Journal of Development Economics 150 (2021) 102615

Contents lists available at ScienceDirect

Journal of Development Economics

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

Regular Article

The Effects of Chess Instruction on Academic and Non-cognitive Outcomes: Field Experimental Evidence from a Developing Country☆

Asad Islam [a,*], Wang-Sheng Lee [b], Aaron Nicholas [c]

[a] Monash University, Australia
[b] Deakin University and IZA, Bonn, Australia
[c] Deakin University, Australia

ARTICLE INFO

JEL classification:
C93
D80
I21

Keywords:
Chess training
Math
Non-cognitive outcomes
Risk
Randomized experiment

ABSTRACT

We conduct a randomized field experiment to investigate the benefits of an intensive chess training program undertaken by primary school students in a developing country context. We examine the effects on academic outcomes, and a number of non-cognitive outcomes: risk preferences, patience, creativity and attention/focus. Our main finding is that chess training reduces the level of risk aversion almost a year after the intervention ended. We also find that chess training improves math scores, reduces the incidence of time inconsistency and the incidence of non-monotonic time preferences. However, these (non-risk preference) results are less conclusive once we account for multiple hypothesis testing. We do not find any evidence of significant effects of chess training on other academic outcomes, creativity, and attention/focus.

1. Introduction

Chess is a popular game played by millions worldwide. Its popularity is at least in part attributable to its perceived effect on cognitive skills in general, and math ability in particular. In recent years, chess coaching for children has become increasingly popular in developed countries.[1] The European Parliament has expressed a favorable opinion on using chess courses in schools as an educational tool (Binev et al., 2011). In 2014, School Library Journal's best education pick of the year was a chess-related product called Yamie Chess, which is backed by Harvard and MIT academics.[2] The benefits of playing chess regularly have been suggested in a documentary that focuses on an inner-city school in New York, and two European countries – Armenia and Poland – have even made chess instruction compulsory in their primary-school curricula.[3] More recently, the city of Bremen in Germany has decided to introduce 1 h of chess per week as a subject in primary schools in 2020, an issue covered widely in the German press.[4]

Parents and teachers generally view chess as a highly regarded

☆ We thank participants at the Labour Econometrics Workshop 2017 in Auckland, the 2017 Asian and Australasian Society of Labour Economics (AASLE) conference in Canberra, the 17th IZA-SOLE Transatlantic Meeting of Labor Economists (2018) in Buch/Ammersee, and workshops at Deakin University, Monash University and Singapore Management University for their comments. Foez Mojumder provided excellent research support. This research would also have not been possible without cooperation from the Department of Primary Education (DPE) in Bangladesh, local school teachers, NGO partner Global Development Research Initiative (GDRI) and our chess coaches. We are grateful to Deakin University and Monash University for their generous research funding.

* Corresponding author. Department of Economics, PO Box 197, Caulfield East, Victoria, 3145, Australia.
E-mail address: Asadul.Islam@monash.edu (A. Islam).

[1] For example, in the US, the Chess Club and Scholastic Center of St. Louis (a 6000-square-foot, state-of-the-art chess center widely recognized as the premier chess facility in the country and one of the best in the world) helps provide chess-coaching services to many elementary and middle schools in St. Louis, Missouri. For the list of schools, see: https://saintlouischessclub.org/education/partners-education (accessed March 27, 2017).

[2] Yamie Chess features an interactive coloring math comic book written by experienced math teachers for K-8 supplemental math learning.

[3] The documentary film Brooklyn Castle (2012) highlights the after-school chess program in an inner-city public school in Brooklyn, New York and how they became the first middle-school team to win the U.S. Chess Federation's national high school championship. For compulsory chess instruction in Armenia, see: https://www.theguardian.com/world/2011/nov/15/armenia-chess-compulsory-schools (accessed March 27, 2017). For Poland, see: http://cis.fide.com/en/chess-news/325-poland-chess-in-all-schools (accessed March 27, 2017).

[4] See https://en.chessbase.com/post/chess-makes-smart-scholastic-tournament-in-bremen-2019 (accessed 2 July 2019).

https://doi.org/10.1016/j.jdeveco.2020.102615
Received 16 October 2019; Received in revised form 16 November 2020; Accepted 26 December 2020
Available online 9 January 2021

extracurricular activity in primary school.[5] However, to date, there is hardly any study rigorously examining the effects of chess instruction. An exception is Jerrim et al. (2018), who report results from a randomized controlled trial (RCT) conducted in the UK to evaluate the impact of teaching children chess on academic outcomes. Contrary to popular belief, they found no evidence that teaching children chess improved their math ability. There were also no impacts on reading and science.

In this paper, we conduct an RCT to examine the effects of intensive chess lessons among grade five students in a developing country. We follow the curriculum approved by the World Chess Federation. We differ from Jerrim et al. (2018) and the literature on the impact of chess training on two counts. First, we study the link between chess and non-cognitive outcomes such as risk preferences, patience, creativity, attention and focus. Second, we examine the effects of chess learning in a developing country context. Children in our experiment come from rural primary schools in Bangladesh who do not have previous experience playing chess. Our setting is particularly well-suited to test the benefits of a chess training program because unlike children in urban areas in a developed country, most children in rural areas in a developing country will never have been exposed to the game of chess before, much less any other cognitively demanding games.[6]

We first examine the effects of chess training on test scores. Our primary outcomes for test scores come from a standardized, compulsory public exam that all fifth-grade students in Bangladesh must take – the Primary School Certificate (PSC) exam – which took place 9–10 months after the completion of chess training. While we are particularly interested in examining the effects on math test scores because of the perceived math benefits from playing chess, we also examine the results for students' first language and science.[7]

Chess is often regarded as a game reflecting real life (Franklin, 1786) and teaching children how to play chess in a prescribed systematic fashion might also help in their development of important non-cognitive outcomes. Therefore, we pay special attention to the collection of extensive data on non-cognitive outcomes to examine the effects of chess training. In particular, we measure risk preferences, patience, creativity and attention/focus.

Chess, through the formation of strategies, can be useful for the conceptualization and calculation of risks.[8] For example, chess players often sacrifice pawns, bishops, knights, rooks, or queens if it helps checkmate the opponent's king and win the game. Such sacrifices are inherently risky because if one's calculations are faulty, the sacrifice could prove to be fatal, eventually leading to a quick loss of the game. Gambits and sacrifices can be made during any of the three phases of a chess game – opening, middlegame, or endgame. Such an association between risk taking and chess playing is, for example, utilized to study the link between risk preferences and attractiveness (Dreber et al., 2013) through behavior in chess.[9] Chess playing styles have also been used as a proxy for differences in risk appetites across civilisations (Chassy and Gobet, 2015). Thus, learning how to play chess and gaining an appreciation of basic chess strategy can help in the development and articulation of risk preferences in children.

The role of risk in chess can be seen in how computer chess software function. Computers are now better at chess than even the world's strongest grandmasters, and we can learn more about how the game is optimally played from studying their games and using them for analysis.[10] When a computer plays chess, there is no element of psychology involved. The computer never gets tired and does not care who it is playing. There is, however, a setting available in many commercial chess programs that allow one to set the "risk level" in the software. This setting changes the style of play of the computer opponent, who might play in a more risky style (i.e. have a higher tendency to sacrifice and attack) or in a less risky style (i.e. have a tendency to focus on longer term strategic objectives). A risky style leads to more wins and losses, with fewer drawn games, whereas a less risky style will lead to relatively more drawn chess games, and fewer wins and losses. It has some similarities to conservative and risky styles in investing, which is why there is anecdotal evidence that many firms in the financial industry view a competitive chess background as a positive factor when hiring.

The chess syllabus used for our experiment (see Online Appendix 1 in the paper) includes coverage of risk related concepts such as using risky openings (the Scholar's mate, otherwise known as the four-move checkmate) and making sacrifices. Going for checkmate early in the game by moving one's queen out early is considered to be a risky strategy because if it does not work, it can backfire and lead to a disadvantage in one's position (e.g. other pieces are undeveloped). However, sacrificing can be an optimal strategy when one is already in a lost position. As there is nothing to lose, one can risk everything to try to checkmate the opponent. Of course, being able to calculate and appreciate risks may either increase or decrease risk aversion: the risk hypothesis we test is therefore two-sided.

Furthermore, chess might help teach children to be more patient, more focused, and have more self-control.[11] It can potentially motivate children to become willing problem-solvers, able to spend hours quietly immersed in logical thinking. Chess can also be a useful tool to teach the importance of forward-looking behavior. An important element in chess is the evaluation process, i.e., one needs to look a few steps ahead during a chess game and consider and evaluate alternative scenarios. Chess can teach children how to focus and visualize by imagining a sequence of events before it happens. The schematic thinking approach in chess resembles trees and branches in sequential-decision analysis and might also be useful and possibly transferable to math skills, as has been emphasized previously (Scholz et al., 2008; Trinchero and Sala, 2016).

In addition to children's risk preferences and time preferences, we also investigate whether undertaking intensive chess lessons can affect

[5] E.g., see the testimonials at: https://www.chessinschools.co.uk/chesstimonials (accessed 14 Oct 2020).

[6] Jerrim et al. (2016, p. 46) report in their study that chess playing activity at their baseline was 48% in treatment schools and 45% in control schools. Such levels are not surprising given that their study was based in an urban developed country setting.

[7] Studies in the education literature (e.g., Scholz et al., 2008; Trinchero and Sala 2016) also suggest that chess improves children's math skills because the game has some elements in common with the mathematical domain and because it promotes suitable habits of mind.

[8] Risk aversion is a trait typically associated with welfare-relevant, later life outcomes. Hence, its detection (and potential manipulation) from an early age may be of policy interest. Davis and Eppler-Wolff (2009) argue that parents need to understand the significance of risk-taking as a teaching experience for children. Higher risk aversion has been shown to be detrimental to key household decisions, such as choice of occupation, portfolio selection and moving decisions (Guiso and Paiella, 2008). On the other hand, higher risk aversion has also been linked to less disciplinary referrals and a higher probability of high school completion (Castillo et al., 2018).

[9] There, risk taking in chess is measured by exploiting a standardized classification of opening moves and expert assessments. As chess players in our setting are beginners who are unlikely to have a well thought out opening repertoire (a regular set of openings they use to start the game), it is not possible to adopt such an approach to measure risk preferences.

[10] The strongest commercially available chess program, Stockfish 12, has an approximate chess rating of 3500, compared to the world chess champion, Magnus Carlsen, who has a current chess rating of 2863.

[11] Becker and Mulligan (1997) suggest that observed differences in time preferences are not innate and that the evolution of these preferences may be endogenous. This implies that children could be taught to be more forward thinking. If patience and other time preference-related characteristics of children vary across gender or demographic groups, different educational paths and career outcomes may occur. For example, Castillo et al. (2011) find that boys are more impatient than girls, and that impatience has a direct correlation with disciplinary referrals – behavior that has been shown to be predictive of economic success.

children's creativity and attention/focus. Although there is some debate over whether creativity is an aspect of intelligence or a personality trait, several studies have shown that creativity can be experimentally manipulated (see Runco and Sakamoto, 1999, for a review). The ability to focus on a task at hand is also a useful non-cognitive outcome that chess might be able to nurture. Attention is considered to be a major part of working memory, responsible for the control of flow of information, switching between tasks and selection of relevant stimuli and inhibition of irrelevant ones (Travis, 1998). The study of the development of attention occupies a central place in cognitive developmental psychology, and we use frequently used tests for focus/attention in our evaluation.

This paper is relevant to several sub-fields of economics. First, there has been much recent interest in the development of non-cognitive skills in children and their importance in later life outcomes in the economics literature. Non-cognitive skills have been shown to be very important for a host of outcomes, including schooling, social behaviors, drugs, smoking, truancy, teenage pregnancy, involvement in crime, and labor market success (Heckman et al., 2006; Carneiro et al., 2007). In addition, although a large literature in experimental economics has focused on the role of risk preferences in explaining life outcomes (e.g. Dohmen et al., 2011; Sutter et al., 2013), surprisingly little is known about differences in risk preferences at an early age and how these preferences are developed, or how they may alter the life paths of students (Andreoni et al., 2019a). Chess may be of particular interest to policymakers who are interested in identifying programs that can provide early stimulation and help develop such important "soft" life skills in children during their formative years. Second, in the program evaluation literature, there is increasing interest in evaluating interventions that have the potential to be scaled up (Banerjee et al., 2017). Given resource and institutional constraints, the effectiveness of scalable interventions that can be deployed which can form the basis of public policy is to date not well explored. As introducing chess as a subject in school will not be very costly, the educational intervention we examine in this paper most certainly has the potential to be scaled up if smaller proof-of-concept studies such as this paper show positive results. Indeed, some countries like Armenia and Poland and cities like Bremen in Germany have already made the decision to scale up despite scant rigorous experimental evidence on the effects of chess instruction. Furthermore, neighbouring India is making progress in introducing chess to the school curriculum. India currently has about 17 million children involved nationwide, especially in the states of Gujarat and Tamil Nadu where chess is part of the curriculum.[12] There is a possibility that chess will be introduced to schools around the country.[13] So far, largely due to the continuing efforts of the All India Chess Federation (AICF), over 1000 schools in the Delhi region in India have already adopted chess as a sport in the past few years.[14]

Overall, the main finding in our paper is that chess training has a significant effect on reducing the level of risk aversion almost a year later. Based on conventional p-values and wild bootstrap p-values, we also find that chess training has a positive impact on math scores in the national exam and reduces the incidence of both time inconsistency and non-monotonic time preferences. However, the results are less conclusive once we account for multiple hypothesis testing using the false discovery rate (FDR). Effects of chess training on the other academic outcomes, creativity, and attention/focus were not statistically significant.

The paper is structured as follows. Section 2 briefly discusses how chess can translate to learning outcomes. Section 3 provides information on the intervention. Section 4 describes the data and the academic and non-cognitive outcomes measured in this study. Section 5 presents the results of the intervention. Section 6 concludes.

2. Chess and learning outcomes

Transfer of learning occurs when a set of skills acquired in one domain generalizes to other domains or improves general cognitive abilities. Little is known about the extent to which chess skills transfer to other domains of learning. Although near transfer (i.e., transfer that occurs between closely related domains, such as math and physics) might be possible, several studies have shown that chess players' skills tend to be context-bound, suggesting that it is difficult to achieve far transfer from chess to other domains. For example, it has been found that memory for chess positions fails to transfer from chess to digits both in adults and children (Schneider et al., 1993), and that chess players' perceptual skills do not transfer to visual memory of shapes (Waters et al., 2002). In the Tower of London task, a well-known test for executive functioning in which participants solve 16 four-, five-, and six-move problems each, chess planning skills did not improve the ability of chess players to solve these tasks (Unterrainer et al., 2011). Levitt et al. (2011) find that the ability to transfer backward induction prowess from the chess board to experimental games is quite sensitive to the particulars of the game in question.

We are not aware of any studies that have explored in depth the link between chess skills and non-cognitive outcomes, although some previous work has focused on the effects of chess on focused attention and metacognition (Scholz et al., 2008), despite an observation made more than two centuries ago from a notable chess enthusiast. The renowned inventor and U.S. founding father Benjamin Franklin wrote the following in a magazine essay, "The Morals of Chess" (1786):

"The game of chess is not merely an idle amusement. Several very valuable qualities of the mind, useful in the course of human life, are to be acquired or strengthened by it, so as to become habits, ready on all occasions. For life is a kind of chess, in which we have often points to gain, and competitors or adversaries to contend with, and in which there is a vast variety of good and ill events, that are, in some degree, the effects of prudence or the want of it."

Franklin goes on to suggest in his essay that by playing chess, one may learn foresight (considering consequences before taking action, i.e., planning chess moves), circumspection (seeing the big picture, i.e., surveying the whole chess board, the relations among pieces and situations, and the dangers the pieces are exposed to) and caution (not to make moves too hastily and to abide by all the consequences of one's rashness). Circumspection implies that a person thinks carefully before doing or saying anything, a quality that is expected to be correlated with patience. Combining foresight and caution implies a person will learn to take calculated risks, thereby linking chess playing style and skill with risk preferences.

3. The program and the data

3.1. The chess intervention

The intervention took place in primary schools in rural communities in two districts- Khulna and Satkhira—in southwest Bangladesh in January–February 2016. Our chess experiment is a clustered randomized controlled trial with randomization at the school level involving fifth grade students (10 years old on average) in 2016 in 16 primary schools.[15] These schools were chosen randomly from a set of more than 200 schools in those regions. The sampling frame included all schools in the sub-

[12] See the October 2015 Financial Times article "Chess can improve children's lives" (https://www.ft.com/content/a7686122-524c-11e5-b029-b9d50a74fd14) (accessed 14 Oct 2020).

[13] See https://chessbase.in/news/International-Chess-in-Education-Conference-Delhi-2019 (accessed 14 Oct 2020).

[14] See https://www.hindustantimes.com/delhi-news/delhi-makes-winning-moves-emerges-as-new-chess-coaching-hub/story-q93uEyjLbvXmJNBQ8Iza8O.html (accessed 14 Oct 2020).

[15] Computer randomization of schools was implemented using a pre-specified seed.

districts where both treatment and control schools were located.[16] The location of the 16 treatment and control schools can be seen in Fig. 1. In general, the treatment schools and control schools were geographically spread out such that no two schools (either treatment or control) are close to each other, with each of them at least 5 km apart. In the context of rural Bangladesh where walking is the predominant mode of transport and where children tend to play with their neighbors, such distance between schools effectively means that program spillovers to control schools is very unlikely.

The schools were randomly divided into two groups: eight in the treatment group and eight in the control group.[17] Students in the treatment schools received 12 days of chess training (spread over three weeks). A pre-program baseline test of chess knowledge suggests that most children in our analysis sample did not know how to play chess. The chess knowledge test comprised a series of four questions. The first question asked: "Do you know how to play chess?" Children who responded "Yes" or "A little bit" were further probed with further specific questions about "which is the most powerful piece on the chess board" and how chess pieces move and capture in two chess positions that were provided in diagrams. Only one child answered all three questions on basic knowledge of the chess rules correctly, and 4.22% in the control group and 2.75% in the treatment group answered at least two out of the three questions correctly. This latter difference was not statistically significant (p-value = 0.514). Training sessions were conducted separately at each school at the beginning of the academic year in January–February of 2016. The chess instruction involved teaching the rules of chess and basic chess strategy.

The lesson plan was based on free instructional chess materials available from the Chess in Schools Commission of the World Chess Federation (FIDE) (see Online Appendix 1 for the syllabus used for the chess lessons). This lesson plan was developed by chess experts specifically for use as course material in primary schools. We hired two instructors to deliver the entire chess program to the eight treatment schools.[18] Both instructors are qualified chess coaches and have extensive experience teaching chess to children. One is a FIDE master and former national champion of Bangladesh, and the other is a seven-time divisional champion and a chess coach by profession. They both also have formally been appointed as trainers by the National Chess Federation in Bangladesh.

The 12-day training program for students in all the treatment schools was spread over three weeks and conducted during regular school hours. The program was first implemented in four treatment schools during three weeks in January 2016, with a further four treatment schools getting exposure to the program in the subsequent three weeks. In the first week of training (three days of training), each instructor conducted one session per day at 8:00 a.m. in the morning. In the second week of training (five days of training), each instructor conducted two sessions per day with the first session at 8:00 a.m. in the morning and the second session at 12:00 p.m. in the afternoon. In the third week of training (four days of training), each instructor continued to conduct two sessions per day with the first session at 8:00 a.m. in the morning and the second session at 12:00 p.m. in the afternoon.

After the 2-h chess lesson for each day was completed, students were allowed to practice chess by playing against each other for an additional 30 min. To carry out the practice sessions, each instructor was supported by several field staff who are amateur chess enthusiasts. During the training sessions each pair of students received a chess set to use in class.[19] The intervention involved providing a total of 24 h of chess instruction (daily 2-h lessons spread over 12 days) and about 6 h of supervised chess practice playing against an opponent, which allowed the students to apply any new skills they had just learned. Thus, the students received approximately 30 h of chess training – above the 25 h Sala and Gobet (2016) report as the threshold above which chess instruction produces substantial effects.

In general, there was little or no disruption to normal academic activities in both the treatment and control schools due to either the program or our elicitation of outcomes from the survey instruments. This was possible due to several factors. First, the school curriculum during the start of the school year (January and February) is relatively light, as contact time with students at the beginning and at the end of the school year is usually dominated by administrative and non-teaching activities. This includes organizing the demanding logistics of registering students, receiving and distributing teaching materials, and understanding new government policies or programs. Throughout January, as part of the annual National Education Week (a government information campaign designed to encourage parents to enroll their children in school), teachers are expected to recruit students by making visits to homes, markets, and other public places to meet parents.

Second, unlike primary schools in developed countries or in urban settings, effective instructional time in rural primary schools in Bangladesh is relatively short (Tietjen et al., 2004; Islam 2019). There are several contributing factors: (i) Teacher absenteeism is a major issue in rural Bangladesh[20]; (ii) Instructional time at rural schools is further reduced by the effective hours of operation. Even if teachers at rural schools are present, they were more likely to arrive late for school or depart before the official end of the school day than their urban counterparts because of domestic chores (predominately female) and income-generating activities (all males). As a result, Tietjen et al. (2004) found that teaching or "instruction" occupied on average 63 percent of the class time in the classes they observed.

Further, given the frequent later than official school start times in rural primary schools, the scheduling of our classes before the start of school day minimized the displacement of day-to-day academic studies. Hence, to the extent that any displacement occurs, the chess training program is most likely displacing idle class time or unstructured play activities that the students in the control group were playing, such as Ekka-dokka (hopscotch), Gulikhela (game of marbles), Ha-du-du (game of tag), and Kanamachi (a game where a blindfolded participant tries to catch other players).

Student feedback on the chess lessons was very positive. Of the 248 students (out of 294) respondents in the treatment group who provided feedback on the chess lessons, all of them said they liked playing chess, and 99.2% said they would like more chess lessons. In addition, 94.5% of the children said that during Week 1, they played or discussed chess with at least one classmate outside the chess program; the percentage remained high in Week 2 (87.5%). The chess sets used in the training program were donated to each respective school at the end of the three-week training program so that the children could continue playing and practising chess after lessons had ended. The students' interest in chess

[16] One of the co-authors (Islam) spent his childhood and attended primary and secondary school in that area. The schools are typical of many parts of rural Bangladesh. The area was chosen because of the author's local knowledge and contacts at the schools and among district-level administrators, who helped facilitate logistics for implementing the intervention.

[17] During the study's design phase, while randomization at the class level was considered and deemed preferable, it was ruled out for several reasons. First, there is the possibility of contamination between treatment and control group classes. For instance, when one class is receiving the intervention, students from other classes might want to join in. Second, most schools in rural Bangladesh only have one class of students for each grade.

[18] One of the co-authors of the paper (Lee) is also a national master in chess and helped ensure the suitability of the syllabus for the intervention.

[19] Some pictures of the field setting can be found in Online Appendix 2, in which normal classrooms have been used to conduct the chess lessons. Some schools have double shifts, where fifth-grade students start classes in the afternoon. We scheduled chess lessons to start later in these schools.

[20] For example, Chaudhury et al. (2006) find that 16 percent of teachers are absent on a given school day, and 23.5 percent were absent once out of two visits in a school.

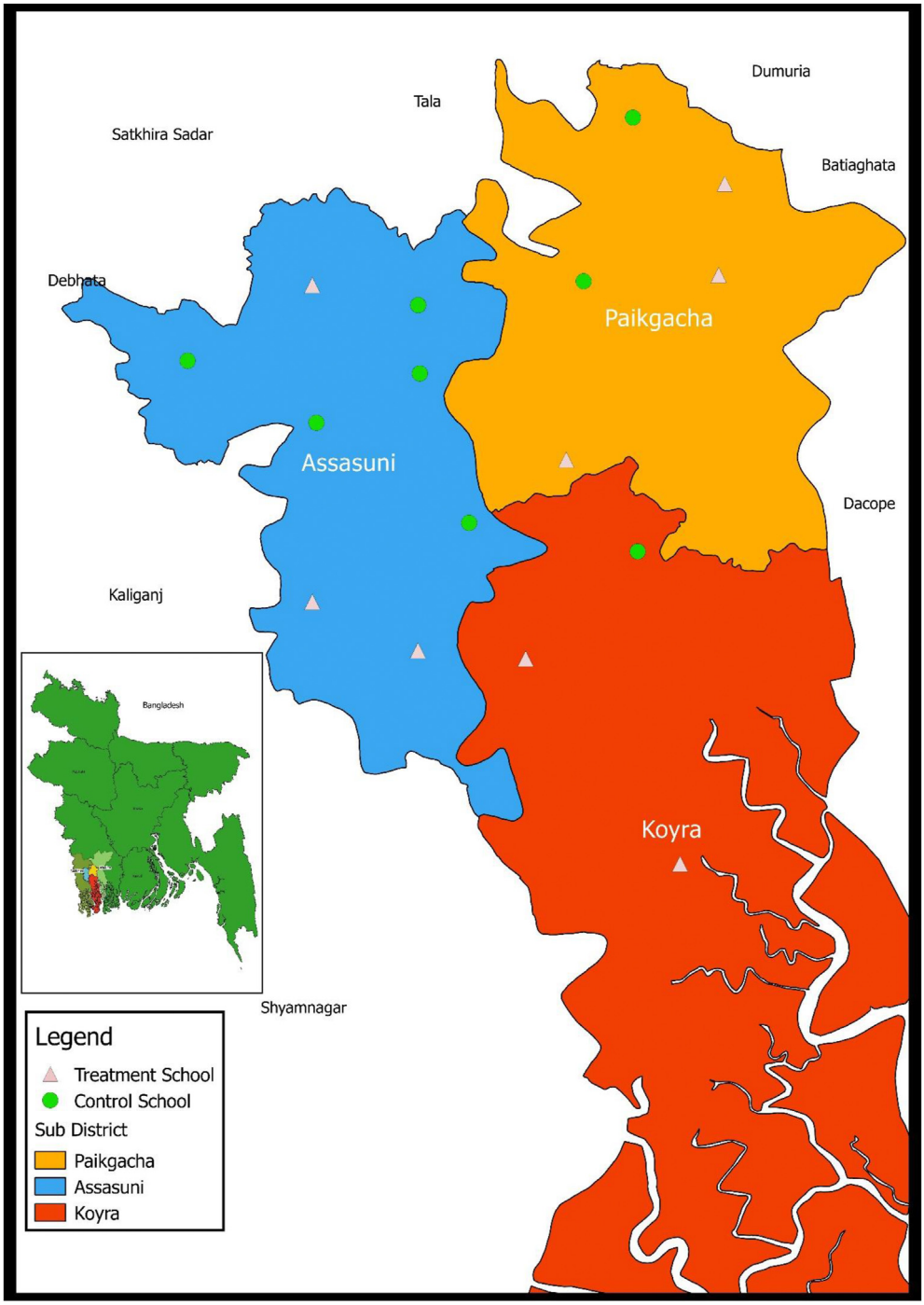


Fig. 1. Location of treatment and control schools in Bangladesh.

does not appear to be transitory. When we checked to see whether treatment-group members were still playing chess 9–10 months later, we found that 94.3% of them had played chess with a classmate during the previous week, and 87.5% of them had played chess with other friends or relatives during the previous week.

Before the chess training program launched, a household survey was carried out in November and December 2015 to collect some basic household information, including demographic profiles of the children and their parents. The respondents were parents of the children participating in the chess experiment. We also tested their pre-program math skills and chess knowledge. At the end of the chess training program, we conducted tests on risk preferences, time preferences, creativity, and math skills. The risk and time preference tests were incentivized as per standard practice in experimental economics.

Fig. 2 describes the project's key timelines. Short-run outcomes (Wave 1) were measured at the end of the three-week chess training program (the day after), and longer-term outcomes (Wave 2) were measured about 9–10 months after training ended – at the end of October 2016. We also assessed whether the program had an impact on academic performance based on results from a national exam that fifth-grade

Data collected prior to the start of the program (Nov 2015 – Jan 2016)	Wave 1 data collected at the end of the three week program (Jan/Feb 2016)	Wave 2 data collected 9-10 months later (Oct/Nov 2016)
• Parent survey (Nov/Dec 2015) • Pre-program chess knowledge test (Jan 2016) • Short personality test (Jan 2016) • Pre-program math test (Jan 2016)	• Time preferences test, Wave 1 • Risk preferences test, Wave 1 • Creativity test • Post-program math test • Post-program chess knowledge test for the treatment group • Network survey for the treatment group	• Time preferences test, Wave 2 (Oct 29/30, 2016) • Risk preferences test, Wave 2 (Oct 29/30, 2016) • Attention/focus test (Oct 29/30, 2016) • Network survey for the treatment group • Primary School Certificate (PSC) national examination (Nov 20-27, 2016)

Note: The chess program was conducted from Saturday to Tuesday over a period of three weeks. Note that Friday is considered the weekly holiday in Bangladesh (equivalent to Sunday in other developed countries) and that the school week runs from Saturday to Thursday. There were a total of 12 program days where chess lessons were provided.

Fig. 2. Intervention Timeline. Note: The chess program was conducted from Saturday to Tuesday over a period of three weeks. Note that Friday is considered the weekly holiday in Bangladesh (equivalent to Sunday in other developed countries) and that the school week runs from Saturday to Thursday. There were a total of 12 program days where chess lessons were provided.

students had to take during November 20–27, 2016.

3.2. Sample and baseline balance

Based on the name list of students provided by the treatment and control schools, 704 families were approached in November and December 2015 in order to collect baseline data for the experiment. The response rate to the parent questionnaire was 594/703 = 84.4%, and a complete set of non-missing covariates were obtained for 281 treatment group members and 288 control group members (n = 569) after accounting for item non-response.

Table 1 presents the differences in means of parental and household characteristics for the treatment and control groups. There are no significant differences between treatment and control groups except for the variable indicating whether the mother is a housewife. The results suggest that the randomization process was well implemented.

The children in our sample are mostly underprivileged, with parents from relatively low socio-economic backgrounds. Approximately a third of parents did not complete primary school. In more than 86% of families, no members of the household have an education higher than 10th grade. About 64% of fathers are engaged in agriculture or day labor, another 29% work in small business activities, and 6% work in services. Almost all the mothers are housewives. The average household size is 4.4, and the monthly income is less than 8500 takas (about US $110).

The sample sizes in our regression adjusted impacts for Wave 1 presented in Tables 2, 4 and 5 are smaller than the baseline sample in Table 1. For example, the sample size for the risk preferences using Wave 1 when we regression adjust controlling for parental and household characteristics is 450/569, which is 79.1% of the grade 4 sample. The main reason for the reduction in sample from baseline to Wave 1 is students dropping out between grades 4 and 5. Note that data from the parent questionnaire was collected at the end of academic year when the students were in grade 4. However, the experiment was conducted when students progressed to the next grade at the start of the following year. Many of these students dropped out from school or could not progress to grade 5. Hence, there was some attrition from our initial baseline sample which happened before our experiment actually started.[21] In addition, a discrepancy in sample size arises when we do and do not use regression adjustment to control for parental and household characteristics as the former requires information from the parent questionnaire, which is not available for all families.[22]

High student absenteeism from schools is a big problem in Bangladesh, with more than a quarter of children aged 7–14 years missing at least one day of school in a six-day school week in the rural areas of Bangladesh (Kumar and Saqib, 2017). Tietjen, Rahman and Spaulding (2004) found based on surprise visits to government primary schools in Bangladesh that the actual percentage of students enrolled who were in attendance on the day of the visit ranged from 43 percent to 67 percent. This explains the variation in sample sizes for the various outcomes we examine.[23] As many outcomes were collected on different school days, whether an outcome was measured largely depended on whether a student attended school that day. In general, however, this attrition did not pose a problem for the integrity of the experimental design. First, we conduct a selective attrition test which determines if the mean of baseline observable characteristics differs across the treatment and control groups conditional on response status. As Tables C.1 and C.2

[21] Ahmed at al. (2007, p.12) report using administrative data that promotion rates in primary schools in Bangladesh have been largely stable over time and were between 75 and 83% for promotion from grade 4 to 5 in 1998–2004. Students need to sit for the PSC exam at the end of grade 5, and the pass rate in this exam is used to evaluate the teachers' performance. Hence, teachers try to not promote students whom they think might fail the PSC exam.

[22] This is why the regression unadjusted sample is larger than the regression adjusted sample in Tables 2–5

[23] Student absenteeism is common in many developing countries – Banerjee et al. (2007) in India for the Balsakhi Program administered by Pratham, and Duflo et al. (2011) on the tracking of students in Kenya found nearly 20% of children were absent on test days. The absenteeism rate in our sample is similar to Islam (2019) who studied schools in the same region as the present study.

Table 1
Treatment/control raw mean differences in household characteristics.

Variable	Treatment Mean	Control Mean	Difference
Household income (in takas)	8377.2	8771.0	−393.8 (544.5)
Number of household members	4.406	4.351	0.055 (0.135)
Sanitary ring latrine in the house	0.626	0.642	−0.016 (0.060)
Drinking water in the house from tube well	0.633	0.816	−0.183 (0.164)
Existence of electricity supply in the house	0.338	0.497	−0.158 (0.171)
Distance of the school from the home (km)	1.115	0.674	0.441 (0.358)
Value of total assets except land (in takas)	68089.0	63041.7	5047.3 (10326.3)
Household religion (Muslim = 1)	0.932	0.938	−0.005 (0.032)
Do any of the parents know how to play chess	0.103	0.066	0.037 (0.030)
Someone with more than grade 10 education in household	0.139	0.132	0.007 (0.029)
Father's years of schooling	4.12	4.37	−0.244 (0.655)
Mother's years of schooling	4.13	4.08	0.048 (0.732)
Father's age	39.96	39.97	−0.011 (0.603)
Mother's age	33.64	33.61	0.029 (0.643)
Father works as labourer/in agriculture	0.676	0.608	0.068 (0.076)
Mother is a housewife	0.986	1.000	−0.014** (0.005)
Two-parent household	0.996	1.000	−0.003 (0.003)
Gender of student (male = 1)	0.430	0.494	0.064 (0.049)
N	281	288	

Notes: Standard errors in parentheses and are clustered at the school level. *p-value<0.1 ** p-value<0.05 *** p-value<0.01.

Table 3
Summary of risk preference tasks in wave 1 and 2.

Lottery	Wave 1 (Items)		Wave 2 (Tokens)	
	Heads	Tails	Heads	Tails
1	4	4	5	5
2	6	3	7	4
3	8	2	9	3
4	10	1	11	2
5	12	0	13	1
6	–	–	15	0

Notes: There are five options to choose from in Wave 1, and six options in Wave 2.

Table 4
Summary of time preference tasks in wave 1 and 2.

Wave 1 (Candy)			Wave 2 (Tokens)			
Choice Set	Tomorrow	Eight Days Later	Choice Set	Alternative	Today	Seven Days Later
1	4	4	1, 2 or 3	1	12	0
2	4	6		2	9	3/4/5
3	4	8		3	6	6/8/10
4	4	10		4	3	9/12/15
5	4	12		5	0	12/16/20

Notes: In Wave 1, for each of the five choice sets, students chose from the earlier or later allocation. In Wave 2, for each of the three decisions (choice sets), students chose from one of five alternatives which determined their allocation across time. Each wave contained an additional task that was identical to the original except that rewards were delayed for an additional seven days.

in Online Appendix 3 show, there were no significant differences in characteristics between the treatment and control groups in any of the samples examined. This suggests that attrition in our sample is not systematically related to any particular set of characteristics and is likely to be unrelated to the process of randomization.

Next, we perform a direct test of differential attrition where the focus is on regressing attrition on the treatment dummy. It determines if attrition rates are different across treatment and control groups. Here, we model attrition relative to using to the largest sample we have in Wave 1,

Table 2
Mathematics (wave 1).

Variable	Control Mean (raw score)	Unadjusted Impact (raw score)	Regression Adjusted Impact (raw score)	Unadjusted Impact (standardized score)	Regression Adjusted Impact (standardized score)
Math pre-marks	18.71	0.506 (3.168) [0.820] {0.999}	1.362 (2.719) [0.608] {0.705}	0.054 (0.335) [0.820] {0.999}	0.144 (0.288) [0.608] {0.705}
N	215	494	445	494	445
Math post-marks	14.38	1.304 (3.019) [0.680] {0.999}	2.072 (2.414) [0.442] {0.648}	0.139 (0.672) [0.680] {0.999}	0.221 (0.258) [0.442] {0.648}
N	209	478	428	478	428

Notes: Standard errors in parentheses and are clustered at the school level, with conventional p-values reported as *p-value<0.1 ** p-value<0.05 *** p-value<0.01. The associated wild bootstrapped p-values are reported in square brackets, while false discovery rate (FDR) sharpened q-values (Benjamini et al., 2006) computed using the procedure in Anderson (2008) are reported in curly brackets. Covariates included in the regression adjustment are: gender, income, size of household, sanitary latrine, tube well, electricity, distance to school, assets, religion, parents play chess, family education level, father labourer, mother housewife, two-parent household. The wild bootstrap p-values are based on 1000 replications. Control means are based on the regression adjusted sample.

Table 5
Primary school certificate (PSC) national exam scores (wave 2).

Variable	Control Mean (raw score)	Unadjusted Impact (raw score)	Regression Adjusted Impact (raw score)	Unadjusted Impact (standardized score)	Regression Adjusted Impact (standardized score)
Bangla	3.76	0.282 (0.224)	0.347* (0.197)	0.308 (0.246)	0.380* (0.217)
		[0.312]	[0.180]	[0.312]	[0.180]
		{0.622}	{0.370}	{0.622}	{0.370}
Math	2.93	0.718* (0.357)	0.705** (0.283)	0.535* (0.266)	0.524** (0.211)
		[0.086]	[0.030]	[0.086]	[0.030]
		{0.520}	{0.161}	{0.520}	{0.161}
Science	3.60	0.341 (0.287)	0.292 (0.294)	0.316 (0.266)	0.271 (0.273)
		[0.282]	[0.426]	[0.282]	[0.426]
		{0.622}	{0.648}	{0.622}	{0.648}
English	2.90	0.457 (0.334)	0.398 (0.330)	0.399 (0.292)	0.349 (0.289)
		[0.222]	[0.338]	[0.222]	[0.338]
		{0.622}	{0.583}	{0.622}	{0.583}
Social Science	3.63	0.240 (0.371)	0.306 (0.319)	0.215 (0.322)	0.273 (0.285)
		[0.612]	[0.434]	[0.612]	[0.434]
		{0.999}	{0.648}	{0.999}	{0.648}
Religious Studies	3.95	0.387 (0.230)	0.405* (0.209)	0.410 (0.243)	0.428* (0.222)
		[0.142]	[0.084]	[0.142]	[0.084]
		{0.520}	{0.283}	{0.520}	{0.283}
Overall GPA	3.45	0.413 (0.242)	0.414* (0.214)	0.452 (0.265)	0.453* (0.235)
		[0.124]	[0.086]	[0.124]	[0.086]
		{0.520}	{0.283}	{0.520}	{0.283}
N	190	434	395	434	395

Notes: Standard errors in parentheses and are clustered at the school level, with conventional p-values reported as *p-value<0.1 ** p-value<0.05 *** p-value<0.01. The associated wild bootstrapped p-values are reported in square brackets, while false discovery rate (FDR) sharpened q-values (Benjamini et al., 2006) computed using the procedure in Anderson (2008) are reported in curly brackets. Covariates included in the regression adjustment are: gender, income, size of household, sanitary latrine, tube well, electricity, distance to school, assets, religion, parents play chess, family education level, father labourer, mother housewife, two-parent household. The wild bootstrap p-values are based on 1000 replications. The conversion from letter grades to scores is as follows: A+ = 5 points, A = 4 points, A- = 3.5 points, B = 3 points, C = 2 points, D = 1 point, F = 0 points. Control means are based on the regression adjusted sample.

which is the risk in Wave 1 sample (n = 450). We do not use the original sample from Table 1 as the base as that sample includes dropouts from Grades 4 to 5 who never had the opportunity to enrol in the chess program. As the actual percentage of students enrolled in school who attend school on any given school day ranges widely, the attrition we capture here is therefore attrition due to variation in daily attendance. These results are presented in Table C.3 in Online Appendix 3.[24]

We observe a statistically significant coefficient of the treatment dummy for the creativity sample. However, as we do not find significant average treatment effects for creativity, this does not affect our conclusion. For the Wave 2 risk sample (which as will be discussed later is the sample which gives rise to the main result in the paper) the coefficient on the treatment dummy is insignificant indicating attrition rates are not different across treatment and control groups.

4. Outcomes

4.1. Academic outcomes

We use exam marks from the Primary School Certificate (PSC), administered nationwide annually in Bangladesh to all fifth-grade students as the primary outcome for cognitive abilities. The PSC is a written exam, administered face-to-face and delivered through paper-and-pencil tests at the end of fifth grade. This exam took place in November 2016, approximately 9–10 months after the conclusion of the chess program. The PSC comprises six mandatory subjects: Bengali, English, science, social science, math, and religion. In the experiment, we focus on examining their results for mathematics, students' first language and science (as in Jerrim et al., 2016, 2018).

The test items consist of multiple-choice questions with three or more response options, open-ended questions requiring short, constructed responses, and essay writing. Student performance is reported by percentage of points scored out of the maximum possible score. The maximum possible score is 600 points (100 points for each subject). The minimum requirement to meet the national standard is 33%.[25]

As we had a particular interest in the potential links between chess and math, two separate math tests were developed to measure students' math skills before and after the chess training sessions. The tests intended to assess problem-solving capacities in math, requiring students to use application and reasoning skills. Both tests included 11 questions to be completed in 1 h. The tests contained two types of items: multiple-choice questions and constructed responses (demonstrating computing ability by solving word problems). To develop the tests, the local math textbook for fourth-grade students in Bangladesh was consulted, as were local school teachers and educators to help develop the test. The tests were conducted to assess students' content and cognitive domains. Content domains include addition, subtraction, multiplication, division (including money and product transactions), fractions, geometric skills, and reading, comparing and interpreting graphical representations of data. As our analysis sample comprised students from rural areas, with students generally coming from poorer socio-economic backgrounds with lower academic knowledge bases than their urban counterparts, we factored in students' backgrounds when designing the tests.

4.2. Risk preferences

Risk preferences were elicited in both waves of the study. Given our

[24] Table C.4 shows what the attrition rates are and which student characteristics (besides treatment status) predict attrition for each sample in our analysis.

[25] Due to privacy reasons, we were unable to access the numerical scores awarded to every student for each of the exams taken. However, we were able to obtain the letter grades awarded to every student for each of the six subjects, as well as an overall grade point average (GPA) score. The conversion from letter grades to scores used in Bangladesh primary schools is as follows: A+ = 5 points; A = 4 points; A- = 3.5 points; B = 3 points; C = 2 points; D = 1 point; and F = 0 points.

sample of young children in a rural environment, the Gneezy and Potters (1997) allocation task was utilized. The single-decision allocation task is also sufficient for our purposes since we are interested in the treatment effects of chess, and not in the estimation of parameters of the utility function.[26] The first-wave task was incentivized by awarding the students stationary items based on their decisions. Different stationary items (e.g. pens, rulers, erasers – see Online Appendix 4 for the precise items) were awarded to reduce diminishing returns in utility associated with receiving multiple instances of the same item.

To aid in the understanding of the task, we present the task as a choice from one of five lotteries (Table 3). The outcome of each lottery is determined by a coin flip. The first lottery is completely risk-free, rewarding four items to a student regardless of the result from the coin flip. The lotteries grow progressively riskier, with each subsequent lottery yielding two additional items from a "heads" but one less item from a "tails". A student who chooses the riskiest lottery reveals as if he is willing to invest four items with a 50% chance of them tripling and a 50% chance of losing the investment. The expected value of the alternatives (in terms of items) increases with the level of risk. Thus, a risk-neutral or risk-loving person always chooses the final lottery, while a risk-averse individual will choose between the first and fourth lottery, depending on the extent of their risk aversion. The instructions are found in Online Appendix 4.

To ensure that students do not discuss or see the choices made by other students during implementation of the task, each student was called up one at a time, then taken to a separate room. A control question was included prior to students making their actual choices to ensure that each student understood the consequences of their decisions. Following their decisions, a coin was flipped in front of them to decide how many stationary items they would receive.

In the second wave, conducted in late October 2016, the same task was used, with two changes. First, to further reduce diminishing returns in utility associated with receiving multiple instances of the same item, we rewarded students with tokens that could be used to purchase several new attractive items (see the second part of Online Appendix 4). Because of the exchange rate between tokens and items, the task in Wave 2 differs marginally from Wave 1: where in Wave 1 students effectively choose how many among four items to invest, in Wave 2, students effectively choose how many among five tokens to invest. Students therefore choose one of six different lotteries in Wave 2, with the riskiest lottery yielding 15 tokens ("heads") or no tokens ("tails") (Table 3). Hence, the rate of return on investment remains the same as in the first wave. Another advantage of having a marginal change in the task is to minimize students simply picking the exact same option as they did in Wave 1 simply due to recalling what they did previously.

Second, since the risk and time preference elicitation tasks were incentivized, and students were 'paid' immediately after each task, earning something in an initial task may influence behavior in a subsequent task. To check for this, in the second wave we switched the orders of the risk and time preference tasks, where the risk preference task was done first in Wave 1. Regression results reveal that the size of actual rewards from the first task does not affect choices in the subsequent task, regardless of which task it was. The added advantage of reversing the order is that students are less likely to anticipate that a risk preference task would occur after the time preference task in Wave 2, since they were not told this beforehand.

4.3. Time preferences

Time preferences were elicited in both waves and at the same time as risk preferences, with the order of the two tasks reversed across waves. In the first wave (January–February 2016), we used a multiple-price-list format popularized by Coller and Williams (1999). Unlike risk preferences, it is less common to find single-decision implementations of elicitation tasks for time preferences.[27] Additionally, it is common for the multiple price list format to be implemented on children.[28]

In this task, students make five decisions. For each decision, they choose between receiving four pieces of candy tomorrow ("earlier"), vs. receiving x pieces of candy in eight days ("later"), where $x \in \{4, 6, 8, 10, 12\}$ (Table 4). This is close to the design adopted by Alan and Ertac (2018), in which the choice was between two gifts today vs. y gifts one week later, where $y \in \{2, 4, 6, 8, 10\}$. We chose candy to differentiate it from the incentives presented in the risk preference tasks in hopes of reducing any diminishing marginal utility associated with potentially obtaining too many stationary items. Candy was also used to incentivize children's time preference elicitation in Andreoni et al. (2017). The design adopts the "front-end delay" found in Harrison et al. (2002) and Castillo et al. (2011), whereby no rewards are presented on the same day the task is performed. In doing so, the aim is to minimize any apparent impatience arising from a lack of trust in the experimenters, or any psychological discontinuities that may arise from imagining payment in the future versus an immediate "now" that may generate a higher level of time inconsistency in the form of present bias.

Following previous studies on time preferences, we attempt to test for time inconsistency by presenting students with an additional five decisions that remain identical to the original, except that they are delayed for seven days (the earlier alternative was paid out in eight days, and the later alternative, in 15 days). This delay resembles the seven-day (earlier) and 14-day (later) implementation that Alan and Ertac (2018) used. Time inconsistency is particularly relevant to our implementation because it often has been tied to self-control, commitment problems, and procrastination (e.g. Frederick et al., 2002). It is unclear a priori whether the effect of chess training will be stronger on patience or on the incidence of time consistency.

The students were paid for only one of the 10 decisions they made for the time preference task. This was determined by having an experimenter (randomly) draw one of 10 numbered pieces of paper from a jar in front of the students (the instructions are found in Online Appendix 5).

The students were extremely patient in Wave 1, with 85% of them choosing the "later" option at an effective interest rate of 50%. Hence, in our Wave 2 time preference task, we adopted the convex time-budget task of Andreoni and Sprenger (2012) in order to increase the granularity and variation in the information elicited from student choices. This also is done in Alan and Ertac (2018) in their follow-up wave. This task differs from the Wave 1 task in the following dimensions: (i) There are only three, rather than five, decisions (choice sets), and each choice set now contains five (instead of only two) alternatives (Table 4); (ii) There is no more front-end delay since this may be making students overly patient in the first wave; and (iii) We rewarded students with tokens that could be used to purchase several new attractive items.

For each decision, the most impatient alternatives result in receiving 12 tokens earlier and no tokens later, while the most patient alternatives result in receiving no tokens earlier and $z = 12 \times (1 + r)$ tokens later, where $r \in \{0, 0.33, 0.66\}$ is the interest rate. The equivalent interest rates in Alan and Ertac (2018) were 0.25 and 0.50. In addition, we continued to test for time inconsistency by including three more decisions that differed only in having the "earlier" outcome in seven days and the "later" outcome in 14 days. Only one of the six decisions was paid out;

[26] For a review of risk-elicitation tasks, see Charness et al. (2013).

[27] The exception to this is Angerer et al. (2015) who effectively implement the time-preference equivalent for the Gneezy and Potters (1997) task. They find that both the multiple price list and simpler single decision task are highly correlated. However, the latter lacks the ability to identify inconsistent behavior (which they find cannot be attributable to mere misunderstanding).

[28] For example, Bettinger and Slonim (2007) study involved children ages 5–16 in the US; Castillo et al. (2011) analysis involved children ages 13–14 in the US; Sutter et al. (2013) study involved children ages 10–18 in Austria; Alan and Ertac (2018) study involved children ages 9–13 in Turkey.

this was determined using the same method as in Wave 1. We included $r = 0$ as an indicator of the concavity of the utility function since any choice to delay receiving tokens in this case can be attributed purely to the diminishing returns to utility of receiving tokens. Since the students could effectively receive everything early and delay their own actual consumption, one can also view choosing to receive tokens later at $r = 0$ as a demand for a commitment device. The tokens earned in this task, together with the tokens earned in the risk task in Wave 2, could be exchanged for several different attractive items (see Online Appendix 4). Instructions for the convex time-budget task are provided in Online Appendix 5.

4.4. Creativity and attention/focus

We also investigate whether undertaking intensive chess lessons can affect children's creativity and attention/focus. For assessing creativity, we use the Torrance Tests of Creative Thinking (Torrance, 1966) and Guilford (1967) alternative uses test. For attention and focus, we employ two frequently used tests for the assessment of attention: the digit-cancellation test (Diller et al., 1974) and the digit-symbol test (Wechsler, 1991). These tests are described in more detail in Appendix A and Appendix B.

5. Empirical approach

With randomization, the identification strategy used is straightforward. The benchmark model used to estimate the intention to treat effects (ITT) – the average treatment effect for children in fifth grade in schools that were randomly assigned to receive chess training – is the following OLS regression:

$$Y_{i,s} = \alpha + \delta treat_s + \beta X_{i,s} + \varepsilon_{i,s} \quad (1)$$

$Y_{i,s}$ denotes outcomes for individual i in school s, and $treat_s$ is whether a school was assigned to treatment group or not. Randomization was done at the school level, and all students in fifth grade in 2016 in the treatment schools were invited to participate in the chess training program.[29] We regression-adjust our results using a set of baseline covariates, $X_{i,s}$ which includes individual and household characteristics of the student to increase the precision of our results. Standard errors are clustered at the school level.

As an alternative way of performing statistical inference due to the clustered nature of the data, p-values using the wild bootstrap proposed by Cameron et al. (2008) are also computed. As many outcomes have been examined, this raises the issue of multiple hypothesis testing. To control for the false discovery rate (FDR), we provide sharpened q-values (Benjamini et al., 2006) using the procedure implemented in Stata by Anderson (2008). The interpretation of q-values is analogous to interpreting p-values – the q-values presented denote the lowest critical level at which a null hypothesis is rejected when controlling for the false discovery rate. Families of related p-values are typically used to estimate q-values. In our study, we take a conservative approach and use all outcomes tested rather than grouping the tests into families based on the domain tested.

6. Results

We present two sets of program impacts – unadjusted and regression adjusted – for the various cognitive and non-cognitive outcomes examined in Tables 2–5 The sample sizes for unadjusted and regression adjusted results vary and depend on whether both baseline data on characteristics and data on the outcome were measured. As data were collected on different days, the variation in sample sizes across outcomes partly reflects the fact that on any given day, student absenteeism is high in primary schools in rural Bangladesh.

Three alternative sets of p-values are presented. First, in the columns for unadjusted and regression adjusted impacts, we present conventional standard errors in parentheses and the associated p-values (using asterisks) from a regression model based on clustered standard errors. Second, p-values using the wild bootstrap (1000 replications) proposed by Cameron et al. (2008) are reported in square brackets. Third, we compute FDR sharpened q-values (Benjamini et al., 2006) using the procedure in Anderson (2008). These q-values are presented using curly brackets.[30]

Table 6
Risk preferences (waves 1 and 2).

Variable	Control Mean	Unadjusted Impact	Regression Adjusted Impact
Wave 1 (Min 1, Max 5), higher value = less risk averse	2.84	0.319* (0.166) [0.084] {0.520}	0.301 (0.175) [0.144] {0.370}
N	225	520	450
Wave 2 (Min 1, Max 6), higher value = less risk averse	2.65	1.647*** (0.437) [0.000] {0.001}	1.752*** (0.442) [0.002] {0.028}
N	191	426	381

Notes: Standard errors in parentheses and are clustered at the school level, with conventional p-values reported as *p-value<0.1 ** p-value<0.05 *** p-value<0.01. The associated wild bootstrapped p-values are reported in square brackets, while false discovery rate (FDR) sharpened q-values (Benjamini et al., 2006) computed using the procedure in Anderson (2008) are reported in curly brackets. Covariates included in the regression adjustment are: gender, income, size of household, sanitary latrine, tube well, electricity, distance to school, assets, religion, parents play chess, family education level, father labourer, mother housewife, two-parent household. The wild bootstrap p-values are based on 1000 replications. Control means are based on the regression adjusted sample. False discovery rate (FDR) sharpened q-values (Benjamini et al., 2006) are computed using the procedure in Anderson (2008).

6.1. Academic results

We consider two types of test scores to measure cognitive ability. The first involves the use of a project-administered math test. The treatment group scored slightly better in the pre-program math test relative to the control group, but the difference was not statistically significant (providing further supporting evidence that the randomization was well-implemented). The gap between the treatment and control groups widened in the post-program test conducted shortly after the intensive chess training had ended. However, the difference was again not statistically significant (see Table 2).

The second measurement of academic achievement involved the use of the PSC exam which took place 9–10 months after the training. The

[29] Unfortunately, student attendance on each day of the chess training was not recorded, thereby not allowing us to measure treatment receipt.

[30] Apart from reporting the FDR, we tried alternative methods to control for multiple hypothesis testing that take into account the important relatedness of outcomes – the Westfall-Young and the Romano-Wolf approaches. As highlighted in Clarke et al. (2019), the Westfall-Young approach assumes a certain subset pivotality condition. However, this assumption can be violated in certain applications and is thus undesirable. Instead, they propose the use of the Romano-Wolf multiple hypothesis correction. Like the Westfall-Young approach, it also uses resampling and step-down procedures to gain additional power by accounting for the underlying dependence structure of the test statistics. However, and crucially, this procedure does not require the subset pivotality condition and is thus more broadly applicable than the Westfall-Young procedure. We find that the Westfall-Young approach gives rise to much more conservative estimates, while the Romano-Wolf approach provides results that are similar to the FDR results provided in the paper (results available upon request).

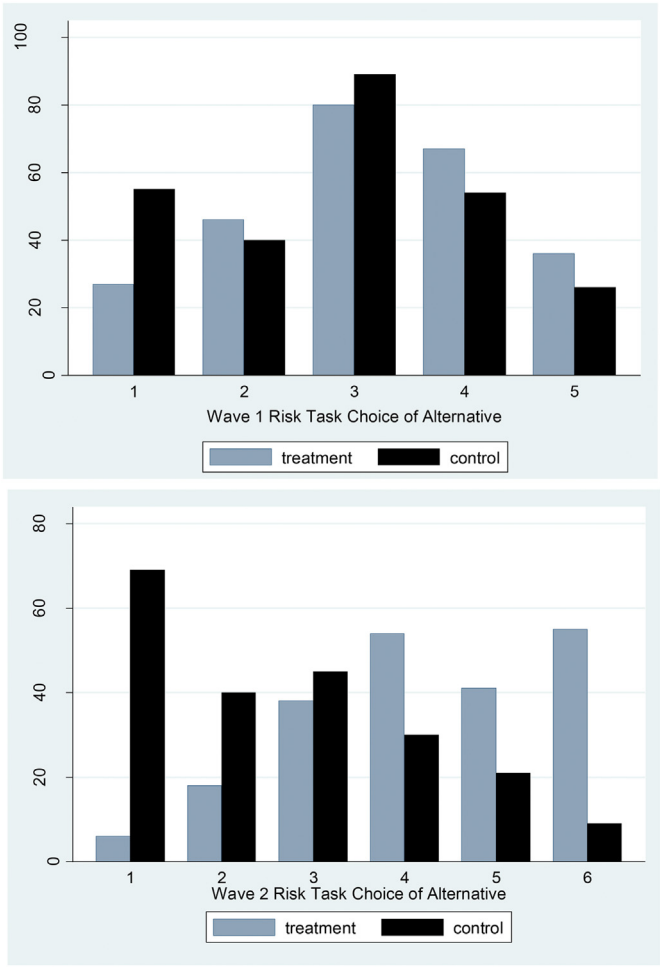


Fig. 3. Distribution of choices across groups, and waves in the risk-elicitation task.

results of the PSC exam are provided in Table 5. We find a significant positive effect from our intensive chess-instruction program on math grades in the PSC exam using both conventional p-values and the wild cluster bootstrap (p-value = 0.030 using the wild cluster bootstrap).[31] The treatment-control difference of 0.71 points is approximately equivalent to between half and a full letter math grade. However, the false discovery rate (FDR) sharpened q-values that account for multiple hypothesis testing suggest that this difference is not significant (q-value = 0.161). Likewise, although the impact on overall GPA (0.41) is statistically significant using conventional clustered standard errors and the wild bootstrap (p-value = 0.086), the FDR sharpened q-values suggest it is not significant.

6.2. Risk preference results

The average value of the alternatives chosen in the risk-elicitation task was used for assessing a treatment effect on risk preferences, in which a higher value indicates a riskier choice. The values range from 1 to 5 in Wave 1, and 1–6 in Wave 2.[32] Results are depicted in Table 6. In Wave 1, treated students invested, on average, 0.3 more items into the risky "asset" (p-value = 0.144). In Wave 2, treated students invested, on average, 1.75 more tokens into the risky asset (p-value = 0.002). Hence, although we find no significant effect on risk preferences in Wave 1, a strong effect (both in terms of size and significance) emerges in Wave 2 – chess training decreases risk aversion.[33] Importantly, this impact remains statistically significant using the FDR q-values and in both the regression adjusted and non-regression adjusted samples.[34]

Fig. 3 breaks down the treatment effects according to each available alternative and highlights the changes between Waves 1 and 2. For both waves, we can see that the largest difference emerges for alternative 1 – the safest alternative. In addition, there is a strong effect in Wave 2 on alternative 6 – the riskiest alternative – suggesting that chess training may have resulted in a significant number of students switching from being risk-averse to either risk-neutral or risk-loving over time.

Another way to analyse the risk results is to look at the same person over time and whether the distribution is shifting to the right or do we have movement both ways. One challenge in looking to see how the distribution shifts over time is that the scale for risk preferences in Wave 1 and 2 are different. In Wave 1, students chose from five different lotteries, while in Wave 2, students chose from six different lotteries with different payoffs. Hence, a simple comparison of the options chosen between Waves 1 and 2 do not allow one to see if risk aversion is increasing or decreasing.

In order to make progress on this, we can first assume that the lotteries in Waves 1 and 2 can be divided into three categories that represent different levels of risk. The first category is one where there is no risk, as the same payoff is obtained regardless of whether the coin shows up

Table 7
Risk preferences transition matrix between waves 1 and 2.

	Treatment Group (n = 181)			
Risk Wave 1	Risk Wave 2			
	Category 1	Category 2	Category 3	Total
Category 1	0	8	9	17
	(0.0%)	(47.1%)	(52.9%)	(100%)
Category 2	3	41	43	87
	(3.5%)	(47.1%)	(49.4%)	(100%)
Category 3	2	42	33	77
	(2.6%)	(54.6%)	(42.9%)	(100%)
Total	5	91	85	181
	(2.8%)	(50.3%)	(47.0%)	(100%)
	Control Group (n = 183)			
Risk Wave 1	Risk Wave 2			
	Category 1	Category 2	Category 3	Total
Category 1	17	22	5	44
	(38.6%)	(50.0%)	(11.4%)	(100%)
Category 2	31	51	7	89
	(34.8%)	(57.3%)	(7.9%)	(100%)
Category 3	14	23	13	50
	(28.0%)	(46.0%)	(26.0%)	(100%)
Total	62	96	25	183
	(33.9%)	(52.5%)	(13.7%)	(100%)

Notes: Sample used is those with non-missing responses to risk preferences in both Waves 1 and 2.
Category 1: Even bet (Wave 1 = lottery 1, Wave 2 = lottery 1).
Category 2: Slight risk (Wave 1 = lottery 2, 3, Wave 2 = lottery 2, 3, 4) – min return is 2 tokens.
Category 3: High risk/All in (Wave 1 = lottery 4, 5 Wave 2 = lottery 5, 6) – min return is 0 or 1 token.

[31] When the number of bootstrap replications is increased from 1000 to 5000, the p-value from the wild cluster bootstrap is very similar (=0.034).
[32] There was one additional alternative in Wave 2 because of the higher granularity of the rewards.
[33] When the number of bootstrap replications is increased from 1000 to 5000, the p-values from the wild cluster bootstrap for the non-regression-adjusted and regression-adjusted impacts are still highly significant (equal 0.0008 and 0.0004 respectively).
[34] The results remain statistically significant when we include pre-program project-administered math test scores as an additional control variable (which we do not use in our general set of controls as it will reduce our sample size).

heads or tails. This is lottery option 1 in both Waves 1 and 2. The second category is where there is a slight risk, and where the minimum return is a payoff of 2 tokens. These are lottery options 2 and 3 in Wave 1, and lottery options 2, 3 and 4 in Wave 2. The third category is the high risk/all in option. These involve choosing lottery options 4 and 5 in Wave 1, and lottery options 5 and 6 in Wave 2. In this case, the minimum return is either 0 or 1 token.

Having classified the lotteries into three categories, we can now examine a simple transition matrix between Waves 1 and 2 for the treatment and control group separately. The results in Table 7 show that the reduction in risk aversion we find in Wave 2 is mainly driven by those in category 1 and category 2 in the treatment group moving up the categories, (i.e. a reduction in risk aversion). While 52.9% shift from category 1 to category 3 in the treatment group (Table 7, top panel), only 11.4% do so in the control group (Table 7, bottom panel). Similarly, while 49.4% shift from category 2 to category 3 in the treatment group, only 7.9% do so in the control group. The transition matrices also show that risk aversion increases in the control group across waves: 34.8% move from category 2 down to category 1, while 74% leave category 3.

The behavior of the control group is consistent with the notion that risk aversion increases over time among children (Schildberg-Hörisch, 2018) and may be tied to the notion of loss-aversion: children respond asymmetrically to experiences of risky losses relative to experiences of risky gains. Accumulated experiences of risky losses (e.g. from the risk preference elicitation task in Wave 1) may increase risk aversion over time and may help explain the behavior of the control group in the Wave 2 risk preference elicitation task. The chess training seems to not only mitigate this increase but is strong enough to result in an overall decrease in risk aversion over time.

One potential problem with our finding for risk preferences is that our treatment may not be directly affecting risk preferences, but rather the student's ability to comprehend and respond to the elicitation task (e.g. whether through the ability to think counterfactually, or to respond consistently) through improvements in cognitive ability as a result of the treatment.

We attempt to check for the effect of the treatment on cognitive ability and subsequently on risk preferences by conducting a formal mediation analysis proposed by Imai et al. (2010) and Imai et al. (2013). First, we use the PSC math score (our proxy for cognitive ability) as a mediator and check if it helps mediate the effect of chess training on Wave 2 risk preferences.[35] The mediation analysis using PSC Math scores as a mediating variable reveals that although the proportion mediated via math is statistically significant, only about 5% of the effect operates through math (see Table F.1 in Online Appendix 6). Thus, it seems unlikely that cognition as proxied through math scores has a significant role in explaining our results on risk preferences.

In addition, we can check more directly for whether students are behaving more consistently as a result of our treatment, and more importantly, whether this mediates the change in risk preference. As the risk preference task involves only a single decision, it is not possible to assess the consistency of their behavior using this task. Instead, we use the 'non-monotonicity' variable in the Wave 2 time preference task as a measure of consistency, which captures the ability of the student to report internally consistent time preferences. The results reveal that the mediation effect of non-monotonicity is even weaker than that of the PSC Math score, with the proportion of the total treatment effect mediated by non-monotonicity being approximately 1% (see Figure F.1 in Online Appendix 6).

Overall, we therefore could not find evidence that cognitive effects mediate the risk preference effects we observe and suggest instead that a compelling interpretation of our result is that exposure to the strategic calculation of risk found in chess training and playing inculcates a better appreciation for risk-taking.[36]

The fact that the effects on risk-preferences are detected only 9–10 months after the initial program was launched also suggests that these effects are possibly linked to changes in habitual and long-term behavior rather than the purely cognitive aspect of having been instructed on how to play chess. The results are consistent with our finding that nine out of ten students were still playing chess 9–10 months after the intervention ended, which allows students enough time to develop a deeper

Table 8
Time preferences (waves 1 and 2).

Variable	Control Mean	Unadjusted Impact	Regression Adjusted Impact
Wave 1			
Impatience (0–5)	1.26	0.038 (0.062)	−0.026 (0.062)
		[0.630]	[0.716]
		{0.999}	{0.723}
Delayed impatience (0–5)	1.32	0.016 (0.065)	−0.040 (0.061)
		[0.810]	[0.540]
		{0.999}	{0.681}
Time inconsistency (binary)	0.28	−0.086** (0.037)	−0.162***
		[0.068]	(0.040)
		{0.464}	[0.008]
			{0.053}
Time inconsistency (0–5)	0.38	−0.091 (0.064)	−0.234***
		[0.210]	(0.060)
		{0.622}	[0.006]
			{0.053}
Non-monotonicity (binary)	0.14	−0.089***	−0.121***
		(0.028)	(0.018)
		[0.010]	[0.002]
		{0.150}	{0.028}
N	224	521	450
Wave 2			
Impatience (2–10)	5.19	−0.338 (0.331)	−0.087 (0.365)
		[0.354]	[0.898]
		{0.622}	{0.951}
Delayed impatience (2–10)	5.27	−0.120 (0.270)	−0.151 (0.257)
		[0.738]	[0.634]
		{0.999}	{0.705}
Time inconsistency (binary)	0.74	−0.073 (0.046)	−0.060 (0.042)
		[0.136]	[0.202]
		{0.520}	{0.389}
Time inconsistency (0–2)	1.13	−0.145 (0.084)	−0.129* (0.065)
		[0.112]	[0.090]
		{0.520}	{0.283}
Non-monotonicity (binary)	0.67	−0.107* (0.052)	−0.126**
		[0.054]	(0.055)
		{0.510}	[0.062]
			{0.283}
N	191	426	381

Notes: Standard errors in parentheses and are clustered at the school level, with conventional p-values reported as *p-value<0.1 ** p-value<0.05 *** p-value<0.01. The associated wild bootstrapped p-values are reported in square brackets, while false discovery rate (FDR) sharpened q-values (Benjamini et al., 2006) computed using the procedure in Anderson (2008) are reported in curly brackets. Covariates included in the regression adjustment are: gender, income, size of household, sanitary latrine, tube well, electricity, distance to school, assets, religion, parents play chess, family education level, father labourer, mother housewife, two-parent household. The wild bootstrap p-values are based on 1000 replications. Control means are based on the regression adjusted sample. False discovery rate (FDR) sharpened q-values (Benjamini et al., 2006) are computed using the procedure in Anderson (2008).

[35] Studies such as Eckel et al. (2012), Benjamin et al. (2013), Sutter et al. (2013) and Andreoni et al. (2019a) also use math scores as a proxy for cognition in regressions on risk preferences.

[36] Chess playing may also decrease risk aversion through increased exposure to competition. Experimental studies by Eriksen and Kvaløy (2017) and Spadoni and Potters (2018), for example, provide evidence that an increase in competitive pressure decreases risk aversion. Future studies may test for this through a treatment that focusses on tournament-play.

understanding of strategy and risk in the game through playing hundreds of games, while also giving them prolonged exposure to interactions involving bilateral and non-physical competition.

6.3. Time preference results

In Wave 1 of the time-elicitation task, students were given five choice sets and indicated in each instance whether they would take the patient alternative ("later") or impatient alternative ("earlier"). For each individual, we assign a count of impatient alternatives chosen. Their sum was used to assess average treatment effects, with higher values indicating more impatience. We also did this for the five choice sets with one week of delay. The results are depicted in the first two rows of the top panel of Table 8. The results for both the standard and delayed choice sets are statistically insignificant (p-values = 0.716 and 0.540, respectively), as well as small in magnitude.

For Wave 2, we utilized two choice sets, with each set containing five alternatives[37] Each alternative is assigned a score 1–5, with a higher score indicating greater impatience. For each student, we summed the scores across the two choice sets. The results (the first two rows of the bottom panel of Table 8), with and without delay, remain statistically insignificant (p-values = 0.898 and 0.634, respectively).

Given that time preferences were elicited using a multiple price-list method, we can conduct two additional tests. The first involves a test for time inconsistency. In both waves, we had students make decisions over an original and one-week-delayed set that differ only in having payoffs in the latter realized seven days later than the original. We consider two possible variables for a test of time inconsistency: (i) a continuous variable that scores a "1" for each decision that fails to match across both the original and the corresponding one-week-delayed decision, and (ii) a binary variable that takes on a value of "1" if at least one decision in the original decisions fails to match their corresponding one-week-delayed decision.

For time inconsistency, there is some evidence that students in the treatment group are less likely to make time inconsistent decisions in Waves 1 and 2 using conventional p-values. The FDR q-values remain significant for time inconsistency in Wave 1, but only for the smaller regression adjusted sample and not for the larger non-regression adjusted sample.

The second additional test we perform on the time preference data involves checking for non-monotonicity of time preferences. Well-defined, monotonic time preferences require that a choice at some interest rate r must be at least as patient as some other interest rate $r' < r$ (e.g. see Harrison et al., 2002). In Wave 1, this translates to students switching from the "earlier" to "later" option at most once. In Wave 2, it requires that a choice at some interest rate r must be of a value at least as high as the choice at some other interest rate $r' < r$. We construct a binary variable that takes the value "1" if such a monotonicity requirement is violated. The results are presented in the last row of each panel in Table 8. Both conventional p-values and wild cluster bootstrap p-values suggest that students in the treatment group are less likely to violate the monotonicity requirement in Waves 1 and 2. However, the insignificance of the FDR q-values suggests that this result might not be robust.

6.4. Results for creativity and attention/focus

Our results do not suggest that there are any short-term effects of chess instruction on creativity, or medium-term effects on focus and attention. Discussion of these additional non-cognitive outcomes are provided in Appendix A and Appendix B.

6.5. Choice of controls

As power is a major issue in our study given the small number of clusters, choosing controls to maximize statistical power can be important. One way power can be improved is to select the control variables in a potentially more principled way, such as by using the double-lasso methodology (Belloni et al. 2014, 2015). This methodology uses the lasso estimator to select the controls and generally achieves a sparse solution, i.e., most coefficients are set to zero. This is because the final choice of control variables to include in the regression is the union of the controls selected from two regressions – one with the outcome variable of interest as the dependent variable and the other with the treatment status as the dependent variable. As a result, the estimated coefficients and associated statistical significance results using the double-lasso are in general close to the results without regression adjustment (results available upon request).

As a second robustness check regarding the choice of controls, we test the sensitivity of our results to using alternative specifications of the regression adjustment model. Specifically, for the two outcomes with statistically significant effects, PSC math score and Risk in Wave 2, we try using all possible combinations of the control variables to determine if varying the choice of controls influences our estimated impacts. We use the p-hacking specification check proposed by Brodeur et al. (2020) to test the use of various combinations of our control variables and see whether the impact remains significant.

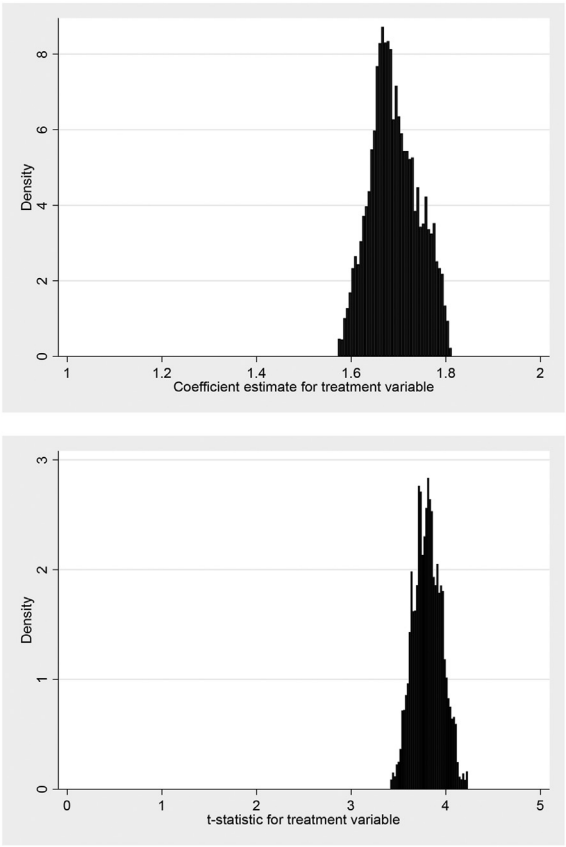
Fig. 4 shows the results of doing a specification check for Wave 2 risk preferences, where we found that treated students invested, on average, 1.75 more tokens into the risky asset (p-value = 0.002). The top panel shows the distribution of the estimated impact. The dispersion of the estimates indicates how much the magnitude of the treatment effect, rather than its statistical significance, varies by various combinations of control variables. The graph shows that the distribution of the estimated impact ranges from 1.58 to 1.81, suggesting that our point estimate of 1.75 is robust. The bottom panel displays a histogram which shows the distribution of t-statistics of the estimated treatment effect. It indicates that regardless of the control variables included, the treatment effect remains statistically significant at conventional levels (t = 1.96 is a conventional threshold for the p = 0.05 level).

A similar exercise is conducted for another key outcome of interest – PSC math grades – where we found a significant positive effect of 0.71 points using both conventional p-values and the wild cluster bootstrap (p-value = 0.030 using the wild cluster bootstrap). The top panel of Fig. 5 shows the distribution of the estimated treatment effect on math grades. The estimates range from 0.53 to 0.84, with about half of the estimates larger than our point estimate of 0.71 and half of them smaller. The bottom panel displays the distribution of t-statistics, which indicate that the estimated treatment effect is significant at the p = 0.05 level in about half the cases (t > 1.96) and insignificant in the other half. Such a result is consistent with the finding in the paper that while conventional p-values and the wild cluster bootstrap p-values suggest significant effects, the false discovery rate (FDR) sharpened q-values that account for multiple hypothesis testing suggest that this difference is not significant. Therefore, while the point estimate for the effect on PSC math grades is not zero, the statistical significance is sensitive to the choice of covariates used in regression adjustment. We therefore cannot make a definitive conclusion on statistical significance based on the point estimate and standard errors for PSC math grades we report in the paper, and our results can be treated as conservative.

6.6. Minimal detectable effects

As our study design is not powered to detect any reasonably-sized effect on many of our outcomes, we need to be careful about not conflating statistically insignificant effects with a zero effect. Therefore,

[37] The third choice set involving $r = 0$ is used to elicit the presence of diminishing returns in utility. If the marginal utility of receiving tokens at any given period of time is non-diminishing, students should choose alternative 1. In our results, only 26% of students chose alternative 1, suggesting that diminishing returns in utility plays a non-trivial role in decisions.



Notes: The figures show the distribution of the coefficient estimates and t-statistics for the treatment variable using all possible combinations of the covariates in the regression adjusted model.

Fig. 4. Varying the Choice of Controls – Impact on Risk in Wave 2. Notes: The figures show the distribution of the coefficient estimates and t-statistics for the treatment variable using all possible combinations of the covariates in the regression adjusted model.

in this section, we present details on what sized effects we cannot rule out given our sample size. This is especially since in the p-hacking exercise we conducted above revealed that in about half of the combinations of control variables used in the model for regression adjustment, significant effects on math would have been found.

One way to summarize the implications regarding statistical power in our study is to compute the minimum detectable effect size (MDES) (see Bloom, 1995) for which it would have adequate statistical power. We follow standard practice and consider a power value of 0.8 (80% power) with a two-sided test at a significance level of $p = 0.05$.

Computation of statistical power in cluster-randomized trials requires knowledge of the intraclass correlation ρ. There is not much information about intraclass correlations appropriate for studies with academic achievement as an outcome. In our study, the fact that we examine multiple outcomes makes it even more difficult to choose the value of ρ. Hedges and Hedberg (2007) provide a comprehensive collection of intraclass correlations of academic achievement on the basis of national representative samples. We therefore compute the MDES for each of the outcomes we examine in the paper using alternative plausible values of the intraclass correlation for grade 5 children provided by them.

In Table G.1 in Online Appendix 7, we provide the MDES for each of the outcomes we examine in the paper using the original units of the outcomes of interest. These represent true impacts with an 80% chance of being identified (producing a significant positive impact estimate at the $p = 0.05$ level). True positive impacts smaller than the figures listed in the table for each outcome will have less than an 80% chance of being identified.

For the PSC math grade, the MDES varies from values of 0.66 ($\rho = 0.1$), 0.86 ($\rho = 0.2$) and 1.16 ($\rho = 0.4$). As our point estimate for PSC math is 0.71, this implies that under larger values of the intraclass correlation, it is below the minimum detectable value and there is less than an 80% chance of the effect being identified, even if it is a true effect.

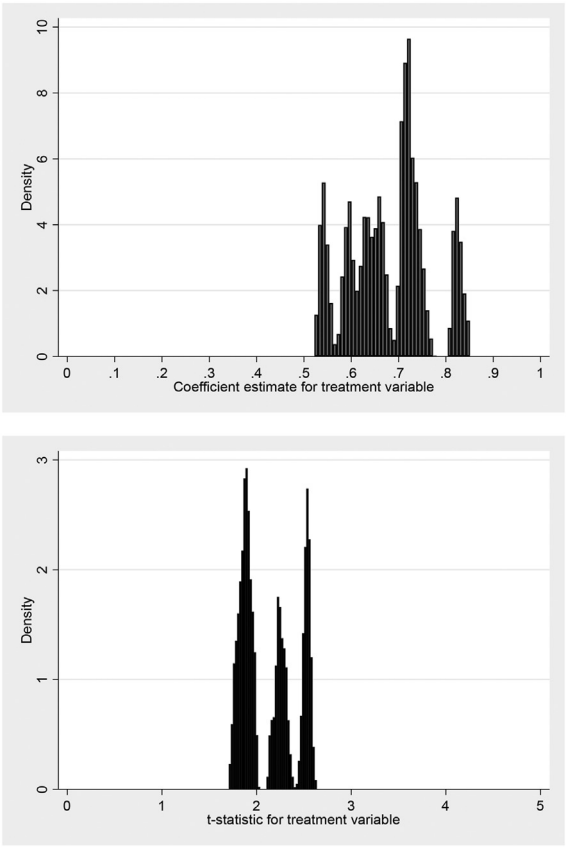
For Wave 2 risk preferences, the MDES varies from values of 0.75 ($\rho = 0.1$), 0.98 ($\rho = 0.2$) and 1.33 ($\rho = 0.4$). These are all less than our point estimate of 1.75, implying that under a variety of plausible values for the intraclass correlations, the effect on risk in Wave 2 has an 80% chance of being identified. This implies that despite our relatively small sample size, the significant finding on Wave 2 risk preferences is detectable under standard assumptions used to define adequate power.

7. Cost effectiveness

Table 9 depicts the cost-effectiveness of our study relative to other studies. The cost-effectiveness of our study appears to be ranked somewhere in the middle when looking at the distribution of cost-effectiveness across studies. In his meta-study of education RCTs on primary schools, McEwan (2015) finds that the cost for raising test scores by 0.1 standard deviations ranges from $0.22 to $45.05 (over 26 studies). Our cost-effectiveness ($4.56) is comparable to the Kremer et al. (2009) study, which examines the effect of awarding merit scholarships to grade 6 girls in Kenya.

We also consider the size of our risk and time preferences findings relative to other studies. Table 10 depicts the summary of these studies. Cost-effectiveness comparisons across studies are not possible here given the lack of uniformity in the reporting of results in these studies. These studies were chosen based on their use of similar elicitation tasks and their involvement of children.

For risk preferences, our Wave 2 result on risk preferences is roughly



Notes: The figures show the distribution of the coefficient estimates and t-statistics for the treatment variable using all possible combinations of the covariates in the regression adjusted model.

Fig. 5. Varying the Choice of Controls – Impact on PSC Math. Notes: The figures show the distribution of the coefficient estimates and t-statistics for the treatment variable using all possible combinations of the covariates in the regression adjusted model.

equivalent to twice the benefit of having avoided domestic violence during childhood (Castillo, 2020), and slightly less than the benefit of having avoided exposure to a flood or earthquake (Cameron and Shah, 2015).[38] For time preferences, our Wave 2 results are similar to those reported in Andreoni et al. (2019b) and Berry et al. (2018) who both also find weak and insignificant effects of their education interventions.

8. Summary and conclusions

This paper evaluates the effects of learning chess using a randomized experiment on grade five students in rural Bangladesh. The intervention comprised of a 30-h training program based on a curriculum approved by the World Chess Federation. By employing a field experiment and collecting a range of academic and non-academic outcomes, we have provided credible estimates of the benefits chess instruction can have for children's cognitive and non-cognitive outcomes. In terms of academic outcomes, we use high-stakes, age-appropriate, and externally marked academic tests for schools to measure the effectiveness of the intervention, meaning our results are unlikely to be influenced by limitations surrounding the outcome test. We examine both short-term effects based on assessments made shortly after the conclusion of the program, as well as medium-term effects based on assessments conducted 9–10 months after the program ended, allowing us to examine whether there is a lasting effect.

One novel contribution of this paper is a focus on the link between chess and non-cognitive outcomes relevant to the labor market: risk, time preferences, patience, creativity, attention, and focus. The previous literature has emphasized potential links between chess and academic outcomes.

Our main finding is that chess training reduces the treatment group's level of risk aversion almost a year after the intervention ended. This finding is robust to correction for multiple hypothesis testing. While our impact estimates based on conventional p-values and wild bootstrap p-values provide some indication of effects on math scores, time inconsistency and non-monotonic time preferences, there is less conclusive evidence after controlling for multiple hypothesis testing using the false discovery rate.

At first glance, it might appear counter-intuitive to argue that it can be beneficial to have a program that can help reduce risk aversion during childhood. For example, adolescence is often perceived as an age of heightened risk taking for many real-world behaviors: consumption of alcohol, drug use, unprotected sex, and driving while distracted. However, empirical evidence on risk preferences in childhood documents systematic changes as children grow. At younger ages, children are more willing to take risks than adults, and a larger share of them behave in a risk-seeking manner. It is only as children grow older that they become less willing to take risks; in adolescence their risk preferences converge to adults (Schildberg-Hörisch and references therein, 2018).

Tymula et al. (2012) suggest that 'risky behaviours' among adolescents may be explained by a willingness to take risks under uncertainty (i.e. ambiguity). Instead, chess may teach students to recognise opportunities to take calculated (rather than unknown) risks, as reflected in the

[38] The effects of natural disasters on risk preferences are far from settled: Hanaoka et al. (2018) and Islam et al. (2020) find that they reduce risk aversion, while Cameron and Shah (2015), Cassar et al. (2017) and Li et al. (2011) find the opposite.

Table 9
Cost-effectiveness in terms of test scores.

	Intervention	Average test score gain (s.d.)	Cost/ obs	Cost/obs per 0.1 s.d.
Banerjee et al. (2007) [BCDL]	Remedial education by young women targeting grade 3 and 4 children in urban India.	0.10	$7.25	$7.25
BCDL	Computer assisted learning to grade 4 students in urban India.	0.10	$40.01	$40.01
Burde and Linden (2013)	Establishing primary schools in villages in Afghanistan.	0.65 (girls) 0.40 (boys)	n/a	n/a
Islam (2019) [IS]	Introducing parent-teacher meetings for primary school children in Bangladesh.	0.38	$3.16	$1.66
Kremer et al. (2009) [KMT]	Merit scholarships to grade 6 girls in Kenya.	0.12	$6.03	$5.02
Muralidharan et al. (2019)	Computer assisted learning for grades 4 to 9 students from low-income households in urban India.	0.37 (math)	$15	$4.05
This study	Chess instruction and play for grade 5 children in Bangladesh.	0.45 (all PSC) 0.52 (math PSC)	$20.50	$4.56 $3.94

Notes: Dollar amounts are in 2015 USD, with BCDL/IS/KMT adjusted by a factor of 1.318/1.054/1.423 to account for inflation in the USA since 2002/2011/1999. Inflation factor calculated from USA CPI data (Federal Reserve Bank of St Louis). KMT estimates do not include accounting for pure transfer of money due to scholarship, which reduces cost per 0.1 s.d. to $2.01 (this also accounts for the deadweight loss from raising government funds for the program). Cost for this study includes only the costs associated with implementing the chess program.

choice of chess opening to use, and in assessing the appropriate time to sacrifice material to attack an enemy king. Both of these topics were touched upon in our chess training program.

These various routes through which risk preferences may be systematically affected during childhood is in line with a standard model of skill formation (e.g. Cunha and Heckman, 2007). Accounting for preference formation enables one to interpret the success of many early childhood programs that do not permanently raise IQ but nonetheless go on to influence a multitude of life outcomes (Cunha and Heckman, 2007: 42). Viewed in this light, knowledge and appreciation of chess strategy can therefore be beneficial in assisting and accelerating children's appreciation of the concept of calculated risk and development in skill formation.

It is often said that chess is an easy game to learn but difficult to master. Our intervention helped to introduce the game of chess to students who had, in general, previously not been exposed to the game. Beyond the rules of how pieces move and how the game is won, strategy and tactics in various phases of the game were also introduced. It was ascertained that approximately nine out of ten students continued to play and practice chess when they were asked 9–10 months after the intensive three-week chess course ended. It is plausible that this repeated playing and honing of their skills could have contributed to a better appreciation for the concept of risk-taking, leading to a reduction in risk aversion. Our findings are consistent with evidence showing a link between cognitive development and a reduction in risk aversion (Frederick 2005; Dohmen et al., 2010; Benjamin et al., 2013; Andreoni et al., 2019a). It also highlights that the skill development potentially offered by chess instruction need not be realized primarily through traditionally recognized cognitive outcomes such as math scores.

Our findings indicate that teaching children basic strategy and tactics in chess has a modest effect on academic outcomes among rural children in a developing country like Bangladesh, but the effects are not strong. These results are not inconsistent with the findings from Jerrim et al. (2018), who did not find significant effects of chess training on academic outcomes for students in an urban setting in UK. However, our results are important as we examine both cognitive and non-cognitive outcomes and uncovered a link between chess training and risk preferences. Further work will need to be done in both developing and developed country settings to better understand more precisely the mechanisms underlying how chess can affect the development of risk preferences.

As some of the outcomes examined in this study are new to this literature, further field experiments can help determine the robustness of our findings. Our intervention is based on data from rural areas of a developing country, and the results obtained do not necessarily have external validity. Nonetheless, by focusing the intervention on a group of children who essentially had no prior experience playing chess and who

Table 10
Effect sizes for risk and time preferences.

	Intervention	Treatment effect (%)	Treatment effect (s.d.)	Treatment effect (other)
Risk Preferences				
Cameron and Shah (2015)	Exposure to an earthquake or flood for families with young children in Indonesia.	n/a	n/a	−41 probability of choosing the two riskiest lotteries
Castillo (2020)	Exposure to domestic violence as a child in Peru.	n/a	−0.66	n/a
Eckel et al. (2012)	High-school students' exposure to other students from low-income families in the USA.	−25	n/a	−11 percentage points
This study (wave 2)	Chess instruction and play for grade 5 children in Bangladesh.	66	1.05	29 percentage points
Time Preferences				
Alan and Ertac (2018) (wave 2)	Teaching 3rd and 4th graders in Turkey to be patient as part of the school curriculum over several months.	23	0.27	n/a
Andreoni et al. (2019b)#	Pre-school and parenting program involving an environment and activities that "promoted patience" of 3–12 year olds in Chicago.	n/a	−0.11 to 0.10	n/a
Berry et al. (2018)#	Financial literacy to students grades 5 and 7 in Ghana.	0.2	n/a	n/a
This study (wave 2)#	Chess instruction and play for grade 5 children in Bangladesh.	2	0.04	n/a

Notes: A positive coefficient indicates increased risk-tolerance (patience). Studies marked with # are insignificant at the 10% level. 'Treatment effect (%)' indicates the treatment effect as a percentage of the control mean; 'Treatment effect (s.d.)' indicates the treatment effect when the dependent variable has been standardized; in 'Treatment effect (other)', percentage points indicates the treatment effect as a percentage of the maximum value the variable can take. Treatment effects for Eckel et al. (2012) are calculated by taking the expected value of the risk instrument, with the most (least) risky option assigned a value of 6 (1). Treatment effects for Berry et al. (2018) are the average across the immediate and delayed tasks.

did not have access to many contemporary toys and games common in developed countries (e.g. board games, computer games, mobile devices, Lego, etc.) that provide mental stimulation, we potentially allow for a fuller impact of chess lessons (if any) to emerge and be realized.

Authors statement

As part of our submission of "The Effects of Chess Instruction on Academic and Non-Cognitive Outcomes: Field Experimental Evidence from a Developing Country" to the Journal of Development Economics, we would like to disclose the following information according to the JDE disclosure requirements:

Asad Islam.

(1) I received a grant from Monash Business School.
(2) I did not receive any payment or personal support from any interested party.
(3) I hold no positions as officer, director, or board member in organizations relevant to this study.
(4) I have no disclosures regarding a relative or partner.
(5) No party had the right to review the paper before circulation.
(6) I have no conflicts of interest to report.

Wang-Sheng Lee.

(1) I received a grant from Deakin University.
(2) I did not receive any payment or personal support from any interested party.
(3) I hold no positions as officer, director, or board member in organizations relevant to this study.
(4) I have no disclosures regarding a relative or partner.
(5) No party had the right to review the paper before circulation.
(6) I have no conflicts of interest to report.

Aaron Nicholas.

(1) I received a grant from Deakin University.
(2) I did not receive any payment or personal support from any interested party.
(3) I hold no positions as officer, director, or board member in organizations relevant to this study.
(4) I have no disclosures regarding a relative or partner.
(5) No party had the right to review the paper before circulation.
(6) I have no conflicts of interest to report.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jdeveco.2020.102615.

References

Ahmed, M., Ahmed, K., Khan, N., Ahmed, R., 2007. Access to Education in Bangladesh - Country Analytic Review of Primary and Secondary Education. Institute of Educational Development, BRAC University.

Alan, S., Ertac, S., 2018. Fostering patience in the classroom: results from randomized educational intervention. J. Polit. Econ. 126, 1865–1911.

Anderson, M., 2008. Multiple inference and gender differences in the effects of early intervention: a reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. J. Am. Stat. Assoc. 103, 1481–1495.

Andreoni, J., Girolamo, A., List, J., Mackevicius, C., Samek, A., 2019a. Risk preferences of children and adolescents in relation to gender, cognitive skills, soft skills, and executive functions. J. Econ. Behav. Organ. (in press).

Andreoni, J., Kuhn, M., List, J., Samek, A., Sokal, K., Sprenger, C., 2019b. Toward an understanding of the development of time preferences: evidence from field experiments. J. Publ. Econ. 177, 104039.

Andreoni, J., Kuhn, M., List, J., Samek, A., Sprenger, C., 2017. Field Experiments on the Development of Time Preferences (No. 00615). The Field Experiments Website.

Andreoni, J., Sprenger, C., 2012. Estimating time preferences from convex budgets. Am. Econ. Rev. 102, 3333–3356.

Angerer, S., Lergetporer, P., Glätzle-Rützler, D., Sutter, M., 2015. How to measure time preferences in children: a comparison of two methods. Journal of the Economic Science Association 1, 158–169.

Banerjee, A., Cole, S., Duflo, E., Linden, L., 2007. Remedying education: evidence from two randomized experiments in India. Q. J. Econ. 122, 1235–1264.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., Walton, M., 2017. From proof of concept to scalable policies: challenges and solutions, with an application. J. Econ. Perspect. 31, 73–102.

Becker, G., Mulligan, C., 1997. The endogenous determinants of time preferences. Q. J. Econ. 112, 729–758.

Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on treatment effects after selection among high-dimensional controls. Rev. Econ. Stud. 81, 608–650.

Belloni, A., Chernozhukov, V., Hansen, C., 2015. High-dimensional methods and inference on structural and treatment effects. J. Econ. Perspect. 28, 29–50.

Benjamin, D., Brown, S., Shapiro, J., 2013. Who is 'behavioral'? Cognitive ability and anomalous preferences. J. Eur. Econ. Assoc. 11, 1231–1255.

Benjamini, Y., Krieger, A., Yekutieli, D., 2006. Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93, 491–507.

Berry, J., Karlan, D., Pradhan, M., 2018. The impact of financial education for youth in Ghana. World Dev. 102, 71–89.

Bettinger, E., Slonim, R., 2007. Patience among children. J. Publ. Econ. 91, 343–363.

Binev, S., Attard-Montalto, J., Deva, N., Mauro, M., Takkula, H., 2011. Declaration of the European Parliament, 0050/2011.

Bloom, H., 1995. Minimum detectable effects: a simple way to report statistical power of experimental designs. Eval. Rev. 19, 547–556.

Brodeur, A., Cook, N., Heyes, A., 2020. A proposed specification check for p-hacking. AEA Papers and Proceedings 110, 66–69.

Burde, D., Linden, L., 2013. Bringing education to Afghan girls: a randomized controlled trial of village-based schools. Am. Econ. J. Appl. Econ. 5, 27–40.

Carneiro, P., Crawford, C., Goodman, A., 2007. The Impact of Early Cognitive and Noncognitive Skills on Later Outcomes. Centre for the Economics of Education Discussion Paper No. 92, UK.

Cameron, A., Gelbach, J., Miller, D., 2008. Bootstrap-based improvements for inference with clustered errors. Rev. Econ. Stat. 90, 414–427.

Cameron, L., Shah, M., 2015. Risk-taking behavior in the wake of natural disasters. J. Hum. Resour. 50, 484–515.

Cassar, A., Healy, A., von Kessler, C., 2017. Trust, Risk and Time Preferences after Natural Disasters. World Development, p. 94, 2017.

Castillo, M., 2020. Negative Childhood Experiences and Risk Aversion: Evidence from Children Exposed to Domestic Violence. IZA Discussion Paper No. 13320.

Castillo, M., Ferraro, P., Jordan, J., Petrie, R., 2011. The today and tomorrow of kids: time preferences and educational outcomes of children. J. Publ. Econ. 95, 1377–1385.

Castillo, M., Jordan, J., Petrie, R., 2018. Discount rates of children and high school graduation. Econ. J. 129, 1153–1181.

Charness, G., Gneezy, U., Imas, A., 2013. Experimental methods: eliciting risk preferences. J. Econ. Behav. Organ. 87, 43–51.

Chassy, P., Gobet, F., 2015. Risk taking in adversarial situations: civilization differences in chess experts. Cognition 141, 36–40.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., Rogers, F., 2006. Missing in action: teacher and health worker absence in developing countries. J. Econ. Perspect. 20, 91–116.

Clarke, D., Romano, J., Wolf, M., 2019. The Romano-Wolf Multiple Hypothesis Correction in Stata. IZA Discussion Paper No. 12845.

Coller, M., Williams, M., 1999. Eliciting individual discount rates. Exp. Econ. 2, 107–127.

Cunha, F., Heckman, J., 2007. The technology of skill formation. Am. Econ. Rev. 97, 31–47.

Davis, S., Eppler-Wolff, N., 2009. Raising Children Who Soar. Teachers College Press, New York.

Diller, L., Ben-Yishay, Y., Gerstman, L., Goodkin, R., Gordon, W., Weinberg, J., 1974. Studies in Cognition and Rehabilitation in Hemiplegia [Rehabilitation Monograph No. 50]. New York University Medical Center Institute of Rehabilitation Medicine, New York.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., 2010. Are risk aversion and impatience related to cognitive ability? Am. Econ. Rev. 100, 1238–1260.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., Wagner, G., 2011. Individual risk attitudes: measurement, determinants, and behavioural consequences. J. Eur. Econ. Assoc. 9, 522–550.

Dreber, A., Gerdes, C., Gränsmark, P., 2013. Beauty queens and battling knights: risk taking and attractiveness in chess. J. Econ. Behav. Organ. 90, 1–18.

Duflo, E., Dupas, P., Kremer, M., 2011. Peer effects, teacher incentives, and the impact of tracking: evidence from a randomized evaluation in Kenya. Am. Econ. Rev. 101, 1739–1774.

Eckel, C., Grossman, P., Johnson, C., de Oliveira, A., Rojas, C., Wilson, R., 2012. School environment and risk preferences: experimental evidence. J. Risk Uncertain. 45, 265–292.

Eriksen, K.W., Kvaløy, O., 2017. No guts, no glory: an experiment on excessive risk-taking. Rev. Finance 21, 1327–1351.

Franklin, B., 1786. The morals of chess. Columbian Magazine 159–161. December issue.

Frederick, S., 2005. On the ball: cognitive reflection and decision-making. J. Econ. Perspect. 19, 25–42.
Frederick, S., Loewenstein, G., O'Donoghue, T., 2002. Time discounting and time preference: a critical review. J. Econ. Lit. 40, 351–401.
Gneezy, U., Potters, J., 1997. An experiment on risk taking and evaluation periods. Q. J. Econ. 112, 631–645.
Guilford, J.P., 1967. The Nature of Human Intelligence. McGraw-Hill, New York.
Guiso, L., Paiella, M., 2008. Risk aversion, wealth, and background risk. J. Eur. Econ. Assoc. 6, 1109–1150.
Hanaoka, C., Shigeoka, H., Watanabe, Y., 2018. Do risk preferences change? Evidence from the Great East Japan earthquake. Am. Econ. J. Appl. Econ. 10, 298–330.
Harrison, G., Lau, M., Williams, M., 2002. Estimating individual discount rates in Denmark: a field experiment. Am. Econ. Rev. 92, 1606–1617.
Heckman, J., Stixrud, J., Urzua, S., 2006. The effects of cognitive and noncognitive abilities on labour market outcomes and social behaviour. J. Labor Econ. 24, 411–482.
Hedges, L., Hedberg, E., 2007. Intraclass correlation values for planning group-randomized trials in education. Educ. Eval. Pol. Anal. 29, 60–87.
Imai, K., Keele, L., Yamamoto, T., 2010. Identification, inference, and sensitivity analysis for causal mediation effects. Stat. Sci. 25, 51–71.
Imai, K., Tingley, D., Yamamoto, T., 2013. Experimental designs for identifying causal mechanisms. J. Roy. Stat. Soc. 176, 5–51.
Islam, A., 2019. Parent-teacher meetings and student outcomes: evidence from a developing country. Eur. Econ. Rev. 111, 273–304.
Islam, A., Mahmud, M., Raschky, P., 2020. Natural disaster and risk-sharing behavior: evidence from a field experiment. J. Risk Uncertain. (forthcoming).
Jerrim, J., Macmillan, L., Micklewright, J., Sawtell, M., Wiggins, M., 2016. Chess in Schools: Evaluation Report and Executive Summary. University College, London: UK.
Jerrim, J., Macmillan, L., Micklewright, J., Sawtell, M., Wiggins, M., 2018. Does teaching children how to play cognitively demanding games improve their educational attainment? Evidence from a randomised controlled trial of chess instruction in England. J. Hum. Resour. 53, 993–1021.
Kremer, M., Miguel, E., Thornton, R., 2009. Incentives to learn. Rev. Econ. Stat. 91, 437–456.
Kumar, A., Saqib, N., 2017. School absenteeism and child labor in rural Bangladesh. J. Develop. Area. 51, 299–316.
Li, J., Li, S., Wang, W., Rao, L., Liu, H., 2011. Are people always more risk averse after a major snow-hit and a major earthquake in China in 2008. Appl. Cognit. Psychol. 25, 104–111.
Levitt, S., List, J., Sadoff, S., 2011. Checkmate: exploring backward induction among chess players. Am. Econ. Rev. 101, 975–990.
McEwan, P., 2015. Improving learning in primary schools of developing countries: a meta-analysis of randomized experiments. Rev. Educ. Res. 85, 353–394.
Muralidharan, K., Singh, A., Ganimian, A., 2019. Disrupting education? Experimental evidence on technology-aided instruction in India. Am. Econ. Rev. 109, 1426–1460.
Runco, M., Sakamoto, S., 1999. In: Sternberg, R. (Ed.), Experimental Studies of Creativity. Cambridge University Press, UK. Handbook of Creativity.
Sala, G., Gobet, F., 2016. Do the benefits of chess instruction transfer to academic and cognitive skills? A meta-analysis. Educ. Res. Rev. 18, 46–57.
Schildberg-Hörisch, H., 2018. Are risk preferences stable? J. Econ. Perspect. 32, 135–154.
Scholz, M., Niesch, H., Steffen, O., Ernst, B., Loeffler, M., Witruk, E., et al., 2008. Impact of chess training on mathematics performance and concentration ability of children with learning disabilities. Int. J. Spec. Educ. 23, 138–148.
Schneider, W., Gruber, H., Gold, A., Opwis, K., 1993. Chess expertise and memory for chess positions in children and adults. J. Exp. Child Psychol. 56, 328–349.
Spadoni, L., Potters, J., 2018. The effect of competition on risk taking in contests. Games 9, 72.
Sutter, M., Kocher, M., Glätzle-Rützler, D., Trautmann, S., 2013. Impatience and uncertainty: experimental decisions predict adolescents' field behavior. Am. Econ. Rev. 103, 510–531.
Tietjen, K., Rahman, A., Spaulding, S., 2004. Bangladesh Educational Assessment Time to Learn: Teachers' and Students' Use of Time in Government Primary Schools in Bangladesh. USAID.
Torrance, E., 1966. The Torrance Tests of Creative Thinking – Norms: Technical Manual Research Edition—Verbal Tests, Forms A and B— Figural Tests, Forms A and B. Personnel Press, Princeton, NJ.
Travis, F., 1998. Cortical and cognitive development in 4th, 8th and 12th grade students: the contribution of speed of processing and executive functioning to cognitive development. Biol. Psychol. 48, 37–56.
Trinchero, R., Sala, G., 2016. Can chess training improve Pisa scores in Mathematics? An experiment in Italian primary schools. Eurasia J. Math. Sci. Technol. Educ. 12, 655–668.
Tymula, A., Belmaker, L., Roy, A., Ruderman, L., Manson, K., Glimcher, P., Levy, I., 2012. Adolescents' risk-taking behavior is driven by tolerance to ambiguity. Proc. Natl. Acad. Sci. Unit. States Am. 109 (42), 17135–17140.
Unterrainer, J., Kaller, C., Leonhart, R., Rahm, B., 2011. Revising superior planning performance in chess players: the impact of time restriction and motivation aspects. Am. J. Psychol. 124, 213–225.
Waters, A., Gobet, F., Leyden, G., 2002. Visuospatial abilities in chess players. Br. J. Psychol. 30, 303–311.
Wechsler, D., 1991. Wechsler Intelligence Scale for Children, third ed. Psychological Corporation, San Antonio, TX.