monetary, ref. cat. = no incentive})\times(\text{fluid intelligence})} = 0.14$ *SD*, $p < .001$). Furthermore, we find evidence that only for boys does sensitivity to an additional status reward depend on fluid intelligence. It is among high-ability boys that we observe a substantial relative increase in effort moving from the piece-rate to the tournament setting ($\beta_{(\text{gender=boy})\times(\text{incentive=status, ref. cat. = monetary})\times(\text{fluid intelligence})} = 0.11$, *SD*, $p = .01$). Still, even after controlling for this differential effect of ability we do not find that boys increase their effort relatively more than girls in the competition rounds overall.

We next investigate whether the incentive-specific gender differences are attributable to potential confounding factors. Regression results in Table 1 shows that boys score on average 0.14 *SD* greater than girls do across all incentive conditions ($p < .001$); that difference is not explained away by controlling for age, frequency of mouse use, nor computer gaming. The base specification, displayed in model (2) of Table 1, controls for differences in sample-standardized measures of fluid intelligence, which does not substantially change the gender gap. We additionally control for need for cognition, risk taking, delay of gratification, conscientiousness, agreeableness, openness, neuroticism, and extraversion in model (4) to see how the differential gender effects may be explained by heterogeneities in these psychological characteristics by gender (referred to throughout the rest of the paper as "personality traits"). Though need for cognition is itself a positive and statistically significant predictor of effort, the addition of these variables does not substantially alter the gender gap.

3.2. Gender differences in real-effort task completion

Next, we test the associations between gender and incentives on the *decision* to complete the real-effort task or not. Mean differences in self-reports of the leisure game likeability reported in Table A1 present evidence for the gendering of interests such that boys, on average, report liking the ball game more than girls ($p < .001$) and girls liking the puzzle more than boys ($p < .001$). There is no large difference between how much boys reported liking the ball game and how much girls reported liking the puzzle game, supporting our assumption that the two games together are equally attractive to both genders. Participants chose the leisure game in 54 % of the rounds in the no-incentive condition (Table A2). Once monetary incentives are introduced, however, the choice of leisure task drops to below 4 % of cases, providing evidence that adding extrinsic rewards increases the relative benefits of effort engagement.

Results from the base specification shown in model (2) of Table 2 show that boys overall gamed more than girls did, and significantly

[6] Linear probability models are used to model task completion. While we recognize that certain assumptions of linear regression are violated by linear probability models (mainly that the predicted probabilities can lie outside of the bounded range of 0 and 1), we are more concerned with the interpretability of the estimated marginal effects rather than prediction, and thus follow rationale from Heckman & Snyder, 1996.

[7] Incentive condition effects are always measured as the effect changes relative to the monetary incentive condition, except when direct comparisons between the no incentive and status incentive conditions are made. In such cases, a Bonferroni correction is made to account for multiple comparisons. Specifically, a significance-level of $\alpha = 0.05/2 = 0.025$ was used.

[8] See Appendix Fig. A2 for the direct comparison of the gender-incentive interaction effect in the no-incentive condition and status-incentive condition.

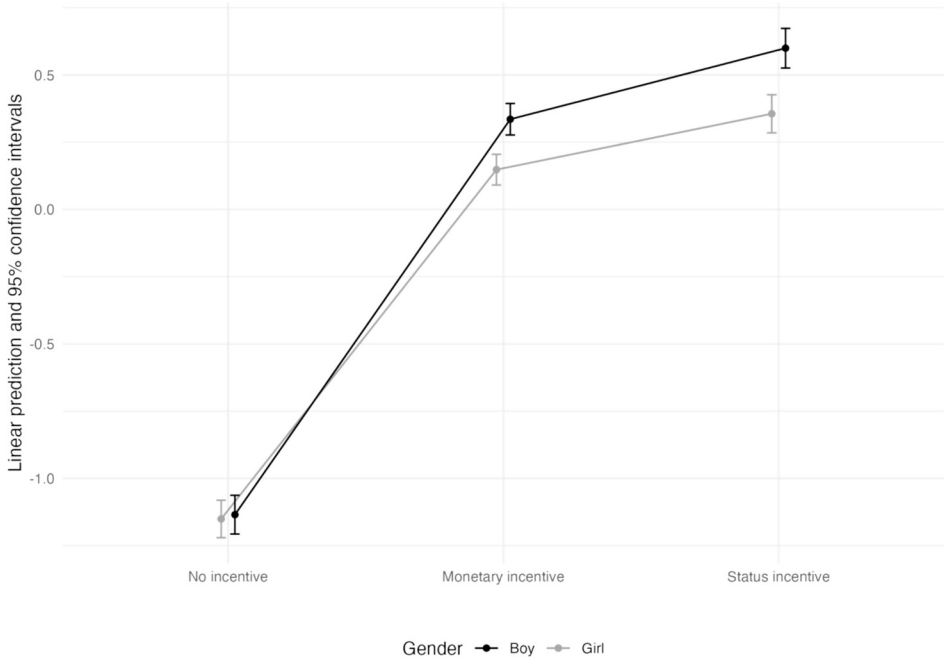


Fig. 3. Linear prediction of real-effort score by gender across incentive conditions.
This figure shows the linear predictions for boys and girls resulting from a three-level hierarchical regression model grouped at the student and class (experimental session) levels, where the dependent variable is the real-effort score per round. Model includes controls for experimental conditions (incentive condition, round, and task), individual-level fixed effects (age, mouse use, computer gaming, and fluid intelligence), and the interaction of gender and incentive condition. For all between-condition comparisons of the gender gap, see Fig. A2 of Appendix.

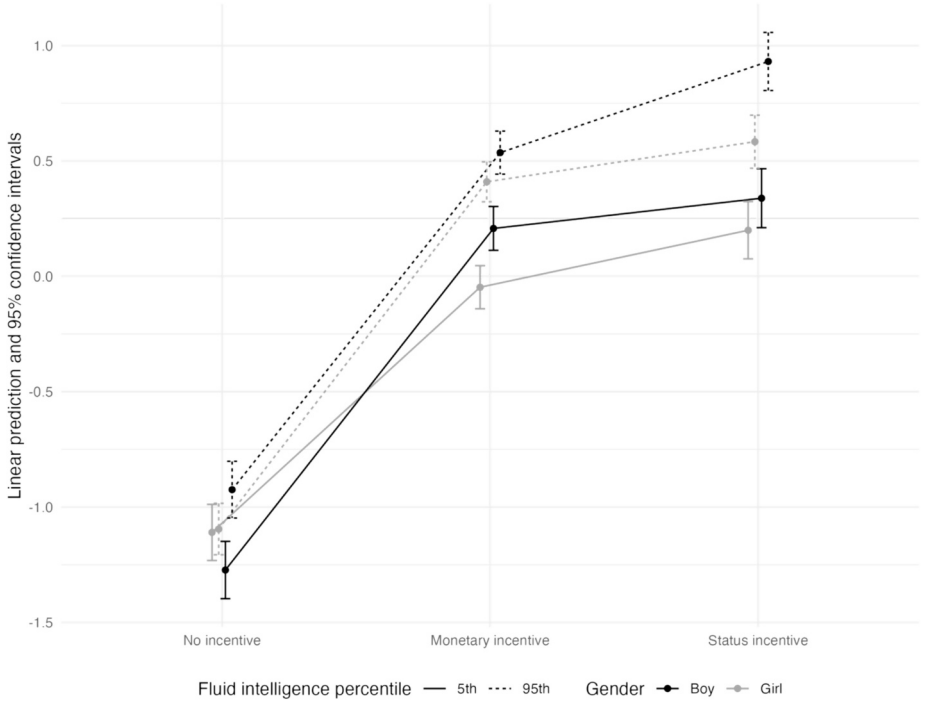


Fig. 4. Linear prediction of real-effort score across incentive condition, by gender and percentile of fluid intelligence (5th and 95th).
This figure shows the linear predictions for boys and girls resulting from a three-level hierarchical regression model grouped at the student and class (experimental session) levels, where the dependent variable is the real-effort score per round. Model includes controls for experimental conditions (incentive condition, round, and task), individual-level fixed effects (age, mouse use, computer gaming, and intelligence), and the interaction of gender, incentive condition, and fluid intelligence. For all between-condition comparisons of the moderating effect of gender on the fluid intelligence and incentive condition interaction, see Fig. A3 of Appendix.

more so in the no-incentive condition ($p < .05$). Specifically, being a boy reduces the probability of completing the task by on average 2 percentage points (PP) ($p = .02$) across all incentive conditions, and by 4 percentage points in the no-incentive condition when compared to tasking rates in the monetary incentive condition ($p = .002$).

While we find no evidence that fluid intelligence affects the

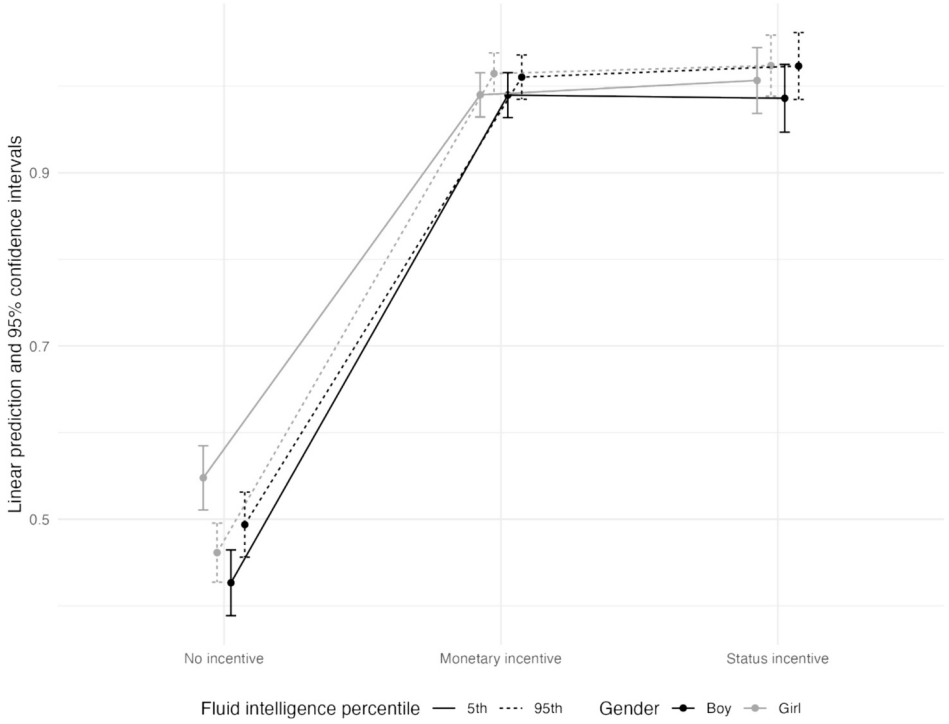


Fig. 5. Linear prediction of the probability of completing the real-effort task across incentive conditions, by gender and percentile of fluid intelligence (5th and 95th). This figure shows the linear predictions for boys and girls resulting from a three-level hierarchical regression model grouped at the student and class (experimental session) levels, where the dependent variable is whether the real-effort task was completed in each round. Model includes controls for experimental conditions (incentive condition, round, and task), individual-level fixed effects (age, mouse use, computer gaming, and intelligence), personality traits, and the interaction of gender, incentive condition, and fluid intelligence.

probability of completing the real-effort task on average, we do find evidence of its differential effects by gender. Results in model (3) of Table 2 are visualized in Fig. 5. Having one *SD* greater fluid intelligence increases boys' probability of completing the real-effort task by about 2 percentage points more than girls' of equal ability across all incentive conditions ($p = .02$). This is driven particularly by tendencies in the no-incentive condition, where lower-ability girls are more likely to engage with the effortful task than girls of high ability, while the opposite effect is observed for boys. Though highlighting within-gender heterogeneities, this differential effect of intelligence does not explain boys' overall tendency to play the game instead of doing the task when effort is unincentivized.

The lack of statistical significance of the gender effect in model (4) suggests that heterogeneities in personality traits by gender may be partially explaining the gender effect on choosing to do the task, with more agreeableness and neuroticism being positively associated with tasking ($\beta_{agreeableness} = 0.01$ PP, $p = .04$; $\beta_{neuroticism} = 0.01$ PP, $p = .04$).

3.3. Sensitivity analyses

Grissom & Reyes, 2019 note that gender effects in EF "depend greatly on the modality of testing and the parameters tested" and suggest that observed differences may reflect differences in strategy and outcome preferences between boys and girls when confronted with an ambiguous task. Therefore, it is prudent to examine if boys achieve this advantage through specific strategic choices or the application of certain subdomains of EF over others. Hence, we investigate whether our findings using a performance score-based index of effort are compatible with other effort-related subindices.

3.3.1. Gender differences in task strategy

In most score-based psychological tasks, subjects organically choose to focus on either accuracy or speed (Westbrook & Braver, 2015). Within the basic dilemma constituted by this tradeoff, recent studies have found that individuals adjust their strategic choices in response to incentives (Otto & Daw, 2019). Therefore, we investigate gender differences in average reaction times under different incentive conditions, measured as the average reaction time in milliseconds for all trials where a correct response was given, and error rates, measured as the percentage of trials where an incorrect answer was given, per round in the AX task and the Simon task, respectively, and standardized within the task-specific distribution of average response times and error rates. Reaction time and accuracy measures are not applicable to the slider task due to its layout. Results in Appendix Table A3 show that boys are on average faster than girls in responding to trials by about 0.40 *SD* ($p < .001$) on the AX task and by about 0.64 *SD* ($p < .001$) on the Simon task, but they are not more accurate on average. Fig. 6 further visualizes how the gender gap in reaction time and error rate for each task changes between incentive conditions. Boys' advantage in reaction time on the AX task does not significantly differ between incentives. On the Simon task, boys' advantage in reaction time is greatest in the absence of rewards ($\beta = -0.78$ *SD*, $p < .001$) but does not change significantly when monetary incentives are introduced (an estimated increase of $\beta = 0.10$ *SD*, $p = .19$). When status incentives are added in the tournament condition, the gender gap in reaction time for the Simon task closes by about 40 % on average as compared to the no-incentive condition gap (increase of $\beta = 0.31$ *SD*, $p = .002$) and by nearly 18 % of the monetary incentive condition gap (increase of $\beta = 0.21$ *SD*, $p = .001$). While these findings diverge somewhat from the overall patterns when predicting real effort, it should be noted that we only observe reaction times for two of the three tasks, and never for the rounds that the leisure task was opted for. For neither task do we detect gender differences in error rates, nor differential effects of incentives on error rates by gender. Thus, boys respond faster without sacrificing accuracy, allowing them to complete more trials than girls within a given time, on average, resulting in higher scores.

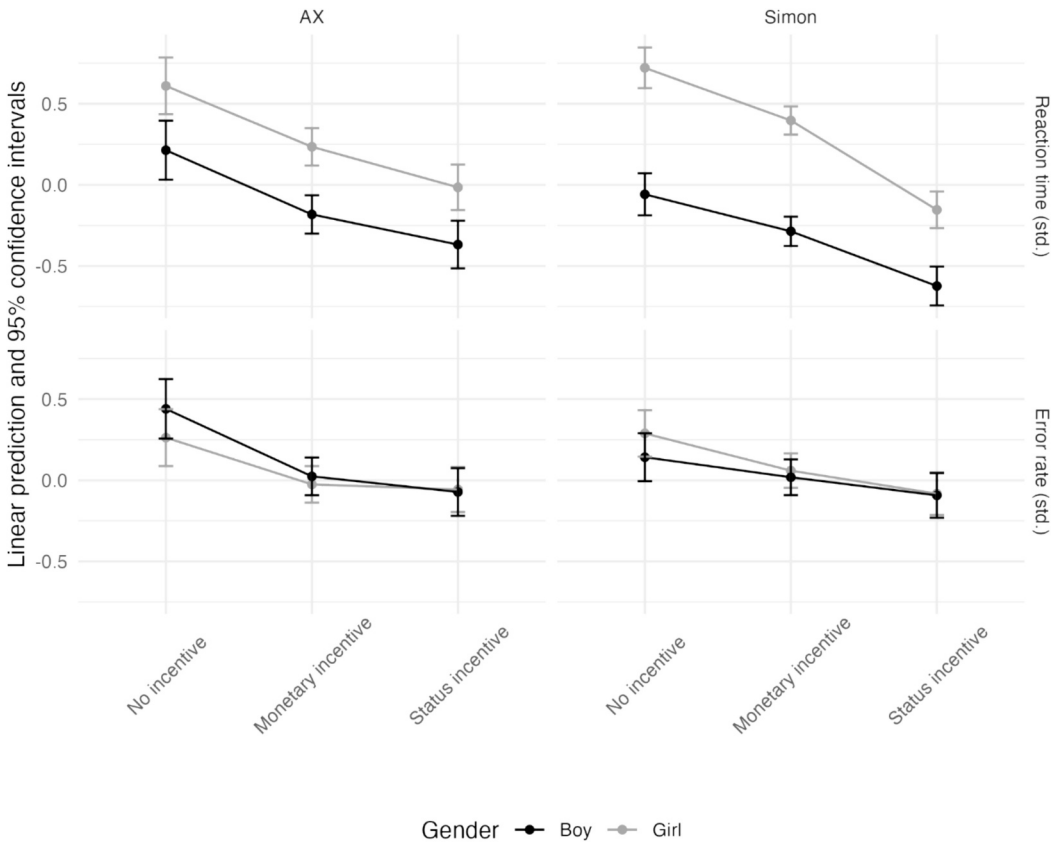


Fig. 6. Linear prediction of standardized average reaction time and error rate by gender across incentive conditions, stratified by real-effort task (AX and Simon). This figure shows the linear predictions for boys and girls resulting from a three-level hierarchical regression model grouped at the student and class (experimental session) levels, stratified by real-effort task. The dependent variables are the standardized average reaction time for correct responses and error rate. All models include controls for experimental conditions (incentive condition, round, and task), individual-level fixed effects (age, mouse use, computer gaming, and intelligence), and the interaction of gender and incentive condition.

3.3.2. Gender differences in inhibition, cognitive flexibility, and proactive and reactive control

To investigate whether the observed strategic choices of boys and girls may instead be the result of the predominance of certain dimensions of cognitive functioning, we examine task-specific indices of speed and accuracy for the Simon and AX tasks that measure inhibition, cognitive flexibility, and cognitive control type (proactive and reactive control).

On the Simon task, subjects tend to respond slower to and commit more errors on incongruent trial conditions, as they must inhibit automatic responses to the conflicting spatial information. To measure this phenomenon, known as the Simon effect, we subtract the average reaction time and error rate on congruent trials per task-condition for each subject from those on incongruent trials. Results in Table A4 indicate that boys tend to be less susceptible to the Simon effect when it comes to reaction time, as indicated by a smaller difference in response time between incongruent and congruent trials than girls ($\beta = -0.23$ *SD*, $p = .001$). Again, we detect no statistically significant gender difference with regards to error rate ($\beta = 0.02$ *SD*, $p = .79$).

The AX-CPT task is used to assess the proactive and reactive dimensions of cognitive control, and the ability to switch flexibly between the two.[9] The Proactive Behavioral Index, or PBI,[10] provides a measure of how much proactive interference one experiences in situations when a reactive approach is required. For average reaction time on correct trials and error rates, we find that being a boy positively predicts PBI for both, indicating that boys tend to engage in proactive control relatively more than girls, while girls tend to engage in reactive control relatively more than boys (Table A4).

Recent studies have argued for the functional independence of proactive and reactive control, rather than operating as two poles of a continuous spectrum (Mäki-Marttunen, Hagen, & Espeseth, 2019). Thus, to test whether these gender-based tendencies in cognitive control type resulted in superior performance for boys in trials where a proactive approach is required and for girls in trials where a reactive approach is required, we model average reaction times and error rates on B-X and A-Y trials separately. The results in Fig. 7 confirm again that boys'

[9] It has been found that in most young, healthy populations, proactive control processes predominate reactive ones (Braver et al., 2009).

[10] The proactive behavioral index (PBI) in reaction times and error rates for the AX task is calculated as *(AY – BX) / (AY + BX)*. Those who experience greater interference from proactive control type will experience greater reaction times and error rates on reactive trials (AY) and thus PBI > 0. PBI < 0 when someone experiences greater interference from reactive control type and thus have greater reaction times and error rates on proactive trials (BX). A correction is made for error rates that are equal to zero such that *(error rate + 0.5)/(frequency of trials + 1)*.

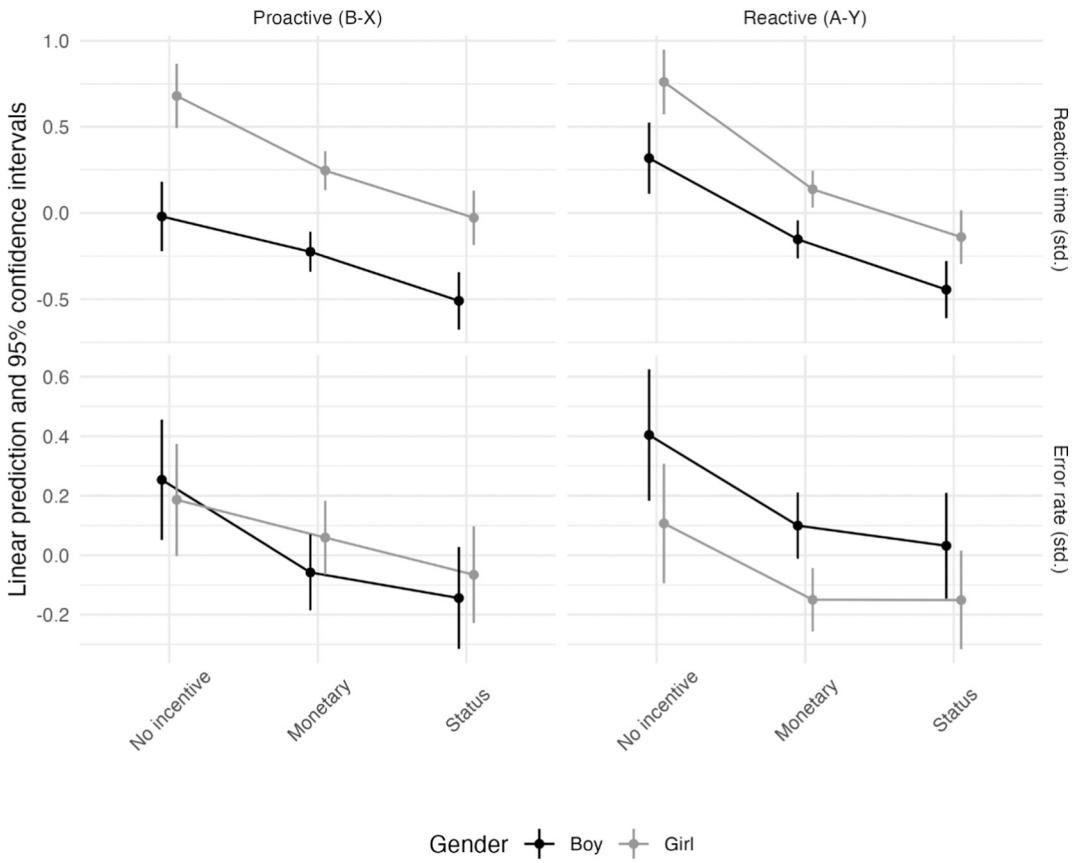


Fig. 7. Linear prediction of standardized reaction time and error rate by gender across incentive conditions and AX-CPT trial condition.
This figure shows the linear predictions for boys and girls resulting from a three-level hierarchical regression model grouped at the student and class (experimental session) levels, where the dependent variable is the mean standardized reaction time for correct responses and error rate for proactive (B-X) and reactive (A-Y) trials on the AX-CPT task, separately. All models include controls for incentive condition, individual-level fixed effects (age, mouse use, computer gaming, and intelligence), and the interaction of gender and incentive condition.

response time is overall faster than girls': boys are significantly quicker than girls in the AX task regardless of the trial condition (B-X condition: $\beta = -0.55$ *SD*, $p < .001$; A-Y condition: $\beta = -0.35$ *SD*, $p < .001$). With respect to error rate, there is no statistically significant difference between boys and girls on proactive trials. Girls, however, are overall more accurate than boys when in reactive mode, the only observed gender gap where girls hold an advantage ($\beta_{\text{girl (accuracy)}} = 0.24$ *SD*, $p = .003$).

4. Discussion

The purpose of this study was to assess how effort levels and changes in effort in response to monetary and status incentives differ by gender. The main results show that when it comes to real-effort scores, boys and girls do not exert significantly different levels on average in the absence of rewards. However, boys are less likely to choose the real-effort task over a leisure task than girls are, with most of that difference arising from the no-incentive condition. Thus, we find partial support of our hypothesis that girls exert more effort than boys in the absence of external rewards (H1a). As hypothesized, the degree to which effort increases with the shift from the no-incentive condition to the monetary condition (with performance-based incentives) is greater for boys than it is for girls, regardless of effort outcome (H2a). We do not find any support, however, that boys increase their effort more so than girls when competing for an additional status incentive placed on top of performance-based monetary incentives, as was argued in behavioral economics studies (H3a).

As suggested under the EVT framework, ability may interact with motivational orientations and affect decisions as to how much effort is worth applying given one's beliefs about his or her probability of success. We find evidence that the degree to which this occurs varies by gender. In the absence of incentives, fluid intelligence has a greater positive impact on determining effort for boys than for girls (H1b). The effect of being a boy on the degree to which fluid intelligence moderates the incentive effect significantly reduces with the introduction of performance-based monetary incentives as compared to the no-incentive condition; however, this is not due to a closing of the gap by fluid intelligence for boys, but rather an opening of the gap for girls (H2b). When an additional status incentive is placed on top of the monetary incentive, the change in effort differential by fluid intelligence is again significantly greater for boys than it is for girls (H3b).

This evidence reaffirms that incentivizing performance can boost average effort levels substantially, especially for low-ability boys who may be particularly prone to disengaging from effortful tasks. However, the results also reveal that gender differences in effort can widen when incentives are introduced, and particularly when material rewards are offered, which low-ability girls are not as sensitive to. Gender gaps that have been documented across academic exams and aptitude tests may therefore be at least partially due to differential motivational responses elicited by the context of the testing situation. While our findings strongly support the effectiveness of monetary and status incentives in boosting achievement via effort, with score-based incentivized effort levels maximized by nearly 1.62 *SD*, even slight gender differences in effort might snowball into relevant inequalities in real-world outcomes. Indeed, Schlosser, Neeman, & Attali, 2019 find that about 4 % more men than women exerted extreme low levels of effort on a low-stakes experimental Graduate Record Examination, a gender difference

Table 1
Effect of gender, incentives, and gender-incentive interaction on real-effort score.

	Dependent variable:			
	Real-effort score (standardized within task)			
	(1)	(2)	(3)	(4)
Gender = boy	0.14***	0.15***	0.15***	0.15***
	(0.03)	(0.03)	(0.03)	(0.03)
Incentive = no incentive (ref. = monetary)	−1.38***	−1.38***	−1.38***	−1.38***
	(0.02)	(0.02)	(0.02)	(0.02)
Incentive = status (ref. = monetary)	0.24***	0.24***	0.24***	0.24***
	(0.02)	(0.02)	(0.02)	(0.02)
Gender = boy × Incentive = no incentive (ref. = monetary)	−0.17***	−0.17***	−0.18***	−0.18***
	(0.04)	(0.04)	(0.04)	(0.04)
Gender = boy × Incentive = status (ref. = monetary)	0.06	0.06	0.06	0.06
	(0.04)	(0.04)	(0.04)	(0.04)
Task = AX (ref. = Slider)	−0.001	−0.002	−0.003	−0.001
	(0.02)	(0.02)	(0.02)	(0.02)
Task = Simon (ref. = Slider)	0.01	0.01	0.01	0.01
	(0.02)	(0.02)	(0.02)	(0.02)
Round 2	−0.10***	−0.09***	−0.09***	−0.09***
	(0.02)	(0.02)	(0.02)	(0.02)
Age (months)	0.003	0.004	0.004	0.005
	(0.003)	(0.003)	(0.003)	(0.003)
Mouse use	0.04**	0.04***	0.04**	0.04**
	(0.01)	(0.01)	(0.01)	(0.01)
Videogaming	0.02	0.02	0.02	0.02
	(0.01)	(0.01)	(0.01)	(0.01)
Fluid intelligence		0.11***	0.11***	0.10***
		(0.01)	(0.02)	(0.02)
Gender = boy × Fluid intelligence			0.04	0.04
			(0.03)	(0.03)
Incentive = no incentive × Fluid intelligence (ref. = monetary)			−0.07**	−0.07**
			(0.02)	(0.02)
Incentive = status × Fluid intelligence (ref. = monetary)			0.03	0.03
			(0.02)	(0.02)
Gender = boy × Incentive = no incentive × Fluid intelligence (ref. = monetary)			0.14***	0.14***
			(0.04)	(0.04)
Gender = boy × Incentive = status × Fluid intelligence (ref. = monetary)			0.11*	0.11*
			(0.04)	(0.04)
Need for cognition				0.04*
				(0.02)
Risk taking				−0.001
				(0.01)
Delay of gratification				0.04
				(0.03)
Conscientiousness				−0.01
				(0.02)
Agreeableness				0.04*
				(0.02)
Openness				−0.01
				(0.02)
Neuroticism				0.02
				(0.02)
Extraversion				0.01
				(0.02)
Constant	−0.58	−0.66*	−0.66*	−0.81*
	(0.35)	(0.33)	(0.33)	(0.33)
Student-level variance	0.11	0.10	0.10	0.10
Class-level variance	0.02	0.01	0.01	0.01
Observation-level variance	0.52	0.51	0.51	0.51
R-squared (marginal)	0.36	0.37	0.37	0.38
R-squared (conditional)	0.48	0.48	0.49	0.49
Students	799	798	798	794
Classes	35	35	35	35
Observations	7874	7866	7866	7834
Akaike Inf. Crit.	18,107.92	17,997.91	17,976.78	17,874.37

This table shows the results of a three-level hierarchical regression model grouped at the student and class (experimental session) levels, where the dependent variable is the real-effort scores standardized within task. Gender and incentive condition are specified as contrasts with mean zero, and fluid intelligence is mean-centered. Therefore, the estimated effect of gender (boy) is the average effect across all incentive conditions and fluid intelligence levels, gender (boy) and fluid intelligence interaction effect is the average estimated effect across incentive conditions, and incentive condition and fluid intelligence interaction effects are the average estimated effects across gender. *P*-values are calculated with the Kenward-Roger approximation to get approximate degrees of freedom. Standard errors are shown in parentheses.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

similar to the one we find in the decision to complete the real-effort task in the no-incentive condition. Additionally, their results show that there is a greater increase in performance for men than for women – a difference of 10 percentile ranks – when scores from the low-stakes test are compared to those from a high-stakes test. A small advantage in percentile ranking can act as the deciding factor during a selective admissions process, showing that even the smallest difference in effort could have disproportionately large repercussions. Given the evidence that high-ability boys may be more motivated than girls of equal ability to compete, these differences may be exacerbated in highly competitive admission processes if gender quotas are not instated.

While the study does not directly identify what underlies the observed motivational differences between boys and girls, previous research has suggested that girls may be more mastery-oriented, while boys are more performance-oriented (Diaconu-Gherasim, Tepordei, Mairean, & Rusu, 2019; Diseth & Samdal, 2014; Dweck & Leggett, 1988). Under these assumptions, girls' valorization of effort and learning for self-improvement may motivate their adoption of a cautious, reactive approach and thus lead to slower reaction times to avoid incorrect responses. Previous studies that have examined gender differences in risk-taking behavior indicate that it is around the age of 10 to 13 that these differences are the largest, with overall risk-taking tendencies growing more similar between men and women with age (Byrnes, Miller, & Schafer, 1999). On the other hand, boys' stronger concern with relative performance may motivate higher-ability boys to compete more intensely, while at the same time causing lower-ability boys to disengage from effortful tasks, particularly when there are no tangible rewards at stake. This is in line with Simzar et al. (2015) who indeed find that performance avoidance goals are more detrimental to lower achieving students, and imply that, given boys' higher rates of performance avoidance, this may mean lower motivation, graduation rates, and overall academic achievement for those at the lowest end of the distribution.[11]

4.1. Limitations

Several limitations of the study should be mentioned. First, while low-skill, monotonous tasks minimize the confounding role of ability and allow us to more accurately measure the impact of incentives on performance-based cognitive effort, they may not accurately capture all dynamics of effortful cognitive processes that determine real-world outcomes such as school grades and educational attainment. Previous studies that have tested the effect of monetary incentives on low-stake testing performance have found a positive effect only for easy- and medium-difficulty questions (O'Neil, Abedi, Miyoshi, & Mastergeorge,

[11] It should be noted that Simzar et al. (2015) examines the relationship between motivation and achievement among high school students in California. Related research on achievement goals suggests that these relationships may be culture-dependent, and that findings using data from Western countries may not apply to more collectivist societies (King, 2016).

Table 2
Effect of gender, incentives, and gender-incentive interaction on real-effort task completion.

	Dependent variable:			
	Real-effort task completion (linear probability)			
	(1)	(2)	(3)	(4)
Gender = boy	−0.02*	−0.02*	−0.02*	−0.02
	(0.01)	(0.01)	(0.01)	(0.01)
Incentive = no incentive (ref. = monetary)	−0.52***	−0.52***	−0.52***	−0.52***
	(0.01)	(0.01)	(0.01)	(0.01)
Incentive = status (ref. = monetary)	0.01	0.01	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Gender = boy × Incentive = no incentive (ref. = monetary)	−0.04*	−0.04**	−0.04**	−0.04**
	(0.01)	(0.01)	(0.01)	(0.01)
Gender = boy × Incentive = status (ref. = monetary)	−0.01	−0.01	−0.01	−0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Task = AX (ref. = Slider)	−0.01	−0.01	−0.01	−0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Task = Simon (ref. = Slider)	0.01	0.01	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
Round 2	−0.04***	−0.04***	−0.04***	−0.04***
	(0.01)	(0.01)	(0.01)	(0.01)
Age (months)	0.0002	0.0002	0.0002	0.0003
	(0.001)	(0.001)	(0.001)	(0.001)
Mouse use	−0.002	−0.002	−0.002	−0.002
	(0.003)	(0.003)	(0.003)	(0.003)
Videogaming	0.0001	0.001	0.001	0.002
	(0.003)	(0.003)	(0.003)	(0.003)
Fluid intelligence		0.005	0.004	0.003
		(0.003)	(0.004)	(0.004)
Gender = boy × Fluid intelligence			0.02*	0.02*
			(0.01)	(0.01)
Incentive = no incentive × Fluid intelligence (ref. = monetary)			−0.01	−0.01
			(0.01)	(0.01)
Incentive = status × Fluid intelligence (ref. = monetary)			0.001	0.001
			(0.01)	(0.01)
Gender = boy × Incentive = no incentive × Fluid intelligence (ref. = monetary)			0.05***	0.05***
			(0.01)	(0.01)
Gender = boy × Incentive = status × Fluid intelligence (ref. = monetary)			0.01	0.01
			(0.01)	(0.01)
Need for cognition				0.01
				(0.004)
Risk taking				−0.0005
				(0.001)
Delay of gratification				−0.002
				(0.01)
Conscientiousness				−0.001
				(0.004)
Agreeableness				0.01*
				(0.004)
Openness				0.002
				(0.004)
Neuroticism				0.01*
				(0.004)
Extraversion				0.005
				(0.004)
Constant	0.80***	0.80***	0.80***	0.79***
	(0.08)	(0.08)	(0.08)	(0.08)
Student-level variance	0.00	0.00	0.00	0.00
Class-level variance	0.00	0.00	0.00	0.00
Observation-level variance	0.06	0.06	0.06	0.06
R-squared (marginal)	0.40	0.41	0.41	0.41
R-squared (conditional)	0.44	0.44	0.44	0.44
Students	799	798	798	794
Classes	35	35	35	35
Observations	7874	7866	7866	7834
Akaike Inf. Crit.	734.24	690.94	683.86	683.49

This table shows the results of a three-level hierarchical linear probability regression model grouped at the student and class (experimental session) levels, where the dependent variable is a binary indicator of whether the student opted to do the task over play the leisure game in a specific round (1 = tasked, 0 = gamed). Gender and incentive condition are specified as contrasts with mean zero, and fluid intelligence is mean-centered. Therefore, the estimated effect of gender (boy) is the average effect across all incentive conditions and fluid intelligence levels, gender (boy) and fluid intelligence interaction effect is the average estimated effect across incentive conditions, and incentive condition and fluid intelligence interaction effects are the average estimated effects across gender. *P*-values are calculated with the Kenward-Roger approximation to get approximate degrees of freedom. Standard errors are shown in parentheses.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

2005), suggesting that expectancy may play a more significant role than task value when the cognitive demands of the task increase. Thus, if the goal is to understand gender gaps in *achievement*, future research should consider whether more demanding tasks invite different patterns of effort engagement.

While a survey of educational literature finds congruent evidence that short-term and sufficiently high incentives matter more for boys than for girls (Levitt, List, Neckermann, & Sadoff, 2016), it would be insightful to moderate incentive schemes to elicit more girl-typed preferences in the form of time-delayed prizes (Angrist, Lang, & Oreopoulos, 2009) or specific rewards that are known to be more attractive to girls than to boys (Sittenthaler & Mohnen, 2020). Employing status incentives under a collaborative rather than competitive environment may better tap into what has been hypothesized as girls'' more prosocial orientations and motivations (Cassar & Rigdon, 2021; Watson & Blanchard-Fields, 1998). Furthermore, what is considered as *motivating* for one individual versus another may depend greatly on context. Thus, the effect of incentives should also be considered across different age groups, locations, and task settings to understand if the evidence of motivational differences by gender that we find among Spanish fifth-graders are also observed under varying circumstances.

It is also important to note that while monetary and status incentives may have a positive effect on effort and performance, adverse effects may arise from high-stakes testing, particularly among students from marginalized or disadvantaged populations (Grant, 2004). The nature of extrinsic incentives may lead to increases in dropout rates, cheating, and mental health issues such as stress and anxiety, and decreases in teaching quality and student engagement with the subject matter (see French, Dickerson, & Mulder, 2023 for a review).

5. Conclusion

One of the toughest dilemmas that educators and policymakers must face is how to achieve an upward shift in effort during learning processes and evaluation without leaving anyone behind nor hindering anyone from advancing forward. Resource limitations in education systems can lead to the implementation of blanket reward schemes that increase overall performance of the class or school, but fail to target equitable relative improvements for subgroups of individuals within the whole (Darling-Hammond, 2007; Thurston, Penner, & Penner, 2016). With gender differences in education remaining a key concern, it thus becomes crucial to understand how boys and girls each respond to different types of incentives so that proper action can be taken to mitigate any arising gender inequalities.

Our findings from real-effort tasks that avoid gender-typing contribute new insights into gender differences in cognitive effort as a result of differential motivational responses to incentives. This evidence can also inform intervention strategies that have been proposed to boost academic achievement through increases in effort. Schools that increase test-taking motivation and classroom participation by compensating effortful behavior with material rewards may do so particularly for low-

achieving boys, while potentially risking an opening of the gender gap. High-ability girls may benefit from knowing that they are well-equipped to compete for resources and status, whether in school or in adult life. At the same time, educators and policymakers should support the implementation of more diverse evaluation methods of students that adjust to the vast array of differences in individuals' preferences and motivations.

Funding

This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 758600). Funding for APC: Universidad Carlos III de Madrid (Agreement CRUE-Madroño 2024).

CRediT authorship contribution statement

Paula Apascaritei: Writing – original draft, Investigation, Formal analysis. **Jonas Radl:** Writing – review & editing, Writing – original draft, Investigation, Conceptualization. **Madeline Swarr:** Writing – review & editing, Writing – original draft, Visualization, Formal analysis.

Declaration of competing interest

None.

Data availability

We confirm that we have reported all measures, conditions, data exclusions, and how we determined the sample sizes. The data underlying this article will be available in the e-ciencia Datos data repository at https://doi.org/10.21950/DEDRIZ after expiration of the embargo on 1 March 2027. The analysis code to replicate this present study and further documentation is already available there. This study is part of a larger project that also uses the same or overlapping data to investigate related issues, but no other manuscript looks specifically at gender differences in effort.

Acknowledgements

We gratefully acknowledge the helpful feedback received from Sílvia Claveria Alias, Victor Gómez Blanco, Daniel Horn, Patricia Lorente, Martin Neugebauer, Alberto Palacios-Abad, Heike Solga and Jan Stuhler.

Appendix A

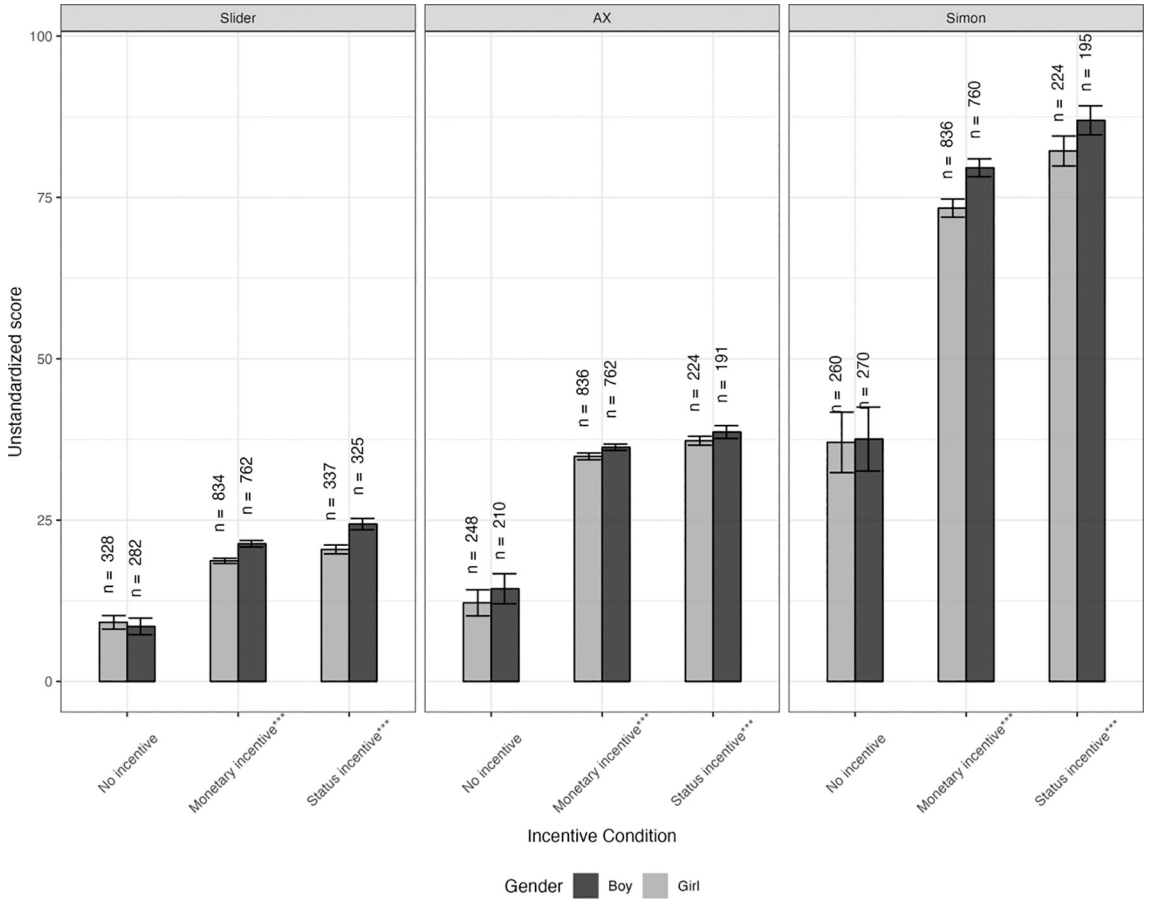


Fig. A1. Gender differences in raw performance scores: Means, confidence intervals, and Wilcoxon Rank Sum Test for each task, gender, and incentive condition. This figure shows performance by gender and incentive condition with 95 % confidence intervals (CI). Statistical significance shown as computed by a two-sided Wilcoxon Rank Sum Test. Performance data by task and counting the leisure choice as zero-score towards the task.
* $p < .05$, ** $p < .01$, *** $p < .001$.

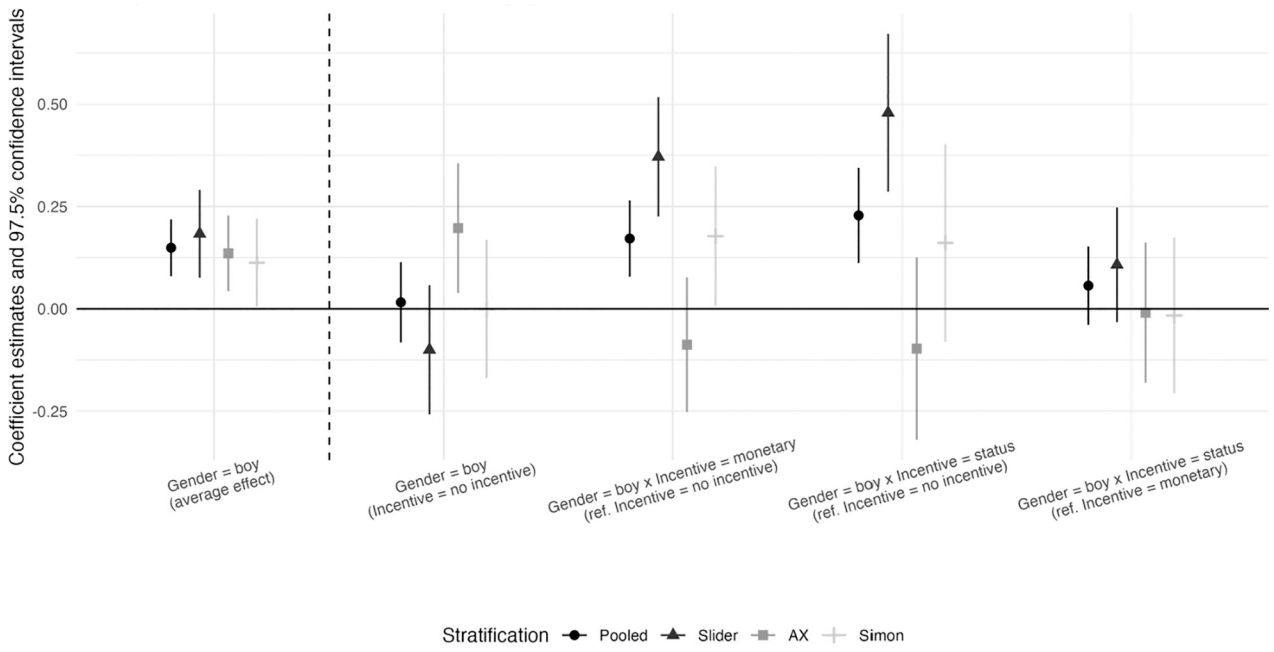


Fig. A2. The gender effect on real-effort score: Average effect and differential effects when changing between incentive conditions, pooled and stratified by task.
This figure shows a coefficient plot of the gender variable indicating boys resulting from a three-level hierarchical regression model grouped at the student and class (experimental session) levels, where the dependent variable is the real-effort score, standardized within task. All models include controls for experimental conditions (incentive condition, round, and task [for pooled data]), individual-level controls (age, mouse use, computer gaming, and fluid intelligence), and the interaction of gender and incentive condition. A Bonferroni correction was made to account for multiple comparisons, i.e. two comparisons of the additional effect of the status incentive by gender (versus no incentive and versus monetary incentive), and thus 97.5 % confidence intervals are calculated.
The average effect of being a boy (first column) is the average main effect across incentive conditions. Effects in the third and fourth columns represent the average change in the effect of being a boy when adding monetary and status incentives, with change relative to the boy effect in the no-incentive condition, which is show in the second column (regression tables not included for sake of brevity, available upon request). Effects in the fifth column represent the average change in the effect of being a boy when adding status incentives, with change relative to the boy effect in the monetary incentive condition.

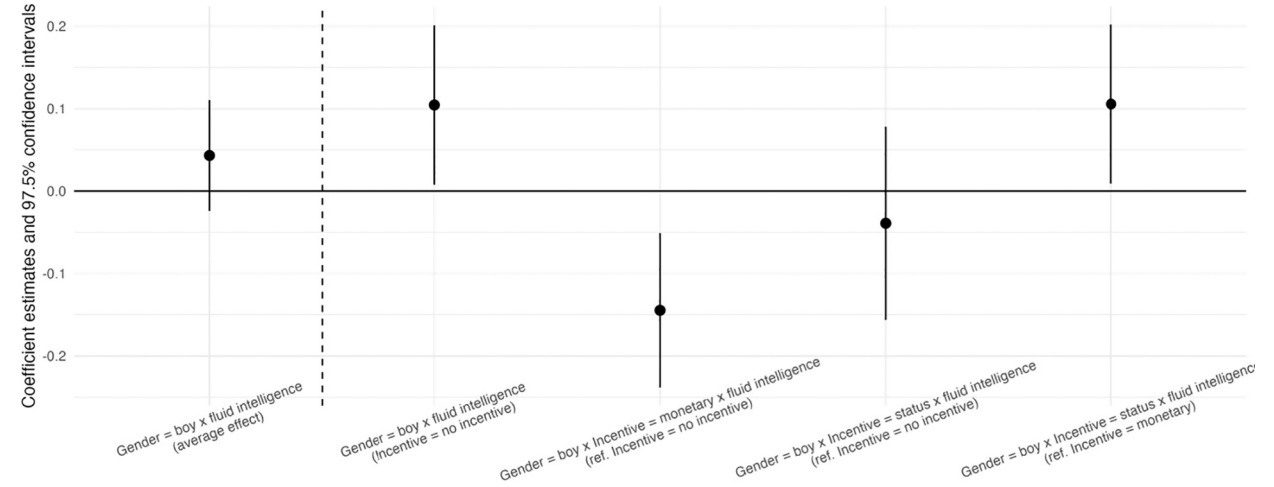


Fig. A3. The moderating effect of gender on the fluid intelligence effect on real-effort score: Average effect and differential effects when changing between incentive conditions.
This figure shows a coefficient plot of the moderating effect of being a boy resulting from a three-level hierarchical regression model grouped at the student and class (experimental session) levels, where the dependent variable is the real-effort score, standardized within task. All models include controls for experimental conditions (incentive condition, round, and task, individual-level controls (age, mouse use, computer gaming, fluid intelligence, personality traits), the interaction of gender and incentive condition, the interaction of gender and fluid intelligence, the interaction of fluid intelligence and incentive condition, and the three-way interaction of gender, fluid intelligence, and incentive condition. A Bonferroni correction was made to account for multiple comparisons, i.e. two comparisons of the additional effect of the fluid intelligence × status incentive by gender (versus no incentive and versus monetary incentive), and thus 97.5 % confidence intervals are calculated.
The average effect of being a boy × fluid intelligence (first column) is the average main effect across incentive conditions. Effects in the third and fourth columns represent the average change in moderating effect of being a boy on the fluid intelligence and incentive interaction when adding monetary and status incentives, with change relative to the boy effect in the no-incentive condition, which is show in the second column (regression tables not included for sake of brevity, available upon request). Effects in the fifth column represent the average change in the moderating effect of being a boy when adding status incentives, with change relative to the moderating effect of being a boy in the monetary incentive condition.

Table A1
Mean differences of self-reported effort, individual task difficulty and individual assessment of preference for real-effort and leisure tasks by gender.

	Boys			Girls			Two-sided Student *t*-test		
	N	Mean	Std. Err.	N	Mean	Std. Err.	P	T-value	DoF
Slider task									
Self-reported effort	314	5.580	1.490	336	5.510	1.570	0.590	0.54	650
Perceived task difficulty	314	3.150	1.080	336	2.980	1.080	0.045	2.00	650
Perceived task likeability	314	4.170	0.910	336	4.270	0.820	0.110	−1.60	650
AX-CPT task									
Self-reported effort	314	5.040	1.800	337	5.040	1.840	0.980	−0.02	650
Perceived task difficulty	314	2.400	0.980	337	2.320	1.010	0.280	1.10	650
Perceived task likeability	314	4.280	0.800	337	4.340	0.840	0.370	−0.90	650
Simon task									
Self-reported effort	314	4.960	1.860	336	5.070	1.820	0.410	−0.82	650
Perceived task difficulty	314	2.390	1.080	336	2.380	1.020	0.870	0.16	650
Perceived task likeability	314	4.360	0.840	336	4.350	0.790	0.850	0.19	650
Leisure task									
Perceived task likeability of the ball game	375	3.660	1.340	416	3.300	1.240	0.000	3.80	790
Perceived task likeability of the puzzle game	375	3.340	1.270	416	3.710	1.130	0.000	−4.30	790

To assess gender neutrality of the real-effort and leisure tasks employed, we asked participants to self-report the effort expended (scale of increasing integers from 1 = *very, very low effort* to 7 = *very, very high effort*), perceived difficulty (scale of increasing integers from 1 = *very easy* to 5 = *very difficult*) and perceived likeability (scale of increasing integers from 1 = *strongly disliked* to 5 = *strongly liked*) of each task. We observe no statistically significant differences by gender in the real-effort tasks for self-reported effort nor perceived task likeability. While there is a statistically significant difference by gender in the perceived difficulty of the slider task, it is only slight, with boys reporting the task as 0.17 of a point more difficult than girls did on average. For the leisure tasks, boys like the ball game more than girls ($p < .001$) and girls like the puzzle better ($p < .001$), on average.

Table A2
Proportion test of the leisure game choice by gender.

	Girls			Boys			Two-sided test of equal proportions	
	Rounds gamed	Total rounds	% Gamed	Rounds gamed	Total rounds	% Gamed	P	χ^2
All tasks								
No incentive: round 1	179	419	42.7	189	381	49.6	0.060	3.54
No incentive: round 2	259	419	61.8	238	381	62.5	0.907	0.01
Monetary incentive: round 1	14	1256	1.1	19	1142	1.7	0.328	0.96
Monetary incentive: round 2	40	1256	3.2	33	1142	2.9	0.763	0.09
Status incentive: round 1	2	393	0.5	6	355	1.7	0.225	1.46
Status incentive: round 2	5	394	1.3	6	354	1.7	0.858	0.03
Slider task								
No incentive: round 1	66	164	40.2	73	141	51.8	0.057	3.61
No incentive: round 2	93	164	56.7	91	141	64.5	0.202	1.63
Monetary incentive: round 1	3	418	0.7	7	381	1.8	0.270	1.22
Monetary incentive: round 2	13	418	3.1	14	381	3.7	0.806	0.06
Status incentive: round 1	1	169	0.6	5	163	3.1	0.200	1.64
Status incentive: round 2	1	170	0.6	3	162	1.9	0.581	0.30
AX-CPT task								
No incentive: round 1	65	124	52.4	54	105	51.4	0.987	0.00
No incentive: round 2	87	124	70.2	65	105	61.9	0.239	1.39
Monetary incentive: round 1	5	419	1.2	6	381	1.6	0.874	0.03
Monetary incentive: round 2	16	419	3.8	9	381	2.4	0.328	0.96
Status incentive: round 1	1	112	0.9	1	94	1.1	1.000	0.00
Status incentive: round 2	1	112	0.9	2	95	2.1	0.886	0.02
Simon task								
No incentive: round 1	48	131	36.6	62	135	45.9	0.158	2.00
No incentive: round 2	79	131	60.3	82	135	60.7	1.000	0.00
Monetary incentive: round 1	6	419	1.4	6	380	1.6	1.000	0.00
Monetary incentive: round 2	11	419	2.6	10	380	2.6	1.000	0.00
Status incentive: round 1	112	112	100.0	98	98	100.0	–	–
Status incentive: round 2	3	112	2.7	1	97	1.0	0.718	0.13

Table A3
Regression results for the effect of gender, incentives, and gender-incentive interaction on standardized average reaction time for correct responses and error rate, by task (AX and Simon tasks) (for Fig. 6).

	Dependent variable:			
	Reaction time (std.)		Error rate (std.)	
	AX	Simon	AX	Simon
Gender = boy	−0.39*** (0.07)	−0.64*** (0.07)	0.07 (0.07)	−0.07 (0.07)
Incentive = no incentive (ref. = monetary)	0.39*** (0.05)	0.28*** (0.04)	0.35*** (0.05)	0.18*** (0.04)
Incentive = status (ref. = monetary)	−0.22*** (0.04)	−0.44*** (0.03)	−0.06 (0.04)	−0.13*** (0.03)
Gender = boy × Incentive = no incentive (ref. monetary)	0.02 (0.11)	−0.10 (0.07)	0.13 (0.11)	−0.11 (0.08)
Gender = boy × Incentive = status (ref. monetary)	0.06 (0.07)	0.21*** (0.06)	−0.06 (0.07)	0.03 (0.07)
Round 2	−0.002 (0.02)	−0.04* (0.02)	0.01 (0.03)	0.02 (0.02)
Age (months)	0.01 (0.01)	−0.01* (0.01)	−0.005 (0.01)	0.003 (0.01)
Mouse use	0.002 (0.03)	−0.05 (0.03)	0.03 (0.03)	−0.02 (0.03)
Videogaming	−0.08** (0.03)	−0.05 (0.03)	0.02 (0.03)	−0.001 (0.03)
Fluid intelligence	−0.20*** (0.03)	−0.21*** (0.03)	−0.13*** (0.03)	−0.13*** (0.03)
Constant	−0.54 (0.67)	1.56* (0.66)	0.59 (0.73)	−0.29 (0.76)
Student-level variance	0.49	0.59	0.63	0.77
Class-level variance	0.05	0.00	0.04	0.02
Observation-level variance	0.33	0.22	0.34	0.25
Students	796	794	796	794
Classes	35	35	35	35
Observations	2204	2237	2205	2239

This table shows the results of a three-level hierarchical regression model grouped at the student and class (experimental session) levels, stratified by real-effort task. The dependent variable is the standardized average reaction time for correct responses and error rates per round. *P*-values are calculated with the Kenward-Roger approximation to get approximate degrees of freedom. The categorical variables representing incentive condition and gender are specified as contrasts centered at zero, so the estimate of the effect of being a boy as well as the effect of incentive scheme on reaction time is the average main effect across incentive conditions. Standard errors are shown in parentheses.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table A4
Effect of gender, incentives, and gender-incentive interaction on the AX-CPT proactive behavioral index (PBI) (standardized) and the Simon effect (standardized) for reaction time for correct responses and error rate.

	Dependent variable:			
	Reaction time	Error rate	Reaction time	Error rate
	PBI index (std.)		Simon effect (std.)	
Gender = boy	0.57*** (0.07)	0.18* (0.08)	−0.23** (0.07)	0.02 (0.07)
Incentive = no incentive (ref. = monetary)	0.04 (0.10)	−0.07 (0.12)	−0.17 (0.11)	0.08 (0.10)
Incentive = status (ref. = monetary)	0.18* (0.08)	0.18 (0.10)	−0.11 (0.11)	−0.19 (0.10)
Gender = boy × Incentive = no incentive (ref. monetary)	−0.18 (0.15)	−0.17 (0.17)	0.33* (0.16)	0.18 (0.15)
Gender = boy × Incentive = status (ref. monetary)	0.13 (0.12)	−0.09 (0.14)	0.02 (0.16)	−0.002 (0.15)
Age (months)	−0.01* (0.01)	−0.01* (0.01)	−0.003 (0.01)	0.0002 (0.01)
Mouse use	0.05 (0.03)	0.01 (0.03)	−0.01 (0.03)	0.02 (0.03)
Videogaming	0.04 (0.03)	0.03 (0.03)	0.06* (0.03)	0.02 (0.03)
Fluid intelligence	0.09** (0.03)	0.13*** (0.03)	−0.05 (0.03)	−0.10** (0.03)
Constant	1.17 (0.70)	1.31 (0.70)	0.34 (0.64)	−0.07 (0.69)
Student-level variance	0.32	0.15	0.00	0.20
Class-level variance	0.04	0.01	0.00	0.01
Observation-level variance	0.49	0.79	0.98	0.78
Students	790	795	793	794
Classes	35	35	35	35
Observations	1152	1166	1201	1205

This table shows the results of a three-level hierarchical regression model grouped at the student and class (experimental session) levels, where the dependent variable is the standardized proactive behavioral index (PBI) or the standardized Simon effect for average reaction time on correct responses and error rate.
A positive PBI value indicates that the subject engages in proactive control, as marked by higher AY interference, whereas a negative PBI value indicates that the subject engages in reactive control, as marked by higher BX interference. A positive Simon effect value indicates that the subject is more prone to slower reactions or errors on incongruent trials than the subject of reference. *P*-values are calculated with the Kenward-Roger approximation to get approximate degrees of freedom. The categorical variables representing incentive condition are specified as contrasts centered at zero, so the estimate of the effect of being a boy is the average main effect across incentive conditions. Standard errors are shown in parentheses.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table A5
Regression results for the effect of gender, incentives, and gender-incentive interaction on standardized average reaction time for correct responses on the AX-CPT task, by trial condition (for Fig. 7).

	Dependent variable:			
	Reaction time (std.), B-X condition		Reaction time (std.), A-Y condition	
	(1)	(2)	(3)	(4)
Gender = boy	−0.55*** (0.07)	−0.54*** (0.07)	−0.35*** (0.07)	−0.34*** (0.08)
Incentive = no incentive (ref. = monetary)	0.43*** (0.09)	0.43*** (0.09)	0.62*** (0.09)	0.62*** (0.09)
Incentive = status (ref. = monetary)	−0.27*** (0.07)	−0.28*** (0.07)	−0.28*** (0.07)	−0.28*** (0.07)
Gender = boy × Incentive = no incentive (ref. monetary)	−0.23 (0.12)	−0.21 (0.13)	−0.15 (0.13)	−0.14 (0.13)
Gender = boy × Incentive = status (ref. monetary)	−0.01 (0.10)	−0.003 (0.10)	−0.01 (0.10)	−0.01 (0.10)
Age (months)	0.01 (0.01)	0.01 (0.01)	0.005 (0.01)	0.01 (0.01)
Mouse use	−0.03 (0.03)	−0.03 (0.03)	−0.001 (0.03)	0.005 (0.03)
Videogaming	−0.08** (0.03)	−0.09** (0.03)	−0.08** (0.03)	−0.10*** (0.03)
Fluid intelligence	−0.20*** (0.03)	−0.18*** (0.03)	−0.24*** (0.03)	−0.23*** (0.03)
Need for cognition		−0.10** (0.04)		−0.10* (0.04)
Risk taking		−0.004 (0.01)		−0.003 (0.01)
Delay of gratification		−0.06 (0.07)		−0.01 (0.07)
Conscientiousness		0.02 (0.03)		−0.02 (0.03)
Agreeableness		−0.02 (0.04)		−0.05 (0.04)
Openness		−0.02 (0.03)		−0.02 (0.04)
Neuroticism		−0.06 (0.03)		−0.07* (0.03)
Extraversion		−0.04 (0.03)		−0.03 (0.03)
Constant	−0.80 (0.71)	−0.70 (0.71)	−0.19 (0.72)	−0.17 (0.72)
Student-level variance	0.45	0.44	0.47	0.46
Class-level variance	0.04	0.04	0.02	0.03
Observation-level variance	0.32	0.32	0.35	0.35
Students	791	788	795	791
Classes	35	35	35	35
Observations	1158	1153	1159	1153

This table shows the results of a three-level hierarchical regression model grouped at the student and class (experimental session) levels, where the dependent variable is the average reaction time for correct responses per incentive condition on either proactive trials (B-X) or reactive trials (A-Y), standardized. *P*-values are calculated with the Kenward-Roger approximation to get approximate degrees of freedom. The categorical variables representing incentive condition are specified as contrasts centered at zero, so the estimate of the effect of being a boy on reaction time is the average main effect across incentive conditions. Standard errors are shown in parentheses.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table A6
Regression results for the effect of gender, incentives, and gender-incentive interaction on standardized error rate on the AX-CPT task, by trial condition (for Fig. 7).

	Dependent variable:			
	Error rate (std.), B-X condition		Error rate (std.), A-Y condition	
	(1)	(2)	(3)	(4)
Gender = boy	−0.04 (0.08)	−0.06 (0.08)	0.24** (0.08)	0.18* (0.08)
Incentive = no incentive (ref. = monetary)	0.13 (0.08)	0.13 (0.08)	0.26** (0.10)	0.25* (0.10)
Incentive = status (ref. = monetary)	−0.12 (0.06)	−0.13 (0.06)	0.0003 (0.08)	0.01 (0.08)
Gender = boy × Incentive = no incentive (ref. monetary)	0.18 (0.12)	0.19 (0.12)	0.05 (0.14)	0.07 (0.14)
Gender = boy × Incentive = status (ref. monetary)	0.04 (0.09)	0.04 (0.09)	−0.07 (0.11)	−0.07 (0.11)
Age (months)	−0.01 (0.01)	−0.01* (0.01)	−0.01 (0.01)	−0.01* (0.01)
Mouse use	0.02 (0.03)	0.01 (0.03)	0.01 (0.03)	−0.01 (0.03)
Videogaming	0.02 (0.03)	0.01 (0.03)	0.04 (0.03)	0.04 (0.03)
Fluid intelligence	−0.16*** (0.04)	−0.16*** (0.04)	−0.06 (0.03)	−0.06 (0.03)
Need for cognition		−0.03 (0.04)		−0.004 (0.04)
Risk taking		0.01 (0.01)		0.03** (0.01)
Delay of gratification		−0.05 (0.07)		−0.07 (0.07)
Conscientiousness		−0.04 (0.04)		0.02 (0.03)
Agreeableness		−0.02 (0.04)		−0.15*** (0.04)
Openness		−0.02 (0.04)		0.02 (0.04)
Neuroticism		−0.04 (0.04)		0.01 (0.03)
Extraversion		0.04 (0.03)		0.06 (0.03)
Constant	1.36 (0.78)	1.91* (0.77)	0.89 (0.74)	1.53* (0.73)
Student-level variance	0.65	0.60	0.47	0.43
Class-level variance	0.05	0.06	0.02	0.01
Observation-level variance	0.28	0.28	0.45	0.45
Students	796	792	795	791
Classes	35	35	35	35
Observations	1167	1161	1166	1160

This table shows the results of a three-level hierarchical regression model grouped at the student and class (experimental session) levels, where the dependent variable is the error rate per incentive condition on either proactive trials (B-X) or reactive trials (A-Y). *P*-values are calculated with the Kenward-Roger approximation to get approximate degrees of freedom. The categorical variables representing incentive condition are specified as contrasts centered at zero, so the estimate of the effect of

being a boy on reaction time is the average main effect across incentive conditions. Standard errors are shown in parentheses.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Appendix B

Table B1
Personality measures.[a]

Dimension	Measure	Items
Need for cognition	Sample standardization of the average of the scores given by each subject on the items, each item measured on a 5-point Likert agreement scale.	1. I like exercises that make me think a lot. 2. I like challenges that I need to think about. 3. I prefer to think the least possible. 4. I just need to know the answer, I don't need to know the reasons.
Risk taking	Sample-standardized score given on a scale from 0 to 10, with 0 indicating that the subject is not willing to take risks and 10 indicating that he or she is very willing to take risks.	1. In general, are you willing to take risks, that means, are you willing to do something that can go well or not?
Delay of gratification	Binary indicator with 0 if the subject answered, "I would prefer receiving one gift today." or 1 if he or she answered, "I would prefer receiving two gifts next week."	1. Imagine someone wants to give you a gift. Would you prefer receiving one gift today or two next week?
Conscientiousness	Sample standardization of the average of the scores given by each subject on the items, each item measured on a 5-point Likert agreement scale.	1. I do my housework willingly. 2. My room is orderly. 3. When I get money from someone, I save it.
Agreeableness	Sample standardization of the average of the scores given by each subject on the items, each item measured on a 5-point Likert agreement scale.	1. When someone in my class needs something, I notice it. 2. When I'm able to help somebody, I do. 3. When I have a new toy, I lend it to others.
Openness	Sample standardization of the average of the scores given by each subject on the items, each item measured on a 5-point Likert agreement scale.	1. When birds are flying, I notice them. 2. When I go on a trip, I like to discover something new (versus relax). 3. I like to learn about new and difficult things.
Neuroticism	Sample standardization of the average of the scores given by each subject on the items, each item measured on a 5-point Likert agreement scale.	1. I go to school worried (versus calm). 2. When something does not work out, I get nervous. 3. I am usually worried.
Extraversion	Sample standardization of the average of the scores given by each subject on the items, each item measured on a 5-point Likert agreement scale.	1. I play with friends (versus on my own). 2. When my friends are playing, I play with them too. 3. When someone jokes, I laugh with my friends (versus I rarely see anything funny about it).

[a] Survey items used to measure the personality dimensions were adapted and translated into Spanish from existing surveys that are cited in Section 3.3. "Measures".

References

Angrist, J., Lang, D., & Oreopoulos, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics, 1*(1), 136–163. https://doi.org/10.1257/app.1.1.136

Apascaritei, P., Demel, S., & Radl, J. (2021). The difference between saying and doing: comparing subjective and objective measures of effort among fifth graders. *American Behavioral Scientist, 65*(11), 1457–1479. https://doi.org/10.1177/0002764221996772

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Beißert, H., Köhler, M., Rempel, M., & Beierlein, C. (2014). *Eine deutschsprachige Kurzskala zur Messung des konstrukts Need for cognition: Die Need for cognition kurzskala (NfC-K)*.

Blossfeld, H. P., Von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: Need, main features, and research potential. *Zeitschrift für Erziehungswissenschaft, 14*, 5–17. https://doi.org/10.1007/s11618-011-0178-3.

Boutyline, A., Arseniev-Koehler, A., & Cornell, D. J. (2023). School, studying, and smarts: Gender stereotypes and education across 80 years of American print media, 1930–2009†. *Social Forces, 102*(1), 263–286. https://doi.org/10.1093/sf/soac148.

Brandts, J., Gërxhani, K., & Schram, A. (2020). Are there gender differences in status-ranking aversion? *Journal of Behavioral and Experimental Economics, 84*, Article 101485. https://doi.org/10.1016/j.socec.2019.101485

Braver, T. S., Paxton, J. L., Locke, H. S., & Barch, D. M. (2009). Flexible neural mechanisms of cognitive control within human prefrontal cortex. *Proceedings of the National Academy of Sciences, 106*(18), 7351–7356. https://doi.org/10.1073/pnas.0808187106

Buchmann, C., DiPrete, T. A., & McDaniel, A. (2008). Gender inequalities in education. *Annual Review of Sociology, 34*, 319–337. https://doi.org/10.1146/annurev.soc.34.040507.134719

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

Butler, R. (2014). Motivation in educational contexts: Does gender matter? *Advances in Child Development and Behavior, 47*, 1–41. https://doi.org/10.1016/bs.acdb.2014.05.001

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin, 125*(3), 367. https://doi.org/10.1037/0033-2909.125.3.367.

Cassar, A., & Rigdon, M. L. (2021). Prosocial option increases women's entry into competition. *Proceedings of the National Academy of Sciences, 118*(45), Article e2111943118. https://doi.org/10.1073/pnas.2111943118.

Chouinard, R., & Roy, N. (2008). Changes in high-school students' competence beliefs, utility value and achievement goals in mathematics. *British Journal of Educational Psychology, 78*(1), 31–50. https://doi.org/10.1348/000709907x197993.

Darling-Hammond, L. (2007). Race, inequality and educational accountability: The irony of 'No Child Left Behind. *Race Ethnicity and Education, 10*(3), 245–260. https://doi.org/10.1080/13613320701503207.

DeAngelo, L., Franke, R., Hurtado, S., Pryor, J. H., & Tran, S. (2011). *Completing college: Assessing graduation rates at four-year institutions*. Los Angeles, CA: Higher Education Research Institute, UCLA.

DeMars, C., Bashkov, B., & Socha, A. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment, 8*, 69–82.

Diaconu-Gherasim, L. R., Tepordei, A.-M., Mairean, C., & Rusu, A. (2019). Intelligence beliefs, goal orientations and children's academic achievement: Does the children's gender matter? *Educational Studies, 45*(1), 95–112. https://doi.org/10.1080/03055698.2018.1443796.

Diseth, Å., & Samdal, O. (2014). Autonomy support and achievement goals as predictors of perceived school performance and life satisfaction in the transition between lower and upper secondary school. *Social Psychology of Education, 17*, 269–291. https://doi.org/10.1007/s11218-013-9244-4

Dreber, A., Von Essen, E., & Ranehill, E. (2014). Gender and competition in adolescence: Task matters. *Experimental Economics, 17*, 154–172. https://doi.org/10.1007/s10683-013-9361-0.

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality, 45*(3), 259–268. https://doi.org/10.1016/j.jrp.2011.02.004.

Duckworth, A. L., & Seligman, M. E. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology, 98*(1), 198. https://doi.org/10.1037/0022-0663.98.1.198.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), 237–251. https://doi.org/10.3102/0013189X15584327.

Beutel, A. M., & Marini, M. M. (1995). Gender and values. *American Sociological Review*, 436–448. https://doi.org/10.2307/2096423

Wolters, C. A., & Benzon, M. B. (2013). Assessing and predicting college students' use of strategies for the self-regulation of motivation. *The Journal of Experimental Education, 81*(2), 199–221. https://doi.org/10.1080/00220973.2012.699901.

Dutcher, G., Salmon, T., & Saral, K.J. (2015). Is 'real' effort more real? Available at SSRN 2701793. https://doi.org/10.2139/ssrn.2701793.

Bates, D., Maechler, M., Bolker, B., & Walker, S.. lme4: Linear mixed-effects models using Eigen and S4. https://doi.org/10.32614/CRAN.package.lme4.

Eccles, J. S. (2005). Subjective task value and the Eccles et al. model of achievement-related choices. Handbook of Competence and Motivation, 105, 121.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*(2), 256. https://doi.org/10.1037/0033-295X.95.2.256.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics, 133*(4), 1645–1692. https://doi.org/10.1093/qje/qjy013.

Falk, A., Becker, A., Dohmen, T., Huffman, D., & Sunde, U. (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science, 69*(4), 1935–1950. https://doi.org/10.1287/mnsc.2022.4455.

French, S., Dickerson, A., & Mulder, R. A. (2023). A review of the benefits and drawbacks of high-stakes final examinations in higher education. *Higher Education*, 1–26. https://doi.org/10.1007/s10734-023-01148-z.

Frömer, R., Lin, H., Dean Wolf, C., Inzlicht, M., & Shenhav, A. (2021). Expectations of reward and efficacy guide cognitive control allocation. *Nature Communications, 12*(1), 1030. https://doi.org/10.1038/s41467-021-21315-z.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights, 1*(3), 291–308. https://doi.org/10.1257/aeri.20180633.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives, 25*(4), 191–210. https://doi.org/10.1257/jep.25.4.191.

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics, 118*(3), 1049–1074. https://doi.org/10.1162/00335530360698496.

Gneezy, U., & Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review, 94*(2), 377–381. https://doi.org/10.1257/0002828041301821.

Grant, C. A. (2004). Oppression, privilege, and high-stakes testing. *Multicultural Perspectives, 6*(1), 3–11. https://doi.org/10.1207/S15327892mcp0601_2.

Grissom, N. M., & Reyes, T. M. (2019). Let's call the whole thing off: Evaluating gender and sex differences in executive function. *Neuropsychopharmacology, 44*(1), 86–96. https://doi.org/10.1038/s41386-018-0179-5.

Heckman, J. J., & Snyder, J. M., Jr. (1996). *Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators* (Vol. 5785). National Bureau of Economic Research Working Paper Series. https://doi.org/10.3386/w5785

Heckman, J. J., Jagelka, T., & Kautz, T. (2021). Some contributions of economics to the study of personality. In *Handbook of personality: Theory and research* (4th ed., pp. 853–892). The Guilford Press.

Heyder, A., & Kessels, U. (2017). Boys don't work? On the psychological benefits of showing low effort in high school. *Sex Roles, 77*, 72–85. https://doi.org/10.1007/s11199-016-0683-1.

Hirnstein, M., Coloma Andrews, L., & Hausmann, M. (2014). Gender-stereotyping and cognitive sex differences in mixed-and same-sex groups. *Archives of Sexual Behavior, 43*, 1663–1673. https://doi.org/10.1007/s10508-014-0311-5.

Hirt, E. R., & McCrea, S. M. (2009). Man smart, woman smarter? Getting to the root of gender differences in self-handicapping. *Social and Personality Psychology Compass, 3*(3), 260–274. https://doi.org/10.1111/j.1751-9004.2009.00176.x.

Horn, D., Kiss, H. J., & Lénárd, T. (2022). Preferences of adolescents–a dataset containing linked experimental task measures and register data. *Data in Brief, 42*, 108088. https://doi.org/10.1016/j.dib.2022.108088.

Inzlicht, M., Shenhav, A., & Olivola, C. Y. (2018). The effort paradox: Effort is both costly and valued. *Trends in Cognitive Sciences, 22*(4), 337–349. https://doi.org/10.1016/j.tics.2018.01.007.

Jackson, C. (2003). Motives for 'laddishness' at school: Fear of failure and fear of the 'feminine. *British Educational Research Journal, 29*(4), 583–598. https://doi.org/10.1080/01411920301847.

James, H. S., Jr. (2005). Why did you do that? An economic examination of the effect of extrinsic compensation on intrinsic motivation and performance. *Journal of Economic Psychology, 26*(4), 549–566. https://doi.org/10.1016/j.joep.2004.11.002.

Jones, S., & Myhill, D. (2004). Troublesome boys' and 'compliant girls': Gender identity and perceptions of achievement and underachievement. *British Journal of Sociology of Education, 25*(5), 547–561. https://doi.org/10.1080/0142569042000252044.

Kesebir, S., Lee, S. Y., Elliot, A. J., & Pillutla, M. M. (2019). Lay beliefs about competition: Scale development and gender differences. *Motivation and Emotion, 43*, 719–739. https://doi.org/10.1007/s11031-019-09779-5.

Khachatryan, K., Dreber, A., Von Essen, E., & Ranehill, E. (2015). Gender and preferences at a young age: Evidence from Armenia. *Journal of Economic Behavior & Organization, 118*, 318–332. https://doi.org/10.1016/j.jebo.2015.02.021.

King, R. B. (2016). Is a performance-avoidance achievement goal always maladaptive? Not necessarily for collectivists. *Personality and Individual Differences, 99*, 190–195. https://doi.org/10.1016/j.paid.2016.04.093.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences, 36*(6), 661–679. https://doi.org/10.1017/S0140525X12003196.

Legewie, J., & DiPrete, T. A. (2012). School context and the gender gap in educational achievement. *American Sociological Review, 77*(3), 463–485. https://doi.org/10.1177/0003122412440802.

Levitt, S. D., List, J. A., Neckermann, S., & Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy, 8*(4), 183–219. https://doi.org/10.1257/pol.20130358.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*(4), 1494–1502. https://doi.org/10.1016/j.actpsy.2019.102891.

Maćkiewicz, M., & Cieciuch, J. (2016). Pictorial personality traits questionnaire for children (PPTQ-C)—A new measure of children's personality traits. *Frontiers in Psychology, 7*, 498. https://doi.org/10.3389/fpsyg.2016.00498.

Mäki-Marttunen, V., Hagen, T., & Espeseth, T. (2019). Proactive and reactive modes of cognitive control can operate independently and simultaneously. *Acta Psychologica, 199*, 102891. https://doi.org/10.1016/j.actpsy.2019.102891.

Mañas Antón, O. (2019). Datos y cifras de la educación 2019-2020. Dirección General de Bilingüismo y Calidad de la Enseñanza. In *Consejería de Educación y Juventud*. https://www.madrid.org/bvirtual/BVCM050013.pdf.

Masclet, D., Peterle, E., & Larribeau, S. (2015). Gender differences in tournament and flat-wage schemes: An experimental study. *Journal of Economic Psychology, 47*, 103–115. https://doi.org/10.1016/j.joep.2015.01.003.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*, 314–324. https://doi.org/10.3758/s13428-011-0168-7.

Meece, J. L., Glienke, B. B., & Burg, S. (2006). Gender and motivation. *Journal of School Psychology, 44*(5), 351–373. https://doi.org/10.1016/j.jsp.2006.04.004

Ministerio de Educación y Formación Profesional. (2021). *Las cifras de la educación en España: Curso 2019-2020. Madrid, Spain*.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science, 21*(1), 8–14. https://doi.org/10.1177/0963721411429458.

Neuenschwander, R., Cimeli, P., Röthlisberger, M., & Roebers, C. M. (2013). Personality factors in elementary school children: Contributions to academic performance over and above executive functions? *Learning and Individual Differences, 25*, 118–125. https://doi.org/10.1016/j.lindif.2012.12.006.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics, 122*(3), 1067–1101. https://doi.org/10.1162/qjec.122.3.1067.

Niederle, M., & Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives, 24*(2), 129–144. https://doi.org/10.1257/jep.24.2.129.

Niederle, M., & Vesterlund, L. (2011). Gender and competition. *Annu. Rev. Econ., 3*(1), 601–630. https://doi.org/10.1146/annurev-economics-111809-125122.

O'Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*(3), 185. https://doi.org/10.1207/s15326977ea1003_3.

Ors, E., Palomino, F., & Peyrache, E. (2013). Performance gender gap: Does competition matter? *Journal of Labor Economics, 31*(3), 443–499. https://doi.org/10.1086/669331.

Otto, A. R., & Daw, N. D. (2019). The opportunity cost of time modulates cognitive effort. *Neuropsychologia, 123*, 92–105. https://doi.org/10.1016/j.neuropsychologia.2018.05.006.

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Radl, J., & Miller, L. (2021). Conceptual and methodological considerations on effort: An interdisciplinary approach. *American Behavioral Scientist, 65*(11), 1447–1456. https://doi.org/10.1177/0002764221996792

Radl, J., Apascaritei, P., Foley, W., Kröger, L., Lorente, P., Palacios-Abad, A., Solga, H., Stuhler, J., & Swarr, M. (2024). How socioeconomic status shapes cognitive effort: A laboratory study among fifth graders. https://hdl.handle.net/10016/43750.

Ratelle, C. F., Guay, F., Vallerand, R. J., Larose, S., & Senécal, C. (2007). Autonomous, controlled, and amotivated types of academic motivation: A person-oriented analysis. *Journal of Educational Psychology, 99*(4), 734. https://doi.org/10.1037/0022-0663.99.4.734.

Raven, J. C., & Court, J. H. (1998). *Raven's progressive matrices and vocabulary scales*. Oxford: Oxford Psychologists Press.

Rios, J. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education, 34*(2), 85–106. https://doi.org/10.1080/08957347.2021.1890741.

Roivainen, E. (2011). Gender differences in processing speed: A review of recent research. *Learning and Individual Differences, 21*(2), 145–149. https://doi.org/10.1016/j.lindif.2010.11.021.

Ryan, R. M., & Deci, E. L. (2000a). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*(1), 54–67.

Ryan, R. M., & Deci, E. L. (2000b). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68.

Schlosser, A., Neeman, Z., & Attali, Y. (2019). Differential performance in high versus low stakes tests: Evidence from the GRE test. *The Economic Journal, 129*(623), 2916–2948. https://doi.org/10.1093/ej/uez015.

Schram, A., Brandts, J., & Gërxhani, K. (2019). Social-status ranking: A hidden channel to gender inequality under competition. *Experimental Economics, 22*, 396–418. https://doi.org/10.1007/s10683-018-9563-6.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science, 58*(8), 1438–1457. https://doi.org/10.1287/mnsc.1110.1509.

Simzar, R. M., Martinez, M., Rutherford, T., Domina, T., & Conley, A. M. (2015). Raising the stakes: How students' motivation for mathematics associates with high- and low-stakes test achievement. *Learning and Individual Differences, 39*, 49–63. https://doi.org/10.1016/j.lindif.2015.03.002

Sittenthaler, H. M., & Mohnen, A. (2020). Cash, non-cash, or mix? Gender matters! The impact of monetary, non-monetary, and mixed incentives on performance. *Journal of Business Economics, 90*(8), 1253–1284. https://doi.org/10.1007/s11573-020-00992-0.

StataCorp.. (2023). *Stata statistical software: Release 18*. College Station, TX: StataCorp LLC.

Sutter, M., Glätzle-Rützler, D., Balafoutas, L., & Czermak, S. (2016). Cancelling out early age gender differences in competition: An analysis of policy interventions. *Experimental Economics, 19*, 412–432. https://doi.org/10.1007/s10683-015-9447-y.

Sutter, M., Zoller, C., & Glätzle-Rützler, D. (2019). Economic behavior of children and adolescents–a first survey of experimental economics results. *European Economic Review, 111*, 98–121. https://doi.org/10.1016/j.euroecorev.2018.09.004.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2009). Skipping the test: Using empirical evidence to inform policy related to students who avoid taking low-stakes assessments in college. *The Journal of General Education, 58*(3), 167–195. https://doi.org/10.1353/jge.0.0043.

Tang, C., & Zhao, L. (2023). Gender social norms and gender gap in math: Evidence and mechanisms. *Applied Economics*, 1–19. https://doi.org/10.1080/00036846.2023.2178631.

Thoman, D. B., Smith, J. L., & Silvia, P. J. (2011). The resource replenishment function of interest. *Social Psychological and Personality Science, 2*(6), 592–599. https://doi.org/10.1177/1948550611402025.

Thurston, D., Penner, A. M., & Penner, E. K. (2016). Membership has its privileges": Status incentives and categorical inequality in education. *Sociological Science, 3*, 264–295. https://doi.org/10.15195/v3.a13.

Vecchione, M., Alessandri, G., & Marsicano, G. (2014). Academic motivation predicts educational attainment: Does gender make a difference? *Learning and Individual Differences, 32*, 124–131. https://doi.org/10.1016/j.lindif.2014.01.003.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140*(4), 1174. https://doi.org/10.1037/a0036620.

Watson, T. L., & Blanchard-Fields, F. (1998). Thinking with your head and your heart: Age differences in everyday problem-solving strategy preferences. *Aging, Neuropsychology, and Cognition, 5*(3), 225–240. https://doi.org/10.1076/anec.5.3.225.613.

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience, 15*, 395–415. https://doi.org/10.3758/s13415-015-0334-y.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1.