Malleability of Preferences for Honesty

Johannes Abeler (University of Oxford, IZA and CESifo)
Armin Falk (briq and University of Bonn)
Fabian Kosse (LMU Munich and briq)

Discussion Paper No. 296

November 15, 2021

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Georg Weizsäcker, Ph.D., Humboldt University Berlin, 10117 Berlin, Germany
info@rationality-and-competition.de

Malleability of preferences for honesty

Johannes Abeler, Armin Falk and Fabian Kosse*

9th August 2021

Abstract

Reporting private information is a key part of economic decision making. A recent literature has found that many people have a preference for honest reporting, contrary to usual economic assumptions. In this paper, we investigate whether preferences for honesty are malleable and what determines them. We experimentally measure preferences for honesty in a sample of children. As our main result, we provide causal evidence on the effect of the social environment by randomly enrolling children in a year-long mentoring programme. We find that, about four years after the end of the programme, mentored children are significantly more honest.

Keywords: honesty, lying, truth-telling, formation of preferences, experiments with children

JEL Codes: C90, D90, D64, D82, H26, J13

*Abeler: University of Oxford, IZA and CESifo (email: johannes.abeler@economics.ox.ac.uk); Falk: briq and University of Bonn (email: armin.falk@uni-bonn.de); Kosse: LMU Munich and briq (email: fabian.kosse@econ.lmu.de). Financial support through Eleven gGmbH, the Jacobs Foundation, 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) and the ESRC (grant ES/K001558/1) is gratefully acknowledged. We thank Kai Barron, Simon Columbus, Uri Gneezy, Agne Kajackaite, Collin Raymond, Chris Roth, Marta Serra-Garcia, Joel Sobel, Erik Snowberg, Matthias Sutter, Bertil Tungodden and Marie Claire Villeval for helpful discussions. Many valuable comments were also received from numerous seminar participants. Ethical approval for the experiment was obtained from the Ethical Review Board of the Department of Economics at the University of Bonn.

Are preferences for honesty malleable? Are they innate to each individual or do they change in response to the social environment? It is well known that there is strong heterogeneity among individuals – while some lie maximally, most have some preference for honesty and lie only a little or not at all (e.g., Gneezy et al. (2013)). Where does this heterogeneity come from? Is it possible to directly affect an individual's preference for honesty?

It is important to answer these questions since honesty plays a crucial role in economics and society. Many economic interactions feature asymmetric information, like a used-car seller describing a car's quality (e.g., Akerlof (1970)) or an expert giving advice (e.g., Crawford and Sobel (1982)). Economists usually assume that the informed party in such situations does not have any intrinsic preference to tell the truth. They just report whatever maximizes their payoff. Over the last decade, this assumption has been challenged by a new, empirical literature studying what people actually do when they have private information (e.g., Gneezy (2005), Mazar et al. (2008), Fischbacher and Föllmi-Heusi (2013); see Abeler et al. (2019) and Gerlach et al. (2019) for reviews of the literature). This literature documents widespread and heterogeneous preferences for honesty, even when offered large monetary incentives to lie (e.g., Kajackaite and Gneezy (2017)).[1] More pressingly, the rise in politicians lying to the public and the increased spreading of false information via social media, and sometimes even traditional media, have brought concerns about honesty and truth-telling to the forefront of the political and academic discussion (e.g., Serra-Garcia and Gneezy (2020), Lazer et al. (2018), Evanega et al. (2020)), with commentators branding persistent lying as a "threat to democracy" (e.g., Brettschneider (2020), Edsall (2021), Boot (2021)).

In this paper, we are interested in how the social environment, including both family and non-family interactions, affects the willingness to tell the truth. Given the importance of honesty for economics and society, it is crucial to understand why some children grow up to become honest adults and others do not. Moreover, understanding which circumstances affect honesty could allow policy makers to implement interventions that increase the level of honesty in societies.[2]

[1] In line with the evidence, more and more theoretical papers build on the assumption of some preference for truth-telling (e.g., Kartik et al. (2007), Matsushima (2008), Ellingsen and Östling (2010), Kartik et al. (2014), Khalmetski and Sliwka (2019), Gneezy et al. (2018), Sobel (2020)).

[2] While it is clear that honesty has strong positive externalities by reducing transaction and audit costs and is thus likely to be desirable on a societal level, it is less clear whether honesty is privately beneficial as this depends on the prevailing institutions and the equilibrium. Several studies, however, relate honesty to positive

To make progress on these questions, we focus on children of primary-school age. This is an important period for the development of honesty as the children's theory of mind is already developed enough to understand the mechanics of dishonest communication (e.g., Talwar and Crossman (2011)) and preferences are potentially less set compared to when they are older (e.g., Cunha and Heckman (2007), Kautz et al. (2014)). We combine experimental and survey data on children, their parents, and the non-family social environment. We first explore the role of parents by correlating parental characteristics with children's honesty. We then study the effects of a random allocation of children to a mentoring programme, thus randomizing critical features of the social environment. We follow participants for several years after the end of the intervention.

We use a modified version of the Fischbacher and Föllmi-Heusi (2013) experimental paradigm to measure preferences for honesty: before rolling a die in private, participants predict in their head what they will roll, they then observe the die roll, and report whether they predicted correctly. If they report having predicted correctly, participants receive a monetary reward. In this setup, the truthfulness of any individual report cannot be determined. The average share of "predicted correctly" reports is, however, a measure of honesty of a group of subjects. This paradigm has become the leading way to measure honesty preferences because it is simple and abstracts from strategic interaction. Moreover, several studies have shown that behavior in this experiment correlates well with honest behaviour outside the lab (Cohn and Maréchal (2018) study children of similar age as we do; also see, e.g., Dai et al. (2018), Potters and Stoop (2016), Hanna and Wang (2017)).

When we correlate this measure of a child's honesty with parental characteristics, we find that children from high socio-economic status (SES) households are more honest. Moreover, we find children are more honest when they experience a warm parenting style and high levels of general trust in their home environment. This suggests that these environmental inputs might be essential for the development for honesty preferences.[3]

To study the causal effect of the social environment on preferences for honesty, we ran-

private outcomes, e.g., in school (Cohn and Maréchal 2018). The desirability of an intervention to strengthen preferences for honesty would need to be decided in a careful welfare analysis, which is beyond the scope of this paper.

[3]We also confirm the previous findings that girls and older children are more honest. Comparing our age results to a meta-study of 35 previous studies, we find that our data fit in smoothly with the overall increase in honesty as participants get older.

domly allocate a sample of low SES children in Germany to a year-long mentoring programme. The volunteer mentors spend an afternoon per week with the children and engage in interactive social activities such as cooking, playing football, or doing handicraft activities. The mentoring programme aims to widen a child's horizon through social interactions with a new attachment person. The mentoring programme thus enriches the social environment of the children by providing inputs and experiences that are potentially scarce in low SES families and, at the same time, essential for the development of honesty preferences.

The main result of the paper is that children who were allocated to the mentoring programme become more honest. If we assume that participants do not lie downwards, 58% of participants in the control group lie while only 44% of participants in the treatment group do so. This is a large effect: the treatment effect is of similar magnitude as the difference between male and female participants, for example. Given that we measure honesty about four years after the mentoring programme, this is evidence of a long-term and persistent change in behaviour. We also show that the treatment effect of the mentoring intervention on honesty is distinct from its treatment effect on prosociality, discussed by Kosse et al. (2020), and that the treatment effect is not due to differential attrition.

We then provide evidence on the hypothesis that the mentoring programme benefits children by providing resources that are scarce in the family environment. We build on our result that parents who use a less warm parenting style or are less trusting have more dishonest children. We find that for children from these backgrounds, the mentoring programme increases the likelihood of being honest particularly strongly. This suggests that mentors can serve as substitutes for parents, in the sense that mentoring is particularly effective when parental teaching, inputs, or role models are limited.

A possible mechanism for the treatment effect is that mentors serve as an honest role model or that they teach the mentees directly. The importance of warm parenting style and maternal trust could be based on the fact that young children often lie to avoid punishment (e.g., Stouthamer-Loeber 1986). A warm parenting style, in which punishment is used only rarely, might reduce the subjective need for the child to lie and thus lead to more honesty. This would be in line with the study by Talwar and K. Lee (2011) who show that children who are exposed to a harsh disciplinary style in school lie more.

Our paper contributes to several literatures. First, we add to the understanding of how an

individual's preferences and attitudes develop by establishing the causal effect of the mentoring programme on honesty. Generally speaking, preferences could be determined by genes, the family environment, or the social environment, or some combination of these. We have little reliable data on any of these channels, partly because establishing the causal determinants of any preference is notoriously difficult. Taking the example of the social environment, only an intense experience could result in a change of preferences and such experiences are usually not randomly allocated (see Callen et al. (2014) for an example on risk preferences). In addition, one needs to demonstrate that behaviour has persistently changed to avoid classifying a short-run effect as a change in preferences. We rely on a year-long mentoring programme that should ex-ante be strong enough to change preferences. We also measure preferences for honesty four years after the intervention and can thus detect long-term effects. A growing literature explores the development of preferences and skills during childhood (e.g., Bettinger and Slonim (2007), Fehr et al. (2013), Alan et al. (2019), Kosse et al. (2020), Alan et al. (2020), Cappelen et al. (2020); for a review see Sutter et al. (2019)). Numerous papers point to the importance of parental investments and parental style for children's skill and preference development (e.g., Cunha and Heckman (2007); Doepke et al. (2019)). We confirm these findings for the case of preferences for honesty.

Second, on the level of communities and nations, our paper is related to the literature on the formation of social capital (e.g., Alesina and La Ferrara (2002), Becker et al. (2020)) and to the literature on the evolution of norms and preferences in society, in partiular, Elias (1969). Lowes et al. (2017) and Heldring (2021) demonstrate that historical institutions can have long-lasting effects on the honesty of the local population. While these papers investigate the long-term emergence of honest behavior over many generations, we focus on the within-person development of honesty. Perhaps most closely related to our paper is the study by Gächter and Schulz (2016) who find that students who grew up in more corrupt countries are more dishonest. One of our contributions is to document a correlation between maternal trust and a child's honesty in addition to the causal effect of mentors (who are more trusting) on a child's honesty, thus pointing to a potential mechanism for the country- and community-level correlations documented in these literatures.

Third, we add to the literature on honesty experiments with children. The effect of parents and the social environment on lying has been studied only very little (but see the correlational

studies by Talwar and K. Lee (2011) and Stouthamer-Loeber and Loeber (1986); for a summary of the psychology literature on the development of honesty see, e.g., Stouthamer-Loeber (1986)). Several economics studies conduct lying experiments with children (e.g., Bucciol and Piovesan (2011), Glätzle-Rützler and Lergetporer (2015), Maggian and Villeval (2016), Tobol and Yaniv (2019)) and find, e.g., that many young children already have a preference for honesty.

The paper is structured as follows. Section 1 explains the honesty experiment and the mentoring intervention. Section 2 describes the sample and the data. Section 3 presents the results and Section 4 concludes.

1 Design

1.1 Measuring preferences for honesty

We use a modified version of the experimental paradigm suggested by Fischbacher and Föllmi-Heusi (2013) ("FFH"). In this paradigm, subjects privately observe the outcome of a random variable (e.g., a die roll), report the outcome and receive a monetary payoff proportional to their report. The FFH paradigm is the leading experimental method to measure preferences for honesty since it is easy to understand for subjects, which is particularly important for experiments with children, and since it abstracts from strategic interaction. Crucially, reports in this experiment have been shown to correlate strongly with non-laboratory cheating behaviour (Potters and Stoop 2016, Gächter and Schulz 2016, Dai et al. 2018, Hanna and Wang 2017, Cohn and Maréchal 2018, Cohn et al. 2015, Kröll and Rustagi 2017). This has lead to an explosive growth in the number of studies using this paradigm: Abeler et al. (2019) identify 90 recent studies based on this design.

We modify the design in line with the "mind game" approach by Jiang (2013) and Greene and Paxton (2009). In our study, before subjects roll a six-sided die in private, they have to predict the number they will roll without telling anybody about their prediction. Only after the roll do they have to report whether they predicted correctly. A participant's report is thus about their own prediction, which is unobserved by the experimenter. In this setup, participants have a $\frac{1}{6}$ chance of predicting correctly. This setup implies the same incentives and probabilities as a more standard "win if you report a 6" die-rolling experiment. We

chose this design since some participants might not believe that the random draw is truly private, even though it is, thus not responding to the actual incentives in place. In our setting, the interviewer withdrew to the other side of the room to do some paper work during the experiment, and the participant was guided only by the computer. Asking participants to report the correctness of their prediction about the (already unobserved) die roll adds a second layer of unobservability. If participants report to have predicted correctly, they are paid 2.50 euros. If not, they receive no monetary reward for this part of the study.[4] Therefore, this experiment yields exactly one outcome: reporting to have predicted correctly or not. In Appendix B, we present a theoretical framework that clarifies how reports in this experiment are linked to preferences for honesty. The full instructions for the experiment can be found in Appendix C. For a discussion on the interpretion of the reporting behavior see section 3.

1.2 Design of the mentoring programme

The mentoring intervention uses a well-established mentoring programme for primary-school aged children in Germany ("Balu und Du", German for "Baloo and You"; for a detailed description, see Müller-Kohlenberg and Drexler (2013)). The programme has been run since 2002 and more than 13,000 children have participated so far. During the programme, participants meet one-to-one with a mentor for about 4 hours per week. The mentors are volunteers and almost all mentors are university students (aged 18–30). The mentoring programme lasts up to one year. For those mentor-mentee pairs that met at least once, the average duration is 9.3 months and the average number of meetings is 22.8 (see Figure A.1 in Appendix A for distributions).[5] Participants thus spend a considerable amount of time with their mentor. During the meetings, the mentor and the mentee engage in interactive social activities such as cooking, visiting a zoo or park, or doing arts and crafts activities. The choice of activities is driven by the individual needs, abilities, and interests of child and mentor. At the start of the programme, participants were on average 7.8 years old (std. dev. = 0.48).

The programme is based on the concept of "informal learning", i.e., it integrates learning processes into everyday activities and does not focus on academic achievements. The idea is

[4] The stake size is in line with many FFH experiments. If anything, the stakes are high, in particular compared to participants' daily "income". For comparison, the children in the sample receive on average 4.57 euros pocket money per week.

[5] If we include the mentor-mentee pairs who never met, the average number of meetings is 16.9.

to widen a child's horizon through social interactions with a new attachment person. The programme aims to strengthen the basic skills and non-academic abilities of participants that increase the likelihood of success in life and school. By enriching the social environment of participants, the mentors allow them to gain new experiences and to acquire these skills and abilities. In doing so, mentors both serve as role models and as motherly or fatherly friends who teach the mentee directly. Building a caring relationship between mentor and mentee is central to the mentoring programme.

Mentors receive professional support. They are overseen by paid coordinators, they fill in weekly online diaries on which they get feedback from the coordinators, and they meet coordinators and other mentors in bi-weekly meetings in which they receive suggestions for activities and discuss potential problems.

2 Data

Within the framework of the briq family panel (for details, see Falk and Kosse (2020)), we recruited participants and their parents from the two cities Cologne and Bonn in Germany. In 2011, we invited all families living in those cities with children born between September 2003 and August 2004 to participate in a mentoring programme, as well as one third of families with children born between September 2002 and August 2003 ($N = 14{,}451$). We informed parents that, due to capacity constraints, participation in the programme was not guaranteed. 1,626 families indicated a willingness to participate and answered a short questionnaire including questions on income, education and whether both parents lived in the same household. We focused on those children whose parents met at least one of the following three criteria: (i) Equivalence income of the household is lower than 1,065 euros, corresponding to the 30^{th} percentile of the German income distribution. (ii) Neither parent has a school-leaving degree qualifying for university studies. (iii) Parents do not live in the same household. We invited these children ($N = 700$) and their parents for a baseline interview conducted in September to October 2011. 590 children and their parents participated in the baseline interview and gave their written consent to allow the transmission of their address to the organization running the "Balu und Du" mentoring programme. This is our main sample. Out of this sample, 212 children were randomly selected to be treated ("treatment group"), the remaining 378 children

form the control group.[6] The actual mentoring intervention took place between October 2011 and January 2013.

We also invited some of the children whose parents did not meet either of the three criteria listed above ($N = 150$ invited, $N = 122$ participated in the baseline interview and gave written consent). None of these children participated in the mentoring programme. We will include this "high SES" comparison group when we correlate parental characteristics with children's reports as it increases the variance in parental characteristics.

Due to an unforeseen shortage of mentors during the intervention period, 18% of participants in the treatment group could not start the mentoring programme. Another 8% of matches were initiated but did not start due to refusals or availability problems (e.g., pregnancy of the mentor or moving). Thus, 74% of participants in the treatment group were actually treated. We will focus on intention-to-treat estimates.

We have information from a baseline survey of mothers[7] and children before the start of the intervention and from yearly follow-up surveys after the intervention. We measured preferences for honesty between September 2016 and February 2017, i.e., about four years after the end of the mentoring programme. Interviews were conducted by the surveying company that also conducts the GSOEP (Wagner et al. 2007) at the homes of participants.[8] The mother received a participation fee of 45 euros. At the time of the interviews, the participating children were on average 12.5 years old.[9]

142 children from the treatment group participated in the honesty experiment, 252 from the control group and 96 from the high SES comparison group, i.e., the re-interviewing rate was about 70%. These rates do not differ significantly across treatment and control group and are not systematically related to baseline honesty (see Table A.1 in Appendix A). Moreover, Table A.2 in Appendix A indicates that the follow-up sample is balanced across treatment

[6] Randomization was stratified by city (Cologne or Bonn), income (above or below the 30^{th} income percentile), education (at least one parent eligible for university studies or not), and parental status (single parent or not), 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 ITT group. Therefore, assignment into treatment was random conditional on city of residence. However, conditioning on city of residence does not affect our results (see Table 2).

[7] More than 95% of participating parents were the biological mother and we thus call the participating parent "mother" regardless of their gender.

[8] The interviewers were full-time employed data collectors with a mean age around 50 and without university education. Mentors and interviewers were thus very different. Moreover, the intervention was not mentioned at any point during the data collection.

[9] We only have one measure of honesty and can thus not study how the treatment effect evolves over time.

and control group regarding all baseline characteristics. Further details and robustness checks are presented in Section 3.3.

In our analysis, we mainly focus on the estimation of the treatment effect since we can establish clear causality. In addition, we explore the role of individual characteristics and the social environment on a descriptive level and thus collected a range of variables that help to measure these aspects. We elicited socio-demograpics, preferences, and beliefs of mothers during the baseline survey before the start of the potential treatment. Parental style was collected after the treatment in 2013.[10] Information from mentors was collected during the treatment period. Mothers' and mentors' preferences and beliefs are measured using validated survey items (Falk et al. 2016).[11] To estimate "warm parenting style", mothers indicated their agreement with eight statements on a 5-point Likert scale from "never" to "always". As in Falk et al. (forthcoming), we use factor analysis to extract one latent parenting style from these items. See Appendix C.3 for details.

3 Results

The dependent variable in all our analyses is a dummy for whether the participant reported to have predicted the number correctly. If everybody told the truth, then $\frac{1}{6}$ of participants, i.e., about 16.7%, should report this. In sharp contrast, overall 60.6% of participants report to have predicted correctly, i.e., a large fraction of participants, though not all, must have lied: they predicted wrongly but report to have predicted correctly and are thus paid the reward. If we assume that no participant lies downwards (as implied by the utility function described in Section 1.1), then 52.7% ($= \frac{60.6-16.7}{100-16.7}$) of actually wrong predictions are falsely reported as correct. In the regressions, any variable that is correlated with *more* honesty will thus have a *negative* coefficient.

To shed light on the malleability of preferences for honesty, we first explore, on a descriptive level, the role of the family environment for the formation of preferences for honesty, before

[10] The psychological literature shows that parenting styles are stable within developmental periods (e.g., Holden and Miller 1999, Forehand and Jones 2002). We also find no treatment effects on parental style, see Table A.2.

[11] While we have data on a range of preferences, we do not have data on mothers' or mentors' preferences for honesty. There is no validated, direct survey measure of honesty and for logistical reasons we were not able to implement incentivized experiments with parents or mentors.

studying the causal effect of the non-family environment by analyzing the impact of the mentoring intervention.

3.1 Parental background and preferences: Descriptive evidence

For the analysis in this section, we restrict the sample to the experimental control group and the high SES comparison group, i.e., we abstract from any effect of the intervention.

Result 1 *Girls, older children, and children from richer households are more honest. Moreover, we find higher levels of honesty for children who experience a warmer parenting style and higher levels of general trust in their family environment.*

We first show that female participants are significantly more honest than boys and that honesty increases with age (column 1 of Table 1, which depicts the results of Probit regressions in form of average marginal effects). This confirms results from the previous literature (e.g., Dreber and Johannesson 2008, Glätzle-Rützler and Lergetporer 2015). This is reassuring, as it shows that our participants, while younger, have similar patterns of behaviour as other subject pools. Figure A.2 in Appendix A demonstrates the effect of age more clearly. The graph combines the data from this paper with data from 35 FFH experiments collected by Abeler et al. (2019) that also contain data on age (N = 16,705). The graph shows that the average level of honesty in our data is very much in line with the previous literature.

We then explore the relation of reporting behaviour and parental socio-economic characteristics. Column 2 adds the three SES categories used in the sampling scheme: income, education, single-parent status (see Section 2). A Wald-test indicates joint significance of the SES variables ($p = 0.036$). More specifically, poorer households have children who are significantly more likely to report to have predicted correctly. This result seems not to be directly driven by available ressources as the effect is robust against controlling for weekly pocket money: the marginal effect of the poor household dummy is 0.135 ($p = 0.012$) conditional on pocket money. The effect of pocket money is positive but not statistically significantly different from zero ($p = 0.195$). [12] There is no significant partial effect of parental education

[12]Moreover, Abeler et al. (2019) show that behavior in this kind of experiment is not affected by the amount of money at stake.

or single parenthood (this result remains if each SES variable is entered separately into the regression).

In the third step of our descriptive analysis, we explore the relation of reporting behaviour and aspects of the social environment. To prepare our analysis of the mentoring intervention, we focus on aspects of the social environment which are changed by the intervention. For example, the *style of interaction* with the child substantially differs between parents and mentors. It is a key idea of the mentoring program to use praise, to avoid punishment and to guide the mentee as "big benevolent friend – reliable, sure, clarifying, leading, secure" (Müller-Kohlenberg and Drexler 2013). In other words, the mentors are instructed to interact with the mentee using what our questionnaire would classify as a warm parenting style (see Appendix C.3 for more details). Therefore, we explore the relation between children's reporting behaviour and the parenting style in their home environment. The results in column 3 indicate that a warmer parenting style of parents is significantly associated with fewer reports of having predicted correctly.

Moreover, as mentors are self-selected volunteers, it is plausible that their *preference and belief structure* differs from the one of parents, especially with regard to social preferences and beliefs. Indeed, mentors are significantly more trusting (31.4% of a standard deviation) and altruistic (20.6% of a standard deviation), while mothers and mentors do not differ regarding time and risk preferences (see Appendix Table A.3 for details). Therefore, we explore the relation between children's reporting behavior and mothers' trust and altruism.[13] The results in column 4 indicate that higher levels of maternal trust are significantly associated with fewer reports of having predicted correctly. The results in column 5 do not indicate a relation between reporting behaviour and altruism. A possible explanation for the relation of children's truth-telling and maternal trust is that a trusting social environment enables children to experience that telling the truth is beneficial in the long run (compare Gächter and Schulz (2016)).

These results set the stage for our main analysis as the mentoring programme changes features of the social environment that are correlated with children's preferences for honesty.

[13]Time and risk preferences of mothers' are not related to their children's reporting behavior ($p > 0.500$).

	Reported to have predicted correctly					
	(1)	(2)	(3)	(4)	(5)	(6)
Child's sex (1 = female)	-0.163*** (0.048)	-0.168*** (0.048)	-0.159*** (0.048)	-0.167*** (0.048)	-0.164*** (0.048)	-0.171*** (0.048)
Child's age (in years) (in years)	-0.122*** (0.043)	-0.111*** (0.043)	-0.118*** (0.043)	-0.118*** (0.043)	-0.123*** (0.043)	-0.110*** (0.042)
Low parental income (dummy)		0.155*** (0.053)				0.138*** (0.053)
Low parental edu. (dummy)		-0.050 (0.053)				-0.066 (0.054)
Single parent (dummy)		-0.025 (0.053)				-0.024 (0.053)
Warm parenting style (standardized)			-0.049** (0.024)			-0.040 (0.024)
Mother's trust (standardized)				-0.047* (0.026)		-0.043 (0.027)
Mother's altruism (standardized)					0.016 (0.027)	0.030 (0.027)
Sample restriction:	Low and High SES control groups					
Observations	348	348	348	348	348	348

Table 1: Correlates of reporting behaviour. Notes: Coefficients are average marginal effects of Probit regressions. Standard errors are in parentheses. The dependent variable is a dummy of whether the participant reported to have predicted the die roll correctly. For further details on the independent variables see Section 2 and Appendix C.3. ***, **, * indicate significance at the 1, 5 and 10 percent level, respectively.

3.2 Treatment effect

The mentoring programme is randomly allocated, which allows for a causal interpretation of the treatment effect. Any effect we find would be long-term: reporting experiments were conducted about four years after the intervention.

Result 2 *The mentoring treatment significantly increases honesty.*

64.7% of participants in the control group report to have predicted correctly but only 53.5% of participants in the treatment group do so (difference: 11.2 percentage points, $p = 0.029$). Assuming no downward lying, this means that 57.6% of control participants lie, and 44.2% of treated participants. Table 2 shows the treatment effect on the probability of reporting to

have predicted correctly in form of average marginal effects of Probit regressions. In column 1 we show the unconditional effect, in column 2 we add controls for gender and age, and in column 3 we also control for interviewer fixed effects and the randomization strata. The results indicate that the treatment effect is robust across these specifications. Column 2 and 3 further show that the treatment effect has a similar size as the difference between genders or the effect of one year of age.[14] Since the treatment take-up is 74%, LATE estimates of the treatment effect are about a third larger than the intention-to-treat effects shown in the table.[15]

	Reported to have predicted correctly		
	(1)	(2)	(3)
Treatment dummy	-0.110**	-0.119**	-0.095**
	(0.050)	(0.049)	(0.048)
Child's sex (1 = female)		-0.115**	-0.126***
		(0.047)	(0.046)
Child's age (in years)		-0.140***	-0.148***
		(0.041)	(0.039)
Additional controls	No	No	Strata & Int. FE
Sample restriction	Treatment & Control Group		
Mean control group:	0.647		
Observations	394	394	394

Table 2: Treatment effect regressions. Notes: Coefficients are average marginal effects of Probit regressions. Standard errors are in parentheses. The dependent variable is a dummy of whether the participant reported to have predicted the die roll correctly. In column 3 we control for strata and interviewer fixed effects. ***, **, * indicate significance at the 1, 5 and 10 percent level, respectively.

Result 3 *The treatment effect is stronger for participants whose parents use a less warm parenting style, and directionally so for parents who are less trusting.*

As discussed in Section 3.1 the mentoring program provides resources that are correlated with children's preferences for honesty. We hypothesize that mentors could serve as (partial)

[14] The treatment effect is stronger for boys but this difference is not significant (p= 0.141).

[15] To explore whether treatment intensity matters, we relate the number of meetings and the length of actual programme participation (i.e., the time between first and last mentoring meeting) to participants' honesty (i.e., the likelihood of not reporting to have predicted correctly). We do not find a significant relationship, neither for the number of meetings (Spearman's $\rho = -0.096$, $p = 0.338$, $N = 102$) nor for the length of actual programme participation (Spearman's $\rho = -0.150$, $p = 0.133$, $N = 102$), though the point estimates suggest a positive dose response-relationship between treatment intensity and honesty.

substitutes for parents and effectively provide these resources in the case of scarcity. We would thus expect the mentoring programme to have a larger effect on honesty in families that lack these characteristics. Table 3 shows OLS coefficients and adds interaction effects of treatment dummy and parenting style or maternal trust to the treatment effect regressions. Both interaction effects go in the hypothesized direction, even though only the coefficient for parenting style is significant. The increased supply of warm parenting in the household (through the mentor) has a similar effect as the parenting style of parents themselves. Figure 1 shows the treatment effect for the three tertiles of parenting style and shows how the treatment effect increases as parents' style gets colder (from left to right). In the treatment group, the level of lying is almost exactly identical in the three parenting style tertiles, as the treatment offsets the effect of parenting style. This suggests that mentors indeed serve as substitutes for parental input.

This result also suggests that the correlations shown in Table 1 between parenting style and trust on the one side and children's honesty on the other side might be due to an underlying causal effect. Mentoring is randomly allocated and it seems to work through similar channels as we identified in the correlational analysis.

	Reported to have predicted correctly	
	(1)	(2)
Treatment dummy	-0.115**	-0.123**
	(0.051)	(0.051)
Warm parenting style (std.)	-0.065**	
	(0.026)	
Treat × warm PS	0.088**	
	(0.044)	
Mother's trust (std.)		-0.067**
		(0.030)
Treat × mother's trust		0.040
		(0.052)
Sample restriction	Treatment and Control Group	
Observations	394	394

Table 3: Treatment effect regressions with interactions. Notes: Coefficients are from OLS regressions. Robust standard errors are in parentheses. The dependent variable is a dummy of whether the participant reported to have predicted the die roll correctly. As in all main specifications we control for gender and age of the child in all three regressions. ***, **, * indicate significance at the 1, 5 and 10 percent level, respectively.

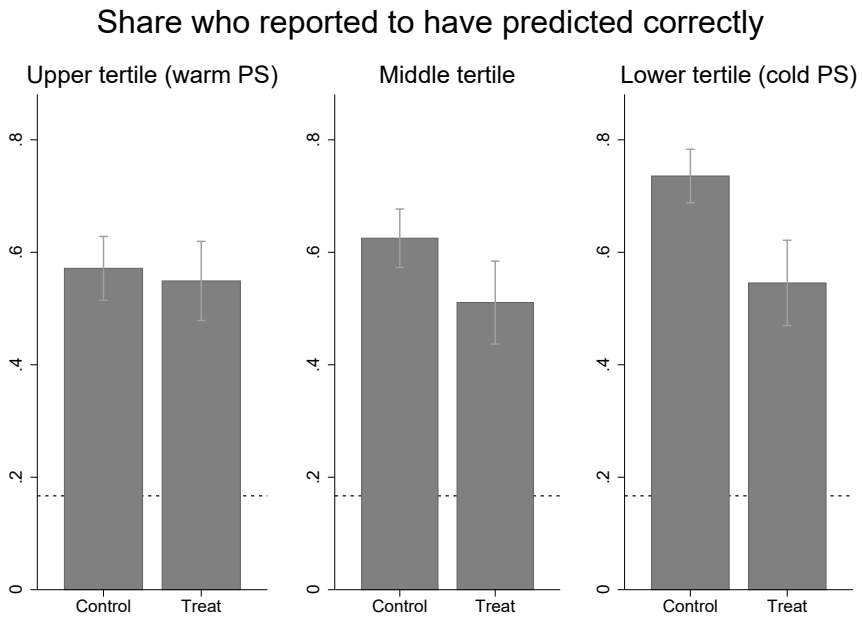


Figure 1: Treatment effect by parenting style. Notes: The vertical axis shows the share of participants who reported to have predicted the die roll correctly. Error bars indicate standard errors. Participants are split by treatment and control group and by the warmth of parenting style of their mother (colder on the right). The dashed horizontal lines mark the expected share of correct predictions without lying (16.7%).

3.3 Robustness of the Treatment Effect

In this section, we check two threats to the validity of our interpretation of the treatment effect.

Result 4 *The treatment effect on reporting behaviour is distinct from the treatment effect on prosociality.*

Kosse et al. (2020) analyze the same intervention and find a causal effect of the mentoring programme on prosociality. Even though prosociality and honesty are distinct concepts, they are arguably related. For example, Maggian and Villeval (2016) show that dictator game giving and truthful reporting are correlated. To find out whether the programme's effect on honesty is distinct from its effect on prosociality, we control for the programme's effect on prosociality when regressing reporting behaviour on the treatment dummy. Kosse et al. (2020) measure prosociality as the equally-weighted score of standardized measures of (i) three

incentivized dictator game experiments that the child played with another child of the same age; (ii) three age-adapted questions on trust; and (iii) parents' answers to the "prosocial scale" questions of the "Strength and Difficulties Questionnaire" (Goodman 1997).

Table 4 repeats the treatment effect regressions of Table 2 controlling for the programme's effect on prosociality. In column 1, we do so by including the within-participant average of the prosociality measures elicited in 2013 and 2015, i.e., about one and three years after the intervention. Prosociality is indeed significantly correlated to honesty. At the same time, the treatment coefficient remains very similar in size and significance. Nevertheless, if the experiments in 2013 and 2015 measure prosociality only with noise, the specification in column 1 will not fully account for the treatment effect on prosociality. To properly control for the treatment effect on prosociality in the presence of measurement error, we use the "obviously related IV" technique suggested by Gillen et al. (2019). This approach eliminates the uncorrelated part of the measurement error in our prosociality measures by using the two measures of prosociality as instrument for each other. We thus duplicate the data, assign the 2013 measure as dependent variable and the 2015 measure as instrument for the top half of the data and vice-versa for the bottom half of the data. To correct for using each observation twice, we cluster standard errors on participant. Column 2 is the second stage of this estimation. It shows that the treatment-effect estimate is again very similar and that significance is only slightly lower. This indicates that the treatement effect on lying is distinct from the treatment effect on prosociality.[16]

Our second robustness check concerns the fact that not all children in the intention-to-treat sample participate in the reporting experiment, which took place four years after the intervention (67.0% of participants in the treatment group participated and 66.7% in the control group). If this attrition is correlated with honesty and the treatment, then our treatment effect estimates could be biased. We find this not to be the case.

Result 5 *The treatment effect is not affected by differential attrition on observables.*

Table A.1 in Appendix A shows that there is no significant effect of treatment status or

[16] Interpreting the effects on the two prosocility measures (from 2013 and 2015) and the reporting behavior (from 2016) as one family of hypotheses and therefore conducting a family-wise error correction confirms our previous results (Romano-Wolf corrected p-values: 0.058 (lying, 2016), 0.010 (prosociality, 2013) and 0.058 (prosociality, 2015), based on 1,000 bootstrap replications).

baseline honesty on the likelihood of participating in our survey. Still, to correct for minor imbalances, we weight observations by the predicted inverse probabilities of participating in the reporting experiment. Our best proxy for honesty preferences measured before the start of the mentoring intervention is the "conduct problems" score of the Strengths and Difficulties Questionnaire. The questionnaire asks, amongst others, for parents' perception of their child's lying and stealing. The Spearman correlation of the score with whether the participant reported to have predicted correctly is 0.110 ($p = 0.015$, $N = 490$). Weights are estimated from a Probit model of a binary selection indicator (indicating participation in the reporting experiment) regressed on baseline "conduct problems" score and treatment assignment and their interaction (as in column 4 of Table A.1). Column 3 of Table 4 shows that the treatment effect is unchanged when we correct for attrition in this way. Columns 4 and 5 show Lee (2009) bounds on the treatment effect ($p = 0.043$ and 0.045). Given the absence of selective attrition, it is not surprising that the bounds are tight and confirm the main result.

	Reported to have predicted correctly				
				Lee bounds	
	OLS	IV	WLS	Upper	Lower
	(1)	(2)	(3)	(4)	(5)
Treatment dummy	-0.108**	-0.092*	-0.112**	-0.109**	-0.114**
	(0.053)	(0.054)	(0.051)	(0.054)	(0.057)
Prosociality (standardized)	-0.072***				
(Av. of 2013 & 2015 measures)	(0.023)				
Prosociality (standardized)		-0.149***			
(ORIV: 2013 & 2015 measures)		(0.050)			
Weights	No	No	IPW	No	No
Sample restriction	Treatment and Control Group				
Observations (cluster)	374	374	394	590	590
Selected observations				394	394

Table 4: Robustness checks: Treatment effects controlling for the programme's effect on prosociality and using inverse probability weights to account for differential attrition. Notes: The dependent variable is a dummy of whether the participant reported to have predicted the die roll correctly. As in all main specifications we control for gender and age of the child in the regressions shown in column 1 to 3. The reduced number of observations in columns 1 and 2 is due to missing observations in the 2015 data collection. Column 1 shows OLS estimates, column 2 shows the second stage of ORIV estimates. In column 3 coefficients are from a weighted least-square (WLS) estimation, weights are predicted inverse probabilities of not being lost to follow-up. Column 1 shows robust standard errors in parentheses (bootstrapped standard errors yield very similar results). Column 2 shows standard errors clustered at the individual level Gillen et al. (2019). Column 3 shows robust standard errors. Columns 4 and 5 show Lee (2009) bounds on the treatment effect (bootstrapped standard errors based on 1,000 bootstrap replications in parentheses). ***, **, * indicate significance at the 1, 5 and 10 percent level, respectively.

4 Conclusion

In this paper, we investigate whether preferences for honesty are malleable and, if yes, what determines them. We find that honesty among children is correlated with having parents who have a warmer parenting style or are more trusting. We provide causal evidence for the influence of the social environment by randomizing children into a year-long mentoring programme. The programme has long-term effects: four year after the end of the programme, mentored children are significantly more honest. Our analyses of heterogeneous treatment effects indicate that the programme especially benefits children with parents who have a less

warm parenting style and are less trusting. The fact that mentors bring in exactly these resources suggests that the programme serves as a substitute for parental input.

We conclude that preferences for honesty are indeed malleable and that they can be changed by an intervention. This has clear policy implications. Our data show that early-childhood interventions cannot just improve a child's achievements but also affect their social and moral behaviour. Whether this is desirable clearly depends on a careful welfare analysis. Our results also imply that preferences for honesty can not just be enhanced but also eroded by the social environment, with potentially long-lasting effects for the working of society. A carefully designed intervention, however, can counteract this effect.

References

Abeler, J., D. Nosenzo, and C. Raymond (2019). "Preferences for truth-telling". *Econometrica* 87.4, pp. 1115–1153.

Akerlof, G. (1970). "The Market for Lemons: Quality Uncertainty And The Market Mechanism". *The Quarterly Journal of Economics* 84.3, pp. 488–500.

Alan, S., T. Boneva, and S. Ertac (2019). "Ever failed, try again, succeed better: Results from a randomized educational intervention on grit". *The Quarterly Journal of Economics* 134.3, pp. 1121–1162.

Alan, S., S. Ertac, and M. Gumren (2020). "Cheating and incentives in a performance context: Evidence from a field experiment on children". *Journal of Economic Behavior & Organization* 179, pp. 681–701.

Alesina, A. and E. La Ferrara (2002). "Who trusts others?" *Journal of Public Economics* 85.2, pp. 207–234.

Becker, A., B. Enke, and A. Falk (2020). "Ancient origins of the global variation in economic preferences". *AEA Papers and Proceedings* 110, pp. 319–23.

Bettinger, E. and R. Slonim (2007). "Patience among children". *Journal of Public Economics* 91.1-2, pp. 343–363.

Boot, M. (Feb. 2021). *In office, Trump was the greatest threat to U.S. democracy. Now it may be Tucker Carlson.* URL: https://www.washingtonpost.com/opinions/2021/02/12/tucker-carlson-conspiracies-fox-news-dangerous/.

Brettschneider, C. (Nov. 2020). *Don't underestimate the threat to American democracy at this moment.* URL: https://www.theguardian.com/commentisfree/2020/nov/04/american-democracy-election-threat-trump.

Bucciol, A. and M. Piovesan (2011). "Luck or cheating? A field experiment on honesty with children". *Journal of Economic Psychology* 32.1, pp. 73–78.

Callen, M., M. Isaqzadeh, J. D. Long, and C. Sprenger (2014). "Violence and risk preference: Experimental evidence from Afghanistan". *American Economic Review* 104.1, pp. 123–48.

Cappelen, A., J. List, A. Samek, and B. Tungodden (2020). "The effect of early-childhood education on social preferences". *Journal of Political Economy* 128.7, pp. 2739–2758.

Cohn, A. and M. A. Maréchal (2018). "Laboratory measure of cheating predicts school misconduct". *The Economic Journal* 128.615, pp. 2743–2754.

Cohn, A., M. A. Maréchal, and T. Noll (2015). "Bad boys: How criminal identity salience affects rule violation". *Review of Economic Studies* 82.4, pp. 1289–1308.

Crawford, V. and J. Sobel (1982). "Strategic information transmission". *Econometrica* 50.6, pp. 1431–1451.

Cunha, F. and J. Heckman (2007). "The technology of skill formation". *American Economic Review* 97.2, pp. 31–47.

Dai, Z., F. Galeotti, and M. C. Villeval (2018). "Cheating in the lab predicts fraud in the field: An experiment in public transportation". *Management Science* 64.3, pp. 1081–1100.

Doepke, M., G. Sorrenti, and F. Zilibotti (2019). "The economics of parenting". *Annual Review of Economics* 11, pp. 55–84.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011). "Individual risk attitudes: Measurement, determinants, and behavioral consequences". *Journal of the European Economic Association* 9.3, pp. 522–550.

Dreber, A. and M. Johannesson (2008). "Gender differences in deception". *Economics Letters* 99.1, pp. 197–199.

Edsall, T. (Jan. 2021). *Have Trump's Lies Wrecked Free Speech?* URL: `https://www.nytimes.com/2021/01/06/opinion/trump-lies-free-speech.html`.

Elias, N. (1969). *Über den Prozeß der Zivilisation*. Suhrkamp Frankfurt.

Ellingsen, T. and R. Östling (2010). "When does communication improve coordination?" *American Economic Review* 100.4, pp. 1695–1724.

Evanega, S., M. Lynas, J. Adams, and K. Smolenyak (2020). "Coronavirus misinformation: quantifying sources and themes in the COVID-19 'infodemic'". *Cornell working paper*.

Falk, A., A. Becker, T. Dohmen, D. Huffman, and U. Sunde (2016). "The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences". *IZA Discussion Paper* 9674.

Falk, A. and F. Kosse (2020). "The briq family panel: An overview". *mimeo*.

Falk, A., F. Kosse, P. Pinger, H. Schildberg-Hörisch, and T. Deckers (forthcoming). "Socio-Economic Status and Inequalities in Children's IQ and Economic Preferences". *Journal of Political Economy*.

Fehr, E., U. Fischbacher, B. Von Rosenbladt, J. Schupp, and G. G. Wagner (2003). "A nation-wide laboratory: Examining trust and trustworthiness by integrating behavioral experiments into representative survey". *CESifo Working Paper*.

Fehr, E., D. Glätzle-Rützler, and M. Sutter (2013). "The development of egalitarianism, altruism, spite and parochialism in childhood and adolescence". *European Economic Review* 64, pp. 369–383.

Fischbacher, U. and F. Föllmi-Heusi (2013). "Lies in disguise—an experimental study on cheating". *Journal of the European Economic Association* 11.3, pp. 525–547.

Forehand, R. and D. J. Jones (2002). "The stability of parenting: A longitudinal analysis of inner-city African-American mothers". *Journal of Child and Family Studies* 11.4, pp. 455–467.

Gächter, S. and J. F. Schulz (2016). "Intrinsic honesty and the prevalence of rule violations across societies". *Nature* 531, pp. 496–499.

Gerlach, P., K. Teodorescu, and R. Hertwig (2019). "The truth about lies: A meta-analysis on dishonest behavior." *Psychological Bulletin* 145.1, p. 1.

Gillen, B., E. Snowberg, and L. Yariv (2019). "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study". *Journal of Political Economy* 127.4, pp. 1826–1863.

Glätzle-Rützler, D. and P. Lergetporer (2015). "Lying and age: An experimental study". *Journal of Economic Psychology* 46, pp. 12–25.

Gneezy, U. (2005). "Deception: The role of consequences". *American Economic Review* 95.1, pp. 384–394.

Gneezy, U., A. Kajackaite, and J. Sobel (2018). "Lying Aversion and the Size of the Lie". *American Economic Review* 108.2, pp. 419–453.

Gneezy, U., B. Rockenbach, and M. Serra-Garcia (2013). "Measuring lying aversion". *Journal of Economic Behavior & Organization* 93, pp. 293–300.

Goodman, R. (1997). "The Strengths and Difficulties Questionnaire: A Research Note". *Journal of Child Psychology and Psychiatry* 38.5, pp. 581–586.

Greene, J. and J. Paxton (2009). "Patterns of neural activity associated with honest and dishonest moral decisions". *Proceedings of the National Academy of Sciences* 106.30, pp. 12506–12511.

Hanna, R. and S.-Y. Wang (2017). "Dishonesty and selection into public service: Evidence from India". *American Economic Journal: Economic Policy* 9.3, pp. 262–90.

Heldring, L. (2021). "The origins of violence in Rwanda". *The Review of Economic Studies* 88.2, pp. 730–763.

Holden, G. W. and P. C. Miller (1999). "Enduring and different: A meta-analysis of the similarity in parents' child rearing." *Psychological Bulletin* 125.2, p. 223.

Jiang, T. (2013). "Cheating in mind games: The subtlety of rules matters". *Journal of Economic Behavior & Organization* 93, pp. 328–336.

Kajackaite, A. and U. Gneezy (2017). "Incentives and cheating". *Games and Economic Behavior* 102, pp. 433–444.

Kartik, N., M. Ottaviani, and F. Squintani (2007). "Credulity, lies, and costly talk". *Journal of Economic Theory* 134.1, pp. 93–116.

Kartik, N., O. Tercieux, and R. Holden (2014). "Simple mechanisms and preferences for honesty". *Games and Economic Behavior* 83, pp. 284–290.

Kautz, T., J. J. Heckman, R. Diris, B. Ter Weel, and L. Borghans (2014). "Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success". *NBER Working Paper Series* 20749.

Khalmetski, K. and D. Sliwka (2019). "Disguising lies—Image concerns and partial lying in cheating games". *American Economic Journal: Microeconomics* 11.4, pp. 79–110.

Kosse, F., T. Deckers, P. Pinger, H. Schildberg-Hörisch, and A. Falk (2020). "The formation of prosociality: causal evidence on the role of social environment". *Journal of Political Economy* 128.2, pp. 434–467.

Kröll, M. and D. Rustagi (2017). "Reputation, Honesty, and Cheating in Informal Milk Markets in India". *mimeo*.

Lazer, D., M. Baum, Y. Benkler, A. Berinsky, K. Greenhill, F. Menczer, M. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. (2018). "The science of fake news". *Science* 359.6380, pp. 1094–1096.

Lee, D. S. (2009). "Training, wages, and sample selection: Estimating sharp bounds on treatment effects". *The Review of Economic Studies* 76.3, pp. 1071–1102.

Lowes, S., N. Nunn, J. A. Robinson, and J. L. Weigel (2017). "The evolution of culture and institutions: Evidence from the Kuba Kingdom". *Econometrica* 85.4, pp. 1065–1091.

Maggian, V. and M. C. Villeval (2016). "Social preferences and lying aversion in children". *Experimental Economics* 19.3, pp. 663–685.

Matsushima, H. (2008). "Role of honesty in full implementation". *Journal of Economic Theory* 139.1, pp. 353–359.

Mazar, N., O. Amir, and D. Ariely (2008). "The dishonesty of honest people: A theory of self-concept maintenance". *Journal of Marketing Research* 45.6, pp. 633–644.

Müller-Kohlenberg, H. and S. Drexler (2013). "Balu und Du ('Baloo and You') - A Mentoring Program: Conception and Evaluation Results". In: *Mentoring: Practices, potential challenges and benefits*. Ed. by M. Shaughnessy. Nova Science Publishers, pp. 107–123.

Potters, J. and J. Stoop (2016). "Do cheaters in the lab also cheat in the field?" *European Economic Review* 87, pp. 26–33.

Serra-Garcia, M. and U. Gneezy (2020). "Mistakes and Overconfidence in Detecting Lies". *mimeo*.

Sobel, J. (2020). "Lying and deception in games". *Journal of Political Economy* 128.3, pp. 907–947.

Stouthamer-Loeber, M. (1986). "Lying as a problem behavior in children: A review". *Clinical Psychology Review* 6.4, pp. 267–289.

Stouthamer-Loeber, M. and R. Loeber (1986). "Boys who lie". *Journal of Abnormal Child Psychology* 14.4, pp. 551–564.

Sutter, M., C. Zoller, and D. Glätzle-Rützler (2019). "Economic behavior of children and adolescents: A first survey of experimental economics results". *European Economic Review* 111, pp. 98–121.

Talwar, V. and A. Crossman (2011). "From little white lies to filthy liars: The evolution of honesty and deception in young children". In: *Advances in Child Development and Behavior*. Vol. 40. Elsevier, pp. 139–179.

Talwar, V. and K. Lee (2011). "A punitive environment fosters children's dishonesty: A natural experiment". *Child Development* 82.6, pp. 1751–1758.

Thönnissen, C., B. Wilhelm, S. Fiedrich, P. Alt, and S. Walper (2015). *Scales Manual of the German Family Panel, Release 6.0*. Tech. rep.

Tobol, Y. and G. Yaniv (2019). "Parents' marital status, psychological counseling and dishonest kindergarten children: An experimental study". *Journal of Economic Behavior & Organization* 167, pp. 33–38.

Wagner, G., J. Frick, and J. Schupp (2007). "The German Socio-Economic Panel Study (SOEP) – Scope, Evolution and Enhancements". *Schmollers Jahrbuch: Journal of Applied Social Science Studies* 127.1, pp. 139–169.

Online Appendix

A Additional analyses

	Lost to follow-up	
	(1)	(2)
Treatment dummy	-0.003	-0.002
	(0.040)	(0.041)
Conduct problems (SDQ, baseline)		0.015
		(0.025)
Conduct problems × treatment		0.012
		(0.039)
Sample restriction	Control Groups	
Observations	590	590
R2	0.000	0.002
p-value F-test	0.938	0.758

Table A.1: Analysis of attrition. Notes: Coefficients are from OLS regressions. Robust standard errors are in parentheses. The dependent variable is a dummy of whether the participant failed to participate in the survey containing the reporting experiment. The proxy for baseline honesty preferences is the "conduct problems" score of the Strengths and Difficulties Questionnaire (Goodman 1997), asking, amongst others, for mothers' perception of the child's lying and stealing. ***, **, * indicate significance at the 1, 5 and 10 percent level, respectively.

Baseline measure	Mean Control Group	Mean Treatment Group	Difference p-value
Family characteristics:			
Low parental income (binary)	0.480	0.443	0.487
	(0.032)	(0.042)	
Low parental education (binary)	0.464	0.493	0.585
	(0.031)	(0.042)	
Single parent (binary)	0.468	0.479	0.840
	(0.031)	(0.042)	
Number of siblings	1.063	1.049	0.891
	(0.063)	(0.080)	
Mother's age (in years)	38.830	39.043	0.723
	(0.368)	(0.460)	
Warm parenting style (standardized)	-0.076	-0.018	0.618
	(0.067)	(0.096)	
Mother's trust (standardized)	0.018	-0.047	0.531
	(0.062)	(0.084)	
Mother's patience (standardized)	-0.101	-0.001	0.363
	(0.068)	(0.081)	
Mother's willingness to take risk (std.)	0.055	0.054	0.989
	(0.063)	(0.084)	
Mother's altruism (standardized)	-0.006	-0.106	0.346
	(0.062)	(0.089)	
Child characteristics:			
Female (binary)	0.484	0.451	0.525
	(0.032)	(0.042)	
Age (in years, at follow-up)	12.504	12.472	0.590
	(0.035)	(0.049)	
Conduct problems (SDQ, std.)	-0.005	-0.071	0.533
	(0.061)	(0.088)	

Table A.2: Baseline balance in the follow-up sample ($N = 394$). 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 (parental style is collected after the treatment in 2013, children's age is the age at follow-up), see Section 2 for details. 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) are used to standardize variables.

Variable (z-scores)	Mentors	Low SES mothers	p-value of difference
Trust	0.314	-0.073	0.001
Altruism	0.206	0.003	0.068
Patience	-0.110	-0.060	0.669
Willingness to take risk	0.007	0.023	0.881

Table A.3: Comparison of mentors and low SES mothers. This table compares mentors and low SES mothers. The measures are standardized. Standardizations are conducted using the baseline distribution of all mothers. For details on the measure see Section 2. The third column indicates results from two-sample two-sided t-tests. The joint number of observations is 680.

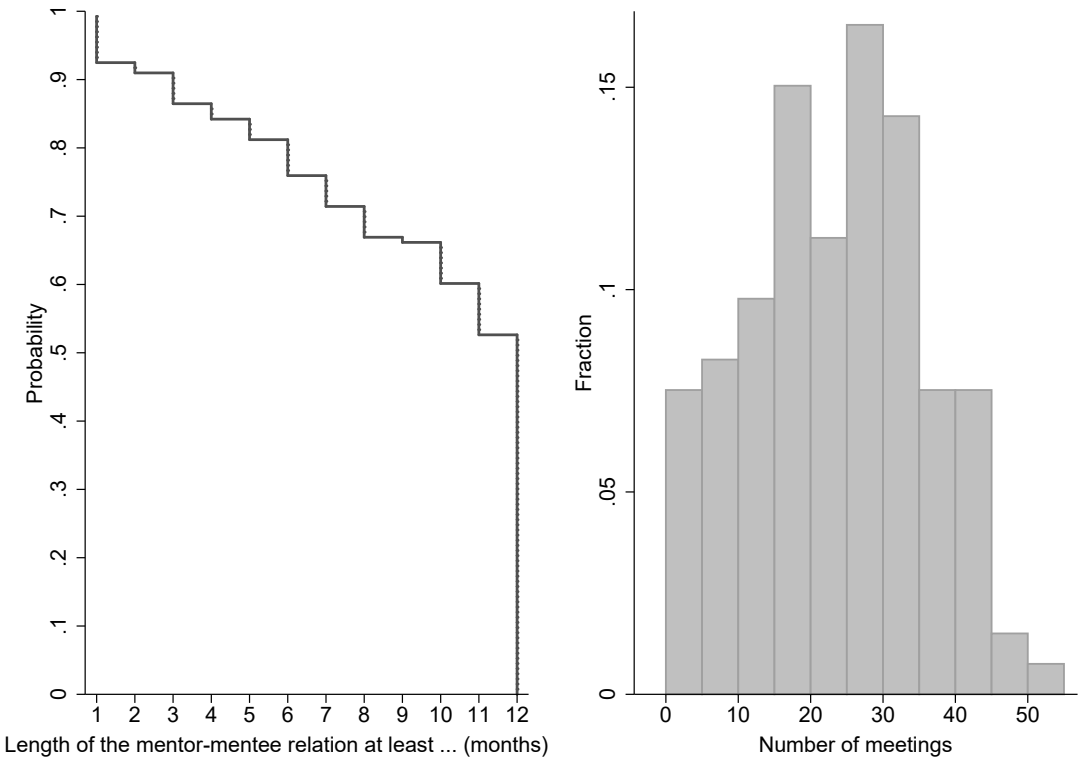


Figure A.1: Length of mentoring relationship. The left panel shows the CDF of the programme duration for those participants in the treatment group who had at least one meeting with their mentor. The right panel depicts a histogram of the number of meetings.

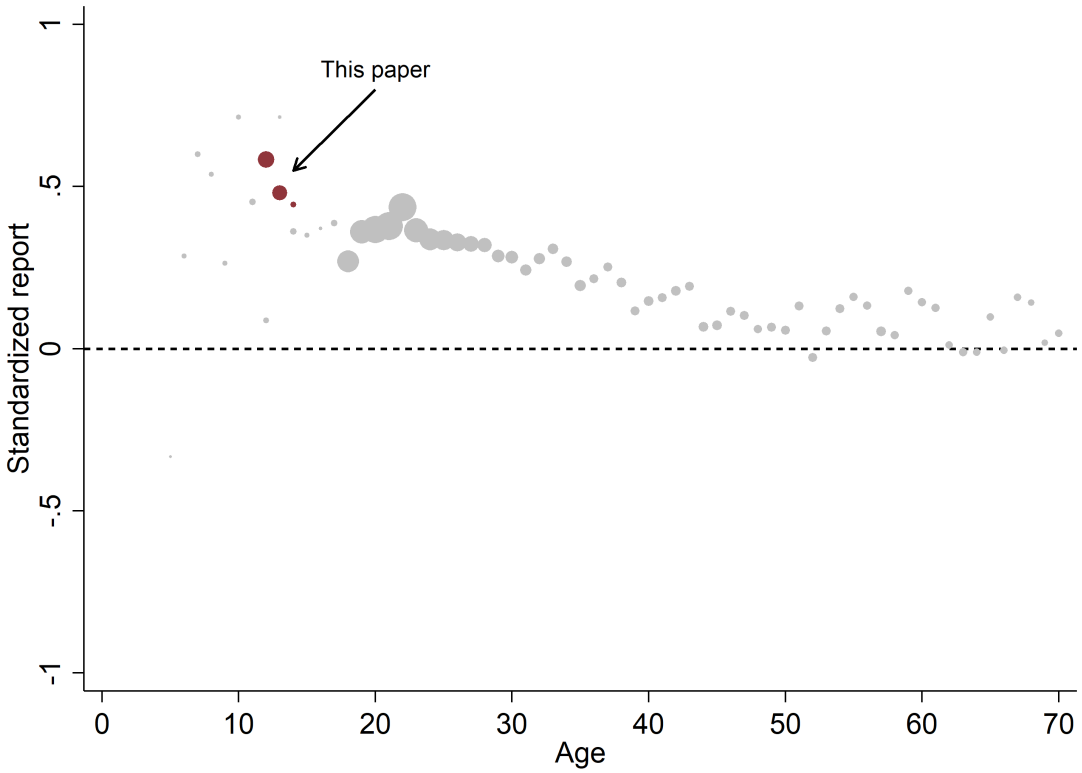


Figure A.2: Average report in FFH experiments by age. The graph combines the data from this paper (in red) with data of 35 other papers using the FFH experiment that also elicited the age of participants, collected by Abeler et al. (2019) (in grey). The x-axis depicts the age of participants. The y-axis depicts the average "standardized report". We pool across all treatments in the 35 papers and across treatments in this paper and show the average standardized report for each age (in years) separately. The size of the bubble is proportional to the number of subjects. The standardized report maps the actual report onto the interval -1 to +1 where -1 signifies the report yielding the lowest payoff and +1 the report yielding the highest payoff. A standardized report of 0 signifies an average report that generates the same payoff as truthful reporting would do (see Abeler et al. (2019) for details).

B Theoretical framework

In this appendix, we aim to clarify how reports in this experiment are linked to preferences for honesty. While we cannot judge individual behaviour in this experiment, we can judge the behaviour of a group of participants. Assume that participant i privately observes an i.i.d. state of nature $\omega_i \in \{0, 1\}$, where 1 means the participant predicted the die roll correctly and 0 means they did not. The participant then reports $r_i \in \{0, 1\}$, i.e., they claim to have predicted correctly ($r_i = 1$) or claim to have not done so. We assume that participants have the following utility function:

$$U_i(\omega_i, r_i) = r_i b - c_i \mathbb{I}(r_i \neq \omega_i)$$

Participants like money and thus prefer to report $r_i = 1$ as this yields the monetary benefit b. They might also prefer to be honest, and thus pay a psychological lying cost $c_i \geq 0$ when they lie. $\mathbb{I}()$ denotes the indicator function.[17] We assume that c_i is heterogeneous in the population and that it is potentially affected by the treatment or by parents.

The utility function implies that no participant will "lie downwards", i.e., all participants who predicted correctly will say so. Lying downwards reduces the monetary payoff and might incur lying costs and is thus never optimal. Participants who did not predict correctly will (dishonestly) claim to have done so if and only if their $c_i < b$, i.e., if lying feels less bad than not obtaining the monetary reward. Since there is a $\frac{1}{6}$ chance to predict correctly, the share of high reports among a group of N participants is $\frac{1}{6} + \frac{5}{6} \sum_{i=1}^{N} \frac{\mathbb{I}(c_i<b)}{N} = \frac{1}{6} + \frac{5}{6} Prob(c_i < b)$. For large N, if a group has a larger share of high reports, then this means that more participants lied in this group (as ω_i is independently drawn for each individual) and this has to come from a left-shift in the distribution of c_i, i.e., lower lying costs.

[17] The functional form of the lying cost does not matter in our setting, as we only have two possible reports. The preference for honesty could, for example, stem from religious or moral reasons, from social norms of honesty, or from self-image concerns. For our model to capture self-image concerns we need to assume that the participant remembers their die roll ω_i and their report r_i. If they lied, they then suffer disutility c_i from knowing they lied. If individuals forget their die roll and only remember their report, even though the two happened essentially at the same time, then a signalling model is more appropriate (see Abeler et al. (2019) for such a model).

C Instructions

C.1 Instructions translated from German

Below are the instructions shown on the screen of the laptop computer used for the questionnaire (translated from German). The original screenshots are shown in the next section. The experimental currency unit is called "stars" and one star is converted into money at the end of the questionnaire (1 star = 0.20 euros).

[New screen]

Finally, something totally different: you can now play a small game, just by yourself, on the computer. The game is called "Predict a number" and you can win stars by playing the game. You will need this die and this dice cup. *[The interviewer hands die and cup to the participant.]*

You can use the die and cup in such a way that nobody can see the result of the die roll. *[The interviewer demonstrates the use of die and cup.]*

How the game works will be explained on the screen. I will prepare something else during that time. You will play the game alone at the computer. Please only come to me if you have any questions. *[The interviewer turns the laptop such that only the participant can see the screen. The interviewer then leaves the interview situation and is told that the participant has to click the answers on the next screen themself.]*

[New screen]

You have received a die and a dice cup. You probably already know dice and cups from other games. You can roll numbers between 1 and 6. Try it a couple of times! Roll in such a way that only you can see the resulting number.

Did this work? *[Button: yes, Button: no]*

[New screen]

We now play a game. The aim is to predict the number the die will show. The game consists of four steps:

Step 1: You think about which number you might roll and keep that number in mind.

Step 2: You roll the die and check whether you predicted the correct number.

Step 3: Roll the die again a couple of times to check that the die is working properly.

Step 4: On the computer, enter whether you guessed the number correctly or not. If you were right, you receive **5 stars**. If you were wrong, you receive nothing.

Please click "next". You can then continue in the questionnaire using the blue arrows or the "Enter" key. *[Button: next]*

[New screen]

Let's now play the game:

Step 1: Predict which number you might roll and keep that number in mind.

Step 2: Roll the die and check whether you predicted the correct number.

Step 3: Roll the die again a couple of times to check that the die is working properly.

When you are done, please click "next". *[Button: next]*

[New screen]

Step 4: Now enter on the computer whether you predicted correctly. If you were right, you receive **5 stars**. If you were wrong, you receive nothing.

[Button: I have predicted correctly. Button: Unfortunately, I have predicted wrongly.]

C.2 Screenshots of the original instructions

Zahlen vorhersagen

--> Würfel und Würfelbecher benötigt!

Zum Schluss noch etwas ganz anderes: Du darfst noch ein kleines Spiel alleine am Computer spielen.
Das Spiel heißt "Zahlen vorhersagen" und Du kannst dabei Sterne gewinnen.

Für das Spiel werden dieser Würfel und dieser Würfelbecher benötigt.

--> Würfel und Würfelbecher dem Kind übergeben!

Damit kann man so würfeln, dass niemand anderes das Würfel-Ergebnis sieht.

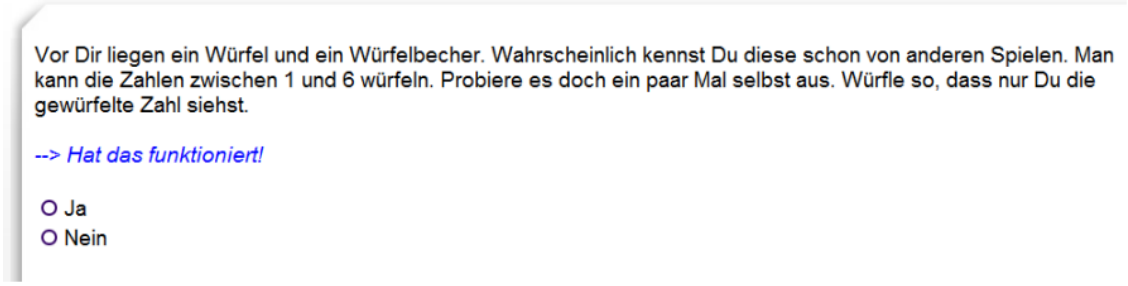
--> Vormachen!

Wie das Spiel genau funktioniert, wird am Bildschirm erklärt.
Ich werde in der Zwischenzeit etwas anderes vorbereiten *(Schülerfragebogen)*.
Du spielst das Spiel allein am Computer. Wende Dich nur an mich, wenn Du Fragen hast.

--> Bitte drehen Sie den Laptop so, dass nur noch die/der Befragte den nächsten Bildschirm sehen kann.

--> VERLASSEN Sie die Interview Situation.

--> Die/der Befragte soll selbst die Antworten auf der nächsten Seite anklicken!



Vor Dir liegen ein Würfel und ein Würfelbecher. Wahrscheinlich kennst Du diese schon von anderen Spielen. Man kann die Zahlen zwischen 1 und 6 würfeln. Probiere es doch ein paar Mal selbst aus. Würfle so, dass nur Du die gewürfelte Zahl siehst.

--> Hat das funktioniert!

O Ja
O Nein

Wir spielen nun ein Spiel, bei dem es darum geht, die gewürfelte Würfelzahl vorher zu sagen.

Das Spiel besteht aus vier Schritten:

Schritt 1: Du überlegst Dir, welche Zahl Du wohl würfeln wirst und merkst Dir diese Zahl gut.

Schritt 2: Du würfelst und schaust nach, ob Du die richtige Zahl vorhergesagt hast.

Schritt 3: Würfel noch ein paar Mal, um zu schauen, dass der Würfel richtig funktioniert.

Schritt 4: Du gibst hier am Computer ein, ob Du die Zahl richtig geraten hast oder nicht.
Falls Du richtig lagst, bekommst Du **5 Sterne**. Wenn Du falsch lagst, bekommst Du nichts.

Bitte klicke "weiter" an. Mit dem blauen Pfeil oder der ENTER-Taste kommst Du anschliessend weiter im Fragebogen.

O weiter

Lass uns nun das Spiel spielen:

Schritt 1: Sag vorher welche Zahl Du wohl würfeln wirst und merk Dir diese Zahl gut.

Schritt 2: Würfele und schaue nach, ob du die richtige Zahl vorhergesagt hast.

Schritt 3: Würfele noch ein paar Mal, um zu schauen, dass der Würfel richtig funktioniert.

Wenn Du fertig bist, klicke bitte auf "weiter".

O weiter

Schritt 4: Gib nun am Computer ein, ob Du richtig vorhergesagt hast.
Falls Du richtig lagst, bekommst Du **5 Sterne**. Wenn Du falsch lagst, bekommst Du nichts.

O Ich habe richtig vorhergesagt.
O Ich habe leider falsch vorhergesagt.

C.3 Questionnaire

Mothers' and mentors' preferences and beliefs are measured using validated survey items. We measure general trust using the two items "As long as I am not convinced otherwise, I always assume that people have only the best intentions" (Falk et al. 2016) and "In general, one can trust people" (Fehr et al. 2003). Responses were given on an eleven-point Likert scale. For patience, we use the measure: "When it comes to financial decisions, how do you assess your willingness to abstain from things today so that you will be able to afford more tomorrow. Please indicate on the scale, where the value 0 means 'not at all willing to abstain today' and the value 10 means 'very willing to abstain today'". To measure willingness to take risk, we ask "How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? Please indicate on the scale, where the value 0 means: 'not at all willing to take risks' and the value 10 means: 'very willing to take risks" (Dohmen et al. 2011). Altruism is measured using the question "How would you assess your willingness to share with others without expecting anything in return, for example your willingness to give to charity?" (Falk et al. 2016).

To estimate "warm parenting style", mothers indicated their agreement with eight statements on a 5-point Likert scale from "never" to "always".[18] As in Falk et al. (forthcoming), we use factor analysis to extract one latent parenting style from these items (+ and – indicate the direction of factor loadings). The items are "I show my child with words and gestures that I like him/her." (+), "I praise my child." (+), "If my child does something against my will, I punish him/her." (–), "I make it clear to my child that he/she is not to break the rules or question my decisions." (–), "I think my child is ungrateful when he/she does not obey me." (–), "I do not talk to my child for a while when he/she did something wrong." (–), "When my child goes out, I know exactly where he/she is." (+) and "When my child goes out, I ask what he/she did and experienced." (+).

[18] For a detailed description of our parenting style measures, see Thönnissen et al. (2015) and the references therein.