Journal of Public Economics 223 (2023) 104908

Contents lists available at ScienceDirect

Journal of Public Economics

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

Self-assessment: The role of the social environment ☆

Armin Falk [a], Fabian Kosse [b,*], Hannah Schildberg-Hörisch [c], Florian Zimmermann [a]

[a] Institute on Behavior and Inequality (briq) and University of Bonn, Germany
[b] University of Würzburg and Institute on Behavior and Inequality (briq), Germany
[c] University of Düsseldorf and Max Planck Institute for Research on Collective Goods, Bonn, Germany

ARTICLE INFO

Article history:
Received 24 September 2021
Revised 17 February 2023
Accepted 2 May 2023
Available online xxxx

JEL-Codes:
D01
C21
C91
I24

Keywords:
Self-assessment
Randomized intervention
Children
Child development
Socioeconomic status

ABSTRACT

This study presents evidence on the role of the social environment in shaping the accuracy of self-assessment. We introduce a new measurement tool to elicit children's accuracy of self-assessment. We use this tool to show that children from high SES families are more accurate in their self-assessment, compared to children from low SES families. We then exploit exogenous variation of participation in a mentoring program designed to enrich the social environment of children. The mentoring program has a causal positive effect on the accuracy of self-assessments and is most effective for children whose parents provide few interactive activities.

1. Introduction

Many decisions of economic relevance entail an element of self-assessment of one's abilities. Should I go to college or pursue an apprenticeship? Should I choose a more or less challenging career path? Should I open a restaurant or not? All these decisions require prior reflection of own abilities, strengths, and weaknesses, and more accurate self-assessments will, on average, yield better decisions. To provide an example, consider the case of selecting a college major. Suppose Jessica is deciding between math and literature. She likes both equally well, but she is relatively better at math than literature and hence would be more successful if she made math her major. Accurate self-assessment would lead Jessica to indeed choose math, while inaccurate self-assessments might induce her to pick the suboptimal literature major (either because she underestimates her math skills or because she overestimates her abilities in literature).[1]

Reaching an accurate self-assessment can be challenging, in particular because self-assessments often need to be based on only few experiences with the task at hand. In order to arrive at an accurate self-assessment, people hence need to be able to effectively use the limited cues about their task-related skill level available to them, and need to be able to draw inferences from different

☆ We thank Johannes Abeler, Benjamin Enke and Chris Roth for very helpful comments. Financial support through the Leibniz Programme, the CRC TR 190 and the CRC TR 224 of the German Research Foundation (DFG), the European Research Council (ERC Advanced Grant 340950), the Jacobs Foundation, and the Benckiser Stiftung Zukunft is gratefully acknowledged.

* Corresponding author.
E-mail addresses: armin.falk@briq-institute.org (A. Falk), fabian.kosse@uni-wuerzburg.de (F. Kosse), schildberg-hoerisch@dice.hhu.de (H. Schildberg-Hörisch), florian.zimmermann@briq-institute.org (F. Zimmermann).

[1] Survey evidence presented in Figs. B1 and B2 corroborates the intuition that accurate self-assessments are often beneficial. We show that individuals who describe themselves as rather accurate in terms of self-assessment have higher income and are healthier, compared to individuals who either over- or underestimate their abilities. Notice that theories of overconfidence have argued that inflated self-assessments can be beneficial, for instance for motivational reasons (Benabou and Tirole (2002)) or to better convince others of own skills (Schwardmann and van der Weele (2019)). While overconfidence can undoubtedly be advantageous, it has been shown to have very aversive consequences in a variety of contexts, ranging from individual investment (Barber and Odean (2000)) to CEO behavior (Malmendier and Tate (2005)) and international relations (Johnson (2004)). Kahneman concludes that overconfidence is the most severe and important cognitive bias and terms it "the mother of all biases" (Kahneman (2011)).

https://doi.org/10.1016/j.jpubeco.2023.104908
0047-2727/

but related prior experiences. The development of the ability to evaluate one's strengths and weaknesses naturally depends on the social environment. Opportunities to make experiences and to learn from experiences tend to be scarce and their frequency likely depends on the social environment in which people grow up and live in. To fix ideas, think of two otherwise identical individuals, Marta and Paul. While Marta grew up in an environment where abundant experiences and feedback were provided, Paul grew up in an environment where experiences and feedback were rare. It is conceivable that Marta developed a higher ability to judge her own skills, strengths, and weaknesses, compared to Paul. Despite their intuitive importance, little is known about the social determinants of self-assessment abilities.

In this paper, we seek to elucidate the role of the social environment for the development of accurate self-assessments. Making progress on this question is challenging for two reasons. First, to establish causality in the relationship between social environment and self-assessment, one would ideally like to observe exogenous changes in the social environment. However, such changes are rare in naturally occurring data. Second, measuring the accuracy of self-assessment is difficult, because many aspects of people's decision problems are unknown to the researcher. We circumvent these challenges by making use of a combination of field and lab-in-the-field experimental evidence.

We focus on children in elementary school, arguably an important time of development where many skills, abilities and preferences are formed (see, e.g., Cunha and Heckman, 2007; Almås et al., 2010; Fehr et al., 2013; Alan et al., 2017; Charness et al., 2019; Samek et al., 2020; Falk et al., 2021). It is also a critical time for the development of metacognitive capacities (see, e.g., Veenman and Spaans, 2005; Perry et al., 2019). Hence, the experiences and feedback provided via the social environment during that period might be a crucial input into the development of self-assessment abilities. In the first step, we exploit naturally occurring variation in children's social environments as captured by the socioeconomic status (SES) of their parents. SES reflects the "social standing" of individuals or families in society and can be summarized as the level of economic, educational, and time resources available at the household level. It is conceivable that high-SES parents can provide their children with a richer social environment, consisting of more frequent and more diverse opportunities to obtain feedback, thus facilitating the development of the ability to assess own strengths and weaknesses. In the second step, we then seek to establish causality in the role of the social environment. For that purpose, we exploit an exogenous enhancement of the social environment for a randomly determined subgroup of low-SES children via an existing social program in Germany (see Kosse et al. (2020)). In the program, children are provided with a mentor for around one year to enrich their social environment. The mentors introduce the children to new activities and generate new experiences and feedback for them.

After the completion of the mentoring program, we conducted controlled experiments and interviews with the children and parents of the treated and non-treated low SES groups, as well as the high-SES group. Our main goal was to obtain a measure of the accuracy of children's self-assessment. This was challenging for several reasons. First, we had to decide on a decision context, i.e., a task for which children had to assess their ability. Ideally, there should be no ability differences as well as no differences in experience with respect to the specific task between the treatment groups. At the same time, the decision context should provide children with cues about their skills such that a higher self-assessment ability can manifest itself. Second, the measure should reflect the accuracy of self-assessment in an incentivized way. Third, we wanted a forward-looking task where children need to be able to predict how well they will do in a future task, resembling the decision problems from our opening examples. Fourth, the measurement exercise should be intuitive and easy to comprehend for children that age. We chose a rather abstract decision environment that children were unlikely to have encountered before. We developed a tailored experimental game that, as we argue, meets all four criteria. The experimental task was to hit a small hole with a marble. In order to familiarize children with the task and to provide them with cues that facilitate self-assessment, children could experience the basic task in a practice round. The key decision problem the children then faced was to select a level of difficulty (the size of the hole) for an upcoming task. The trade-off we implemented is that higher difficulty levels yielded a higher reward, but at the same time came with a greater risk of not mastering the task, in which case the reward was zero. The key idea underlying the design of the game is that ceteris paribus, more accurate judgments of own skill levels allow children to achieve higher earnings in expectation, through the choice of more appropriate difficulty levels. Hence, controlling for other factors, rewards in the task serve as a measure of the accuracy of self-assessment. At the same time, the game is simple to understand and intuitive for children at that age.

Our first result is that children from families with high SES demonstrated more accurate self-assessments (as measured by higher earnings in the self-assessment game) compared to children from families with low SES. We then establish causality in the relationship between social environment and accurate self-assessments by exploiting the random variation in whether low-SES children participated in the mentoring program. Our second and main result is that the enhanced social environment substantially and significantly improves the accuracy of self-assessment. Both results are robust to using different empirical specifications and controlling for selective attrition, ability, as well as risk preferences. Taken together, these findings highlight the importance of the social environment as a causal determinant of self-assessment abilities.

We proceed by delving into the underlying mechanisms of our key result. Specifically, we seek to understand which environmental factors are missing in low-SES families that affect accurate self-assessments. Intuitively, feedback and experiences are prime candidates. The repeated exposure to new experiences alongside with the provision of feedback might help children to better assess their ability in a novel task. This intuition is bolstered by the literature on the development of metacognition, which emphasizes that feedback and learning opportunities are crucial inputs for the development of metacognitive capacities (Flavell (1979)). Specifically, frequent experiences and feedback might allow children to develop the ability to efficiently use cues that can help them evaluate their task-related skill level. A richer set of prior experiences from different but related tasks should also facilitate the assessment of abilities in new tasks. To make progress, we look at self-reports obtained from parents about the nature of the social environment they provide to their children. We proxy the opportunities to learn about oneself provided by the social environment by looking at the number of highly interactive activities undertaken with the children (e.g. having a conversation, playing board or card games, doing handicrafts, having a snack together, playing music together). We find that these activities are correlated with the accuracy of self-assessment as measured by our paradigm and we find that the effect of participation in the mentoring program is more pronounced for children whose parents provided few

highly interactive activities. These results suggest that highly interactive activities and the associated richness of feedback are a key aspect of the social environment that determines accurate self-assessment. It also appears to be an aspect that is missing in low-SES environments. Crucially, the mentoring program appears to fill this gap and thereby improves the belief accuracy of low-SES children.

In a final step, we empirically gauge the relation between our measure of self-assessment ability and this ability in other domains. A possible concern might be that our rather abstract measure is not associated with self-assessment abilities in other, perhaps more relevant domains of decision-making. For this purpose, we focus on self-assessments about school performance, arguably a very important domain for both children and parents. In fact, self-assessment regarding school performance has been shown to be an important driver of long-run school success (Goux et al. (2017)). We measure children's subjective beliefs about their school performance and contrast these beliefs with actual school performance. Two basic patterns emerge: (i) Self-assessment as measured in the marble game predicts self-assessment regarding school performance ($p = 0.059$); (ii) the mentoring program positively affects the accuracy of self-assessment regarding school performance ($p = 0.053$). Taken together, we provide evidence that our measure of self-assessment directly relates to self-assessment abilities in an arguably highly relevant domain, and that the mentoring program also positively affects accuracy of self-assessment in that domain.

It is well-documented that children from families with high and low SES differ in important life outcomes, such as educational attainment and labor market success (e.g. Bradley and Corwyn, 2002; Heckman et al., 2006; Duncan et al., 2011). Prevalent explanations for differences in educational or labor market attainment largely focus on differences in children's cognitive and non-cognitive skills, such as IQ, persistence, and patience (see e.g. Heckman and Vytlacil, 2001; Duckworth et al., 2007; Hanushek and Woessmann, 2008; Humphries and Kosse, 2017). Our paper relates to this literature and reveals that SES predicts the ability to assess one's strengths and weaknesses, arguably a key determinant of the quality of economic decision-making.

More specifically, the findings from this paper relate to an active literature that analyzes the causal effect of the social environment on skills and preferences (e.g. Charness et al., 2019; Alan et al., 2019; Kosse et al., 2020; Sorrenti et al., 2020; Berger et al., 2020; Cappelen et al., 2020). Our paper contributes to this literature by focusing on the self-assessment of skills. In other words, while existing work has highlighted the crucial role of a broad set of social and environmental factors for the development of both cognitive and non-cognitive skills, our work sheds light on the role of the social environment for the ability to self-assess these skills.[2]

In the next section, we present details about the study design and our main outcome variables. Section 3 summarizes our findings and Section 4 concludes.

2. Study design and data

2.1. Recruitment and randomization

Fig. B3 in Appendix B presents a flow chart of the timing, sampling, and procedural details of the study (see also Kosse et al. (2020) and Falk and Kosse (2021) for further details). Recruitment started in the summer of 2011. We used official registry data to obtain the addresses of families (with children aged from seven to nine) living in the German cities of Bonn and Cologne. Families were contacted via postal mail and informed about the possibility to take part in the mentoring program and the interviews. We informed parents that participation in the mentoring program was not guaranteed due to limited capacity. The interested families were asked to fill out and return a short questionnaire concerning the socioeconomic characteristics of the household and to sign a non-binding letter of intent to take part in the interviews and the mentoring program. We received 1,626 complete responses and, based on the questionnaire, we categorized respondents as either high or low-SES households.[3]

All low-SES families that expressed interest were invited to take part in the study. To take part, families had to participate in a baseline wave of interviews (fall 2011) and provide written consent to allow the transmission of their addresses to the mentoring program. Importantly, the mentoring program could only accommodate 212 families; hence, out of 590 low-SES families who participated in the baseline wave and gave consent, 212 were *randomly* selected and constitute our treatment group (Treatment Low SES). The remaining 378 families form the control group (Control Low SES).[4]

Notice that the actual assignment of mentors to children in Treatment Low SES was conducted by the mentoring program. Each child in the treatment group could potentially be matched, but not all selected children were matched in the end. A mentor–mentee match was successfully implemented for 151 of the 212 children. For the remaining 61, matches could not be realized due to a local shortage of mentors, mentor refusals, or coordination problems between mentors and families (e.g. pregnancy of the mentor or moving of mentor or family). In the analysis, we hence focus on intent-to-treat effects (ITT) between Treatment and Control Low SES.

We also invited 150 randomly-chosen high-SES families to take part in the study (not the mentoring program). 122 took part in the baseline wave of interviews and serve as an additional benchmark group (Control High SES).

After the one-year mentoring program, all families that participated in the baseline wave (Treatment Low SES, Control Low SES, and Control High SES) were invited to take part in the post-treatment interviews and experiments (post-treatment wave) in which all of our main outcome variables were elicited.[5]

[2] As such, our findings also relate to the literature that studies beliefs about own skills and abilities in the lab (e.g. Eil and Rao, 2011; Möbius et al., 2022; Zimmermann, 2020) and in the field (e.g. Huffman et al., 2022; Malmendier and Tate, 2005, 2008). A common finding in this literature is that people on average tend to be overconfident, although people sometimes also appear underconfident. Moore and Don (2008) provide evidence that the incidence of over- and underconfidence crucially depends on task difficulty. Importantly, thus far this literature has not considered the long-run effects of the social environment on the development of beliefs about own skills and abilities. The developmental psychology literature has focused on notions of self-esteem and self-confidence and their development in children and adolescents (e.g. Twenge and Campell, 2001; Wigfield et al., 1991). Apart from the focus on different outcome measures, this literature does not study the causal role of the social environment on these outcomes.

[3] SES reflects the level of resources available at the household level, i.e., material, educational, and time resources. Accordingly, a household was classified as low SES if at least one of the three following criteria was met (see Kosse et al. (2020) for further details): (i) *Low income*: Equivalence income of the household is lower than 1,065 Euro. This corresponds to the 30% quantile of the German income distribution. (ii) *Low education*: Neither the mother nor the father of the child has a school-leaving degree qualifying for university studies. (iii) *Single-parent status*: A parent is classified as a single parent if he/she is not living together with a partner.

[4] Randomization was stratified by city (Cologne or Bonn) and SES criteria, for a total of 14 strata. Given the larger relative supply of mentors in Bonn, we assigned a higher share of children in Bonn to the treatment group. Thus, the assignment into treatment was random conditional on location. Therefore, we condition on location for the analyses.

[5] 85.3% of the families took part in this second wave of interviews and experiments. See Sections 3.1 and 3.4 for discussions of sample balance and attrition.

2.2. The mentoring program

We exogenously enhanced the social environment of the treated low-SES families with the help of an existing and well-established non-profit mentoring program in Germany, "Balu und Du".[6] In this program, elementary school children are provided with a mentor for up to one year. It is a one-to-one mentoring program which means that every mentor is assigned to only one child. The mentors are predominantly university students (aged from 18 to 30) who volunteer to serve as a mentor for a child. About 80% of the mentors are females. A mentor typically spends one afternoon per week in one-to-one interactions with his/her mentee.

The mentoring program is not targeted toward specific learning goals (such as improved school grades), but rather to enrich the social environment of children. A key component of the program is to introduce children to new activities and experiences such as cooking, sports, handicraft work, or visiting a zoo, museum, or playground. The broad goal is hence to expose children to new experiences, and provide feedback; possibly exactly the inputs that are needed for them to develop an accurate sense of their abilities and that might be missing in low-SES families.

To date, "Balu und Du" has arranged and supervised more than 15,800 mentor-child relationships in more than 50 different locations in Germany. The mentoring program is embedded in a tightly organized structure. Every week, mentors complete an online report in which they document their activities with the child. Program coordinators offer support whenever necessary and provide coaching and advice to mentors. They also organize bi-weekly monitoring meetings in which mentors receive suggestions for new activities to enrich the environment of the child and where potential problems can be discussed.

The mentoring program is designed to last up to 12 months. In our sample, the average duration of mentor–mentee relationships was 9.3 months. Variation in duration is mainly due to unforeseeable events such as moving decisions of parents or mentors due to a job change. On average, treated children met their mentor 22.8 times (std. dev. 11.9), typically for an entire afternoon (amounting to an average total of around 92 h).

2.3. Setting of experiments and procedures

In both waves of the experiment, the child was accompanied by one parent. In 95% of the cases, the interviewed parent was the biological mother. Therefore, for convenience, we use the term "mother" for the adult who was interviewed. The interviews took place at central locations in Bonn and Cologne. The interviews and experiments were conducted according to a detailed protocol. During the interviews and experiments, the interviewer, the mother and the child were in the same room. However, a standardized seating plan ensured that the mother and child did not have eye contact and could not communicate otherwise. The interviews lasted about one hour and, for participation in the interview, mothers received 35 Euros at baseline and 45 Euros in the post-treatment wave.

The children participated in several experiments and intelligence tests and answered a brief questionnaire. The experiments were incentivized using toys. We introduced an experimental currency called "stars". At the end of the experiment, children could exchange their stars for toys. Toys were arranged in four categories that increased in objective value and subjective attractiveness to children. Children were told that more stars would result in the option of choosing toys from a higher category.[7]

We took great care to create a pleasant interview situation. One experimenter ran experiments with only one child at a time. During the experiments, mothers completed a comprehensive survey covering topics such as basic information about the child, assessments of personality and attitudes of the child, the socioeconomic background of the family, details on how the parent(s) spend time with the child including joint activities, as well as economic preferences, personality, and attitudes of the mother.

Several measures were implemented to mitigate potential concerns related to biased reporting and experimenter demand effects (DeQuidt et al. (2018)): (i) mentors received no information about the elicited measures, to avoid any form of "training to the test"; (ii) experimenters were not informed about the purpose of the study or the treatment assignment of the participating families; (iii) the intervention was not mentioned during the data collection phase, and (iv) the research team never interacted directly with the children or their parents.

2.4. Main variables

In the following, we summarize our key outcome measures.[8]

Accuracy of Self-Assessments: We designed an experimental paradigm with three main goals in mind. The paradigm should: (i) provide children with cues about their skills such that a higher self-assessment ability can manifest itself. At the same time, there should be no ability differences or differences in experience with respect to the specific paradigm between treatment groups; (ii) provide an incentivized measure that reflects the accuracy of self-assessments; (iii) be forward-looking in the sense that children need to try to predict future performance in different scenarios, and (iv) be intuitive and easy to comprehend for children of age eight or nine. The latter goal, in particular, posed a challenge. Arguably, the sophisticated state of the art belief elicitation paradigms that pervade modern experimental economics are unsuitable for children of that age. Hence, we opted for the following more intuitive paradigm.

The basic experimental task was to hit a small hole with a marble (see Fig. B4 in Appendix A). The children could first experience the task in a practice round (10 trials) where the level of difficulty was fixed at a medium level. This allowed them to become acquainted with the task and to collect cues with respect to their task-related ability. In addition, it served as an individual-level "marble lane ability" measure which we use as a control variable in the analysis. At the same time, the abstract nature of the task made it unlikely that children from either treatment group have actually experienced that specific task before. Indeed, as we verify below, there is no treatment effect on marble lane ability (see Section 3.4).

After, the practice round, the key decision problem we implemented was that children had to select a level of difficulty (the size of the hole). The basic trade-off we implemented was that higher difficulty levels yielded a higher reward, but at the same time came with a greater risk of not mastering the task, in which case the reward was zero. Specifically, the difficulty levels were varied via the size of the hole (21, 18, 15, 12, 9, 6, or 3 cm diameter) that needed to be hit with the marble to score. Hitting fewer than five times in ten attempts resulted in zero earnings. Scoring at least five times in ten attempts on the chosen marble lane resulted in positive earnings, and earnings linearly increased with the difficulty of the chosen lane. For scoring at least five times on the easiest marble lane, a child earned one unit of the experimental currency; for scoring at least five times on the most difficult marble lane, a child earned seven units of the experimental currency. The key idea underlying the design of the game was that, ceteris paribus, more accurate self-assessment of skill levels allow children to obtain

[6] More details about the mentoring program can be found on www.balu-und-du.de.

[7] We ensured that each additional star that would not result in a higher category was nevertheless valuable: these stars were exchanged into "Lego" bricks.

[8] Kosse et al. (2020) summarize additional measures unrelated to this paper, such as measures of prosociality.

higher rewards, in expectation, due to more appropriate lane choices. At the same time, the game is simple to understand and intuitive for children at that age. Hence, controlling for other factors, the rewards in the task serve as a measure of the accuracy of self-assessment.

To see this, consider a child with a given ability level. We can conceptualize this ability as a function $f(.)$ that translates each difficulty level (each lane) into a success probability (i.e. the probability of scoring at least five times in 10 attempts). Given this true ability, there exists an optimal lane choice that maximizes expected earnings. For simplicity, let's assume risk neutrality.[9] Of course, children might not know their true ability. Self-assessment of their ability can be captured by function $b(.)$ which translates each difficulty (each lane) into a *perceived* success probability. A perfectly accurate self-assessment now implies that $b(.) = f(.)$ for all lanes. As a consequence, a child with perfect self-assessment skills will pick the lane that maximizes expected earnings. Instead, a child with an imperfect self-assessment ($b(.) \neq f(.)$ for some lanes), might not pick the lane that maximizes expected earnings and hence in expectation will have lower earnings.[10]

The downside of implementing our more intuitive paradigm is that other factors, namely ability in the marble game, risk preferences as well as motivation, and effort, potentially might play a role and, jointly with the accuracy of self-assessments, determine rewards in the task. Therefore, we make extensive use of our measures of ability, risk preferences, motivation, and effort (see below) to account for their potential roles in various specifications. Section 3.4 summarizes how ability, risk preferences, motivation and effort could affect earnings in the task and how we address this empirically.

The measure was elicited in the post-treatment wave. Before the start of the game, the experimenter asked several control questions to carefully check the child's understanding of the game and, if necessary, explained the rules again. During the actual game, we took care to minimize the role of the experimenter in order to avoid potential social image concerns of the children vis a vis the experimenter.[11] Appendix C.1 contains a translated version of the exact wording of the instructions given to experimenters and children.

Risk preferences: Similar to our belief measure, measuring risk preferences among children poses a challenge. While one would ideally want to implement standard price lists or BDM mechanisms, one needs to ensure that children intuitively understand the risk-return trade-off. With this in mind, we implemented the following risk elicitation task (see Falk et al. (2021)).[12] We measured children's risk attitudes in the post-treatment wave by presenting two coins to children (situation A): one with three stars printed on each side, the other with seven stars on one side and zero on the other. The children had to choose which coin would be tossed. The safe value of three was also "determined" by a coin toss to ensure that children do not prefer the risky option for its higher entertainment value. After children made their decision, but before actually tossing the chosen coin, the experimenter presented two additional coins in another color (situation B): one with four stars on each side, the other, as before, with seven stars on one side and zero on the other. Again, children had to choose which coin would be tossed and the interviewer then tossed the two chosen coins. The order in which the two variations of the game (situation A and situation B) were played was randomized. In the analyses, we use the number of risky choices to control for risk preferences. While this approach certainly only delivers a coarse measure of risk attitudes, we argue that the elicitation method is appropriate for children that age.

Effort and motivation: In order to get a measure of willingness to exert effort or general motivation in the experiments, we measured willingness to exert effort in an independent task. The children had to work on a tedious and non-incentivized real-effort task,[13] checking for mistakes in sequences of letters and numbers, for four minutes. The children could stop at any time without any consequences. Our measure of effort and motivation takes the value of one if a child voluntarily worked on the task for four minutes and zero if they stopped before.

Social interaction patterns: The main goal here was to understand which aspects that are being offered by the mentoring program might be missing in low-SES environments. The literature suggests that feedback and learning opportunities are important inputs for the development of metacognitive capacities (Flavell (1979)).

To make progress, we asked mothers in structured interviews at baseline how they spend time with their child. We focused on activities that might be missing in low-SES families and that typically entail learning opportunities and feedback (Flavell (1979)). We also focused on activities that we knew are typically part of the mentoring program. These activities were: having a conversation, playing board or card games, having a snack together, playing music together, or going to music lessons. For each activity, the precise wording of the question was: "How many times during the last 14 days have you or the main caregiver done the following activities together with your child?". We created a variable for "intensity of social interactions" as the average daily frequency of these highly interactive activities.

3. Results

3.1. Preliminaries and data description

Recall that after the one-year treatment period, all families who had participated at baseline were invited to take part in the post-treatment wave. 85.3% (607 out of 712) took part in this second wave of interviews. 596 answered the control questions correctly and constitute our core sample.[14]

Our main treatment comparison is between Treatment Low SES and Control Low SES. Columns 1 and 2 in Table 1 reveal no detectable sample imbalance regarding our key variables at baseline between the two groups. Columns 3 and 4 indicate no evidence for selective attrition. Attrition is not significantly related to treatment status, performance at baseline, the intensity of interaction at baseline, nor to the respective interactions. Moreover, Table A1 in the Appendix indicates that the follow-up sample is balanced across treatment and control group regarding further characteris-

[9] We cover risk preferences in Section 3.4.

[10] Notice that this holds irrespective of the nature of the imperfect self-assessment. Take the case of of overconfidence, which we conceptualize as $b(.) > f(.)$ for difficult lanes. If this misperception is big enough, a lane more difficult than the optimal lane is chosen. Similarly, think of underconfidence as a case where $b(.) < f(.)$ for difficult lanes. In that case, a lane easier than the optimal is likely to be selected. Importantly, in both cases, the misperception entails a risk of picking a suboptimal lane, i.e. a lane that does *not* maximize expected earnings.

[11] We excluded 11 observations from the analysis (about 2% of observations for each of the three groups, Control High SES, Control Low SES, Treatment Low SES) because these children did not completely understand the rules of the game even after three repeated explanations by the interviewer.

[12] Instructions are shown in Appendix C.2.

[13] Instructions are shown in Appendix C.3.

[14] As we already alluded to in Section 2, as part of this study several measurement exercises of different variables were implemented. We want to emphasize, however, that the results presented in this paper followed a very clear ex-ante hypothesis. We argue that this is both visible from how our hypothesis derives from existing work, as well as from the design of our measure of self-assessment skills, whose purpose follows rather clearly from how we designed it. Therefore, we abstain from employing methods to correct for multiple hypothesis testing in our empirical analysis.

Table 1
Analyses of treatment assignment and attrition. In columns 1 and 2, we test for baseline balance. The dependent variable is one if a child was selected into the Treatment Low SES group and zero if selected into the Control Low SES group. In columns 3 and 4, we test for selective attrition. The dependent variable is one if a child is lost to follow-up, i.e. did not take part in the post-treatment interview, and zero otherwise. All baseline measures were collected before the treatment assignment took place. Coefficients are OLS estimates, standard errors in brackets. ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Sample: Low SES Treatment & Control	Assigned to treatment		Lost to follow-up	
	(1)	(2)	(3)	(4)
Gained stars (baseline, std.)	0.008	0.003		−0.005
	(0.020)	(0.022)		(0.020)
Intense interaction (baseline, std.)	−0.000	−0.006		−0.012
	(0.020)	(0.022)		(0.020)
Treatment dummy			−0.013	−0.012
			(0.033)	(0.033)
Gained stars x Treatment dummy				0.015
				(0.033)
Intense interaction x Treat. dummy				0.017
				(0.033)
Sample restriction	No	Wave 2	No	No
Observations	590	485	590	590
R^2	0.000	0.000	0.000	0.001
p-value F-test (all indep. vars. = 0)	0.925	0.956	0.699	0.977

tics. Nevertheless, we also include inverse probability weighting methods (IPW) to adapt for minor imbalances as a robustness check. We estimated the weights from a linear probability model of a binary selection indicator (indicating whether the self-assessment measure is available for the post-treatment wave), regressed on baseline measures of self-assessment, social interaction and ability, treatment and high-SES dummies, and the interaction of the baseline measures and the group dummies.

Fig. B5 displays histograms for chosen lane difficulty and earnings over all three groups of children (Control Low SES, Treatment Low SES, Control High SES). On average, children chose a lane difficulty of 4.74 (standard deviation: 1.22) and 24% of children failed the task for the chosen difficulty level. Average earnings from the self-assessment task were 3.30 (standard deviation: 2.08) units of the experimental currency (“stars”). For earnings, Fig. B5 reveals substantial left-censoring. Hence, we estimate a Tobit model (lower limit at zero). For robustness, we also provide OLS and Poisson (count data) estimates in Appendix A.

We focused on children in elementary school because the literature suggests that this might be the time in which abilities to accurately judge one’s own strengths and weaknesses are formed. Table A2 in the Appendix corroborates this view and shows positive correlation between the accuracy of self-assessments and the age of the children. This pattern holds irrespective of controlling for ability and risk preferences.

3.2. Socioeconomic status and self-assessment

We begin our analysis of the role of the social environment for the accuracy of self-assessments (measured post-treatment) by comparing children from a low-SES background to children with a high-SES background. While we do not claim causality for this comparison, we nevertheless view it as a useful benchmark exercise that allows us to document the extent to which the accuracy of self-assessment is associated with the social environment in which children grow up.

In all main regressions, we condition on current marble lane ability (dummies for each ability level), interviewer FEs, location, gender, and age. As discussed above, given the nature of our data we report Tobit estimates. Table 2 displays the results of regressing earnings in the self-assessment game on a SES dummy (High SES versus Low SES Control). The results indicate that SES is significantly associated with earnings in the self-assessment game. In column 1, without using further controls, we show that children from high-SES families earn, on average, more than an additional half of a star. In columns 2 and 3 we check for robustness and show that the gap is largely unaffected by systematic attrition (column 2) and heterogeneities in risk preferences (column 3).[15] The gap is sizable and corresponds to about 25% to 30% of a standard deviation and to about 15% to 20% of average earnings.

Table 2
Coefficients are Tobit estimates, standard errors in brackets. All regressions also include a constant, age, gender, location fixed effects (see sampling), interviewer FEs, and marble ability dummies. Marble ability is the performance in the trial round. Willingness to take risk is the number of risky choices (lottery over safe amount). IPWs account for potential selective attrition and are estimated from a linear probability model of a binary selection indicator (indicating whether the self-assessment measure is available for the post-treatment wave) regressed on baseline measures of self-assessment, social interaction, and ability, treatment and high SES dummies, and the interaction of baseline measures and the group dummies. ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Sample: Low & High SES Control	Gained stars (# 0–7)		
	(1)	(2)	(3)
Base: Low SES Control			
High SES dummy	0.612*	0.624**	0.511*
	(0.327)	(0.304)	(0.302)
Inverse probability weighting	no	yes	yes
Controlling for risk pref.	no	no	yes
Mean Low SES Control	3.146	3.146	3.146
Observations	420	420	420

3.3. The mentoring program and self-assessment

We now move to our main analysis and shed some light on the *causal* role of the social environment. For this purpose, we exploit the randomization of low-SES children into the mentoring program. To do so, we follow the same estimation approach as before and regress earnings in the self-assessment game from the post-treatment wave on a treatment dummy (Treatment Low SES versus Control Low SES). In all main regressions, we again condition on ability dummies, interviewer FEs, location, gender, and age.

The results presented in Table 3 (column 1) reveal a pronounced and significant positive causal effect of the enrichment of the social environment on the accuracy of self-assessment. The enrichment of the social environment through the mentoring program increased children’s earnings in the self-assessment task by more than 0.5 stars. There is no evidence that the treatment effect is

[15] The slight decrease in effect size in column 3 relates to a small high-to-low SES gap in risk preferences, see e.g. Falk et al. (2021).

Table 3
Coefficients are Tobit estimates, standard errors in brackets. All regressions also include a constant, age, gender, location fixed effects (see sampling), interviewer FEs, and marble ability dummies. Marble ability is the performance in the trial round. Willingness to take risk is the number of risky choices (lottery over save amount). IPWs account for potential selective attrition and are estimated from a linear probability model of a binary selection indicator (indicating whether the self-assessment measure is available for the post-treatment wave) regressed on baseline measures of self-assessment, social interaction, and ability, treatment and high SES dummies and the interaction of baseline measures and the group dummies. ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Sample: Low SES Treatment & Control	Gained stars (# 0–7)		
	(1)	(2)	(3)
Base: Low SES Control			
Treatment dummy	0.650**	0.639**	0.621**
	(0.283)	(0.284)	(0.279)
Inverse probability weighting	no	yes	yes
Controlling for risk pref.	no	no	yes
Mean Low SES Control	3.146	3.146	3.146
Observations	485	485	485

Table 4
Coefficients are inverse probability weighted (IPW) OLS estimates, standard errors in brackets. All regressions also include a constant, age, gender, location fixed effects (see sampling), interviewer FEs, marble ability dummies and standardized willingness to take risk. IPWs account for potential selective attrition and are estimated from a linear probability model of a binary selection indicator (indicating whether the self-assessment measure is available for the post-treatment wave) regressed on baseline measures of self-assessment, social interaction, and ability, treatment and high SES dummies and the interaction of baseline measures and the group dummies. The are two missings in our dataset for the chosen lane difficulty, due to experimenter misreporting. ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Sample: All children	Chosen lane (# 1–7)	Failure (0/1)
	(1)	(2)
Base: Low SES Control		
Treatment dummy	−0.310***	−0.121***
	(0.113)	(0.042)
High SES dummy	−0.453***	−0.111**
	(0.121)	(0.046)
Mean Low SES Control	4.935	0.294
Observations	594	596

biased by selective attrition (column 2) or the effects of risk preferences (column 3). Relating these effects to the SES gap documented in Table 2, it seems that the mentoring program has the potential to close this gap.[16] These results are robust to various changes in the specification. In Table A3, we show estimates of similar size and significance for estimations without control variables. In Tables A4 and A5, we confirm these results by estimating Poisson and OLS regressions.

To shed light on the underlying choice pattern generated by the mentoring program, Table 4 analyses children's lane choice and their failure rate. To this end, the table shows OLS coefficients of regressing the chosen difficulty level and the failure probability on high-SES and treatment dummies. Column 1 shows that untreated low-SES children selected more difficult lanes compared to treated low-SES children and high-SES children. This lane choice pattern might indicate overly confident self-assessments of the untreated low-SES children. Better developed self-assessment abilities seem to allow treated children to overcome such overconfident tendencies. As a consequence of choosing more difficult lanes, non-treated low-SES children failed more frequently (see column 2 of Table 4) which (as we saw in Tables 2 and 3) leads to lower earnings on average.

3.4. Discussion

Next, we discuss the role of ability, risk preferences, as well as effort and motivation. The main concern with ability is that it obviously relates to rewards in the game and that it might also be affected by the intervention. To address this concern, Table A6 in the Appendix verifies that we do not detect any differences in marble ability between Treatment Low SES and Control Low SES. Moreover, Tables 3, A3, A4, and A5 indicate that our results are robust to including ability dummies in various specifications.

The concern with risk preferences might be less obvious, but risk preferences can affect lane choice because children may opt for a "safer" lane and thereby be willing to forego expected earnings. In addition, risk preferences might also be affected by the treatment. To address the role of risk preferences, similar to ability, Table A6 in the Appendix verifies that we do not detect any differences in risk preferences between Treatment Low SES and Control Low SES. Furthermore, Tables 3, A4, and A5 show robustness when we include controls for risk preferences. Table A7 estimates the treatment effect separately for different risk groups (no risky choices versus at least one risky choice). We obtain similar and statistically significant treatment effects for both subsamples.

Another potential concern one might have is that our measure of accuracy of self-assessment might to some extent capture willingness to exert effort to find out the ideal marble lane, rather than an actual ability to assess own skills. While the two notions are probably often related, we seek to address this concern with our independent measure of willingness to exert effort. Table A8 in the Appendix verifies that we do not detect any differences in motivation between Treatment Low SES and Control Low SES. Our findings also indicate that our main results are robust to including our effort proxy as control variable.[17]

Taken together, we find no evidence that ability, risk preferences, or willingness to exert effort drive our results or change the interpretation of our findings in any meaningful way.

We notice here that there might be other factors that affect lane choice that might be more difficult to rule out. One possibility might be that Control Low SES children simply have a stronger preference for selecting difficult lanes compared to High SES or Treatment Low SES children. Such a preference channel, by its very nature seems difficult to rule out empirically. While it is not immediately clear why such a preference might arise and why it might differ by socioeconomic status, we acknowledge that it is conceivable in principle. In our view, the strongest evidence against this channel is coming from the relation of our self-assessment task with self-assessments about school performance that we identify and summarize in Section 3.6.

Another possibility is that the choice of lane might have an effect on performance that goes beyond a mechanic effect. This could for instance arise due to "choking under pressure", where children select a difficult lane but then get very nervous when they start performing. Alternatively, picking a difficult lane could have a motivational effect where for instance the high difficulty level increases focus on the task and hence improves performance. Importantly, in our view these factors are natural features of forward-looking self-assessments. In other words, people that are good at self-assessment and know themselves well must be able to take such effects into account. In turn, people that do not have great self-assessment skills might fail to take such effects into account. We think that this reflects self-assessments in real life

[16] Using the full sample, including dummies for Control High SES and Treatment Low SES (i.e. using Control Low SES as the base) and testing for equality of the treatment and high-SES coefficients, yields p-values greater than 0.6 for all specifications.

[17] The results are also robust if we use the share of correctly solved tasks as an alternative measure of effort and motivation.

situation. A person who knows that she will choke under pressure in, say, a very competitive environment, should try not to select into such an environment. A person with poor self-assessment skills will fail to do so because she might fail to take her own choking under pressure into account.

It is also conceivable that different lanes are more or less informative about children's marble skills. If the desire to learn about own skills differs by socioeconomic status, then this could in principle explain some of our patterns. In our view, this interpretation is unlikely given the incentive structure as well as the framing of the task. At the point where children were selecting a lane, there were no incentives to further experiment to maximize learning about marble abilities. Also the framing of the task was such that it was clear to children that they should focus on deciding which lane will maximize their expected earnings. Furthermore, the association of our self-assessment task with self-assessments about school performance that we identify and summarize in Section 3.6 speaks against this interpretation.

3.5. Heterogeneous treatment effects

We proceed by delving into the underlying mechanisms of our treatment effect. Specifically, we seek to shed light on the environmental factors that appear to be missing in low-SES families that influence accurate self-assessment. We take our measure of social interaction patterns as a proxy for the opportunities to learn about oneself that the social environment provides.

Following the estimation approach as before, column 1 of Table 5 reveals that for Control Low SES children, there is a pronounced positive relationship between the mother-reported intensity of social interactions and the accuracy of self-assessment of the child. Crucially, if the intensity of social interactions is a critical element in the relationship between social environment and the accuracy of self-assessment, and if this is the input that the mentoring program delivers, then we should see that the effect of the mentoring program is more pronounced for families with fewer intense social interactions. Table 5 column 2 shows this to be the case for our sample. We regress earnings in the self-assessment game on our measure of the intensity of social interactions (measured at baseline), a treatment dummy, and an interaction term of treatment status and intensity of social interactions. The results indicate a significant negative interaction effect, which means that the mentoring program is less effective for children that already experience relatively more intense interaction in their families and is pronounced for children that experience fewer intense interactions in their families. The same pattern is found when we look at the effects of lane choice and the probability of failure, see Table A9. This suggests that the mentoring program provides resources that are scarce in low-SES family environments, namely intense social interactions that allow children to have new experiences and obtain feedback that sharpens their sense of their strengths and weaknesses.

3.6. Relation to self-assessment of school performance

How does our measure of self-assessment ability obtained from the marble lane paradigm relate to self-assessment in other domains? A possible concern one might have is that our measure is not associated with self-assessment abilities in other, perhaps more relevant domains of economic decision-making. To address this concern, we focus on self-assessment about school performance. In a 6-year follow-up interview we collected information about children's objective school performance and their self-rated performance in school.[18]

We begin by investigating to what extent our measure of accuracy in self-assessment is associated with self-assessment accuracy about school performance in our control groups (Low and High SES). Column 1 of Table 6 indicates that self-rated performance

Table 5
Coefficients are inverse probability weighted (IPW) Tobit estimates, standard errors in brackets. All regressions also include a constant, age, gender, location fixed effects (see sampling), interviewer FEs, marble ability dummies, and standardized willingness to take risk. IPWs account for potential selective attrition and are estimated from a linear probability model of a binary selection indicator (indicating whether the self-assessment measure is available for the post-treatment wave) regressed on baseline measures of self-assessment, social interaction, and ability, treatment and high SES dummies, and the interaction of baseline measures and the group dummies. ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

	Gained stars (# 0–7)	
	(1)	(2)
Intense interaction (baseline, std.)	0.417**	0.378**
	(0.161)	(0.154)
Treatment dummy		0.613**
		(0.277)
Treatment x intense interaction		−0.460**
		(0.227)
Sample:	Control Low SES	T & C Low SES
Mean Low SES Control	3.146	3.146
Observations	309	485

Table 6
Coefficients are OLS estimates, standard errors in brackets. All regressions also include a constant and location fixed effects (see sampling). Subjective and objective performance in school are elicited six years after the end of intervention, see also Falk et al. (2020). ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Sample: Low & High SES Control	Self-rated performance in school (std.)	
	(1)	(2)
Objective performance in school (grades, std.)	0.466***	0.475***
	(0.049)	(0.049)
Gained stars (standardized)	−0.041	−0.034
	(0.047)	(0.047)
Objective performance x gained stars		0.086*
		(0.045)
Observations	327	327

Table 7
Coefficients are OLS estimates, standard errors in brackets. All regressions also include a constant and location fixed effects (see sampling). Subjective and objective performance in school are elicited six years after the end of intervention, see also Falk et al. (2020). ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Sample:	Self-rated performance in school (std.)		
	Low SES Control (1)	Low SES Treatment (2)	High SES Control (3)
Objective performance (grades, std.)	0.406***	0.596***	0.606***
	(0.063)	(0.074)	(0.071)
Observations	235	133	92

[18] To measure self-rated performance in school, participants indicated on a 7-point Likert scale how much the following statements apply to them: "I am good in the subject Mathematics" and "I am good in the subject German". We use the average rating as our measure of self-rated performance at school. Our measure of objective performance in school is the average grade in the subjects Mathematics and German (coded such that higher values indicate better performance).

in school is correlated with objective performance indicated by grades. Importantly, column 2 reveals that this association is stronger for children who also showed more accurate self-assessment in the marble lane task. The interaction effect is sizable and significant ($p = 0.059$). A one standard deviation increase in self-assessment as measured in the marble game is related to an about 20% higher correlation between self-rated and objective performance in schools.

We proceed by analyzing the effect of the mentoring program on self-assessment accuracy about school performance. Table 7 reveals that self-assessment regarding school performance (i.e. the correlation between self-rated and objective performance) is substantially higher in Treatment Low SES compared to Control Low SES (compare columns 1 and 2, $p = 0.053$)[19] as well as in Control High SES versus Control Low SES (compare columns 1 and 3, $p = 0.068$). Taken together, the mentoring program positively affects self-assessment abilities about school performance and closes the gap between high and low SES children (compare columns 2 and 3, $p = 0.888$). Since the outcomes presented in Table 7 are measured 6 years after the intervention, these results also underscore the persistence of the effects of the mentoring program.

4. Concluding remarks

In this paper, we introduced a novel tool to measure the accuracy of self-assessment among children. We employed this tool to study the role of the social environment in shaping the accuracy of self-assessment. We showed that: (i) children from high-SES families are more accurate in their self-assessments compared to children from low-SES families; (ii) an exogenous enrichment of the social environment via a mentoring program has a causal positive effect on low-SES children's self-assessments; (iii) the mentoring program is most effective for children whose parents provide fewer social and interactive activities for their children, and (iv) our measure of self-assessment ability relates to self-assessment about school performance in meaningful ways and the mentoring program also positively influences the latter.

The skill to accurately assess one's strengths and weaknesses is arguably a key determinant of good decision-making in many contexts of economic relevance, e.g. educational or career choices. The literature on the development of metacognition posits that metacognition and related skills are malleable and shaped by feedback and experiences. Our results bolster this view and provide causal evidence for the importance of feedback and experiences for the development of an accurate sense of self.

Our results suggest that low-SES children are more likely to select lanes that are too difficult, consistent with an account of overconfident beliefs. Delving into the sources of differences in overconfidence might help to develop an understanding which precise types of feedback and more generally which aspects of the social environment are crucial in shaping self-assessment skills. The literature on overconfidence distinguishes between the demand and the supply side of overconfident beliefs (Benabou and Tirole, 2002). The demand side captures reason for why people are overconfident, whereas the supply studies how people reach overconfident beliefs. Both the demand and the supply side seem rather natural candidates for why belief differences by socioeconomic status might arise. In terms of the demand side, it seems for instance conceivable that low SES children have a higher need for overconfident beliefs as an ego booster. This would suggest that more external validation and support from their social environment would be effective in reaching more accurate self-assessments. In terms of the supply side, it could also be that all children start with a certain level of overconfidence, which is then mitigated over time by feedback from the social environment. This would highlight a direct role of targeted and context-specific feedback for the development of accurate self-assessments.

[19] Tests on the equality of coefficients are based on interaction effects in joint regressions.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Additional Tables

Tables A1,A2,A3,A4,A5,A6,A7,A8,A9.

Table A1
Baseline balance in the follow-up sample ($N = 485$). Notes: The values in columns 1 and 2 are means in control and treatment groups, standard errors are in parentheses. Measures are collected at baseline, see Section 2.4 for descriptions. Column 3 lists p-values of t-tests on the null hypotheses that the differences in means between treatment and control group are zero. The full follow-up sample (including high SES) is used to standardize variables.

Baseline measure	Mean Control Group	Mean Treatment Group	Difference p-value
Family characteristics:			
Low parental education (binary)	0.476	0.460	0.743
	(0.028)	(0.038)	
Low parental income (binary)	0.508	0.466	0.373
	(0.028)	(0.038)	
Single parent (binary)	0.463	0.449	0.768
	(0.028)	(0.038)	
Intense Interaction (std.)	−0.016	−0.040	0.799
	(0.056)	(0.075)	
Child characteristics:			
Female (binary)	0.476	0.455	0.654
	(0.028)	(0.038)	
Age (in months, at follow-up)	108.75	109.02	0.629
	(0.327)	(0.476)	
Marble ability (0–10)	5.184	4.801	0.066
	(0.128)	(0.160)	
Chosen lane (0–7)	4.810	4.649	0.285
	(0.091)	(0.118)	
Gained stars (0–7)	2.608	2.642	0.872
	(0.126)	(0.167)	
Willingness to take risk (std.)	0.101	0.023	0.425
	(0.058)	(0.077)	
Effort and motivation (binary)	0.726	0.679	0.273
	(0.025)	(0.036)	

Table A2
Coefficients are inverse probability weighted (IPW) Tobit estimates, standard errors in brackets. All regressions also include a constant, gender, location fixed effects (see sampling), interviewer FEs, and dummies indicating high SES and treatment group. Marble ability is the performance in the trial round (dummies for each ability level). Willingness to take risk is the number of risky choices (lottery over safe amount). ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Sample: All children	Gained stars (# 0–7)		
	(1)	(2)	(3)
Age (in years)	0.525**	0.493**	0.405*
	(0.231)	(0.231)	(0.227)
Marble ability	no	yes	yes
Controlling for risk pref.	no	no	yes
Observations	596	596	596

Table A3
Coefficients are inverse probability weighted (IPW) Tobit estimates, standard errors in brackets. All regressions also include a constant and location fixed effects as treatment probabilities differ by location (see sampling). IPWs account for potential selective attrition and are estimated from a linear probability model of a binary selection indicator (indicating whether the self-assessment measure is available for the post-treatment wave) regressed on baseline measures of self-assessment, social interaction, and ability, treatment and high SES dummies, and the interaction of baseline measures and the group dummies. Marble ability is the performance in the trial round (dummies for each ability level). ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Sample: Low SES Treatment & Control	Gained stars (# 0–7)				
	(1)	(2)	(3)	(4)	(5)
Base: Low SES Control					
Treatment dummy	0.543**	0.592**	0.503*	0.625**	0.639**
	(0.275)	(0.278)	(0.273)	(0.284)	(0.284)
Marble ability	no	yes	no	no	yes
Age & gender	no	no	yes	no	yes
Interviewer FEs	no	no	no	yes	yes
Observations	485	485	485	485	485

Table A4
Coefficients are average marginal effects after Poisson regressions, standard errors in brackets. All regressions also include a constant, age, gender, location fixed effects (see sampling), interviewer FEs, and marble ability dummies. Marble ability is the performance in the trial round. Willingness to take risk is the number of risky choices (lottery over safe amount). IPWs account for potential selective attrition and are estimated from a linear probability model of a binary selection indicator (indicating whether the self-assessment measure is available for the post-treatment wave) regressed on baseline measures of self-assessment, social interaction, and ability, treatment and high SES dummies, and the interaction of baseline measures and the group dummies. ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Poisson regressions	Gained stars (# 0–7)		
Sample: Low SES Treatment & Control	(1)	(2)	(3)
Base: Low SES Control			
Treatment dummy	0.440**	0.436**	0.413*
	(0.192)	(0.213)	(0.211)
Inverse probability weighting	no	yes	yes
Willingness to take risk	no	no	yes
Observations	485	485	485

Table A5
Coefficients are OLS estimates, standard errors in brackets. All regressions also include a constant, age, gender, location fixed effects (see sampling), interviewer FEs, and marble ability FEs. Marble ability is the performance in the trial round. Willingness to take risk is the number of risky choices (lottery over safe amount). IPWs account for potential selective attrition and are estimated from a linear probability model of a binary selection indicator (indicating whether the self-assessment measure is available for the post-treatment wave) regressed on baseline measures of self-assessment, social interaction, and ability, treatment and high SES dummies, and the interaction of baseline measures and the group dummies. ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

OLS regressions	Gained stars (# 0–7)		
Sample: Low SES Treatment & Control	(1)	(2)	(3)
Base: Low SES Control			
Treatment dummy	0.435**	0.429**	0.419*
	(0.219)	(0.217)	(0.214)
Inverse probability weighting	no	yes	yes
Willingness to take risk	no	no	yes
Observations	485	485	485

Table A6
No treatment effects on marble ability and risk preferences. Coefficients are inverse probability weighted (IPW) OLS estimates, standard errors in brackets. Marble ability is the number of scores in the trial round. Willingness to take risk is the number of risky choices (lottery over safe amount). All regressions also include a constant, location and interviewer fixed effects. IPWs account for potential selective attrition and are estimated from a linear probability model of a binary selection indicator (indicating whether the self-assessment measure is available for the post-treatment wave) regressed on baseline measures of self-assessment, social interaction, and ability, treatment and high SES dummies, and the interaction of baseline measures and the group dummies. ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

	Marble ability (std.)	Willingness to take risk (std.)
Sample: Low SES Treatment & Control	(1)	(2)
Treatment dummy	0.016	−0.024
	(0.099)	(0.102)
Observations	485	485

Table A7
Coefficients are inverse probability weighted (IPW) Tobit estimates, standard errors in brackets. Column 1 regards the sub-sample of children who did not make any risky choice in the coin toss experiment. Column 2 regards the sub-sample of children who did make at least one risky choice in the coin toss experiment. All regressions also include a constant, age, gender, location fixed effects (see sampling), interviewer FEs, and marble ability dummies. Marble ability is the performance in the trial round. ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Sample: Low SES Treatment & Control	Gained stars (# 0–7)	
	(1)	(2)
Base: Low SES Control		
Treatment dummy	0.713*	0.743*
	(0.369)	(0.390)
Regarded sub-sample:	Zero risky choice	At least one risky choice
Observations	156	329

Table A8
Coefficients are inverse probability weighted (IPW) OLS (column 1) and Tobit (column 2) estimates, standard errors in brackets. All regressions also include a constant, age, gender, location fixed effects (see sampling), interviewer FEs, marble ability dummies, and standardized willingness to take risk. Effort and motivation is a dummy variable being one if the participant worked for four minutes on a tedious task and zero if he or she stopped before. Four observations are missing due to missing information on the effort and motivation variable. IPWs account for potential selective attrition and are estimated from a linear probability model of a binary selection indicator (indicating whether the self-assessment measure is available for the post-treatment wave) regressed on baseline measures of self-assessment, social interaction, and ability, treatment and high SES dummies, and the interaction of baseline measures and the group dummies. ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

	Effort and motivation (binary)	Gained stars (# 0–7; std.)
Sample: Low SES Treatment & Control	(1)	(2)
Treatment dummy	−0.023	0.619**
	(0.040)	(0.282)
Effort and motivation dummy		0.017
		(0.328)
Observations	481	481

Table A9
Coefficients are inverse probability weighted (IPW) OLS estimates, standard errors in brackets. All regressions also include a constant, age, gender, location fixed effects (see sampling), interviewer FEs, marble ability dummies, and standardized willingness to take risk. IPWs account for potential selective attrition and are estimated from a linear probability model of a binary selection indicator (indicating whether the self-assessment measure is available for the post-treatment wave) regressed on baseline measures of self-assessment, social interaction, and ability, treatment and high SES dummies, and the interaction of baseline measures and the group dummies. In columns 1 and 2, two observations are missing due to experimenter misreporting. ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

	Chosen lane (# 1–7)		Failure (0/1)	
	(1)	(2)	(3)	(4)
Intense interaction	−0.067	−0.045	−0.057**	−0.054**
(baseline, std.)	(0.064)	(0.062)	(0.024)	(0.024)
Treatment dummy		−0.316***		−0.122***
		(0.115)		(0.043)
Treatment x intense interaction		0.191*		0.082**
		(0.113)		(0.034)
Sample:	Control Low SES	T & C Low SES	Control Low SES	T & C Low SES
Observations	307	483	309	485

Appendix B. Additional Figures

Figs. B1,B2,B3,B4,B5.

Fig. B1. The relation of self-assessment and income. Sample: 744 parents of children in the study, for details see Section 2.1. Income is the self-reported net household income in Euro. Self-assessment is measured using the survey item "In general, do you tend to underestimate your own abilities or do you tend to overestimate your own abilities?" Answers were given on a 11-point scale (0 = strongly underestimate, 10 = strongly overestimate). The size of the circles represents the relative share of the sample. The dashed line indicates the quadratic fit.

Fig. B2. The relation of self-assessment and health. Sample: 744 parents of children in the study, for details see Section 2.1. Health is the subjective health status measured using the following item "How would you describe your current health status?". Responses are given on a 5-point scale ranging from "very good" to "bad". Self-assessment is measured using the survey item "In general, do you tend to underestimate your own abilities or do you tend to overestimate your own abilities?". Answers were given on a 11-point scale (0 = strongly underestimate, 10 = strongly overestimate). The size of the circles represents the relative share of the sample. The dashed line indicates the quadratic fit.

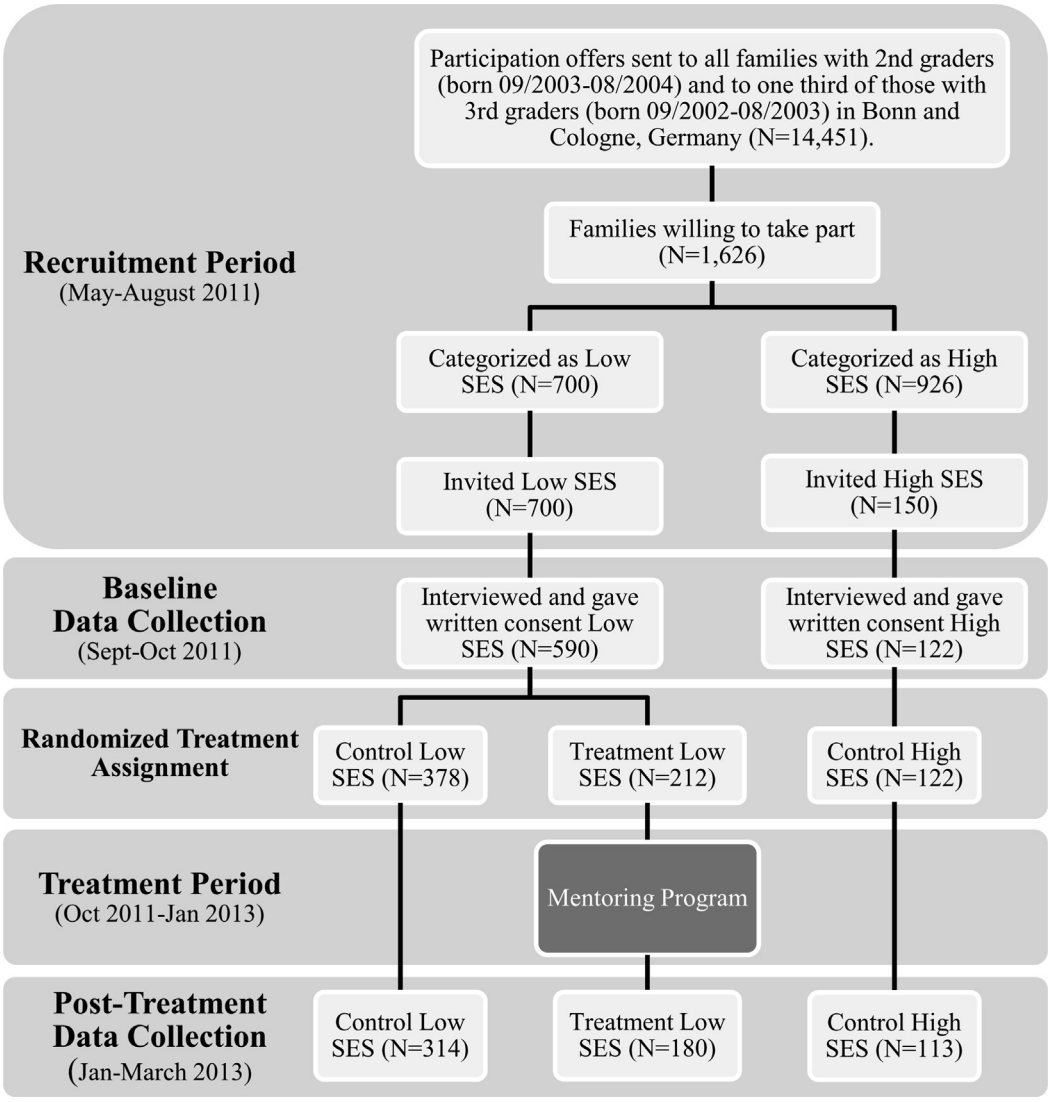

Fig. B3. Flow chart of sampling and procedural details.

Fig. B4. The marble lanes.

Appendix C. Translated version of instructions

C.1. Accuracy of self-assessment

1st round

Rules of the game

"Look, here I have a marble lane. At the end there is a squared hole. That is where you are supposed to get the marble in. You have to stay here at this end of the lane and are not allowed to touch the lane while rolling your marble. If the marble does not stay in the hole, we cannot count it as a success. You can now try ten times. Just see how often you can hit the hole out of these ten times."

Results

The child scored _ times (0–10).

2nd round

"That worked out fine. Please come over here. Here are seven more marble lanes with round holes which have different sizes.

You can now win stars with your marbles. However, you can only win stars if at least five of your ten marbles drop into the hole. So, only if you either score 5, 6, 7, 8, 9 or 10 times. If you score less than five times (which means 4, 3, 2 or 1 time or never), you won't get any stars.

All marble lanes have different levels of difficulty. That is why you can win different amounts of stars playing on them. On the easiest lane

with the biggest hole you can win one star; on the hardest lane with the smallest hole you can win seven stars. How many stars you can win on each lane is written on the lanes themselves. On this lane one star, here two, here three, here four, here five, here six and here seven."

→ Point to the lanes while stating the amounts of stars.

"But remember: You will only win stars at all if you score at least five out of ten times!

You can now pick one of the seven lanes on which you would like to play. Think carefully about your choice. Okay, now we try to recapitulate the rules together."

Testing how well the rules were understood

→ Pick the two-stars lane.

"Please tell me, if three of your marbles drop into this lane's hole, how many stars will you get?"

Correct answer: 0.

Answer to "Three-marbles question" correct □.

Answer to "Three-marbles question" false □.

"And if eight of your marbles drop into this lane's hole, how many stars will you get?"

Correct answer: 2.

Answer to "Eight-marbles question" correct □.

Answer to "Eight-marbles question" false □.

→ Pick the five-stars lane.

"Please tell me, if four of your marbles drop into this lane's hole, how many stars will you get?"

Correct answer: 0.

Answer to "Four-marbles question" correct □.

Answer to "Four-marbles question" false □.

"And if six of your marbles drop into this lane's hole, how many stars will you get?"

Correct answer: 5.

Answer to "Six-marbles question" correct □.

Answer to "Six-marbles question" false □.

→ If the child does not understand the rules of the game, thus if it doesn't answer the control questions correctly, please briefly repeat the rules and pose a new control question. If the answer is wrong again, repeat again. Repeat the rules at most three times. If the child does not grasp the rules at all, play the game anyhow to avoid disappointing the child, except for the case that the child is so frustrated that it does not want to play the game.

The child has understood the rules of the game at once □.

The child has understood the rules of the game after _ (1, 2 or 3) repetitions □.

The child has not understood the rules after three repetitions □.

"Well done, you have understood the game very well. So please decide now on which marble lane you would like to play."

Results

Time until the decision was made: _ seconds [Start measuring time at the end of the phrase].

The child has decided to play on lane _ (1–7).

The child scored _ times (0–10). [In any case let the child play all ten rounds].

→ If the child **scored at least 5 times:**

Hand out the stars; put them into a NEW bag labeled with the child's name and put the bag next to the table close to the child.

"That was great. You have scored [#score] times at the []-stars lane, therefore you get [] stars."

→ If the child **scored less than 5 times:**

"Unfortunately, less than five of your marbles have dropped into the hole, hence you won't get any stars. But never mind, later on you can win more stars."

Remarks:

C.2. Risk preferences

→ **Note:** The order of choice A and B is randomized.
→ Label a new bag with name.

Choice A: *"Here are two blue coins. This coin has three stars on each side.*"

→ Please show both sides, count the stars out loud and then let the child hold the coin.

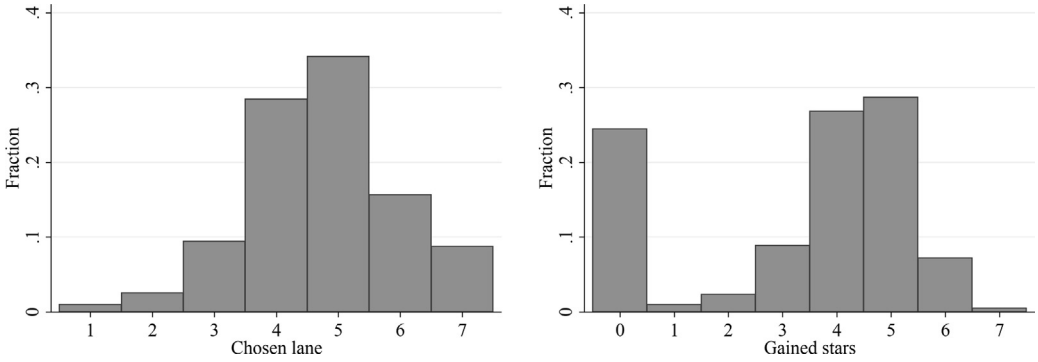

Fig. B5. Histograms for chosen lane difficulty and earnings over all three treatment conditions.

"The other coin has 7 stars on one side and no stars on the other side."

→ Please show both sides, count the stars out loud and then let the child hold the coin.

"You may now decide which coin I should toss. When I toss a coin, the coin always lands randomly with one side on the ground. The other side faces up and we will be able to see the stars on the upper side. You will then receive as many stars as we see on the top side.

So if you choose the coin with three stars on each side, you will get three stars in any case. If you choose the coin with 7 stars on one side and zero stars on the other side, chance will decide whether you get 7 stars or none at all."

Decision:

3–3 □.

7–0 □.

[If choice A is done before choice B: *"Before I toss the coin, you will have to make another decision."].*

Choice B: *"Here are two red coins. This coin has four stars on each side."*

→ Please show both sides, count the stars out loud and then let the child hold the coin.

"The other coin has 7 stars on one side and no stars on the other side."

→ Please show both sides, count the stars out loud and then let the child hold the coin.

"You get to decide again which coin I should toss. Again, you will then receive as many stars as we see on the side that happens to be facing up. So if you choose the coin with four stars on each side, you will get four stars in any case. If you choose the coin with 7 stars on one side and zero stars on the other side, chance will decide whether you get 7 stars or none at all."

Decision:

4–4 □.

7–0 □.

This is coin toss No. 1 □ No. 2 □.

[If choice B is done before choice A: *"Before I toss the coin, you will have to make another decision."].*

C.3. Effort and Motivation

→ Here, the child should voluntarily decide for how long to work on the tasks. The child itself may quit the task; if not, the interviewer ends the task after four minutes by saying: ''You have done a very good job, we will now proceed to our last game."

[Turn the protocol sheet 90 degrees such that it can be seen by both the child and the interviewer] *"Now let us do something completely different. As you can see there are boxes with letters and numbers on this sheet. The letters and numbers from the left box have been copied into the right box."*

→ Point to the top left line of the task sheet corresponding to the text.

"Sometimes a mistake has been made while copying, but not always. You are now to check whether the two boxes in a line are the same or whether they are different. If they are the same, tick the line. If there is a mistake and the two combinations are not the same, mark the line with a 'minus sign'."

"Let us have a look at the first two lines:"

→ Point to the first line. Please read out loud the number-letter-combination inside the left and right box: *"m-f-8–4-j-i-1"*

"The two boxes in the first line are exactly the same. In the second line a mistake was made, the two boxes are not the same."

→ Tick the first line, mark the second line with a 'minus sign'.

"It is important that you do not make any mistakes. Please keep checking if two boxes are the same for as long as you want. When you want to stop, just put the pencil down onto the picture of the pencil [point to the picture] *at the bottom of the sheet and* **raise your hand**."

A	mf84ji1	mf84ji1	
B	3ghd46a	3ghe46a	

1	4svt3p9q	4svt2p9q	
2	Lk2m4n	Lk2m4n	
3	P0lk8g4	P0lkrg4	
4	Mr5v36q	Mr5v26q	
5	lf0mä22	lf0mä22	

6	3m9kkb5	3m9kkp5	
7	3v8mx21	3x8mx21	
8	9fa39m9	9fa29m9	
9	2n4m6n9	2n4m6n9	
10	1n3f5h1b	1n3f3h1b	

11	1b14x9y	1b14x9y	
12	8m99xy2	8m99xx2	
13	913n6xx	913n6yx	
14	wqu663g	wqu263g	
15	663jskjg	663jskjg	

16	hdsf6d8j	hdsf6f8j	
17	dgfdjkd8	dgfdjkd8	
18	gsöjsz8hh	gsöjsz8hh	
19	gavbsuhjd	gavbgghjd	
20	hdsjds65s	hggjds65s	

21	636gafdd2	636gafhd2	
22	hdk9dkd7	hdk9dkd7	
23	gajndvy6d	gajndvy6d	
24	hd635hhs	hd635jks	
25	shsgfs5hd	shjgfs5hd	

26	6mb0fg2	6ab0fg2	
27	zt9k0lm	zt9q0lm	
28	2m0vb8	2m0vb8	
29	0lkj69b2	0lkj69b2	
30	9j4k5n2m	9j4k5m2m	

31	8t8jk02m	8t8jk02m	
32	4n6v21p	4n2v21p	
33	9üm4ö2	9üm4ö2	
34	8k2w81nx	8k2w81nx	
35	23mk239	23mk2b9	

36	8i922mz	8i922mz	
37	4nb9we1	4nb9qe1	
38	m122mya	m122mya	
39	q3c1abn2	q2c1bbn2	
40	9m3cx5	9m3cx5	

41	821m53x	821v53x	
42	99944wq	99944wq	
43	92ngjk3fr	92mgjk3fr	
44	hud77ezd	hud71ezd	
45	625gsje7	625fsje7	

46	g5s5d67d	g5s5d64d	
47	hdhjie6e6	hdhjie6e6	
48	hskfbsgf3	hskfpsgf3	
49	45j56j6h	45j56j6h	
50	hjsdhsh37	hjsdhvh37	

51	fg5hwj2g3	fg5hwj1g3	
52	hdh6kjsh	hdh6kjsh	
53	545g4g4f	545g4g4f	
54	h44h5jk43	h44h5jg43	
55	hdggsa8jf	hdggxa8jf	

References

Alan, Sule, Baydar, Nazli, Boneva, Teodora, Crossley, Thomas, Ertac, Seda, 2017. Transmission of risk preferences from mothers to daughters. J. Econ. Behav. Organ. 134 (1), 60–77.

Alan, Sule, Boneva, Teodora, Ertac, Seda, 2019. Ever failed, try again, succeed better: results from a randomized educational intervention on grit. Quart. J. Econ. 134 (3), 1121–1162.

Almås, Ingvild, Cappelen, Alexander W., Sørensen, Erik Ø., Tungodden, Bertil, 2010. Fairness and the development of inequality acceptance. Science 328 (5982), 1176–1178.

Barber, Brad, Odean, Terrance, 2000. Trading is hazardous to your wealth: the common stock investment performance of individual investors. J. Finance 55 (2), 773–806.

Benabou, Roland, Tirole, Jean, 2002. Self-confidence and personal motivation. Q. J. Econ. 117 (3), 261–292.

Berger, Eva M, Fehr, Ernst, Hermes, Henning, Schunk, Daniel, Winkel, Kirsten, 2020. "The Impact of Working Memory Training on Children's Cognitive and Noncognitive Skills". IZA Disussion Paper No. 13338.

Bradley, Robert H., Corwyn, Robert F., 2002. Socioeconomic status and child development. Annu. Rev. Psychol. 53, 371–399.

Cappelen, Alexander, List, John, Samek, Anya, Tungodden, Bertil, 2020. The effect of early-childhood education on social preferences. J. Polit. Econ. 128 (7), 2739–2758.

Charness, Gary, List, John A, Rustichini, Aldo, Samek, Anya, Van De Ven, Jeroen, 2019. Theory of mind among disadvantaged children: evidence from a field experiment. J. Econ. Behav. Organ. 166, 174–194.

Cunha, Flavio, Heckman, James J., 2007. The technology of skill formation. Am. Econ. Rev. 97 (2), 31–47.

DeQuidt, Jonathan, Haushofer, Johannes, Roth, Christopher, 2018. Measuring and bounding experimenter demand. Am. Econ. Rev. 108 (11), 3266–3302.

Duckworth, Angela L., Peterson, Christopher, Matthews, Michael D., Kelly, Dennis R., 2007. Grit: perseverance and passion for long-term goals. J. Pers. Soc. Psychol. 92 (6), 1087–1101.

Duncan, Greg J., Morris, Pamela A., Rodrigues, Chris, 2011. Does money really matter? Estimating impacts of family income on young children's achievement with data from random-assignment experiments. Dev. Psychol. 47 (5), 1263–1279.

Eil, David, Rao, Justin, 2011. The good news-bad news effect: asymmetric processing of objective information about yourself. Am. Econ. J., Microecon. 3, 114–138.

Falk, Armin, Kosse, Fabian, 2021. "The briq family panel: An overview." Mimeo.

Falk, Armin, Kosse, Fabian, Pinger, Pia, 2020. "Mentoring and Schooling Decisions: Causal Evidence." IZA Disussion Paper No. 13387.

Falk, Armin, Kosse, Fabian, Pinger, Pia, Schildberg-Hörisch, Hannah, Deckers, Thomas, 2021. Socio-economic status and inequalities in children's IQ and economic preferences. J. Polit. Econ. 129 (9), 2504–2545.

Fehr, Ernst, Rützler, Daniela, Sutter, Matthias, 2013. The development of egalitarianism, altruism, spite and parochialism in childhood and adolescence. Eur. Econ. Rev. 64, 369–383.

Flavell, John, 1979. Metacognition and cognitive monitoring. Am. Psychol. 34 (10), 906–911.

Goux, Dominique, Gurgand, Marc, Maurin, Eric, 2017. Adjusting your dreams? High school plans and dropout behaviour. Econ. J. 127 (602), 1025–1046.

Hanushek, Eric A., Woessmann, Ludger, 2008. The role of cognitive skills in economic development. J. Econ. Literat. 46 (3), 607–668.

Heckman, James J., Stixrud, Jora, Urzua, Sergio, 2006. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. J. Labor Econ. 24 (3), 411–482.

Heckman, James J., Vytlacil, Edward, 2001. Identifying the role of cognitive ability in explaining the level of and change in the return to schooling. Rev. Econ. Stat. 83 (1), 1–12.

Huffman, David, Raymond, Colin, Shvets, Julia, 2022. Persistent Overconfidence and Biased Memory: Evidence from Managers. Am. Econ. Rev. 112 (10), 3141–3175.

Humphries, John Eric, Kosse, Fabian, 2017. On the interpretation of non-cognitive skills–What is being measured and why it matters. J. Econ. Behav. Organ. 136, 174–185.

Johnson, Dominic, 2004. Overconfidence and War: The Havoc and Glory of Positive Illusions. Harvard University Press, Cambridge, MA.

Kahneman, Daniel, 2011. Thinking, Fast and Slow. Straus and Giroux.

Kosse, Fabian, Deckers, Thomas, Pinger, Pia, Schildberg-Hörisch, Hannah, Falk, Armin, 2020. The formation of prosociality: causal evidence on the role of social environment. J. Polit. Econ. 128 (2), 434–467.

Malmendier, Ulrike, Tate, Geoffrey, 2005. CEO overconfidence and corporate investment. J. Finance 60 (6), 2661–2700.

Malmendier, Ulrike, Tate, Geoffrey, 2008. Who Makes Acquisitions? CEO overconfidence and the market's reaction. J. Financ. Econ. 89, 20–43.

Möbius, Markus M., Niederle, Muriel, Niehaus, Paul, Rosenblat, Tanya S., 2022. Managing self-confidence: theory and experimental evidence. Manage. Sci.

Moore, Don, Healy, Paul, 2008. The trouble with overconfidence. Psychol. Rev. 115 (2), 502–517.

Perry, John, Lundie, David, Golder, Gill, 2019. Metacognition in schools: what does the literature suggest about the effectiveness of teaching metacognition in schools? Educ. Rev. 71 (4), 483–500.

Samek, Anya, Cowell, Jason M., Cappelen, Alexander W., Cheng, Yawei, Contreras-Ibáñez, Carlos, Gomez-Sicard, Natalia, Gonzalez-Gadea, Maria L., Huepe, David, Ibáñez, Agustin, Lee, Kang, et al., 2020. The development of social comparisons and sharing behavior across 12 countries. J. Exp. Child Psychol. 192, 104778.

Schwardmann, Peter, van der Weele, Joel, 2019. Deception and self-deception. Nat. Human Behav. 3, 1055–1061.

Sorrenti, Giuseppe, Zölitz, Ulf, Ribeaud, Denis, Eisner, Manuel, 2020. "The Causal Impact of Socio-Emotional Skills Training on Educational Success." IZA Disussion Paper No. 13087.

Twenge, Jean, Campell, Keith, 2001. Age and birth cohort differences in self-esteem: a cross-temporal meta-analysis. Personal. Soc. Psychol. Rev. 5 (4), 321–344.

Veenman, Marcel, Spaans, Marleen, 2005. Relation between intellectual and metacognitive skills: age and task differences. Learn. Individual Diff. 15, 159–176.

Wigfield, Allan, Eccles, Jacquelynne, Iver, Douglas Mac, Reuman, David A., Midgley, Carol, 1991. Transitions during early adolescence: Changes in children's domain-specific self-perceptions and general self-esteem across the transition to junior high school. Dev. Psychol. 27 (4), 552–565.

Zimmermann, Florian, 2020. Dynamics of motivated beliefs. Am. Econ. Rev. 110, 337–361.