Material incentives drive 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

*Corresponding author:

Jonas Radl
Address: Universidad Carlos III de Madrid, Calle Madrid 135, 28903 Getafe, Spain
Phone number: +34 916249675
Email: jonas.radl@uc3m.es

Author Contributions: Authors contributed equally to this work. **Paula Apascaritei:** Investigation, Formal analysis, Writing – Original Draft. **Jonas Radl:** Conceptualization, Investigation, Writing – Original Draft, Review & Editing. **Madeline Swarr:** Formal analysis, Visualization, Writing – Original Draft, Review & Editing.

Declarations of interest: None

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.

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)

Material incentives drive gender differences in cognitive effort among children.

Abstract

Academic performance relies on effort and varies by gender. However, it is not clear at what age nor under what circumstances gender differences in effort arise. Using behavioral real-effort measures from 806 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. While boys achieve superior performance through more proactive control and faster reaction speed, we find no gender differences in overall accuracy. Girls' preferences for a more prudent approach pay off only when reactive control is elicited. These findings are robust to controlling for key personality traits and cognitive ability (fluid intelligence). The results have important implications for understanding gender divides in education and learning.

Keywords: Cognitive effort, gender, incentives, children, laboratory experiments

Educational Relevance and Implications Statement

Effort is essential for young students as it boosts learning and achievement. Educators and parents often use incentives 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' effort edge. To attenuate gender differences in performance-based achievement, girls may benefit from incurring calculated risks, knowing that they are well-equipped to compete with boys.

MAIN TEXT

1. INTRODUCTION

Much concern has been raised surrounding the ability of low-stakes academic examinations to accurately represent learning achievement. Many studies have argued that when the stakes of a test are not sufficiently motivating, low performance scores may actually be measuring lower effort rather than lower ability (see Wise & DeMars (Wise & DeMars, 2005) for a review). Therefore, what is observed as under-achievement may reflect lower levels of test-taking motivation. In order to boost effort through motivation and thus improve the construct validity of test scores, researchers have suggested offering external incentives to test-takers, such as performance-based monetary pay or academic awards and certificates (Gneezy et al., 2019; Levitt et al., 2016). One concern, however, is the differential impact that these rewards may have on motivation depending on heterogenous preferences.

In many Western industrialized countries, girls have generally shown greater performance on indicators of learning attainment in relation to their male counterparts (Buchmann et al., 2008; DeAngelo et al., 2011; Voyer & Voyer, 2014). Gender differences in test-taking motivation have been proposed to explain this gap, supported by the consistent finding that boys score significantly lower on self-reported measures of academic effort than girls (DeMars et al., 2013). However, when the stakes of the exams involve competing for accolades or limited admission placements, some studies have shown a shrinking or even reversal of the gender gap favoring boys and men (Ors et al., 2013; Schlosser et al., 2019). Similarly, there is robust evidence in the behavioral economics literature according to which females are less sensitive and perhaps even averse to competition (Niederle & Vesterlund, 2007), while males thrive on it (Gneezy et al., 2003).

The study of gender differences in cognitive effort must confront the complex methodological challenge of neutralizing the influence of differential abilities and preferences. The type of task is crucial. Indeed, the most prominent effort advantages for girls are seen in subjects such as reading and language, for which girls tend to exhibit greater competence (Roivainen, 2011). In experimental studies, similarly, girls have been shown to exhibit greater effort when female-typed tasks such as rope-skipping were used (Khachatryan et al., 2015), whereas boys displayed superior effort when studies employed male-typed tasks such as solving mazes (Gneezy et al., 2003). Not only may the difficulty of such tasks vary by gender, they may also have varying appeal for girls and boys. Indeed, there is evidence that effort investments are more strongly determined by self-perceived affect than by self-perceived competence (Akhtar & Firdiyanti, 2023; Arens & Hasselhorn, 2015). Thus, our ability to understand the independent effect of incentives on effort by gender is limited when effortful tasks considerably favor the abilities or interests of either boys or girls.

In this study, we draw on a balanced sample of 806 fifth-grade students and use an experimental setting to understand how gendered reactions to different types of incentive schemes shape cognitive effort investments among children. We employ three different non-gender-typed real-effort tasks of low difficulty that tap into different executive functions. The evidence drawn from our study will help inform policy and educational design about effort-boosting incentives and illuminate their potential side-effect of driving open gender gaps in achievement.

1.1 Gender differences in effort and motivation

Cognitive effort is generally aversive because it draws on limited resources and binds attention (Inzlicht et al., 2018). Therefore, individuals need motivation to cover the costs that come with engaging in effortful tasks. Faced with learning tasks, children have to make decisions about

how to expend effort in function of their intrinsic and extrinsic motivation. Intrinsic motivation refers to the propensity of pursuing an activity for its own sake, and might reflect inherent intellectual curiosity, sense of purpose, need for cognition, or alignment with self-concept that comes along with a particular behavior (Inzlicht et al., 2018; Ryan & Deci, 2000). Research has shown that individuals can allocate more cognitive effort to a task while feeling less depleted based on their intrinsic motivation (Segal, 2012; Thoman et al., 2011). Extrinsic motivation, in contrast, operates through behaviors that are performed as a means to an end, and not for their own sake (Ryan & Deci, 2000). The rewards by which extrinsic motivation is elicited are often external, i.e. prizes given or symbolic recognition awarded contingent upon a particular behavior, and have been proven to increase performance in a variety of settings (see Rios (2021) for a meta-analysis).

Most studies show that girls report higher levels of intrinsic motivation than boys in school (Ratelle et al., 2007; Vecchione et al., 2014). Gender norms cultivate differences in academic culture and motivation that may lead to gender gaps in effort (Boutyline et al., 2023; Butler, 2014; Jones & Myhill, 2004; Legewie & DiPrete, 2012). While boys may be socially rewarded for their “effortless” talent often associated with inherent ability or intelligence, girls are praised for their hard work as the reason for their success (Heyder & Kessels, 2017). 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). Girls, on the other hand, tend to value effort more, both in themselves and others (Hirt & McCrea, 2009). They also make fewer excuses for under-performance because they perceive higher social consequences for failure attributed to lack of effort rather than lack of ability, as evidenced by psychological research on self-handicapping (McCrea et al., 2008). 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 indicate a disposition towards effort exertion (Neuenschwander et al., 2013). Given the evidence that support the notion that girls are more intrinsically motivated than boys and more inclined to value effort in and of itself, we formulate our first hypothesis:

H1: Girls exert more effort than boys in the absence of external rewards.

Standard economic theory suggests that incentivizing with money will boost performance as it compensates for the cost of effort. To what extent and for whom financial incentives motivate, though, depends on *how* and *for what* performance is rewarded. Masclet et al. (2015) find that women do better when reward is not contingent on task performance, e.g. in a flat-wage scheme. In line with this finding, 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 et al., 2014; Niederle & Vesterlund, 2010; Sutter et al., 2016). Analyzing within-individual behavior changes when incentive schemes are experimentally modulated, Levitt et al. (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. Therefore, we posit the following hypotheses:

H2: Boys increase their effort more than girls do in response to the introduction of performance-based monetary incentives.

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), a preference that is associated with differential beliefs about the benefits of competing, such as that it enhances performance, builds character, and leads to innovative solutions (Kesebir et al., 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 et al., 2020). In terms of performance under competition, studies report that girls tend to do worse than boys in contest settings where only top performers are rewarded (Gneezy et al., 2003; Horn et al., 2022; Schram et al., 2019; Sutter et al., 2019). Given what has been shown to be boys' greater preference for status and competition, we hypothesize the following:

> H3: Boys increase their effort more so than girls when competing for an additional status incentive placed on top of performance-based monetary incentives.

1.2 Difficulties of measuring cognitive 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 and incentives. 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). Although performing cognitive control tasks requires effort, performance-dependent incentives are not consistently used in (neuro)psychology. Therefore, the effort observed in these studies could be a lower bound of effort exertion, reflecting intrinsic motivation, rather than peak performance given tangible rewards. Economists, on the other hand, have traditionally assumed that effort is only exerted in the presence of incentives that outweigh the cost to effort. To explicitly induce preferences in experiments, "chosen effort" designs incorporate the role of incentives by asking participants to

allocate effort to gain monetary rewards (Fehr et al., 1993; Nalbantian & Schotter, 1997). However, these designs do not require actual mental or physical exertion from the subject, nor do they capture intrinsic utility gained from the process of working itself (Van Dijk et al., 2001).

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 et al., 2015). Neuroscientific research has increasingly implemented real-effort tasks under differing incentive conditions to evoke more accurately the motivation-effort relationship present in work or school settings (Buser et al., 2014; Frömer et al., 2021). However, according to expectancy-value theory, effort may also be function of one’s beliefs about their ability to succeed (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000). Therefore, differences in self-efficacy make it plausible that those who possess more ability towards a demanding task but perhaps are less motivated by its rewards will exhibit similar effort levels as someone with less ability but more motivation. Certain tasks, e.g. solving mazes (Gneezy et al., 2003) or word-search tasks (Dreber et al., 2014), may favor the competencies or interests of one gender over those of the other. While the intentional use of gender-typed tasks is useful for recreating certain real-world contexts (Buser et al., 2014; Khachatryan et al., 2015) and for understanding how related social self-perceptions affect effort investments (Dreber et al., 2014), they are unable to successfully unconfound effort from ability. Many studies to-date have failed to robustly examine differences in the effect of incentives by gender using tasks that are not gender-typed.

2. METHOD & MATERIALS

To overcome the difficulties of measuring effort, we adopt a novel approach comparing performance on three distinct real-effort tasks across three different incentive conditions, while holding ability (i.e. fluid intelligence) constant. Additionally, we test the mediating effects of key

non-cognitive skills (i.e. personality traits). Our multidimensional measure of cognitive effort has strong claims to validity, drawing on three distinct real-effort tasks that target different dimensions of cognitive effort: 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. The tasks require minimal skill but engage multiple executive functions, while avoiding gender-typing. More details on the real-effort tasks are described in Section 2.2.

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, 806 fifth grade students (mean age in months = 126.2; SD = 6.2) from 35 classes representing 19 schools visited the campus between October 2019 and March 2022. Each child was classified by gender as either a boy or girl as indicated on their self-reported surveys (420 girls, 382 boys, 4 unreported). All students participated only if they had their parents' written informed consent and signed data protection agreement, in accordance with stipulations of the ethics board and data protection officer at the university.

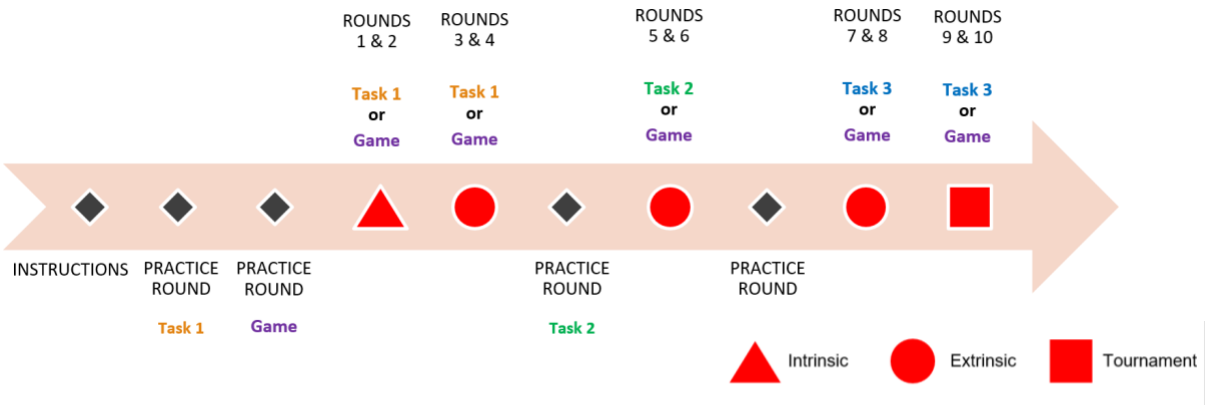
All laboratory sessions took place in the experimental economics laboratory on campus that is equipped with standard desktop computers and cubicles. In each session, students performed one task in the absence of incentives. Then, all three tasks were performed under a piece-rate scheme where points for correct answers were "cashed in" at the end of the session for prizes selected from "the market", 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 the well-known “crowding-out” effects on intrinsic motivation that occurs once extrinsic rewards for task performance are removed (Gneezy et al., 2011; James Jr, 2005). Tasks, however, were allocated to incentive conditions counterbalancing randomly by experimental session.

The structure of the experiment is detailed in Figure 1. The first task assigned was performed for two rounds of the no-incentive condition, then two rounds of the monetary incentive condition. Next, the second task was performed for two rounds of the monetary incentive condition. Finally, the third task was performed for two rounds of the monetary incentive condition and two rounds of the status incentive condition. This resulted in a total of 10 observations per student (10 “rounds” per student, each round lasted 2 minutes with a short break in between). Before starting each round, children had a choice between completing the task or playing a computer-based leisure activity (either a puzzle or a ball game) with the understanding that their scores would be recorded as zeros for any rounds that the task was not performed (more details provided in Section 3.3.1).

After the experimental sessions, children were given a break to have a snack, relax, and take a guided campus tour. They then returned to the laboratory and answered a survey with questions on their personality, opinions, and expectations. After the survey sessions, they had lunch followed by an educative game on brain functions. Before leaving the campus, children received their chosen toys, a medal, and the tournament winners received their diploma. In total, the campus visit lasted about 6 hours. All interactions happened in Spanish.

Figure 1. Experimental set-up.



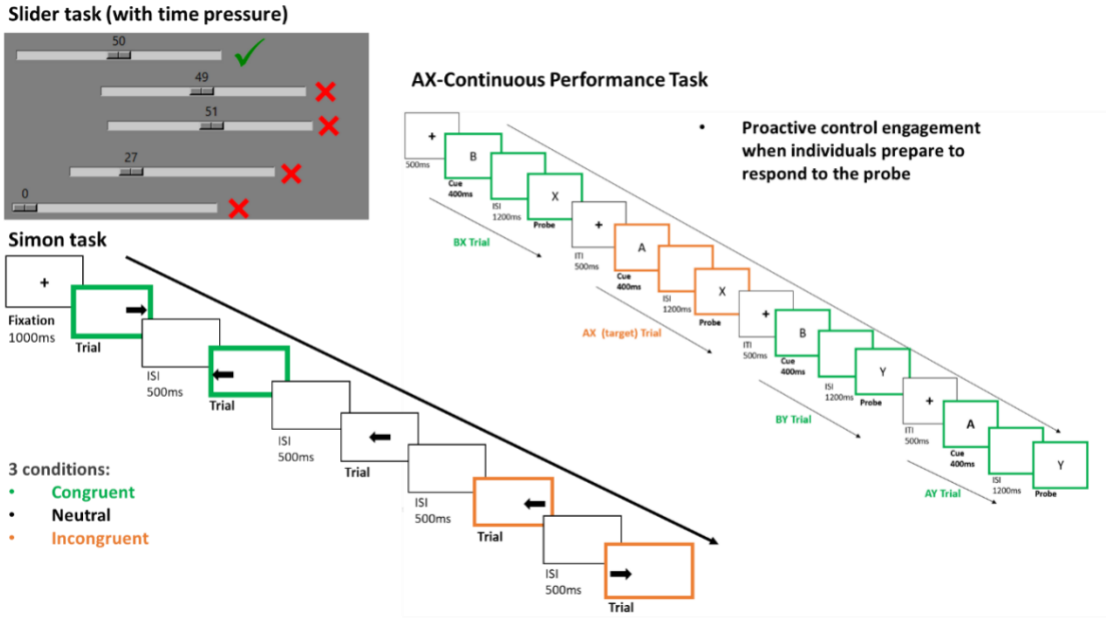
The diagram shows design and structure of the experiment using the three real-effort tasks – the slider task, AX-CPT task, and the Simon task. Which task came first, second, and third varied across sessions to avoid order effects. Before each round the student could choose to do the task or play the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

2.2 The real-effort tasks

Each task engages a different component of executive function (EF), which refers to the set of 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. The slider task primarily covers the information processing and updating subdomain. In this task, the participants are presented with 48 horizontal lines. There is a dial on each line and the participant must click and drag the dial so it is exactly at the midpoint, which corresponds to 50 on a scale from 0 to 100. Participants could gauge the midpoint, as the current position of the

slider was indicated with a number. 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 and control 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 omitted 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. The layouts of each task are illustrated in Figure 2.

Figure 2. Illustration of the real-effort tasks.



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. All tasks as well as the survey were programmed in OpenSesame (Mathôt et al., 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 et al., 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 based on task characteristics considering the non-familiar and rather mundane nature of the tasks. Finally, our claim that these tasks are not

gender-typed is supported by the fact that there are no significant gender differences in the self-reported measures of task effort engagement or likeability (see Table A1).

2.3 Measures

2.3.1 *Real effort*

Real effort is estimated as the number of correct responses per two-minute round, standardized within the distribution of all round scores from the same task, and always controlled for ability (fluid intelligence).

2.3.2 Fluid intelligence

Fluid intelligence is measured via an adapted version of the Raven's progressive matrices test, and the total number of correct answers given by a subject are standardized within the distribution of all other scores given in the Madrid sample.

2.3.3 Personality measures

The personality scales used in the study measure need for cognition, risk preferences, and delay of gratification, as well as the Big Five traits -- conscientiousness, agreeableness, openness, neuroticism, and extraversion. These were selected on theoretical grounds in order to capture potential heterogeneity that may be mediating the relationship between gender and effort. Need for cognition, conscientiousness, agreeableness, openness, neuroticism, and extraversion are all measured as the sample standardization of the average of the scores given by each subject on a series of items measuring each personality dimension, with each individual item measured on a 5-point Likert agreement scale. Risk preferences are measured according to the sample-standardized score given on a 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.

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. See Table B1 of Appendix B for a complete description of each personality dimension and their corresponding items/measures.

2.3.4 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

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). Scores were assigned according to the number of correct answers given per 2-minute round, with scores recorded as zeros for any rounds for which the task was not performed.

Given the non-normal distribution of task performance scores, two-sided Wilcoxon rank sum tests were used to compare the distributions of boys and girls within specific task-incentive conditions (Figure 3).

All regression models in the main analyses are two-level random intercept hierarchical models grouped at the subject level to account for random variation in baseline real effort, decisions to task, and reaction time/error rate between individual students (Figures 4-8, Tables 1-2, Tables A3-S4, Tables S7-S10). Unless otherwise mentioned, all variables that are interacted with

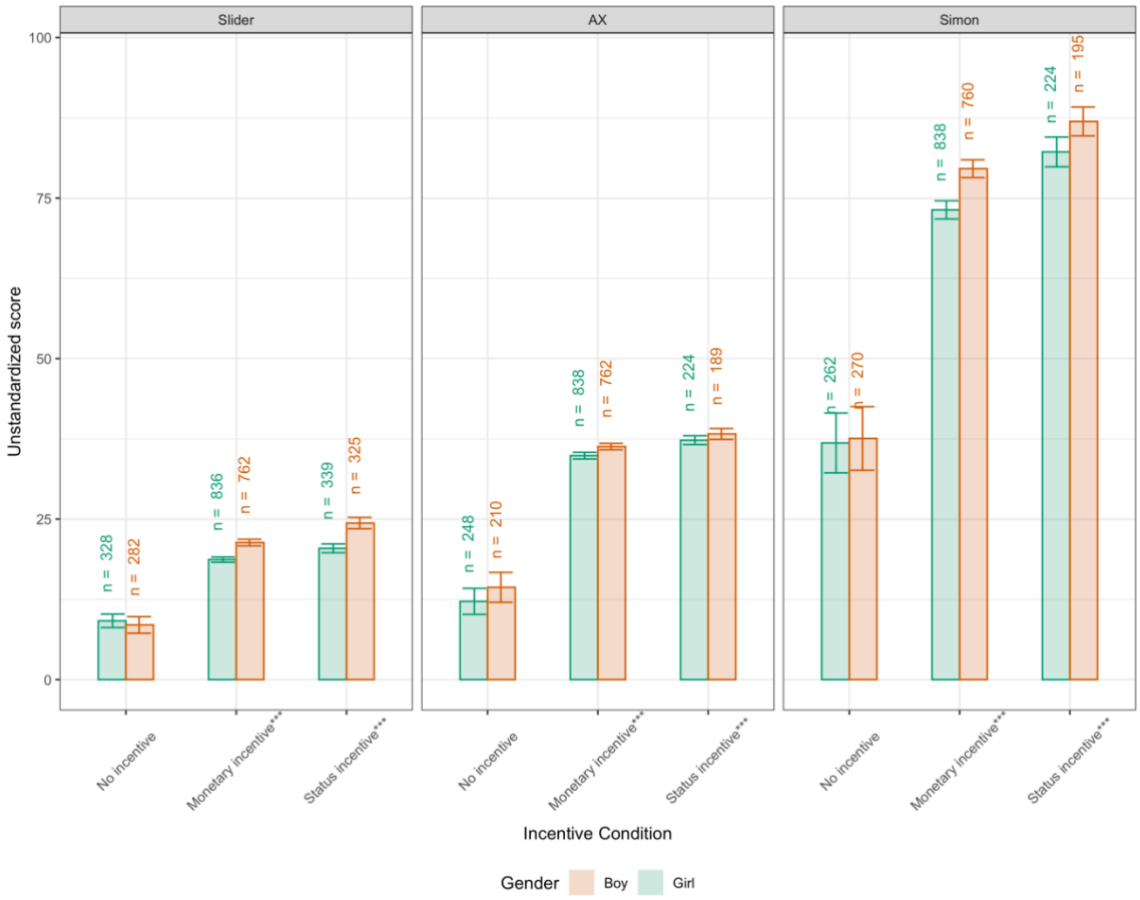
gender, i.e. incentive condition and task, are specified as contrasts with zero mean such that the estimated effect of gender (boy) is the average effect across all incentive conditions and tasks. Gender is also specified as a contrast with zero mean such that the estimated effect of incentive condition also represents the average effect across boys and girls. Hierarchical models were run using the *lmer* function in the R (v.4.2.2) package *lme4* v.1.1.31. P-values are calculated using the Kenward-Roger approximation to get approximate degrees of freedom.

3. RESULTS

3.1 Gender differences in raw performance scores

Figure 3 shows the mean performance results by task, incentive condition, and gender. In the "no-incentive" condition, the differences in performance between girls and boy are slight and lack statistical significance, as indicated by the Wilcoxon Rank Sum Test: about 0.6 completed sliders favoring girls ($P = 0.263$), 2.2 correct trials favoring boys in the AX task ($P = 0.138$) and 0.7 correct trials favoring boys for the Simon task ($P = 0.430$). Yet in the monetary incentive condition, boys outperform girls by 2.7 sliders ($P < 0.001$), 1.4 correct trials in the AX task ($P < 0.001$) and 6.4 correct trials in the Simon task ($P < 0.001$). In the status-incentive scheme, gender differences in performance also favor boys, but the gap only increases for the slider task, with on average 3.9 more trials completed for boys ($P < 0.001$). In the AX and Simon tasks, boys score on average 1.0 ($P < 0.001$) and 4.8 more trials correct ($P < 0.001$), respectively.

Figure 3. Means, confidence intervals, and Wilcoxon Rank Sum Test for each task, gender, and incentive condition.



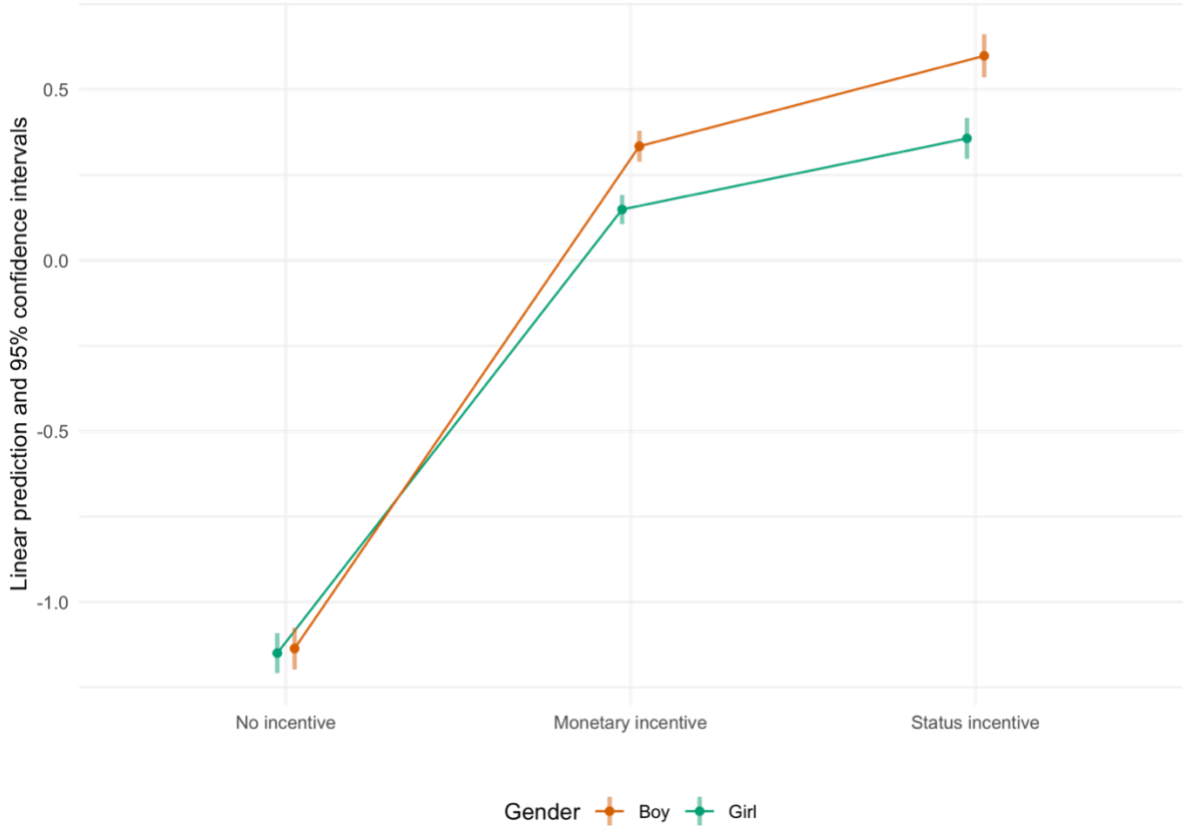
This figure shows performance by gender and incentive condition with 95% confidence intervals (CI). Statistical significance shown as computed by a two-sided Wilcoxon Rank Sum Test. Performance data by task and counting the leisure choice as zero-score towards the task.
*** $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$**

3.2 Gender differences in real effort

To gauge the overall magnitude of gender differences and more fully abstract from task specificities, we calculate for each round the individual's score standardized within the distribution of all other scores from the same task. To distinguish effort from performance, all

models of real effort using these standardized scores are controlled for differences in sample-standardized measures of fluid intelligence to account for the effect of cognitive ability.

Figure 4. Linear prediction of real effort by gender across incentive conditions.



This figure shows the linear predictions for boys and girls resulting from a two-level hierarchical regression model grouped at the student level, where the dependent variable is the real effort measure 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), group-level fixed effects (class), and the interaction of gender and incentive condition.

Figure 4 displays gender gaps in real effort. There is virtually no difference in cognitive effort between boys and girls in the absence of rewards (effect of being a boy β = 0.013 standard deviations, P = 0.759). Thus, we do not find support of our hypothesis that girls exert more effort than boys in the absence of external rewards (H1). In the monetary-incentive condition, average

effort across all students increases by 1.385 standard deviations (SD), but boys outwork girls on average by an additional 0.172 SD ($P < 0.001$). This evidence supports our hypothesis that boys increase their effort more than girls do in response to the introduction of performance-based monetary incentives (H2). Similarly, in the status-incentive condition (which also includes piece-rate payoffs), average effort increases by 1.621 SD as compared to the no-incentive condition, with boys again outworking girls by an additional 0.228 SDs ($P < 0.001$) (see Figure A1 of Appendix A for the direct comparison of the gender effect in the no-incentive condition and status-incentive condition). However, the gender gap in effort in the monetary incentive condition does not widen significantly with the addition of a status incentive ($\beta = 0.057$ SD, $P = 0.193$). We therefore do not find support that boys increase their effort more so than girls when competing for an additional status incentive placed on top of performance-based monetary incentives (H3). Overall, gender differences in effort thus emerge from varying the incentive condition, and particularly from offering material rewards.

Next, we estimate the overall gender gap in effort and investigate whether the incentive-specific gender differences are attributable to potential confounding factors. Table 1 shows results from two-level hierarchical multivariate regressions grouped at the student level with session (class)-level fixed effects. As expected, effort increases substantially as incentives are added. Boys show overall greater effort than girls across all incentive conditions; that difference is not explained away by controlling for age, frequency of mouse use, nor computer gaming. The magnitude of this gender gap is considerable: across all incentive conditions, boys score on average 0.136 SD greater than girls do ($P < 0.001$). The base specification is displayed in column (2) of Table 1, a model that controls for differences in sample-standardized measures of fluid intelligence, which does not substantially change the gender gap ($\beta = 0.147$ SD, $P < 0.001$). We additionally control for need for cognition, risk preferences, delay of gratification,

conscientiousness, agreeableness, openness, neuroticism, and extraversion to see how the differential gender effects by incentive condition observed in Figure 4 are mediated by these psychological characteristics (referred to throughout the rest of the paper as “personality traits”). Though need for cognition is itself a positive and significant predictor of effort (β = 0.039 SD, P < 0.020), the addition of these variables does not substantially alter the gender gap.

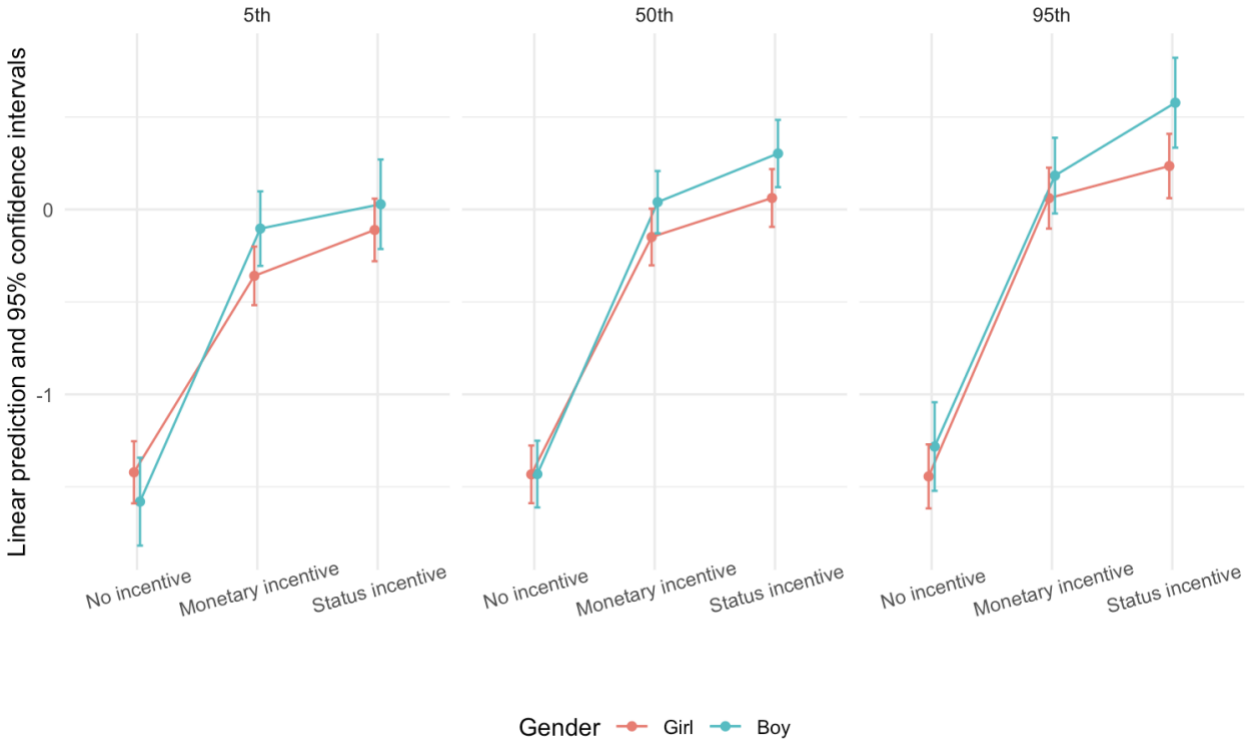
Table 1. Effect of gender, incentives, and gender-incentive interaction on real effort score.

	Dependent variable:			
	Real effort task scores (standardized within task)			
	(1)	(2)	(3)	(4)
Gender = boy	0.136*** (0.032)	0.147*** (0.031)	0.147*** (0.032)	0.146*** (0.032)
Incentive = no incentive (ref. = monetary)	−1.383*** (0.021)	−1.385*** (0.021)	−1.385*** (0.021)	−1.381*** (0.021)
Incentive = status (ref. = monetary)	0.237*** (0.022)	0.236*** (0.021)	0.236*** (0.021)	0.239*** (0.021)
Gender = boy x Incentive = no incentive (ref. = monetary)	−0.169*** (0.042)	−0.172*** (0.042)	−0.172*** (0.042)	−0.179*** (0.042)
Gender = boy x Incentive = status (ref. = monetary)	0.058 (0.043)	0.057 (0.043)	0.056 (0.043)	0.058 (0.043)
Task = AX (ref. = Slider)	0.0001 (0.020)	0.00002 (0.020)	0.002 (0.020)	0.001 (0.020)
Task = Simon (ref. = Slider)	0.010 (0.020)	0.011 (0.020)	0.011 (0.020)	0.011 (0.020)
Round 2	−0.095*** (0.016)	−0.093*** (0.016)	−0.092*** (0.016)	−0.092*** (0.016)
Age (months)	0.004 (0.003)	0.004 (0.003)	0.005* (0.003)	0.005 (0.003)
Mouse use	0.035** (0.013)	0.037** (0.013)	0.037** (0.013)	0.037** (0.013)
Videogaming	0.020 (0.013)	0.019 (0.013)	0.024 (0.013)	0.024 (0.013)
Fluid intelligence		0.108*** (0.015)	0.102*** (0.015)	0.098*** (0.016)
Need for cognition			0.039* (0.016)	0.039* (0.016)
Risk-loving			−0.0001 (0.005)	−0.00003 (0.005)
Delay of gratification			0.047 (0.031)	0.047 (0.031)
Conscientiousness			0.009 (0.017)	0.009 (0.017)
Agreeableness			0.030 (0.016)	0.030 (0.016)
Openness			−0.004 (0.016)	−0.004 (0.016)
Neuroticism			0.024 (0.016)	0.024 (0.016)
Extraversion			0.013 (0.015)	0.013 (0.015)
Gender = boy x Fluid intelligence				0.041 (0.030)
Incentive = no incentive (ref. = monetary) x Fluid intelligence				−0.067** (0.021)
Incentive = status (ref. = monetary) x Fluid intelligence				0.029 (0.021)
Gender = boy x Incentive = no incentive x Fluid intelligence				0.142*** (0.042)
Gender = boy x Incentive = status x Fluid intelligence				0.105* (0.043)
Constant	−0.950** (0.356)	−0.993** (0.344)	−1.172*** (0.347)	−1.167*** (0.347)
Session fixed effects	Yes	Yes	Yes	Yes
$sd(u_j)$	0.34	0.32	0.31	0.31
$sd(e_{ij})$	0.72	0.72	0.72	0.71
Students	799	798	794	794
Observations	7,874	7,866	7,834	7,834

This table shows the results of a two-level hierarchical regression model grouped at the student level, where the dependent variable is the real effort task scores standardized within task. The estimated effect of gender (boy) is the average effect across all incentive conditions and tasks. All variables that are interacted with gender, i.e. incentive condition and task, are specified as contrasts with mean zero. P-values are calculated with the Kenward-Roger approximation to get approximate degrees of freedom. Standard errors are shown in parentheses.
*** *p* < 0.05, ** *p* < 0.01, *** *p* < 0.001**

It has been suggested that gender sensitivity to competition depends on ability, as evidenced by an array of 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). Therefore, we test the three-way interaction of gender, incentives, and fluid intelligence to see if effort investments by gender differ depending on ability level. While the impact of the addition of a status incentive on effort is insignificant for girls regardless of fluid intelligence, we find evidence that the extent to which boys are sensitive to competition does indeed depend on it. It is only amongst high-ability boys that we observe a substantial relative increase in effort as compared to girls under the tournament condition, suggesting that (self-perceived) ability may matter more for boys when the stakes involve status hierarchy. 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.

Figure 5. Linear prediction of real effort by gender and incentive, by percentile of fluid intelligence (5th, 50th, 95th)



This figure shows the linear predictions for boys and girls resulting from a two-level hierarchical regression model grouped at the student level, where the dependent variable is the real effort measure per 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, group-level fixed effects (class), and the interaction of gender, incentive condition, and fluid intelligence.

3.3 Effort-related indices

The overall finding thus far is that boys exert more task-related effort than girls do (net of fluid intelligence as well as personality traits), but only if incentivized by monetary rewards. Without external incentives, boys do not provide more task-related effort than girls do, and adding a status incentive beyond the monetary incentive does not significantly enlarge the gender gap. Yet it is also of interest if boys achieve this advantage through specific strategic choices or the

application of certain subdomains of EF over others. Finally, it is worth investigating whether our findings using a performance score-based index of effort are compatible with other effort-related indices so that we may better contextualize the observed gender gap and apply our findings appropriately to real-world learning contexts.

3.3.1 Gender differences in real effort engagement

To further elucidate the determinants of effort provision, we investigate the associations between incentives, gender, and personality traits on the *decision* to exert oneself on the real-effort task or not. Before starting each experimental round, each child was presented with the option to play a leisure game instead of completing the task. In the monetary and status-incentive conditions this was with the understanding that they would earn no points if they played the game. 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 et al., 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. 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 ($P < 0.001$) and girls liking the puzzle more than boys ($P < 0.001$). 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. Girls gamed less than boys in the first round of the no-incentive condition, though this difference is not statistically significant at the $\alpha = 0.05$ level ($P = 0.060$), and rates of gaming (tasking) slightly increased (decreased) in the second round of the no-incentive condition, though more so for girls.

Fitting a hierarchical linear probability model of choosing to complete the task over the game in Table 2, we find that boys overall gamed more than girls did, and significantly more so in the no-incentive condition ($P < 0.05$). Specifically, being a boy reduces the probability of completing the task by 1.9% on average ($P = 0.016$), and by 4.1% in the no-incentive condition when compared to tasking rates in the monetary incentive condition ($P = 0.002$). Personality traits seem to mediate only slightly the positive effect of being a girl on choosing to do the task, with more agreeableness being positively associated with tasking ($\beta = 0.008$, $P = 0.041$).

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

	Dependent variable:			
	Tasked instead of gamed			
	(1)	(2)	(3)	(4)
Gender = boy	−0.019* (0.008)	−0.018* (0.008)	−0.016* (0.008)	−0.016* (0.008)
Incentive = no incentive (ref. = monetary)	−0.501*** (0.010)	−0.501*** (0.010)	−0.501*** (0.010)	−0.500*** (0.010)
Incentive = status (ref. = monetary)	0.012 (0.010)	0.013 (0.010)	0.013 (0.010)	0.013 (0.010)
Gender = boy x Incentive = no incentive (ref. = monetary)	−0.037* (0.014)	−0.038** (0.014)	−0.037** (0.014)	−0.038** (0.014)
Gender = boy x Incentive = status (ref. = monetary)	−0.007 (0.015)	−0.008 (0.015)	−0.008 (0.015)	−0.008 (0.015)
Task = AX (ref. = Slider)	−0.009 (0.007)	−0.009 (0.007)	−0.009 (0.007)	−0.009 (0.007)
Task = Simon (ref. = Slider)	0.007 (0.007)	0.007 (0.007)	0.007 (0.007)	0.006 (0.007)
Round 2	−0.044*** (0.006)	−0.043*** (0.006)	−0.043*** (0.006)	−0.043*** (0.006)
Age (months)	0.0002 (0.001)	0.0002 (0.001)	0.0003 (0.001)	0.0003 (0.001)
Mouse use	−0.002 (0.003)	−0.002 (0.003)	−0.002 (0.003)	−0.002 (0.003)
Videogaming	0.0002 (0.003)	0.001 (0.003)	0.002 (0.003)	0.003 (0.003)
Fluid intelligence		0.005 (0.004)	0.004 (0.004)	−0.005 (0.005)
Need for cognition			0.007 (0.004)	0.007 (0.004)
Risk-loving			−0.0005 (0.001)	−0.0004 (0.001)
Delay of gratification			−0.002 (0.007)	−0.002 (0.007)
Conscientiousness			0.0003 (0.004)	0.0004 (0.004)
Agreeableness			0.008* (0.004)	0.008* (0.004)
Openness			0.003 (0.004)	0.003 (0.004)
Neuroticism			0.007 (0.004)	0.007 (0.004)
Extraversion			0.004 (0.004)	0.004 (0.004)
Gender = boy x Fluid intelligence				0.016* (0.008)
Incentive = no incentive (ref. = monetary) x Fluid intelligence				−0.035*** (0.010)
Incentive = status (ref. = monetary)				−0.002 (0.010)
Gender = boy x Incentive = no incentive x Fluid intelligence				0.048*** (0.014)
Gender = boy x Incentive = status x Fluid intelligence				0.008 (0.015)
Constant	0.742*** (0.083)	0.746*** (0.083)	0.729*** (0.084)	0.730*** (0.084)
Session fixed effects	Yes	Yes	Yes	Yes
$sd(u_j)$	0.05	0.05	0.05	0.05
$sd(e_{ij})$	0.25	0.25	0.25	0.25
Students	799	798	794	794
Observations	7,874	7,866	7,834	7,834

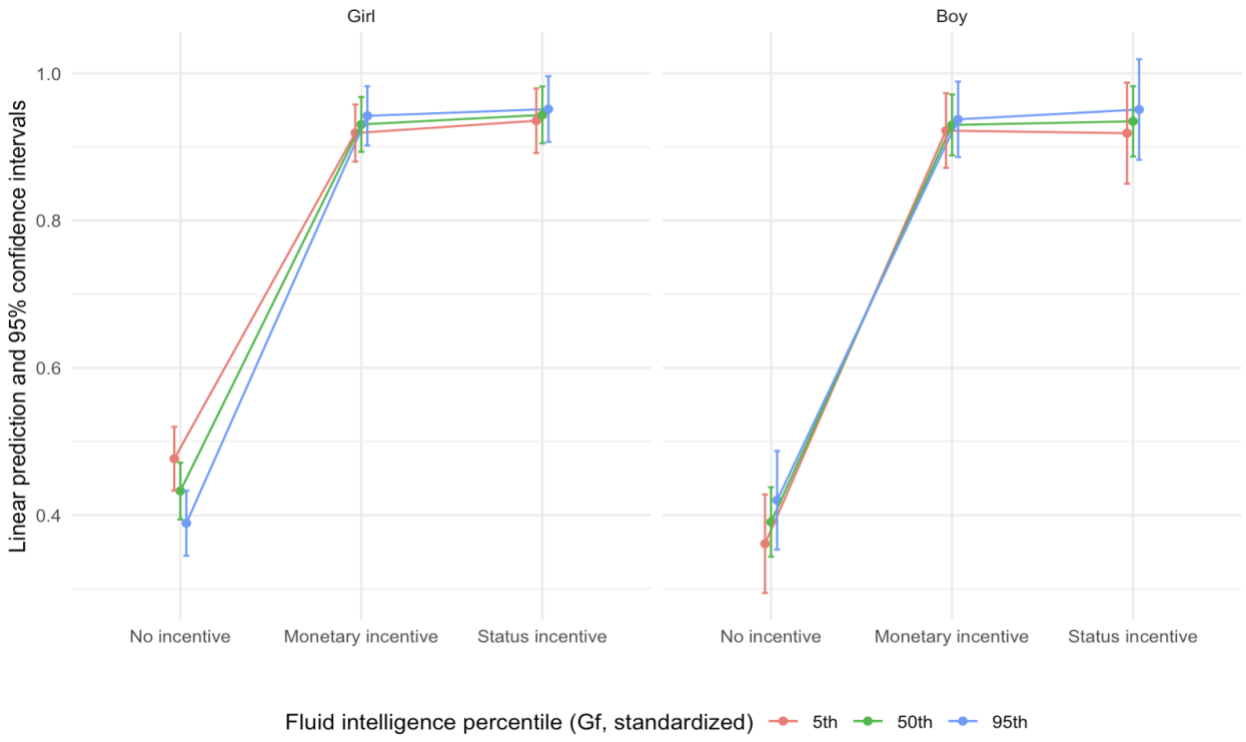
This table shows the results of a two-level hierarchical regression model grouped at the student level, where the dependent variable is a binary indicator of whether the student opted to do the task over the leisure game in a specific round (1 = tasked, 0 = gamed). The estimated effect of gender (boy) is the average effect across all incentive conditions and tasks. All variables that are interacted with gender, i.e. incentive condition and task, are specified as contrasts with mean zero. P-values are calculated with the Kenward-Roger approximation to get approximate degrees of freedom. Standard errors are shown in parentheses.

*** *p* < 0.05, ** *p* < 0.01, *** *p* < 0.001**

What we observe as boys' greater task avoidance parallels research on gender differentials in test non-compliance, where male-identifying students are often less likely to attend testing sessions for low-stakes exams (Swerdzewski et al., 2009). Jackson (2003), however, argues that boys may be more likely to disengage from effortful tasks as a form of self-preservation in response to lower perceived ability – a sort of "don't try, can't fail" mentality. To find out whether the gender differences in tasking versus gaming are attributable to gender differences in

motivation or to gender differences in the role that (self-perceived) ability plays in decisions about effort investments, we test the three-way interaction effect of gender, incentives, and fluid intelligence on the probability of task completion. Results in column (4) of Table 2 are visualized in Figure 6 and indicate that having one SD greater fluid intelligence increases boys' probability of completing the real-effort task by about 1.6% more than girls who are of equal ability. This is driven particularly by gaming patterns in the no-incentive condition, where lower-ability girls are actually more likely to engage with the effortful task than girls of high ability, while the opposite 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 nor their substantially greater rates of gaming in the no-incentive condition.

Figure 6. Linear prediction of the probability of tasking by percentile of fluid intelligence (5^{th}, 50^{th}, 95^{th}) and incentive condition, by gender.

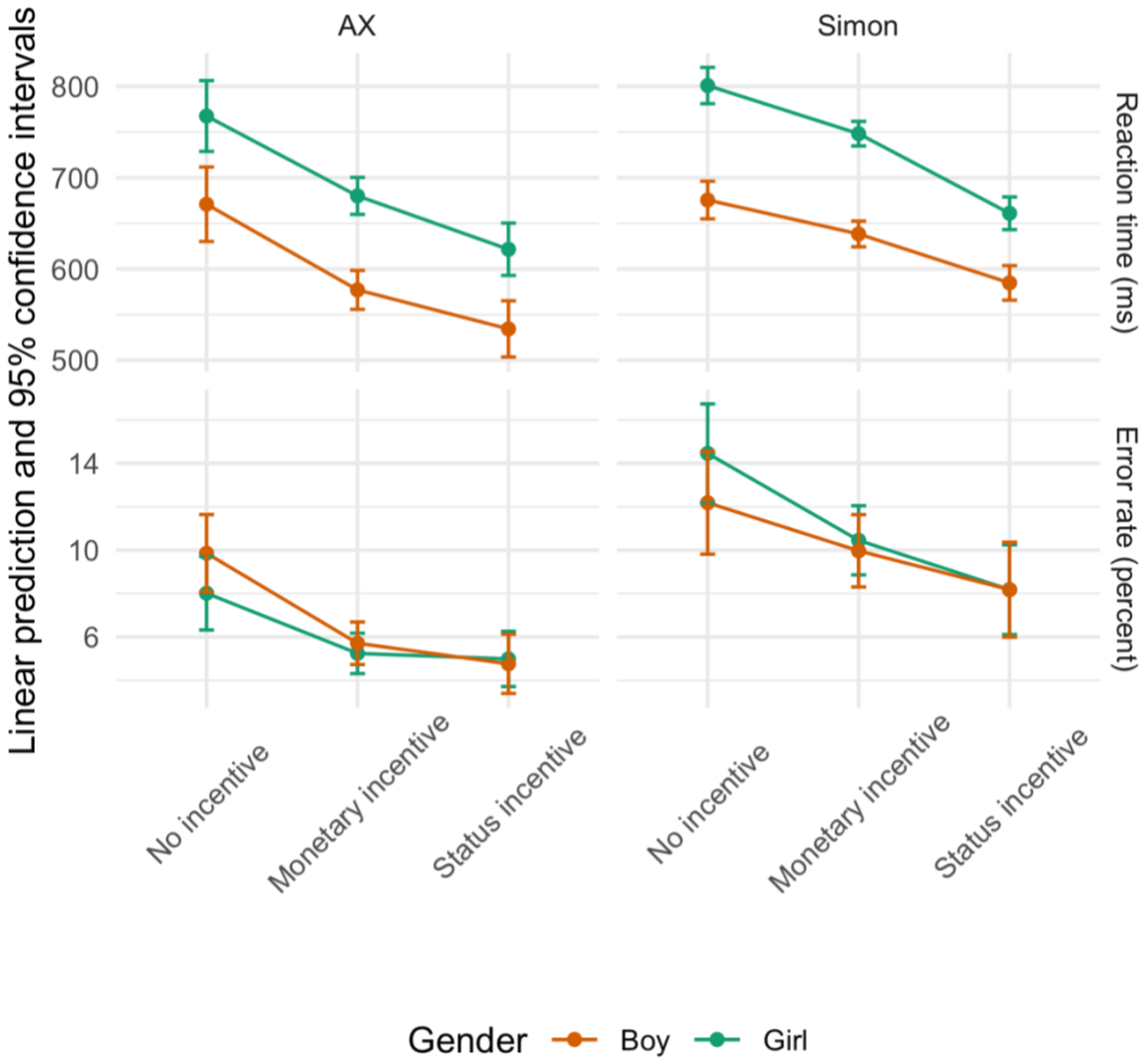


3.3.2 Gender differences in task strategy

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. Reaction time and accuracy measures are not applicable to the slider task due to its layout. Results in Appendix Tables A3 and A4 show that boys are on average faster than girls in responding to trials by about 95.6 milliseconds (ms) ($P < 0.001$) on the AX task and by about 103.8ms SD ($P < 0.001$) on the Simon task, but they are not more accurate on average. We find evidence that heterogeneities in personality traits only partially mediate the gender gap in reaction time, more so for the Simon task than the AX-CPT task. Figure 7 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 (β = -125.4ms, $P < 0.001$) and shrinks slightly but non-significantly when monetary incentives are introduced (an estimated increase of β = 15.6ms, $P = 0.189$). When status incentives are added in the tournament condition, the gender gap in reaction time for the Simon task closes by about 39% as compared to the no-incentive condition gap (increase of β = 49.2ms, $P = 0.002$) and by nearly 31% from the monetary incentive condition gap (increase of β = 33.6ms, $P = 0.001$). While these findings diverge somewhat from the overall patterns when predicting real effort, it should be taken into account 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. 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.

Figure 7. Linear prediction of reaction time and error rate by gender across incentive conditions, AX and Simon tasks.



This figure shows the linear predictions for boys and girls resulting from a two-level hierarchical regression model grouped at the student level, where the dependent variable is the mean standardized reaction time for correct responses and error rate, respectively. All models include controls for experimental conditions (incentive condition, round, and task), individual-level fixed effects (age, mouse use, computer gaming, and intelligence), group-level fixed effects (class), and the interaction of gender and incentive condition.

3.3.3 Gender differences in cognitive flexibility and inhibition

Suboptimal adoption of testing strategies has been linked with heterogeneities in personality traits such as greater risk aversion and neuroticism (Förster et al., 2003). Neuropsychological research suggests that personality and cognition are closely related when it comes to their phenotypical and genetic determinants (Williams et al., 2010). Therefore, we investigate whether the observed strategic choices of boys and girls may instead be the result of the predominance of certain dimensions of cognitive functioning, and how personality traits correlate with these dimensions. Specifically, we examine task-specific indices of speed and accuracy for the AX and Simon tasks that measure cognitive flexibility, cognitive control type, and inhibition (see Tables S3-S4 for descriptive statistics).

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. Special attention is given to the A-Y condition, that is, when the cue indicates a possible target, but the probe is different. In this case, participants that tend towards proactive control could have a higher failure rate if they respond before correctly assessing the probe. Participants who engage more in reactive control may have a higher failure rate in the B-X condition as they incorrectly react to perceiving the target probe. It has been found that in most young, healthy populations, proactive control processes predominate reactive ones (Braver et al., 2009). The Proactive Behavioral Index, or PBI[1], provides a measure of how much proactive interference one experiences in situations when a reactive approach is required. For both average reaction time on correct trials and error rates,

[1] 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)*.

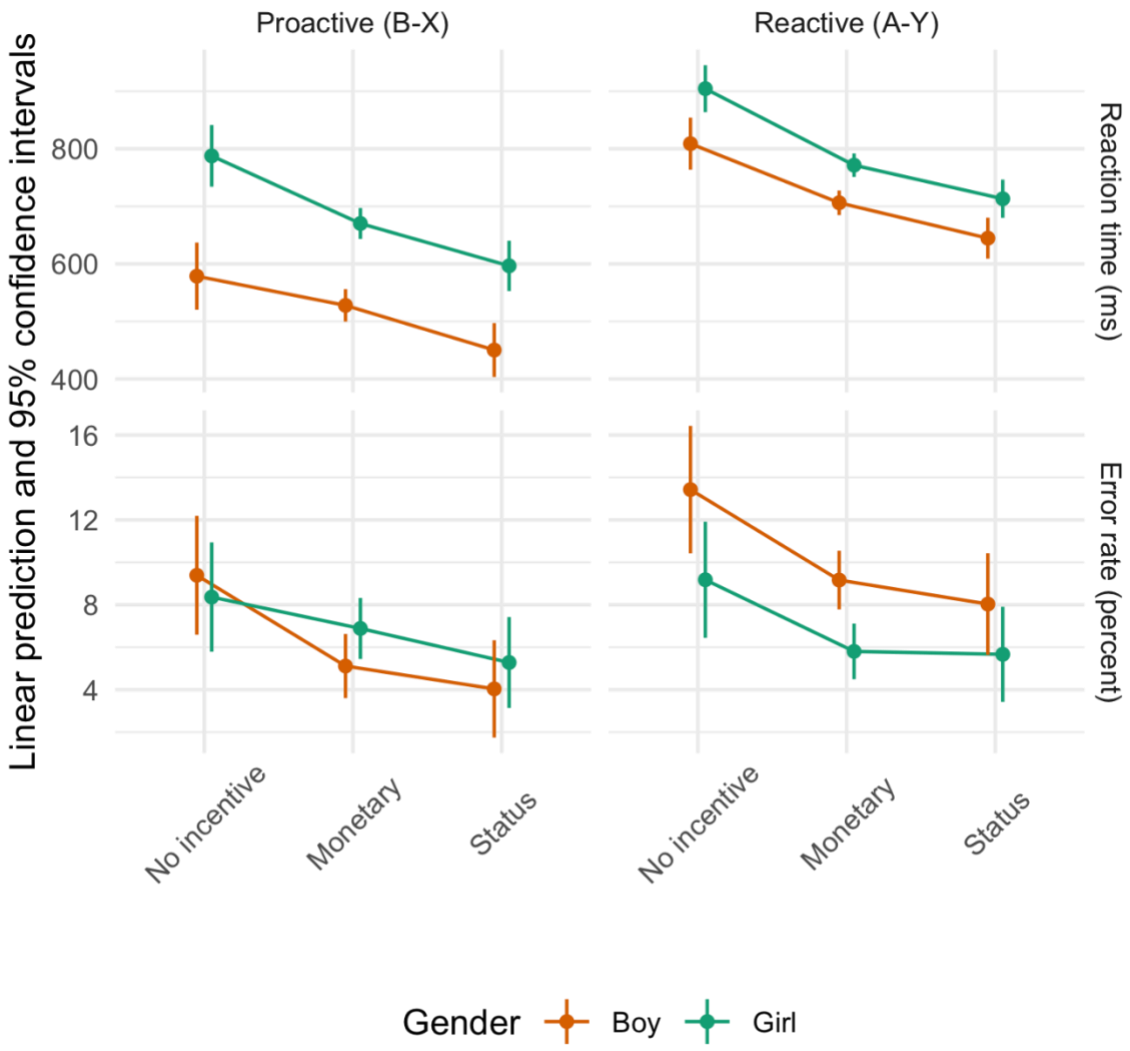
we find that being a boy positively predicts PBI for both speed and accuracy, indicating that boys tend to engage in proactive control relatively more than girls, and thus experience greater proactive interference, whereas being a girl is a negative predictor, indicating that girls tend to engage in reactive control relatively more than boys (Table A7). However, after controlling for personality traits, the gender effect on PBI error rate is no longer significant at the $\alpha = 0.05$ level ($\beta = 0.158$ SD, P = 0.054), with higher agreeableness associated with more reactivity ($\beta = -0.101$ SD, P = 0.005).

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 et al., 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 Figure 8 confirm again that boys' response time is overall faster than girls', but the magnitude of this difference depends on the type of trial: boys are substantially quicker than girls in the AX task when a proactive approach is required (B-X condition, $\beta = -165.9$ms, P < 0.001) than they are when a reactive approach is required (A-Y condition, $\beta = -76.5$ms, P < 0.001). Need for cognition and neuroticism are both positively correlated with quicker reaction time regardless of trial type, though these personality traits do not help explain much of boys' advantage in speed (Table A8). With respect to error rate, there is no significant difference between boys and girls on proactive trials. Girls, however, are overall more accurate than boys when in reactive mode ($\beta = 3.2\%$, P = 0.003). Some of this advantage is mediated by personality traits, with higher agreeableness and lower risk preferences associated with greater accuracy when reactive approaches are elicited (Table A9).

On the Simon task, it is well known that people respond slower to and commit more errors on incongruent trial conditions, as subjects 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 subject-round from those on incongruent trials and standardize the differences. Results in Table A10 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 than girls (-22.0ms, $P < 0.001$). Again, we detect no significant gender difference with regards to error rate ($P = 0.681$).

These findings confirm that much of girls' disadvantage in effort stems from their slower reaction times. While our findings suggest that girls' strategic tendencies partially stem from their greater reactivity and risk aversion, allowing them to switch more flexibly between different cognitive control types, this prudent approach leads to slower speed that only gives them an accuracy advantage when a reactive approach is required.

Figure 8. Linear prediction of reaction time (milliseconds) and error rate (percentage) by gender across incentive conditions and AX-CPT trial condition.



This figure shows the linear predictions for boys and girls resulting from a two-level hierarchical regression model grouped at the student level, 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), group-level fixed effects (class), and the interaction of gender and incentive condition.

4. DISCUSSION

One of the toughest dilemmas that educators and policymakers must face is how to achieve an upward shift in effort exertion for academic performance 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 et al., 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. This evidence can also inform intervention strategies that not only boost academic achievement through increases in effort, but also help to attenuate the gender gap that arises later on in life. Schools could increase test-taking motivation and classroom participation, especially amongst low-achieving boys, by compensating effortful behavior with material rewards, while helping orient young girls more towards an achievement-focused mindset that emphasizes the material benefits of incurring calculated risks. High-ability girls may also 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.

4.1 Limitations and future directions

Several limitations of the study should be mentioned. First, while the use of low-skill, monotonous tasks helps to minimize the confounding role of ability and allows us to more

accurately measure the impact of incentives on performance-based cognitive effort, they are not perfectly representative of all tasks that are encountered in the real world and that determine outcomes such as school grades and educational attainment. Thus, to further advance our understanding of gender gaps in effort, future research should consider whether more demanding tasks invite different patterns of engagement.

Furthermore, the strategy of faster reaction time that boys utilized during the experiment to optimize performance and gain an advantage over girls does not necessarily translate into beneficial long-term strategies that enhance performance in school or work settings, such as diligence or perseverance. It could also be that even after controlling for fluid intelligence, other unobserved variables related to ability were at play, such as boys' faster simple visual reaction times (Silverman, 2006).

While a survey of educational literature finds congruent evidence that short-term and sufficiently high incentives matter more for boys than for girls (Levitt et al., 2016), it would be insightful to moderate incentive schemes to elicit more female-typed preferences in the form of time-delayed prizes (Angrist et al., 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 girl's more prosocial orientations and motivations (Cassar & Rigdon, 2021; Watson & Blanchard-Fields, 1998). While our experimental design does not allow for the direct comparison of the effect by gender of status incentives alone versus that of a no-incentive or monetary incentive scheme on cognitive effort, the inclusion of a status incentive in addition to a monetary incentive more accurately captures the coupling of motivation schemes used in traditional schooling environments.

Finally, our results refer to a single country, one specific age range, and one specific order in which incentives are introduced, whose effects could be confounded with differential practice or fatigue effects by gender. However, the desire to avoid motivational crowding-out effects as well as logistical constraints of the experiment limited our ability to test these potential order effects.

6. CONCLUSION

The results from this study show that while the addition of external rewards boosts overall cognitive effort, boys increase their effort more than girls when materially incentivized. However, we do not find significantly greater gender differences in effort following the inclusion of a status incentive. The findings regarding gender differences in intrinsic motivation are somewhat ambiguous: when there are no external incentives present, we find evidence of greater cost of effort for boys, in particular for those with low ability, than for girls as witnessed by boys' greater rates of gaming in the no-incentive condition. Yet, among those who opted for the real-effort task in the no-incentive condition, boys engage more cognitive resources via greater speed and proactive control engagement and thus perform faster while not sacrificing accuracy, though they do not exhibit a clear advantage over girls in terms of inhibition of false responses. In fact, these strategic preferences lead to lowered accuracy for boys when reactive approaches are required. While girls adopt more prudent strategies, survey-based personality measures such as risk aversion could not completely explain the gender gap in cognitive effort or task performance strategy, nor could they meaningfully account for why girls were more likely than boys to complete the task rather than play the leisure game.

Previous studies have emphasized lower preferences of women towards competition and status-ranking. That is, while previous accounts stress gender differences in the *direction* of effort (Bonner & Sprinkle, 2002) that manifests in the self-selection into tournament versus piece-rate compensation schemes (Niederle & Vesterlund, 2007), our study successfully

identifies gender differences in the *intensity* of effort under the two conditions given that nearly all students selected into doing the real-effort task when material or status incentives are present. The results suggest that in environments where both types of incentives are at play, gender differences in effort provision are no more different than in environments where only monetary incentives are present, despite differential effects of status competition by gender for high-ability students. Thus, according to our findings, the key differentiating feature for girls' and boys' effort provision is the presence of material rewards.

Data availability and usage statement: The data underlying this article will be available in the "e ciencia Datos" Repository of UC3M at https://doi.org/10.21950/DEDRIZ after expiration of the embargo on 1 September 2026. 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.

References

Akhtar, H., & Firdiyanti, R. (2023). Test-taking motivation and performance: Do self-report and time-based measures of effort reflect the same aspects of test-taking motivation? *Learning and Individual Differences*, *106*, 102323.

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.

Arens, A. K., & Hasselhorn, M. (2015). Differentiation of competence and affect self-perceptions in elementary school students: extending empirical evidence. *European Journal of Psychology of Education*, *30*, 405-419.

Beutel, A. M., & Marini, M. M. (1995). Gender and values. *American Sociological Review*, 436-448.

Bonner, S. E., & Sprinkle, G. B. (2002). The effects of monetary incentives on effort and task performance: theories, evidence, and a framework for research. *Accounting, Organizations and Society*, *27*(4-5), 303-345.

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*, soac148.

Brandts, J., Gërxhani, K., & Schram, A. (2020). Are there gender differences in status-ranking aversion? *Journal of Behavioral and Experimental Economics*, *84*, 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.

Buchmann, C., DiPrete, T. A., & McDaniel, A. (2008). Gender inequalities in education. *Annual Review of Sociology*, *34*, 319-337.

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.

Cassar, A., & Rigdon, M. L. (2021). Prosocial option increases women's entry into competition. *Proceedings of the National Academy of Sciences*, *118*(45), e2111943118.

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.

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

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

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The Role of Gender in Test-Taking Motivation under Low-Stakes Conditions. *Research & Practice in Assessment*, *8*, 69-82.

Dreber, A., Von Essen, E., & Ranehill, E. (2014). Gender and competition in adolescence: task matters. *Experimental Economics*, *17*, 154-172.

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.

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.

Dutcher, G., Salmon, T., & Saral, K. J. (2015). Is' Real'Effort More Real? *Available at SSRN 2701793*.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109-132.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *The Quarterly Journal of Economics*, *108*(2), 437-459.

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.

Förster, J., Higgins, E. T., & Bianco, A. T. (2003). Speed/accuracy decisions in task performance: Built-in trade-off or separate strategic concerns? *Organizational Behavior and Human Decision Processes*, *90*(1), 148-164.
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.
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.
Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, *118*(3), 1049-1074.
Gneezy, U., & Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review*, *94*(2), 377-381.
Heckman, J. J., Jagelka, T., & Kautz, T. (2021). *Some contributions of economics to the study of personality*. 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.
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.
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.
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.
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.
Jackson, C. (2003). Motives for 'laddishness' at school: Fear of failure and fear of the 'feminine'. *British Educational Research Journal*, *29*(4), 583-598.
James Jr, H. S. (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.
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.
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.
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.
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.
Legewie, J., & DiPrete, T. A. (2012). School context and the gender gap in educational achievement. *American Sociological Review*, *77*(3), 463-485.
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.
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.
Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*, 314-324.
McCrea, S. M., Hirt, E. R., & Milner, B. J. (2008). She works hard for the money: Valuing effort underlies gender differences in behavioral self-handicapping. *Journal of Experimental Social Psychology*, *44*(2), 292-311.
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.
Mäki-Marttunen, V., Hagen, T., & Espeseth, T. (2019). Proactive and reactive modes of cognitive control can operate independently and simultaneously. *Acta Psychologica*, *199*, 102891.

Nalbantian, H. R., & Schotter, A. (1997). Productivity under group incentives: An experimental study. *The American Economic Review*, 314-341.
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.
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.
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.
Niederle, M., & Vesterlund, L. (2011). Gender and competition. *Annu. Rev. Econ.*, *3*(1), 601-630.
Ors, E., Palomino, F., & Peyrache, E. (2013). Performance gender gap: does competition matter? *Journal of Labor Economics*, *31*(3), 443-499.
Otto, A. R., & Daw, N. D. (2019). The opportunity cost of time modulates cognitive effort. *Neuropsychologia*, *123*, 92-105.
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.
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.
Roivainen, E. (2011). Gender differences in processing speed: A review of recent research. *Learning and Individual Differences*, *21*(2), 145-149.
Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54-67.
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.
Schram, A., Brandts, J., & Gërxhani, K. (2019). Social-status ranking: a hidden channel to gender inequality under competition. *Experimental economics*, *22*, 396-418.
Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science*, *58*(8), 1438-1457.
Silverman, I. W. (2006). Sex differences in simple visual reaction time: A historical meta-analysis. *Sex Roles*, *54*(1-2), 57.
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.
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.
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.
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.
Tang, C., & Zhao, L. (2023). Gender Social Norms and Gender Gap in Math: Evidence and Mechanisms. *Applied Economics*, 1-19.
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.
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.
Van Dijk, F., Sonnemans, J., & Van Winden, F. (2001). Incentive systems in a real effort experiment. *European Economic Review*, *45*(2), 187-214.
Vecchione, M., Alessandri, G., & Marsicano, G. (2014). Academic motivation predicts educational attainment: Does gender make a difference? *Learning and Individual Differences*, *32*, 124-131.
Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: a meta-analysis. *Psychological Bulletin*, *140*(4), 1174.

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.
Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience*, *15*, 395-415.
Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68-81.
Williams, P. G., Suchy, Y., & Kraybill, M. L. (2010). Five-factor model personality traits and executive functioning among older adults. *Journal of Research in Personality*, *44*(4), 485-491.
Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1-17.
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.

APPENDIX A

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

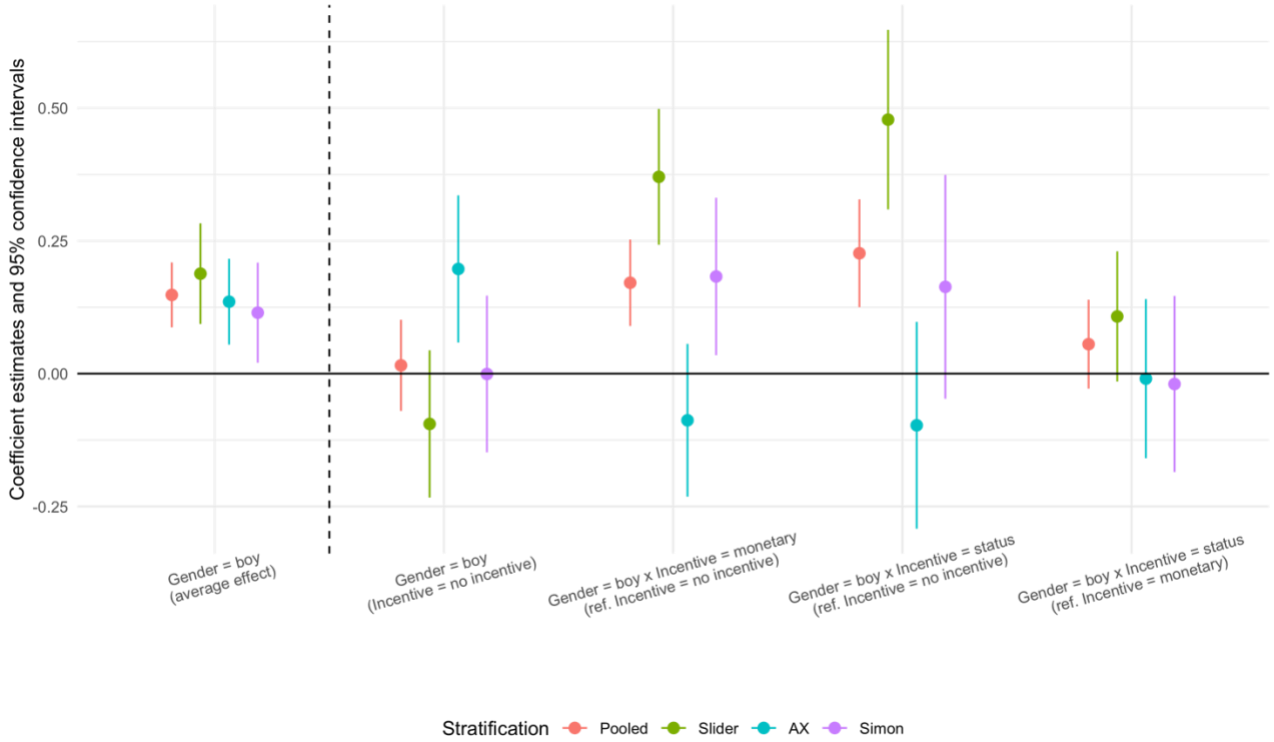
	Boys			Girls			Two-sided Student t-test		
	N	Mean	Std. Err.	N	Mean	Std. Err.	P	T-value	DoF
Slider task									
Self-reported effort	314	5.580	1.490	336	5.510	1.570	0.590	0.54	650
Perceived task difficulty	314	3.150	1.080	336	2.980	1.080	0.045	2.00	650
Perceived task likeability	314	4.170	0.910	336	4.270	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 as “very, very low effort” to 7 as “very, very high effort”), perceived difficulty (scale of increasing integers from 1 as “very easy” to 5 as “very difficult”) and perceived likeability (scale of increasing integers from 1 as “strongly disliked” to 5 as “strongly liked”) of each task. We observe no significant differences by gender in the real-effort tasks for self-reported effort nor perceived task likeability. While there is a 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 ($P < 0.001$) and girls like the puzzle better ($P < 0.001$), 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

Figure A1. The gender effect on effort: 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 two-level hierarchical regression model grouped at the student level, where the dependent variable is the real effort measure, standardized within task. All models include controls for experimental conditions (incentive condition, round, and task [for pooled data]), individual-level fixed effects (age, mouse use, computer gaming, and intelligence), group-level fixed effects (class), and the interaction of gender and incentive condition.

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.

TABLE A3: Regression results for the effect of gender, incentives, and gender-incentive interaction on average reaction time for correct responses by task (AX and Simon) (for Figure 7)

	Dependent variable:			
	Reaction time (ms)			
	AX	AX	Simon	Simon
Gender = boy	−95.604*** (16.821)	−92.541*** (17.413)	−103.823*** (10.573)	−94.898*** (10.884)
Incentive = no incentive (ref. = monetary)	87.492*** (18.296)	87.970*** (18.231)	52.737*** (8.345)	53.084*** (8.350)
Incentive = status (ref. = monetary)	−58.472*** (12.560)	−59.127*** (12.510)	−87.217*** (7.079)	−87.149*** (7.083)
Gender = boy x Incentive = no incentive (ref. monetary)	6.362 (26.217)	9.824 (26.348)	−15.564 (11.842)	−16.170 (11.847)
Gender = boy x Incentive = status (ref. monetary)	15.760 (18.064)	17.158 (18.000)	33.635** (10.210)	33.443** (10.212)
Round 2	−0.493 (6.119)	0.048 (6.109)	−7.079* (3.227)	−7.174* (3.235)
Age (months)	1.178 (1.371)	0.792 (1.378)	−2.041* (0.913)	−2.111* (0.915)
Mouse use	1.699 (6.385)	2.488 (6.424)	−7.883 (4.274)	−6.152 (4.281)
Videogaming	−20.034** (6.557)	−23.261*** (6.676)	−6.753 (4.358)	−8.796* (4.419)
Fluid intelligence	−45.828*** (7.426)	−42.793*** (7.556)	−30.749*** (4.968)	−26.494*** (5.033)
Need for cognition		−20.715* (8.227)		−21.935*** (5.469)
Risk-loving		−0.354 (2.737)		−2.329 (1.823)
Delay of gratification		−18.202 (15.631)		−14.919 (10.399)
Conscientiousness		−12.341 (8.884)		−5.514 (5.859)
Agreeableness		−10.646 (8.081)		1.232 (5.329)
Openness		−1.220 (8.162)		0.358 (5.463)
Neuroticism		−16.062* (8.023)		−0.368 (5.335)
Extraversion		−4.595 (7.851)		0.466 (5.190)
Constant	617.684*** (173.757)	689.973*** (175.255)	1,011.095*** (115.693)	1,041.133*** (116.191)
Session fixed effects	Yes	Yes	Yes	Yes
$sd(u_j)$	175.66	174.27	123.1	121.39
$sd(e_{ij})$	142.05	141.45	75.12	75.18
Students	796	792	794	790
Observations	2,204	2,193	2,237	2,229

This table shows the results of a two-level hierarchical regression model grouped at the student level, where the dependent variable is the average reaction time for correct responses 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.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE A4: Regression results for the effect of gender, incentives, and gender-incentive interaction on error rate by task (AX and Simon) (for Figure 7)

	Dependent variable:			
	Error rate (percent)			
	AX	AX	Simon	Simon
Gender = boy	0.694 (0.765)	0.252 (0.760)	−0.919 (1.247)	−0.627 (1.299)
Incentive = no incentive (ref. = monetary)	2.769*** (0.782)	2.835*** (0.781)	4.004*** (0.932)	3.998*** (0.929)
Incentive = status (ref. = monetary)	−0.255 (0.536)	−0.274 (0.536)	−2.273** (0.791)	−2.272** (0.788)
Gender = boy x Incentive = no incentive (ref. monetary)	1.379 (1.123)	1.450 (1.129)	−1.790 (1.325)	−1.792 (1.321)
Gender = boy x Incentive = status (ref. monetary)	−0.693 (0.774)	−0.648 (0.771)	0.482 (1.143)	0.476 (1.140)
Round 2	0.132 (0.260)	0.106 (0.261)	0.401 (0.360)	0.367 (0.360)
Age (months)	−0.058 (0.063)	−0.117 (0.060)	−0.034 (0.108)	−0.054 (0.110)
Mouse use	0.456 (0.294)	0.191 (0.282)	−0.167 (0.507)	−0.129 (0.514)
Videogaming	0.220 (0.302)	0.073 (0.293)	−0.139 (0.517)	−0.264 (0.531)
Fluid intelligence	−1.183*** (0.343)	−1.207*** (0.331)	−2.186*** (0.589)	−2.242*** (0.604)
Need for cognition		−0.204 (0.361)		0.067 (0.657)
Risk-loving		0.312** (0.120)		−0.097 (0.219)
Delay of gratification		−0.430 (0.686)		−1.128 (1.248)
Conscientiousness		−0.007 (0.390)		−1.231 (0.703)
Agreeableness		−1.035** (0.354)		−0.499 (0.640)
Openness		0.109 (0.358)		0.232 (0.656)
Neuroticism		−0.088 (0.352)		0.514 (0.640)
Extraversion		0.340 (0.344)		−0.208 (0.623)
Constant	12.213 (8.010)	18.946* (7.685)	14.347 (13.708)	18.491 (13.945)
Session fixed effects	Yes	Yes	Yes	Yes
$sd(u_j)$	8.25	7.69	14.71	14.73
$sd(e_{ij})$	6.05	6.05	8.39	8.36
Students	796	792	794	790
Observations	2,205	2,194	2,239	2,231

This table shows the results of a two-level hierarchical regression model grouped at the student level, where the dependent variable is the error rate 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.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE A5: Mean differences of the AX-CPT reaction time (milliseconds) for correct responses and accuracy rate (percentage) by gender

	Boys		Girls		Welch's unequal variances t-test		
	N	Mean (SD)	N	Mean (SD)	P	T-value	DoF
Reaction time by condition							
No incentive: condition AY	63	853.36 (504.70)	76	932.41 (306.15)	0.279	-1.09	97.7
No incentive: condition BX	63	683.25 (388.72)	76	845.97 (385.18)	0.014	-2.50	135.6
Monetary incentive: condition AY	379	705.23 (221.42)	411	772.58 (213.96)	<.001	-4.35	782.2
Monetary incentive: condition BX	379	519.23 (301.31)	411	677.71 (280.09)	<.001	-7.65	771.8
Status incentive condition AY	106	609.63 (129.85)	124	708.72 (183.37)	<.001	-4.78	220.7
Status incentive: condition BX	106	370.51 (206.47)	124	568.70 (259.48)	<.001	-6.45	226.9
Error rate by condition							
No incentive: condition AY	65	12.50 (20.78)	78	9.79 (18.11)	0.413	0.82	128.1
No incentive: condition BX	65	10.34 (15.36)	78	11.41 (20.52)	0.724	-0.35	139.5
Monetary incentive: condition AY	382	9.45 (13.80)	416	5.55 (11.71)	0.000	4.29	750.1
Monetary incentive: condition BX	382	5.27 (13.19)	416	6.55 (17.19)	0.237	-1.18	771.9
Status incentive condition AY	106	9.74 (14.44)	124	4.91 (7.91)	0.002	3.08	156.9
Status incentive: condition BX	106	2.80 (8.21)	124	2.89 (9.41)	0.938	-0.08	227.9

Boys are faster than girls on average across all AX conditions within the monetary and status incentives, as well as under no incentives for the BX conditions. However, girls are significantly more accurate than boys in the AY condition under monetary and Status incentives.

TABLE A6: Mean differences of the Simon task congruency for reaction time for correct responses and error rate by gender

	Boys		Girls		Welch's unequal variances t-test		
	N	Mean (SD)	N	Mean (SD)	P	T-value	DoF
Reaction time by condition							
No incentive: congruent	98	642.18 (126.82)	106	775.36 (166.55)	<.001	-6.45	195.0
No incentive: incongruent	98	702.51 (165.48)	106	838.08 (181.90)	<.001	-5.57	202.0
No incentive: neutral	98	633.42 (126.45)	106	756.91 (157.73)	<.001	-6.17	196.6
Monetary incentive: congruent	380	624.71 (116.83)	414	725.97 (149.80)	<.001	-10.67	772.2
Monetary incentive: incongruent	380	670.36 (123.03)	414	800.16 (183.34)	<.001	-11.80	727.0
Monetary incentive: neutral	380	611.02 (122.03)	414	722.80 (171.09)	<.001	-10.67	750.7
Status incentive: congruent	98	590.80 (185.81)	111	627.50 (131.53)	0.105	-1.63	172.1
Status incentive: incongruent	98	627.52 (172.53)	111	709.58 (249.80)	0.006	-2.79	196.1
Status incentive: neutral	98	554.79 (116.92)	111	615.10 (137.80)	<.001	-3.43	207.8
Error rate by condition							
No incentive: congruent	98	7.30 (14.12)	106	10.97 (19.88)	0.127	-1.53	189.7
No incentive: incongruent	98	14.37 (16.16)	106	17.32 (19.79)	0.244	-1.17	199.0
No incentive: neutral	98	9.15 (14.65)	106	11.96 (21.32)	0.271	-1.11	186.9
Monetary incentive: congruent	380	7.42 (15.11)	416	8.04 (18.48)	0.606	-0.52	784.5
Monetary incentive: incongruent	380	13.34 (16.02)	416	14.12 (19.59)	0.536	-0.62	784.6
Monetary incentive: neutral	380	8.08 (15.21)	416	8.86 (18.63)	0.516	-0.65	784.4
Status incentive: congruent	98	5.57 (13.66)	112	5.59 (16.43)	0.993	-0.01	207.5
Status incentive: incongruent	98	10.13 (15.67)	112	10.25 (17.22)	0.958	-0.05	207.7
Status incentive: neutral	98	5.52 (14.24)	112	5.83 (16.44)	0.886	-0.14	208.0

Boys are, on average, faster than girls across all Simon congruency conditions, but not more accurate. This confirms that boys use a speed strategy in the Simon task, as there are no statistically significant differences in error rate by congruency.

TABLE A7: Regression results for the effect of gender, incentives, and gender-incentive interaction on the AX-CPT proactive behavioral index (PBI) for reaction time for correct responses and error rate

	Dependent variable:			
	Reaction time (standardized)		Error rate (standardized)	
	(1)	(2)	(3)	(4)
Gender = boy	0.552*** (0.074)	0.527*** (0.077)	0.183* (0.079)	0.158 (0.082)
Incentive = no incentive (ref. = monetary)	0.128 (0.104)	0.122 (0.104)	0.054 (0.125)	0.022 (0.126)
Incentive = status (ref. = monetary)	0.122 (0.085)	0.126 (0.085)	0.107 (0.105)	0.114 (0.105)
Gender = boy x Incentive = no incentive (ref. monetary)	−0.193 (0.147)	−0.213 (0.149)	−0.175 (0.172)	−0.148 (0.174)
Gender = boy x Incentive = status (ref. monetary)	0.127 (0.116)	0.121 (0.117)	−0.085 (0.140)	−0.099 (0.140)
Age (months)	−0.008 (0.005)	−0.007 (0.005)	−0.006 (0.005)	−0.008 (0.005)
Mouse use	0.043 (0.027)	0.036 (0.027)	0.009 (0.027)	0.004 (0.028)
Videogaming	0.047 (0.028)	0.049 (0.028)	0.025 (0.028)	0.028 (0.029)
Fluid intelligence	0.077* (0.033)	0.061 (0.034)	0.115*** (0.034)	0.107** (0.035)
Need for cognition		0.097* (0.039)		0.040 (0.039)
Risk-loving		0.0001 (0.011)		0.002 (0.012)
Delay of gratification		0.044 (0.065)		−0.039 (0.067)
Conscientiousness		0.017 (0.037)		0.038 (0.037)
Agreeableness		−0.025 (0.036)		−0.101** (0.036)
Openness		−0.004 (0.034)		−0.003 (0.035)
Neuroticism		0.068* (0.033)		0.012 (0.034)
Extraversion		−0.001 (0.032)		−0.011 (0.033)
Constant	0.280 (0.660)	0.225 (0.666)	0.526 (0.681)	0.693 (0.686)
Session fixed effects	Yes	Yes	Yes	Yes
$sd(u_j)$	0.57	0.56	0.39	0.38
$sd(e_{ij})$	0.7	0.7	0.89	0.89
Students	792	789	796	792
Observations	1,156	1,151	1,167	1,161

This table shows the results of a two-level hierarchical regression model grouped at the student level, where the dependent variable is the proactive behavioral index (PBI) for average reaction time on correct responses and error rate, standardized. 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. 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.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A8: Regression results for the effect of gender, incentives, and gender-incentive interaction on average reaction time for correct responses (milliseconds) on the AX-CPT task, by trial condition (for Figure 8)

	Dependent variable:			
	Reaction time (ms), B-X condition		Reaction time (ms), A-Y condition	
	(1)	(2)	(3)	(4)
Gender = boy	−165.853*** (22.018)	−157.911*** (22.862)	−76.484*** (16.755)	−73.212*** (17.399)
Incentive = no incentive (ref. = monetary)	117.461*** (26.605)	118.324*** (26.623)	132.907*** (20.268)	133.033*** (20.313)
Incentive = status (ref. = monetary)	−73.608*** (21.417)	−75.028*** (21.421)	−58.321*** (16.282)	−58.855*** (16.310)
Gender = boy x Incentive = no incentive (ref. monetary)	−66.682 (38.236)	−61.538 (38.573)	−30.172 (29.447)	−28.652 (29.764)
Gender = boy x Incentive = status (ref. monetary)	−3.993 (30.135)	−1.093 (30.156)	−3.223 (22.962)	−2.081 (23.012)
Age (months)	2.252 (1.777)	2.034 (1.785)	0.776 (1.342)	0.825 (1.351)
Mouse use	−7.847 (8.276)	−5.264 (8.338)	0.687 (6.248)	2.405 (6.301)
Videogaming	−24.410** (8.494)	−27.522** (8.673)	−18.281** (6.406)	−21.241** (6.537)
Fluid intelligence	−54.749*** (10.300)	−49.232*** (10.509)	−48.482*** (7.791)	−45.838*** (7.946)
Need for cognition		−31.824** (11.911)		−22.067* (8.997)
Risk-loving		−1.957 (3.549)		−1.511 (2.680)
Delay of gratification		−20.832 (20.220)		−3.252 (15.288)
Conscientiousness		−7.989 (11.293)		−12.914 (8.495)
Agreeableness		−1.756 (10.932)		−8.652 (8.293)
Openness		−6.377 (10.626)		−3.185 (8.045)
Neuroticism		−21.170* (10.265)		−18.401* (7.770)
Extraversion		−8.886 (9.874)		−3.659 (7.477)
Constant	495.398* (225.300)	545.781* (227.417)	734.667*** (169.991)	748.901*** (172.032)
Session fixed effects	Yes	Yes	Yes	Yes
$sd(u_j)$	204.66	203.47	154.58	153.28
$sd(e_{ij})$	174.17	174.12	132.8	133.01
Students	791	788	795	791
Observations	1,158	1,153	1,159	1,153

This table shows the results of a two-level hierarchical regression model grouped at the student level, 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). 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.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE A9: Regression results for the effect of gender, incentives, and gender-incentive interaction on error rate (percentage) on the AX-CPT task, by trial condition (for Figure 8)

	Dependent variable:			
	Error rate (percent), B-X condition		Error rate (percent), A-Y condition	
	(1)	(2)	(3)	(4)
Gender = boy	−0.663 (1.165)	−0.634 (1.188)	3.211** (1.076)	2.602* (1.100)
Incentive = no incentive (ref. = monetary)	1.479 (1.220)	1.651 (1.214)	3.400* (1.380)	3.365* (1.374)
Incentive = status (ref. = monetary)	−1.604 (0.981)	−1.627 (0.976)	−0.125 (1.125)	−0.116 (1.121)
Gender = boy x Incentive = no incentive (ref. monetary)	2.798 (1.773)	2.839 (1.778)	0.949 (1.972)	1.160 (1.978)
Gender = boy x Incentive = status (ref. monetary)	0.527 (1.401)	0.576 (1.394)	−1.020 (1.571)	−1.042 (1.564)
Age (months)	−0.180 (0.096)	−0.239* (0.095)	−0.123 (0.085)	−0.176* (0.084)
Mouse use	0.498 (0.449)	0.338 (0.444)	0.338 (0.397)	0.049 (0.394)
Videogaming	0.322 (0.461)	0.114 (0.461)	0.523 (0.408)	0.466 (0.409)
Fluid intelligence	−2.138*** (0.561)	−2.123*** (0.560)	−0.658 (0.494)	−0.737 (0.495)
Need for cognition		−0.433 (0.634)		−0.117 (0.562)
Risk-loving		0.178 (0.189)		0.352* (0.167)
Delay of gratification		−0.909 (1.078)		−0.958 (0.954)
Conscientiousness		−1.161 (0.599)		0.216 (0.532)
Agreeableness		0.039 (0.584)		−1.969*** (0.516)
Openness		−0.288 (0.568)		0.228 (0.502)
Neuroticism		−0.789 (0.548)		0.139 (0.485)
Extraversion		0.647 (0.527)		0.890 (0.466)
Constant	34.365** (12.217)	43.433*** (12.117)	18.594 (10.805)	27.011* (10.752)
Session fixed effects	Yes	Yes	Yes	Yes
$sd(u_j)$	12.04	11.66	9.32	8.99
$sd(e_{ij})$	7.92	7.89	9.23	9.2
Students	796	792	795	791
Observations	1,167	1,161	1,166	1,160

This table shows the results of a two-level hierarchical regression model grouped at the student level, 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.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE A10: Regression results for the effect of gender, incentives, and gender-incentive interaction on the Simon effect on average reaction time and error rate on the Simon task

	Dependent variable:			
	Reaction time (ms)		Error rate (percent)	
	(1)	(2)	(3)	(4)
Gender = boy	−22.002** (7.084)	−21.720** (7.313)	0.214 (0.519)	0.239 (0.536)
Incentive = no incentive (ref. = monetary)	−10.467 (11.835)	−10.839 (11.874)	1.268 (0.766)	1.272 (0.768)
Incentive = status (ref. = monetary)	−4.767 (11.937)	−4.199 (11.971)	−1.068 (0.761)	−1.031 (0.763)
Gender = boy x Incentive = no incentive (ref. monetary)	32.989* (15.449)	33.989* (15.514)	1.215 (1.027)	1.218 (1.030)
Gender = boy x Incentive = status (ref. monetary)	1.262 (15.303)	−0.112 (15.364)	−0.076 (1.012)	−0.155 (1.016)
Age (months)	−0.259 (0.547)	−0.340 (0.555)	−0.006 (0.041)	−0.006 (0.042)
Mouse use	−0.425 (2.562)	−0.785 (2.610)	0.168 (0.192)	0.127 (0.195)
Videogaming	5.869* (2.589)	5.319* (2.679)	0.051 (0.195)	0.007 (0.200)
Fluid intelligence	−4.468 (3.211)	−3.791 (3.308)	−0.611* (0.241)	−0.634* (0.247)
Need for cognition		−4.806 (3.732)		−0.246 (0.279)
Risk-loving		1.614 (1.121)		0.090 (0.083)
Delay of gratification		−2.306 (6.387)		0.729 (0.475)
Conscientiousness		−0.429 (3.546)		−0.178 (0.264)
Agreeableness		−2.277 (3.432)		−0.126 (0.256)
Openness		2.478 (3.405)		0.410 (0.253)
Neuroticism		3.496 (3.300)		0.286 (0.245)
Extraversion		3.701 (3.065)		0.226 (0.229)
Constant	91.122 (70.291)	106.581 (71.703)	7.176 (5.243)	7.246 (5.331)
Session fixed effects	Yes	Yes	Yes	Yes
$sd(u_j)$	0	0	3.12	3.09
$sd(e_{ij})$	97.37	97.6	6.12	6.13
Students	793	789	794	790
Observations	1,201	1,197	1,205	1,201

This table shows the results of a two-level hierarchical regression model grouped at the student level, where the dependent variable is the Simon effect for average reaction time on correct responses and error rate per incentive condition. A positive 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.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

APPENDIX B

Table B1. Personality measures (continued)

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 preferences	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).*