Theory of Mind among Disadvantaged Children: Evidence from a Field Experiment*

August 14, 2019

GARY CHARNESS
Department of Economics
University of California, Santa Barbara
charness@econ.ucsb.edu

JOHN A. LIST
Department of Economics
University of Chicago
jlist@uchicago.edu

ALDO RUSTICHINI
Department of Economics
University of Minnesota
rusti001@umn.edu

ANYA SAMEK
CESR and Department of Economics
University of Southern California
anyasamek@gmail.com

JEROEN VAN DE VEN
Tinbergen Institute and Amsterdam School of Economics University of Amsterdam j.vandeven@uva.nl

Abstract. Theory of Mind (ToM), the ability to correctly attribute mental states to others, is important in social interactions. We evaluate the development of ToM in about 800 mostly disadvantaged young children. We next conduct a field experiment with about 160 children in which we find that the low ToM rates for these disadvantaged children improve substantially in environments where the presence of other children is made salient. We see that ToM performance increases for both younger and older children in the treatment with strong salience, but that the treatment with weaker salience seems to be only effective in improving the ToM rates for older children.

Keywords: Theory of Mind, Social salience, Children, Field experiment, Disadvantaged

JEL Codes: B49, C70, C93, I24

· We thank Isabel Brocas, Juan Carrillo, Giorgio Coricelli, Carsten de Dreu, Uri Gneezy, Dan Houser, Cristine Legare, and Theo Offerman for very valuable comments. We also thank seminar and workshop participants at the Institute for Advanced Studies Toulouse, the HCEO Conference on Networks (Chicago), the "Understanding cognition and decision making by children" conference (University of Southern California), the Experimental Methods in Policy Conference (2018), and "Field Days" (Rotterdam, 2019).

1 Introduction

Children differ tremendously in their skills. Differences across individuals persist and even widen with age (Heckman, 2000, 2006; Heckman et al., 2006). A major social concern is that the socio-economic background of children has important influence on their skills, and so differences in socio-economic backgrounds are potentially associated with large inequalities in economic success later in life. Since achievements later in life build on the foundations that are laid down earlier (Heckman, 2006), it is important to study the development of skills among children in their early years.

One skill that is particularly useful is the ability to understand the mental states of other people, or *Theory of Mind* (Premack and Woodruff, 1978). This ability stems from two essential skills. The first is the ability to recognize that the thinking of others is a process independent of one's own (and cannot use, for example, information that is only available to the person who is reconstructing the thought of the others). The second is the recognition that the thinking process is common to all human beings, so one can effectively try to mimic the thinking process of others, once the difference in information is taken into account. The ability to see things from someone else's point of view is essential in almost every social interaction, as not having Theory of Mind (ToM) reduces the ability to understand and foresee the motivation of others, and thus makes effective cooperation more difficult to achieve. This, in turn, makes it more difficult to achieve personal economic success, and creates high social costs.[1]

Not having ToM is often associated with an inability to think strategically. While there is considerable research on cognitive development in children, there are only a handful of studies in economics on strategic thinking by young children. ToM has been associated with the ability to

[1] For instance, bullying behavior has been linked to the absence of ToM (Randall, 1997; Hazler, 1996), because bullies fail to understand the feelings and intentions of others. In its extreme form, a lack of ToM is a core feature of autism, a severe developmental disorder associated with the inability to comprehend the social environment (Baron-Cohen et al., 1985).

think in a non-myopic, far-sighted way in strategic situations (Perner, 1979; Shultz et al., 1981; Sher et al., 2014). Two studies by Brocas and Carrillo (2018, forthcoming) test strategic thinking by children between four and seven years of age with a variety of simpler and more complex tasks. The first study shows that young children can reason strategically in simple individual decisions, they are much less able to solve more complex problems and to interact effectively with others. While "children understood the need to apply logical reasoning", they are often unable to anticipate future events and to make choices where this ability is needed. The second study considers equilibrium strategies in four iterative games that vary the degree of complexity, requirements in perspective-taking (own or rival), and action symmetry. It appears that iterative complexity, while important, is not the only factor, since perspective and action symmetry help to provide short cuts for the reasoning process.

Research has demonstrated that ToM typically develops at around three to four years of age, and that by the age of five to six almost all children pass standard tests of ToM (e.g., Wellman et al., 2001). The evidence for ToM development is very strong, but there appears to be relatively little research on children from parents with lower socio-economic levels. An understanding of how a skills gap among children of different backgrounds may arise requires studying the development of ToM among children from disadvantaged backgrounds, and an examination of differences in the developmental progress of ToM (Holmes et al., 1996; Cutting and Dunn, 1999; Cicchetti et al., 2003; Yagmurlu et al., 2005.[2] Devine and Hughes (2018) concludes that children from lower socioeconomic status families lag behind children from higher socioeconomic status families. They find a modest but significant association between ToM and socioeconomic status.

[2] Cutting and Dunn (1999) state: "Family background has a significant impact on the development of Theory of Mind." Most of the studies they include in their analysis focus on a very specific age range and rely on relatively small samples. Notable exceptions in these respects include the studies by Cicchetti et al. (2003), Cole and Mitchell (2000), Hughes and Ensor (2005) and Pears and Moses (2003).

We contribute to this area by collecting data about ToM among children from mostly low socioeconomic-status (SES) families. Our data collection was conducted at the Chicago Heights Early Childhood Center (CHECC; see Fryer et al., 2015). This data set is unique in that it is unusually large (827 children) and spans a wide age range (38-131 months), enabling us to track the development of ToM across a relatively large group of children. This approach has benefits and some costs. We are able to provide a unique glimpse from amongst disadvantaged children, but we have reduced coverage among higher SES children. Thus, we do not have a non-SES control sample of our own and can only compare to results in previous studies. In particular, we largely rely on the meta-analyses in Wellman et al. (2001) and Devine and Hughes (2018), whose results are discussed later.

In agreement with this earlier literature, our results provide support for the hypothesis that the group of children in our sample has considerably lower ToM scores compared to the average score of a more diverse sample including children from low, middle and high SES (as in the meta-analysis of Wellman, 2001). We acknowledge that several factors can explain the difference, such as differences in procedures. Nevertheless, given that we use a very standardized task, we believe that the most plausible candidate is that our sample consists of mostly disadvantaged children whereas children in other studies often come from middle- and higher SES families. This differential persists over the entire age range of our sample when we compare our results to the received literature. Given our results, a critical issue is whether, and how, one can increase effective ToM rates among low-SES kids.

A first approach to this problem is to stimulate the development of ToM by letting children experience differences in perspectives, for instance by engaging them in role-playing (e.g., Lillard et al., 2013). The empirical evidence of the effectiveness of such interventions is mixed. In addition, many of the existing studies have methodological issues. For example, these often document correlations rather than causal mechanisms. In addition, the interventions often last for a long period and embody other factors beyond role-playing even in the control conditions, making it unclear what interpretations can be made (Lillard et al., 2013). A different approach is to indicate explicitly the intention to deceive others by making children very aware of the fact that those others may have different beliefs. Wellman et al. (2001) found that such an

intervention improved performance in the ToM task. However, pointing towards the intention to deceive is such a strong hint to the correct answer that it is unclear if the task still measures ToM.

To explore an approach to increase ToM, we instead use a related, less-explicit, approach with a very simple manipulation. Our manipulation consists of stimulating children, in a subtle way, to think about the interaction with another child. After completing an unrelated memory task, we asked children to guess how well they performed on the task. Children were randomly assigned into one of three treatments, in a design adapted from Charness, Rustichini, and van de Ven (2018). Children's guesses in the *Baseline* treatment were kept private. Children in the *Social* treatment were told that their guess would be shared with one other child. The *Social* treatment is meant to stimulate children to think about how their guess will be perceived by other children. In our third treatment, the *Contest* treatment, children were also told that their guess would be shared with another child, and that upon learning this guess the paired child could decide whether to enter into a contest with him or her. The *Contest* treatment adds another step to the reasoning process, and it is meant to stimulate children to think about how their guess will be perceived by other children and how those other children will use this as an input to their decision to enter into a contest or not. Sher et al. (2014) refer to this capacity as Strategic Theory of Mind (SToM).

Our principal finding is that the rate of ToM is quite sensitive to the treatment that a child experienced. For example, subtle cues that increase the salience of a variety of social factors (such as reputation, social norms, and other children's perspectives, all induced by the presence of other children) substantially increase ToM rates. Yet, there seems to be heterogeneity amongst children—the treatment effect is most sizeable for children over 80 months (six and a half years) old. We estimate for that group that the rate of ToM is on average a remarkable 56 percentage points lower compared to other studies with middle- and high-income children. The treatment with the largest effect narrows this gap by a full 25 percentage points, representing nearly half of the gap.

The fact that subtle cues can increase the rate of ToM so dramatically may seem puzzling. Children lacking ToM should not suddenly gain this capacity. The way in which our intervention operates at least in part, we believe, is that some children who *appear* to lack ToM actually do possess ToM, but that this ability needs to be *activated*. This resonates with the findings in

Gneezy et al. (2017), in which test scores are responsive to incentives, even though the incentives are only announced just before the test, when new knowledge can no longer be acquired. In that study, children already had the ability to perform well, and increased effort explains the increased performance. In our case, the capacity to see things from someone else's perspective may be latent for some children in our sample and appears to be stimulated by subtle cues. In this way, measures of ToM or other skills might not reflect foundational differences, but rather simply show transient differences that can be washed away, at least in part, with simple manipulations. If there is insufficient stimulation in a small child's environment, one might see this as providing useful exercise to a mind that has not previously had much useful exercise. Of course, our study is merely a first step because we do not measure how such subtle manipulations affect ToM scores of high SES children. This is an important future direction holding promise for a deeper understanding in this literature.

The remainder of study proceeds as follows. The next section discusses our methods and design. Section 3 summarizes our results. Sections 4 and 5 provide an interpretation of our data and concluding remarks.

2 Methods and Design

Participants. The study was conducted at the Chicago Heights Early Childhood Center (CHECC).[3] CHECC is a large-scale intervention study with nearly 2,000 children on the role of different early education programs on schooling outcomes of disadvantaged children conducted in 2010-2014 (Fryer et al., 2015; 2018). Households who participated in CHECC originated from the surrounding area of Chicago Heights, Illinois. Chicago Heights is an ethnically diverse (41% African American, 34% Hispanic) and generally low-income area (29% of persons below poverty level, $18,121 per capita money income). To recruit participants for the overall

[3] CHECC was called the Griffin Early Childhood Center (GECC) between 2010 and 2012, and was renamed to CHECC in 2012.

program, CHECC ran a local marketing campaign each year, including direct mailings, automated phone calls to families with children enrolled in the district, and information booths at community events. Program information was also distributed to district leadership staff in the school districts, and administrative assistants at schools were asked to collect and submit registration forms for CHECC. The main goal of CHECC was to investigate the role of early childhood programs on educational attainment; therefore, households who signed up for the program were randomized each year (during four years 2010-2013) into one of several different treatment arms or to a control group, which included a preschool, a parenting program, or a control group that did not receive educational interventions. Fryer et al. (2015) and Fryer et al. (2018) report on the impact of the programs on cognitive abilities and executive functions.[4]

We use data from four CHECC studies in which a Theory of Mind (ToM) task was administered. Those studies took place between October 2010 and May 2014, with 991 children participating in one or more studies. Children took the test in their preferred language (English or Spanish). We exclude children for whom we do not have information about age (15 observations). Of the remaining sample, we have a valid ToM measure for 827 children. Most of those children (685) participated only once, 121 participated twice, and 21 participated three times. Children were between 38 and 131 months (about three years to eleven years) old at the time they took part in the study (mean 62.5, s.d. 15.0), and 51 percent are female. Figure C.1 in

[4] Children participating in CHECC are often asked to complete cognitive and non-cognitive assessments, including the assessments of ToM described here. Children are also often asked to come in for other experimental studies. Children are typically recruited for these studies by either asking parents to bring children to the study on an evening or a weekend, or conducting the study during school time, in which case parents are given the opportunity to opt out if they prefer their child not to participate. Thus far, CHECC children have participated in studies of risk preferences (Andreoni et al., 2018a), time preferences (Andreoni et al., 2018b), social preferences (Cappelen et al., 2016; Ben-Ner et al., 2017; List and Samak, 2013, List et al., 2018) and competitiveness (Samak, 2013). Castillo et al. (2018) evaluates the associations of economic preferences, cognitive abilities and executive functions at an early age with disciplinary referrals later in life. Cowell et al. (2015) uses a smaller sample of ~100 CHECC children to evaluate the association of a different ToM task with dictator game giving. No other study has reported on the ToM tasks we report on here.

the Appendix shows the distribution of age. Most children are between 38 and 78 months (about three years to six and a half years) old, but a non-negligible number of children (97) are over 78 months old.

The first three studies are non-experimental in the sense that the ToM task was administered before the children worked on any other task. In our experimental study ($N = 159$), children were randomized into one of three different treatments with varying degrees of interaction with other children. The ToM task in the experimental study was administered after the children had completed the other tasks. Fifty-six children are in the *Baseline* treatment, 52 in the *Social* treatment, and 51 in the *Contest* treatment. Table C1 in Appendix C reports more detailed summary statistics. Importantly, we conducted the experimental study by inviting parents from all treatment groups to bring their children to the center on a weekend day. Because CHECC is so large, it is unlikely that most children in this study knew one another.

Experimental design. Here we summarize the main features of the field experiment, a detailed description can be found in Appendix A. We randomly assigned children to one of three treatments. In every treatment, children were first "paired" and were told that each of the pair would be working on the task separately (out of sight or in a different room) with an activity leader. Children were able to visually see who was in their pair, and we tried to match children up by roughly the same age. We also asked children if they knew one another and did not assign any child to a child whom he/she already knew. Again, since CHECC is so large and includes multiple different programs, there were few cases where children knew one another. Children in every treatment first worked on a memory task. The experimenter showed the child a sheet with animals, each in its own house, and asked children to remember which animal belonged in which house. The child then received a sheet with blank houses and a set of animal cards, and had three minutes to put the animals back in their own houses. After completing the memory task, we asked children to guess how well they performed on the task.

Treatments differed depending on whether a child's guess was kept private or told to another child and whether or not the paired child could enter into a contest. In the *Baseline* treatment, children knew that their guess would be kept private (even though they were also told they were "paired"). In the *Social* treatment, children were told that their guess would be shared with another child. In the *Contest* treatment, children were also told that their guess would be

shared with another child, and that upon learning this guess the paired child could decide whether to enter into a contest with the person making the guess. Only the paired child could make the entry decision, so that his or her decision determined whether a contest took place.

Children were rewarded for accurate guesses. In the Contest treatment, the reward partly depended on whether the paired child entered into a contest. In the case of a contest, the child with the best performance received five prizes, while the other child received nothing. When the paired child did not enter into a contest, the first child (who had to enter the contest) received seven prizes and the paired child received three prizes. Children could strategically report a high confidence level to attempt to deter the other child from entering into a contest, thereby securing a high payoff. Thus, in this treatment the children were prompted to think about strategic considerations in relation to the paired child.[5]

Measurement of Theory of Mind.

We chose to use a short ToM test, adapted from the "unexpected contents" paradigm (Perner et al., 1987; also used in Holmes et al., 1996; Cole et al., 2000).[6] We showed the child a bag that purported to contain candies and asked the child what he or she thought was inside the bag. The experimenter then opened the bag and showed the child that, in fact, there were crayons inside. The experimenter then asked the child a control question about what the child thought

[5] One might wonder why we thought that young children would be sophisticated enough to make sense of the Contest treatment. In fact, we observed that many university students inflated reported confidence in a Contest setting in a more complex version in an earlier study (Charness, Rustichini, and van de Ven, 2018) and we were curious concerning whether this effect would be present with children in a simplified setting. However, we found no evidence of this strategic behavior, as discussed briefly later.

[6] We use this particular task because it has been widely used in the literature, thus facilitating a comparison with other studies. There are many other tasks that have been used to measure ToM. For example, a perspective-taking task (Flavell, Botkin, Fry, Wright, and Jarvis, 1968) considers whether two people see different things if there are differences in their physical positions or physical barriers to prevent them from seeing what the other sees. In an emotion-recognition task (Denham, 1986), people are shown four faces depicting happy, sad, angry, and scared emotions; the experimenter points to each face in turn and asked, 'How does s/he feel?' matching the sex of the face to the child's sex. A desire task (Wellman and Wooley, 1990) assesses one's understanding of the effect of fulfilled or unfulfilled desires on another character's feelings.

was actually in the bag.[7] The experimenter put the crayons back in the bag and closed the bag. Then the experimenter pulled out a teddy bear and told the child that Teddy had been sleeping and did not see what was in the bag. The experimenter then asked the child what Teddy would guess was in the bag.[8] Children who answered that Teddy would guess "candies" are considered to have ToM, while those who answered "crayons" are considered to not have ToM. Children who did not know what Teddy would guess, or who expected Teddy to guess something else than candies or crayons are also considered to not have ToM. Instances where the experimenter recorded that the child did not understand the task, or where the child originally guessed something else than candies (or variations thereof), are recorded as missing values. Because some children had already seen this task, we sometimes used a slight variation of the task in which a paper doll (Sally) and a bag of crackers (goldfish) were used instead of the Teddy Bear and candies. Specifically, the first three studies typically used Teddy while the last (experimental) study used Sally.

Background characteristics. We collected several socio-economic indicators and administered standardized tests (vocabulary, math, memory, and self-regulation/inhibitory control). The socio-economic indicators include (self-reported) household income, and the parent's education level and employment status. For some children we do not have all background characteristics, reducing the number of observations for this part of the analysis. A detailed description of the variables and sample sizes can be found in Appendix B. As explained earlier, children in the Chicago Heights area predominantly come from a low-income background. The mean income in our sample is $29,456 (median $20,500). About 30 percent of families in our sample have an

[7] Only very few children answered the control question incorrectly (less than two percent). Most studies in the meta-study by Wellman et al. (2001) either do not ask the control question or do not exclude children that fail the question (Sobel and Austerweil, 2016). For comparability, we therefore keep them in our sample. Our results remain almost identical if we exclude them from the sample.

[8] While the use of a doll or toy may seem awkward, Wellman et al. (2001) have shown that children answer similarly when asked about real persons, dolls, or toys, at least for false-belief tasks.

income above $35,000. Twenty-nine (16) percent of mothers (fathers) has a college degree, and 35 (68) percent of mothers (fathers) had a full-time job at the time of testing.

3 Gaps in Theory of Mind

3.1 Theory of Mind by Age Group

Figure 1A shows the fraction of children that displays ToM across different age groups (indicated by the solid squares). The outcome variable (fraction of children with ToM) is transformed by taking the logarithm of the odds-ratio.[9] As expected, there is a clear upward trend; older children are more likely to display ToM. The percentage of children displaying ToM increases from 10-15 percent for the youngest children to about 70 percent for the oldest children in our sample.

The children in our sample are primarily from disadvantaged families, and one consideration is whether there is indeed a gap with children from other socio-economic backgrounds. We do not have a control group, in that the subjects available through CHECC are primarily from low-SES backgrounds; thus, there is little variability in the SES related variables. Instead, we compare our results to those from a large set of previous studies using more standard subject pools including children from middle- and upper SES families.

Figure 1A plots a fitted curve based on the results from a meta-study (Wellman et al, 2001) using a population of "normally developing children".[10] Studies included in the meta-analysis also use the false-belief paradigm, as is the task that we use. The curve is well above the ToM

[9] We do this for comparison purposes with other studies, notably the meta-study by Wellman et al. (2001). Figure E.1 in the Appendix shows the untransformed means.

[10] It appears that the meta-study of Wellman et al. (2001) mostly (though not exclusively) uses children from middle and upper SES. We were not able to retrieve exactly how many studies in the meta-study included children from lower socio-economic backgrounds. However, according to Cutting & Dunn (1999, p. 855), the majority of published studies until around that time investigated ToM among children from middle and upper-middle class families. Moreover, in their meta-study, Devine and Hughes (2016) report only nine studies predating the Wellman et al. meta-study that systematically reported SES and used a false-belief task.

rate in our sample for all age groups, and ToM rates in our sample are always outside the 95 percent confidence interval of the meta-study. To get a sense of the magnitude of the differences consider that on average, other studies have found that half of the children have ToM at around 45 months, and at 70 months over 90 percent of children have ToM (Wellman et al, 2001).[11] By contrast, in our study, less than half of the children have ToM even at 78 months, and the ToM rate never gets close to 100 percent, even for the highest age range. Figure 1B shows the individual studies reported in Wellman et al. (2001), including only those that use the same task that we used (unexpected-contents task). These results are the more natural ones for comparison; however, we cannot compute confidence intervals since we could not obtain the original data from those studies. We again see that virtually all the other studies using this task have found higher ToM rates than in our sample of disadvantaged children.[12]

Of course, there could be other differences between our study and those included in the meta-study, and we cannot say with certainty that SES is the only responsible factor. Nevertheless, our interpretation that SES is driving the difference is certainly consistent with the evidence reported by Devine and Hughes (2018). In their systematic review of the literature (these studies use a wide range of tasks to identify ToM), they identified 50 effect sizes for the correlation between ToM and SES. They find a strongly statistically-significant ($p < 0.001$) overall positive relationship in a meta-analysis. Their Figure 1 illustrates nicely that 44 studies report a positive relationship and only six reported a negative relationship (a binomial test gives $Z = 7.60$, $p < 0.001$). Of the 44 positive relationships, 22 were statistically-significant at the 5% level; none of the six negative relationships were close to statistical significance. Building on

[11] Many studies administer several false-belief tasks per child. This is not driving the higher ToM rates, however. Children's performance is very consistent across the first and subsequent trials (see Wellman et al., 2001, p. 675).
[12] A major exception includes studies measuring ToM among autistic children. Those children have much lower passing rates (e.g., Baron-Cohen et al., 1985; Girli and Tekin, 2010). For instance, in Baron-Cohen et al. (1985), 80 percent of normal children with a mean age of 53 months passed the test, which is a bit above the mean of the meta-study. By contrast, only 20 percent of autistic children with a mean age of 143 months passed the test.

their findings, we demonstrate that an appropriate treatment can at least in part overcome the effect of SES; thus ToM is a skill that can be enhanced by environment and past experiences.

There is little difference between the patterns in the two panels of Figure 1, suggesting that the results are not overly sensitive to the specific nature of the task and that the results with children who are not from disadvantaged backgrounds are largely robust. Related work has also observed that while disadvantaged children have the same basic progress of false belief mastery as their higher income counterparts, the average rate of development is slower (Holmes et al., 1996; Cole and Mitchell, 2000).

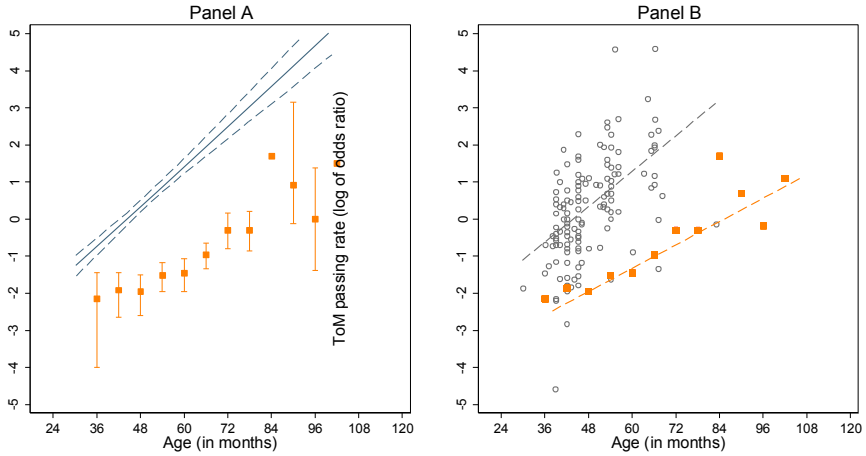

Figure 1: Proportion of children passing the ToM test. Mean passing rates of ToM by age category (grouped in bins of 6 months). Rates are transformed as $y = \ln(x/(1-x))$, where x is the fraction of children with ToM. All children older than 100 months are grouped into a single bin (102). Only bins with at least 5 observations are included (972 observations left, 18 observations dropped). Panel A (left): Solid squares are from our sample. Error bars indicate the 95 percent CI based on bootstrapped standard errors of the untransformed means. The solid blue line is the fitted line from the Wellman et al., (2001) meta-study. The dashed lines indicate the 95 percent CI. Panel B (right): Solid squares are from our sample. Open circles are data points from individual studies in Wellman et al. (2001) using the unexpected-contents task. The dashed lines are linear fits.

Table 1 shows the estimation results of a logistic regression of age on ToM rates. The reported coefficients are in terms of odds ratios. Column (1) shows that the coefficient of age is significantly above 1, indicating that ToM increases with age. In terms of marginal effects, an increase in age by one year increases the rate of ToM by 9.5 percentage points. Column (2) adds

the square of age as an explanatory variable. We do not detect a nonlinear relationship between age and the rate of ToM: the age coefficient is still significant but the age-squared coefficient is insignificant and close to 1. As will become evident, the age coefficient is very stable across many different specifications controlling for different background characteristics.

Table 1: ToM and age.

Sample:	(1) All	(2) All	(3) All	(4) Single test	(5) Single test, Non-treated††
Age (in months)	1.052***	1.077**	1.045***	1.041***	1.029***
	(0.007)	(0.040)	(0.006)	(0.007)	(0.010)
Age squared		1.000			
		(0.000)			
Took ToM test before†			2.202***		
			(0.419)		
Experimental Sample (Baseline treatment)					1.391
					(0.650)
Constant	0.012***	0.006***	0.016***	0.021***	0.039***
	(0.005)	(0.008)	(0.006)	(0.010)	(0.024)
Observations	990	990	990	685	628

Notes: Logistic regressions reporting odds ratios. Sample: columns (1)-(3): All children; (4) children that took the ToM test only once. †† Sample in column (5) are children in the non-experimental sample and children in the Baseline treatment of the experimental sample (including only children that took the ToM test only once). Robust standard errors (clustered at the child level) in parentheses. †Dummy variable (1 if at the time of the test the child had taken the test previously). ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

Children who already took the test at some previous date show a marked increase in ToM; the odds ratio is more than twice as high compared to children who took the test for the first time, controlling for age (column 3). These children may have learnt from previous experience or simply recalled the correct answers from last time. A selection effect may have contributed to produce this difference: these children are from parents who brought them multiple times to the CHECC program. In column (4) we restrict the sample to children that took the ToM test only once. The age coefficient is very similar to that in column (3).

3.2 Background characteristics

Table 2 includes several background characteristics as covariates. Column (1) includes gender and ethnicity as variables. There is no significant gender effect: males and females perform about equally well on the ToM task. Taken over the entire sample, the difference in ToM rates is three percentage points (25.5 for males versus 22.6 for females, *p*-value 0.287, two-sided *t*-test). Figure E.2 in the Appendix shows the gender difference and 95 percent confidence interval for each age bin. With one exception, there is no significant gender difference in any of the age bins.

Being non-white is associated with a substantially lower rate of ToM, reducing the odds-ratio by one half. In terms of marginal effects, the rate of ToM is 12 percentage points lower for non-whites than for whites. Distinguishing between different ethnicities, we observe that the effect is comparable for Blacks and Hispanics (see also Table E.1). For other ethnicities the difference is even larger, but not significantly different from 1, which can be attributed to the low number of observations we have for that category (only seven children fall into that category). The difference cannot be primarily attributed to language comprehension; among non-whites, children from families with English as the home language show a ToM rate less than four percentage points higher than those from families where the home language is not English (19.3 versus 23.1, *p*-value 0.200, two-sided *t*-test).[13]

Column (2) of Table 2 shows the association with several test scores. All test scores are standardized to have mean zero and a standard deviation 1. While the math and vocabulary scores have the predicted sign, working memory is associated with a slightly lower ToM. However, none of the test scores is significantly correlated with ToM. The same holds for the measure of self-regulation, which includes measures of impulse control and attention. Note that

[13] Adding to column (1) of Table 2 a control that specifies whether the home language is English or not does not meaningfully affect the estimated coefficients.

the estimated coefficients of test scores will include an age effect, since test scores will increase with age.[14]

Estimates of socio-economic indicators are reported in column (3) of Table 2. Income is positively but not significantly associated with ToM. Of course, one must keep in mind that our sample consists mostly of low-income families; 70 percent of families in our sample fall into the household income bracket of $35,000 or below (see also Figure C.2 for the distribution). Children are more likely to display ToM when their mother has a college degree, but less likely when their father has a college degree. The reverse is true for the employment status of the parents. However, none of the coefficients is significant and the effects are modest in terms of marginal effects (always less than six percentage points).

The final column in Table 2 includes all covariates into a single regression. Most estimated coefficients remain very similar. The coefficient of female is now marginally significant, but the effect size is modest (the marginal effect is six percentage points). The ethnicity coefficient becomes only marginally significant, but the coefficient remains stable.

The lack of strong correlations between the different background characteristics and ToM within our sample of low-SES subjects may seem surprising but is not inconsistent with some existing studies. While the Devine and Hughes (2018) meta-study found a significant correlation between socioeconomic status and ToM, note that several of the studies included (e.g., Garner et al., 2005) even find a slight negative correlation.

Our result of no correlation between ToM performance and test scores within our sample of low-SES subjects is not surprising if one adopts the Simulation Theory of ToM (Goldman, 1989; Gordon, 1986). This theory posits that people use their own experiences to simulate the

[14] The test scores were not always administered at the same time as the ToM task. In Appendix E, we report the results when we restrict the sample to cases where the two tests were taken within 90 days of each other. The results remain similar. The coefficient of the math score is a bit higher, but none of the coefficients is significant. We also included the age at the time of cognitive assessments as a control (columns 2 and 3 in Table E.2) and this also does not change our results.

mental states of others, by imagining how they would feel themselves. In this approach, people do not go through a reasoning or cognitive process in which they form theories about other people's mental states by observations and hypothesis testing. The low correlation is also in line with some other studies: Carlson et al. (2004) find a significant correlation between a child's vocabulary (measured with the Pearson Peabody Vocabulary Test) and a composite measure of ToM, but the correlation is not significant if they look at false belief measures of ToM, as in our study. Mutter et al. (2006) find that working memory only makes a small positive contribution to performance on a ToM task.

3.3 Out-of-sample predictions

The overall mean ToM rate in our sample is 24 percent, whereas based on the Wellman et al. (2001) meta-study we should expect to see 76 percent of children displaying ToM given the age distribution in our sample. Although we find few significant covariates within our sample, they may jointly explain a substantial part of the gap in ToM rates between the different studies. Based on the most parsimonious model of Table 2, we made out-of-sample predictions for children of different ages, imputing values for the different background characteristics.

In particular, we compute the predicted ToM rate for a white child, with test scores that are half of a standard deviation higher than the children in our sample, and whose parents' background characteristics reflect the median for the US population (see Appendix E for details). Under these assumptions, the predicted ToM rate is still only 35 percent, explaining only 21 percent (11/52) of the gap. We suspect that there are strong nonlinearities in the relation between ToM and the background characteristics, or there are other important factors correlated with ToM for which we are not accounting.

Table 2: ToM and background characteristics

	(1)	(2)	(3)	(4)
Child characteristics				
Female	0.826			0.678*
	(0.139)			(0.155)
Non-white	0.499***			0.514*
	(0.134)			(0.176)
Test scores				
Math		1.064		1.116
		(0.184)		(0.243)
Vocabulary		1.015		0.871
		(0.147)		(0.176)
Memory		0.862		0.855
		(0.103)		(0.138)
Self-regulation		0.911		1.142
		(0.119)		(0.211)
Socio-economic status indicators				
Household income (logs)			1.222	1.202
			(0.205)	(0.216)
Father has college degree			0.727	0.643
			(0.228)	(0.223)
Mother has college degree			1.108	1.107
			(0.298)	(0.314)
Father is full-time employed			1.256	1.240
			(0.328)	(0.325)
Mother is full-time employed			0.788	0.814
			(0.210)	(0.221)
Controls	Yes	Yes	Yes	Yes
Observations	942	910	525	515

Notes: Logistic regressions reporting odds ratios. Controls: Age, Took test before, Experimental sample. Test scores are standardized (mean 0, standard deviation 1). Robust standard errors (clustered at the child level) in parentheses. ***p < 0.01, **p < 0.05, *p < 0.1.

4 Social Saliency Improves Theory of Mind

A unique design feature of our study, different from related work exploring ToM, is that prior to completing the ToM task children are randomly exposed to different treatments which differ in how salient is the presence of another child. A reasonable conjecture is that children are more stimulated to think about the paired child in the *Social* treatment than in the *Baseline* treatment, and even more so in the *Contest* treatment. In the *Baseline* treatment, the paired child

has no role to play. In the *Social* treatment, the paired child will learn the child's guess. In the *Contest* treatment, the paired child will also learn the child's guess, and might condition his or her decision to enter the contest based on that information.

Table 3: Theory of Mind by treatment and age range

Sample	Baseline	Social	Contest
All (N = 159)	0.411 (0.066)	0.500 (0.069)	0.627 (0.068)
Pre-K/Kindergarten (N = 88)	0.323 (0.085)	0.313 (0.083)	0.480 (0.102)
Grade 1+ (N = 71)	0.520 (0.102)	0.800 (0.092)	0.769 (0.084)

Note: Standard errors are in parentheses.

Table 3 shows the ToM rate for each treatment. The rate of ToM is indeed highest in the *Contest* treatment (63 percent) and lowest in the *Baseline* treatment (41 percent). This difference is highly significant (two-sided test of proportions, $p = 0.025$). Applying the correction for multiple-hypothesis testing (multiple treatments) as developed in List et al. (2015), we obtain a *p*-value of 0.048. The rate of ToM in the *Social* treatment is somewhere in between those treatments (50 percent) and not significantly different from either one of the other treatments (compared to *Baseline*, $p = 0.352$; compared to *Contest*, $p = 0.192$). The increase in ToM between *Contest* and *Baseline* (22 percentage points) has roughly a similar magnitude as 28 extra months of age.

One might expect the ToM rate in the *Baseline* treatment to be comparable to that of children in our non-experimental studies. Because children in our non-experimental sample are on average slightly younger, we need to correct for the age difference. We used the estimated coefficients of specification (1) in Table 1 to predict the ToM rate for children with the same mean age as children in the experimental sample. We indeed find that the actual and predicted ToM rates are close (0.41 versus 0.36) and not significantly different ($p = 0.442$, two-sided *t*-test). Column (5) of Table 1 includes a dummy for children that were in *Baseline*, excluding

children who took the ToM test before. We reach the same conclusion: children in the experimental sample perform better on ToM, than children in the non-experimental sample, but the difference is not significant (the marginal effect is 8.5 percentage points).

The treatment effects on ToM rates seem to differ across age groups. We classify children into younger or older depending on their eligibility to be in Pre-K/Kindergarten (less than six years old on September 1 of the test year) or first grade and up (at least six years old on September 1 of the test year)[15]. In both age groups, the ToM rate is higher in *Contest* compared to *Baseline* (Table 3). The difference is nearly significant for older children ($p = 0.063$, two-tailed test) but not so for younger children ($p = 0.231$). However, while there is no difference for the younger group across *Baseline* and *Social* ($p = 0.932$), there is a substantial one for the older group ($p = 0.051$). Note that the effects for older children are not robust to multiple hypothesis testing (p-values above 0.10) and therefore should be taken as suggestive.

It is interesting that the *Social* condition seems sufficient to trigger an increase in ToM rates for older children, but that only the more challenging *Contest* condition triggers an increase for younger children. These results are robust to adding gender and performance on the memory task as controls. Table 4 reports the results from a linear probability model.[16]

Consistent with the earlier results, when we pool all children then only the coefficient of *Contest* is significant (column 1). The treatment coefficients are not significant for younger children (column 3). For the older children, the odds ratios of the treatment coefficients are large and significant (column 5). The marginal effects are 28 and 34 percentage points for the *Contest* and *Social* treatments, respectively. The estimates are robust to adding other controls (columns 2,

[15] This age split is roughly a median split. Our results are robust to alternative age categorizations.
[16] The linear model is used to make the interpretation easier. Alternative specifications yield similar results: Table E.3 in Appendix E reports the results from logistic-regression models, with the reported coefficients reflecting odds ratios.

4, and 6), except that the coefficient of *Contest* becomes insignificant when we pool all data (the coefficient remains similar in size).

Table 4. Influences on Theory of Mind

	(1)	(2)	(1)	(2)	(3)	(4)
	All		Pre-K/Kindergarten		Grade 1+	
Social	0.093	0.037	0.009	-0.088	0.275**	0.337**
	(0.096)	(0.134)	(0.123)	(0.174)	(0.125)	(0.134)
Contest	0.238**	0.208	0.147	0.200	0.339**	0.427**
	(0.093)	(0.127)	(0.137)	(0.179)	(0.129)	(0.165)
Performance memory game	0.026**	0.028	-0.018	-0.010	0.044**	0.062**
	(0.013)	(0.020)	(0.021)	(0.030)	(0.020)	(0.027)
Female	0.042	0.094	0.046	0.083	0.146	0.198
	(0.078)	(0.105)	(0.101)	(0.133)	(0.099)	(0.126)
Took ToM test before	0.171**	0.173	0.136	0.074	0.265***	0.123
	(0.078)	(0.111)	(0.105)	(0.129)	(0.098)	(0.156)
Constant	0.172	-0.256	0.292**	-0.094	0.053	0.675
	(0.110)	(0.575)	(0.138)	(0.149)	(0.193)	(0.907)
Controls:	No	Yes	No	Yes	No	Yes
Observations	157	94	87	60	70	34

Notes: Linear probability model. We have 2 missing observations for performance level. Controls are: test scores (math, vocabulary, memory, self-regulation), household income (log), ethnicity (dummy for non-white). Robust standard errors (clustered at the child level) in parentheses. *** p < 0.01, ** p < 0.05, * p < 0.1.

We interpret treatment differences as induced by the salience of the other child. It is worthwhile to point out that there are several other differences between the treatments. First, the contest treatment is arguably more cognitively demanding than the others. If so, we believe that this would work against finding higher ToM passing rates in *Contest*, as those children are more fatigued. Second, the incentives are different between the treatments, with higher possible payoffs for children in the contest treatment. We do not find any effect of payoffs earned in this stage on ToM performance. The correlation between the number of prizes won and ToM performance is low and insignificant ($\rho = 0.108$, $p = 0.465$, Pearson's correlation coefficient). Neither do we find that winning the competition effects ToM; children who won the contest score are indistinguishable from those who lost, with an average difference of 0.014 on the ToM task ($p = 0.919$, t-test). The fact that for the older children we already see an increase in ToM in *Social*, and no difference between *Social* and *Contest*, provides some additional evidence that

different incentives or outcomes of the contest are not driving the results. Third, since children in the contest treatment had incentives to overall report, they may differ in their average reported guess. This turns out not to be the case. Average guesses are indistinguishable across treatments (7.9 in *Baseline*, 8.0 in *Social* and *Contest*, all *p*-values from pairwise *t*-tests are above 0.900).

4.1 Discussion

What does it mean that children are more likely to pass the ToM test when the presence of another child is more salient? The actual ability to understand the mental states of other people might not be affected in such a substantial way by this simple intervention. As mentioned, children lacking theory of mind should not suddenly gain this capacity during the course of the experiment. Rather, our results suggest that this capacity to see things from someone else's perspective may be latent for some children in our sample. Our intervention is a stimulus to activate theory of mind.

This interpretation is in line with current views of theory of mind as a complex set of competencies. There is not one single consensus measure of theory of mind. The false-belief task measures an explicit understanding of theory of mind, but studies have shown that younger children already have some implicit understanding of mental states of others (Airenti, 2015; Clements and Perner, 1994). The fact that very young children (under the age of four) have difficulties with the false-belief test might just reflect that other necessary skills to pass the test (such as working memory and language skills) are not sufficiently developed at that age (Baillargeon et al., 2010; Bloom and German, 2000).

Theory of mind should be more properly seen as a skill that has many dimensions and levels of understanding. The work by Wellman and Liu (2004) suggests that children first acquire an understanding that other people can have different tastes before they can understand that others can have false beliefs. Only at a later stage do they understand that emotions expressed by others do not necessarily reflect the actual emotions of these others. This does not imply that any single task does not capture important aspects of this process. To the contrary, a failure on any theory of mind measure is likely to be diagnostic of a child's processes regarding "mind-reading" at that stage, and should be a concern. From this perspective, the *level* of performance on the false-belief task is of secondary importance. It is the fact that children in our sample score lower than

the average score of a more diverse sample - including children from low, middle and high SES (as in the meta-analysis of Wellman, 2001) – that is worrying. The Thompson (2017) paper conforms to our findings and supports our view; these lags appear to begin at home.

5 Conclusion

In existing studies, most children display ToM by the age of five to six years. In our sample of children from disadvantaged backgrounds, it seems that ToM develops at a much slower rate. ToM is an important skill, so the gap with children from more advantaged backgrounds is a major concern that may be an important contributor to the large income inequality later in life between people from different socio-economic backgrounds. An intriguing question is whether children from disadvantaged backgrounds are also slower to acquire Strategic Theory of Mind (SToM). This refers to a capacity to reason recursively and predict the behavior of other people, on top of ordinary ToM (Sher et al, 2014). SToM is particularly relevant in situations in which children strategically interact with others. Measuring SToM requires a different set of tasks (Sher et al., 2014).

A key challenge is to narrow the skills gap, for which a deeper understanding of the determinants of ToM should be useful. The reader can easily consider that various factors (some more interesting/novel than others, at least from a psychological point of view) could be responsible for the effects of this manipulation. While researchers have been very interested in the causal factors that shape the development of ToM, much of the work to date has nevertheless been correlational. Hughes et al. (2005) report that environmental factors play a major role in ToM variation, though the mechanisms are less clear. For instance, children with siblings have been shown to have higher ToM scores than their single-child counterparts (McAlister and Peterson, 2007), yet the correlation is not observed among disadvantaged children, possibly because of the potential contaminating issue of socio-economic status (Cole and Mitchell, 2000). Similarly, researchers have investigated the role of pretend play on the development of ToM, but found inconsistent correlational results, potentially due to confounding factors of child, parent and environment characteristics (Lillard et al., 2013).

To the best of our knowledge, our study is the first to use a treatment assignment to infer a causal effect of thinking about others on ToM, which was measured immediately following the

intervention. Our finding is that the saliency of the presence of another child affects ToM rates. This suggests that *having* ToM is not the same as *showing* ToM. Many children that appear not to have ToM may in fact have this ability, but this ability needs to be *activated*. In our study, subtle cues about the presence of another child were sufficient to activate ToM in a substantial percentage of largely low-SES children. It would be most interesting to see how these treatments play out in a sample of high-SES children.

The treatment effect on ToM rates is equivalent to 28 extra months of age. While this is a result that should be replicated, it may well be that richer environments, whether social or even strategic (e.g., poker), can help to foster such skills. Since the absence of ToM can result in costly mistakes in many types of social interactions, activating ToM with the help of simple social cues might be very beneficial to those children. In the longer run, the activation of ToM might become automatic, and might help close skills gaps between children. This seems an exciting area for new research.

References

Airenti, G. (2015). Theory of mind: a new perspective on the puzzle of belief ascription. *Frontiers in Psychology*, *6*, 1184.

Andreoni, J. Di Girolamo, A., List, J.A., Mackevicius, C. & Samek, A. (2019). *Risk Preferences of Children and Adolescents in Relation to Gender, Cognitive Skills, Soft Skills and Executive Functions.* Working paper.

Andreoni, J., Kuhn, M., List, J.A., Samek, A., Sokal, K. and Sprenger, C. (2019). *Toward an understanding of the development of time preferences: Evidence from field experiments.* Working paper.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in cognitive sciences*, 14(3), 110-118.

Baron-Cohen S., Leslie A., Frith U. (1985). Does the autistic child have a "theory of mind"?. *Cognition* **21** (1): 37–46.

Ben-Ner, A., List, J. A., Putterman, L., & Samek, A. (2017). Learned generosity? An artefactual field experiment with parents and their children. *Journal of Economic Behavior & Organization*, *143*, 28-44.

Blair, C.B., & Willoughby, M.T. (2006). Measuring Executive Function in Young Children: Operation Span. and Spatial Conflict II: Arrows. Chapel Hill, NC: The Pennsylvania State University and The University of North Carolina at Chapel Hill.

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1): B25-B31.

Brocas, I. and Carrillo, J. (2018). The Determinants of Strategic Thinking in Preschool Children. *PLOS One*, May 31.

Brocas, I. and Carrillo, J. (forthcoming). Iterative Dominance in Young Children: Experimental Evidence in Simple Two-Person Games. *Journal of Economic Behavior and Organization*.

Cappelen, A. W., List, J. A., Samek, A., & Tungodden, B. (2016). *The effect of early education on social preferences*(No. w22898). National Bureau of Economic Research.

Castillo, M., List, J.A., Petrie, R. & Samek, A. (2019). Dectecting drivers of behavior at an early age: Evidence from a longitudinal field experiment. Working paper.

Carlson, S. M., Moses, L. J., & Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *Journal of Experimental Child Psychology*, *87*(4), 299–319.

Charness, G., Rustichini, A., & van de Ven, J. (2018). Self-confidence and strategic behavior. *Experimental Economics*, *21*(1), 72-98.

Cicchetti, D., Rogosch, F. A., Maughan, A., Toth, S. L., & Bruce, J. (2003). False belief understanding in maltreated children. *Development and Psychopathology*, 15, 1067 - 1091.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, *9*(4), 377–395.

Cole, Kristina, and Peter Mitchel (2000). Siblings in the development of executive control and a theory of mind. *British Journal of Developmental Psychology* 18.2: 279-295.

Cowell, J. M., Samek, A., List, J., & Decety, J. (2015). The curious relation between theory of mind and sharing in preschool age children. *PLoS One*, *10*(2), e0117947.

Cutting, Alexandra L., and Judy Dunn (1999). Theory of mind, emotion understanding, language, and family background: Individual differences and interrelations. *Child development*, 70, 853-865.

Denham, S.A. (1986), Social cognition, prosocial behaviors, and emotion in preschoolers: Contextual validation. *Child Development*, 57, 194–201.

Devine, R., & Hughes, C. (2018). Family Correlates of False Belief Understanding in Early Childhood: A Meta-Analysis. *Child Development*, 89, 971-987.

Dunn, M. and L. M. (1997). PPVT-III: Peabody picture vocabulary test. Circle Pines, MN: American Guidance Service.

Flavell, J. H., Botkin, P. J., Fry, C. L., Wright, J. W., and Jarvis, P. E. (1968). *The development of role-taking and communication skills in children*. New York: Wiley & Sons.

Fryer, R, Levitt, S.D., List, J.A. (2015). Parental Incentives and Early Childhood Achievement: A Field Experiment in Chicago Heights, NBER working paper 21477.

Fryer, R, Levitt, S.D., List, J.A., Sadoff, S. (2018). Enhancing the Efficacy of Teacher Incentives through Framing: A Field Experiment, mimeo.

Garner, P. W., Curenton, S. M., & Taylor, K. (2005). Predictors of mental state understanding in preschoolers of varying socioeconomic backgrounds. *International Journal of Behavioral Development*, *29*(4), 271-281.

Girli, A., & Tekin, D. (2010). Investigating false belief levels of typically developed children and children with autism. *Procedia-Social and Behavioral Sciences*.

Gneezy, U, J.A. List, J.A. Livingston, X. Qin, S. Sadoff, Y. Xu, (2017). Measuring Success in Education: The Role of Effort on the Test Itself, working paper.

Goldman, A. (1989). In Defense of the Simulation Theory. *Mind & Language*, *7* 104–119.

Gordon, R. (1986). Folk psychology as simulation. Mind & Language 1.2: 158-171.

Hazler, R. J. (1996). Breaking the cycle of violence: interventions for bullying and victimization. Washington DC: Accelerated Development.

Heckman, J. (2000). Policies to foster human capital. *Research in Economics*, 54(1), 3-56.

Heckman, J. J., Stixrud, J. and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, 24(3), pp. 411-82.

Heckman, James J. (2006) Skill formation and the economics of investing in disadvantaged children. *Science* 312.5782: 1900-1902.

Holmes, Heather A., Cherice Black, and Scott A. Miller (1996). A cross-task comparison of false belief understanding in a Head Start population. *Journal of Experimental Child Psychology* 63.2: 263-285.

Hughes, C., & Ensor, R. (2005). Executive function and theory of mind in 2 year olds: A family affair?. *Developmental neuropsychology*, *28*(2), 645-668.

Hughes, C., & Ensor, R. (2007). Executive function and theory of mind: Predictive relations from ages 2 to 4. *Developmental psychology*, *43*(6), 1447.

Lillard, Angeline S., Matthew D. Lerner, Emily J. Hopkins, Rebecca A. Dore, Eric D. Smith, and Carolyn M. Palmquist (2013). The impact of pretend play on children's development: A review of the evidence. *Psychological Bulletin* 139, no. 1): 1-34.

List, J. A., & Samak, A. C. (2013). Exploring the origins of charitable acts: Evidence from an artefactual field experiment with young children. *Economics Letters*, *118*(3), 431-434.

List, A. J., List, J. A., & Samek, A. (2017). Discrimination among pre-school children: Field experimental evidence. *Economics Letters*, *157*, 159-162.

List, John A., Azeem M. Shaikh, and Yang Xu (2015). *Multiple hypothesis testing in experimental economics*. No. w21875. National Bureau of Economic Research.

McAlister, Anna, and Candida Peterson (2007). A longitudinal study of child siblings and theory of mind development. *Cognitive Development* 22.2: 258-270.

Mutter, B., Alcorn, M. B., & Welsh, M. (2006). Theory of mind and executive function: working-memory capacity and inhibitory control as predictors of false-belief task performance. *Perceptual and Motor Skills*, *102*(3), 819–835.

Pears, K. and Moses, L. (2003), "Demographics, Parenting, and Theory of Mind In Preschool Children," *Social Development*, 12, 1-20.

Perner, J. (1979). Young children's preoccupation with their own payoffs in strategic analysis of 2 × 2 games. *Developmental Psychology, 15*(2), 204-213.

Perner, J. et al. (1987) Three-year-olds' difficulty with false belief: the case for a conceptual deficit. Br. J. Dev. Psychol. 5, 125 – 137

Premack, D. G., Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and Brain Sciences* **1** (4): 515–526.

Randall, P. (1997). Adult bullying: Perpetrators and victims. London: Routledge.

Samak, A. C. (2013). Is there a gender gap in preschoolers' competitiveness? An experiment in the US. *Journal of Economic Behavior & Organization*, *92*, 22-31.

Sher, Itai, Melissa Koenig, and Aldo Rustichini (2014). Children's strategic theory of mind. *Proceedings of the National Academy of Sciences* 111.37: 13307-13312.

Smith-Donald, Radiah, C. Cybele Raver, and Tiffany Hayes (2007). Preliminary construct and concurrent validity of the Preschool Self-Regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly* 22.2: 173-187.

Sobel, D. M., & Austerweil, J. L. (2016). Coding choices affect the analyses of a false belief measure. *Cognitive Development*, *40*, 9-23.

Shultz, T. R., & Cloghesy, K. (1981). Development of recursive awareness of intention. *Developmental Psychology, 17*(4), 465-471.

Thompson, O. (2017). Determinants of Racial Differences in Parenting Practices. *Journal of Political Economy* (2017), forthcoming.

Wellman, Henry M., David Cross, and Julianne Watson (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*: 655-684.

Wellman, Henry M. and David Liu, (2004). Scaling of Theory-of-Mind Tasks. *Child Development Child Development*, *75*(2), 523–541.

Wellman, Henry M. and Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, 35, 245–275.

Woodcock, Richard W., Kevin S. McGrew, and Nancy Mather (2001). Woodcock-Johnson III. Itasca, IL: Riverside Publishing.

Yagmurlu, Bilge, Sibel Kazak Berument, and Seniz Celimli (2005). The role of institution and home contexts in theory of mind development. *Journal of Applied Developmental Psychology* 26.5: 521-537.

Appendix A: Detailed Description of Experimental Treatments

The experimental treatments follow the design of that in Charness et al. (2018) adapted for children.

Experimental design. In all treatments, children started with a memory game, similar to the task used in Lipko et al. (2008). The experimenter showed the child a sheet with 12 animals, each in its own house, and asked the child to remember which animal belonged in which house. After the child named all the animals, the experimenter put the sheet away. The child then received a sheet with blank houses and a set of 20 animal cards (the 12 original animals and eight new animals). Children had three minutes to put the animals back in their own houses. After they completed the memory task, children were asked to guess how many animals they correctly put back in their original house. They reported their confidence level by selecting one of 13 cards that were numbered 0 to 12. The corresponding number was also visualized with the help of stickers. Before making their guesses, children were asked to count the number of houses out loud, to make sure that they understood that the maximum guess was 12. They were also told that this game is hard, and that many children do not get all animals right. The reason for this was to avoid children all reporting the maximum possible score due to wishful thinking.

Children participated in the memory game twice. They first participated in what we call the practice round, in which their reported confidence was not incentivized. They then participated in the game round, in which their reported confidence was incentivized. In the game round, they received three prizes for a correct guess, two prizes for a guess that was at most two houses away from their actual performance, and one prize for a guess that was at most four houses away from their actual performance. Children received feedback about their performance in the practice round. The set of animals was the same, but their houses were different in the two rounds.

Treatments differed in whether or not a child's guess in the game round was kept private (*Baseline* treatment) or told to another child (*Social* and *Contest* treatments), and whether the paired child was inactive (*Baseline* and *Social*) or made a decision to enter into a contest or not (*Contest*). Table A.1 summarizes the treatments. Children were informed about whether or not their guess would be kept private before they reported their guess to the experimenter. Children were also told that their actual performance would never be shared with any other child. The

purpose of the *Social* treatment was to learn how the social component of confidence manifests itself in children. We also conducted a *Contest* treatment in which the child was told that his or her guess would be told to another child, after which the paired child (from the *Baseline* treatment) had to decide whether or not to enter into a contest with the child. We will refer to the child making that guess as the 'sender,' and the other child as the 'receiver.' After hearing the sender's guess, the receiver had to decide whether or not to enter into a contest with the sender. The payoffs were as follows. In case of a contest, the child with the best performance got 5 prizes, the other child nothing. When the paired child did not enter into a contest, the sender got 7 prizes and the receiver got 3 prizes. The purpose of this treatment was to test whether children would strategically report a high confidence to deter the other child from entering into a contest, thereby securing a high payoff.

Table A.1: Overview of Experimental Treatments

Treatment	Paired child observes guess?	Paired child decides whether to enter contest?
Baseline ($N = 56$)	No	No
Social ($N = 52$)	Yes	No
Contest ($N = 51$)	Yes	Yes

Procedures. The study was conducted at the Chicago Heights Early Childhood Center in May of 2014. Upon arrival, each child was paired with another child of about the same age, and they could see the other child as they walked towards the experiment room. This was done so that children knew that they were paired with another real child. Besides knowing the gender, children were not given any other information about the paired child. In the classroom, children were seated at different tables and out of each other's sight. Each child participated in the experiment one-on-one with a trained experimenter and performed the tasks individually (see Figure A.1). The study was conducted on a single day with the help of 14 trained experimenters, using a standard script. The experiment took approximately 20 minutes per child. The prizes used were stickers. At the end of the session, we administered a short Theory of Mind (ToM) test (see the main text for details).

Figure A.1: The experimental environment and illustration of the materials.

Appendix B: Data Collection and Variables

Recruitment. The Griffin Early Childhood Center (GECC) accepted children who were 3 years of age or older and younger than 5 years old on September 1 in each year of the program. Participation was not restricted to the school district (Illinois School District 170), and children from nearby school districts were eligible. Recruitment coordinators were hired to solicit applications from eligible families in the school district and in surrounding areas. Recruitment efforts included setting up and staffing informational booths at school open houses, at community events, outside of grocery stores, in churches, food pantries, and libraries. Parents who were interested in participating could sign up either immediately at the table, via mail, via the website, or by dropping off applications at a nearby elementary school within the school district. Recruitment efforts began during the fall of the previous school year and concluded in the summer immediately prior to the randomization. Parents wishing to enroll their child in a program completed a registration form, which included basic contact and demographic information. The registration form also included a consent form.

Household income. Our household income variable is based on self-reported survey data. The categories are: below $6,000; $6,000-$15,000; $16,000-$25,000; $26000-$35,000; $36,000-$45,000; $46,000-60,000; $61,000-$75,000; and above $75,000. The variable "Household Income" is constructed by taking the midpoint of a category. For the category "above 75,000" we set the income level to $82,000.

Employment status. Employment status is based on self-reported survey data. The possible categories are "disabled or unable to work", "full time student", "homemaker", "retired", "unemployed or temporarily laid off", "work full time", "work part time", and "other".

Education status. Highest achieved education level is based on self-reported survey data. The possible categories are "Less than high school", "Some high school but no diploma", "High school diploma", "some college", "college degree", "other".

Ethnicity. Ethnicity is based on self-reported survey data. The categories are "Black", "Hispanic", "White", and "Other".

Test scores

Children took standardized tests of abilities. Some children took the test multiple times. In that case, we use the score obtained at the date closest to the date on which the ToM task was administered. The median difference in test dates between the cognitive score and the ToM task was 90 days, and it was less than 180 days in 78 percent of the cases. Our results are robust to excluding cases where the difference was larger than 90 days. Bilingual children completed the assessment in either English or Spanish, depending on which was their dominant language, as determined by an assessor together with the parent or teacher at the time of assessment.

Children's verbal ability is measured using the Peabody Picture Vocabulary Test (Dunn and Dunn, 1997). PPVT-III is a leading measure of receptive vocabulary for standard English (Spanish) and a screening test of verbal ability. This is a norm-referenced standardized assessment that can be used with subjects ages 2-90+. The test is not timed, and takes approximately 5 to 20 minutes to complete.

Additional verbal and ability scores were measured with the Woodcock Johnson III Tests of Achievement (Woodcock et al, 2001). The WJ-III is a normed set of tests for measuring general intellectual ability, specific cognitive abilities, oral language, and academic achievement. This is a norm-referenced standardized assessment that can be used with subjects 2-80+. Of the 12 tests in the standard battery, 4 were selected that suit our needs. The test is not timed, and each sub-test takes approximately 5-10 minutes. We used the following sub-tests:

1. Letter Word Identification: Measures ability to identify letters and words.
2. Spelling: Measures ability to draw shapes and trace lines, and in older ages, write orally presented letters and words.
3. Applied Problems: Measures ability to analyze and solve math problems.
4. Quantitative Concepts: Measures knowledge of mathematical concepts and symbols.

We used two sub-tests from a newly developed battery of executive function tasks for use in early childhood (Blair and Willoughby, 2006). These include:

1. Operation Span (OSPAN): This task measures the construct of working memory, asking children to identify and remember pictures of animals.
2. Spatial Conflict II (SPAT): Arrows: This task measures the construct of inhibitory control, asking children to match 37 arrow cards in sequence.

At the end of the assessment, the assessor completed a Preschool Self-Regulation Assessment (PSRA) report, which is designed to assess self-regulation in emotional, attentional and behavioral domains (Smith-Donald et al., 2007).

Construction of Indices:

- Vocabulary: the standardized average of the scores from PPVT, WJ-III Letter-Word, WJ-III Spelling;
- Math: the standardized average scores from WJ-III Applied Problems, and WJ-Quantitative Concepts;
- Memory: the standardized score on OSPAN;
- Cognitive: the standardized average of vocabulary, math, and memory scores.
- Self-regulation/Non-cognitive score: the standardized average of SPAT and PSRA.

Appendix C: Descriptive Statistics

Table C1 provides some summary statistics. The rest of the Appendix provides more details about the background statistics.

Table C1: Summary Statistics

	Mean	Std. Dev.	Min.	Max.	N
Non-experimental Sample**:**					
Age (months)	58	10	39	101	668
Girl	0.509	0.500	0	1	668
ToM pass rates	0.195	0.396	0	1	668
Number of ToM tests	1.103	0.305	1	2	668
Household income	30,080	23,917	2,500	82,000	501
Experimental Sample**:**					
Age (months)	80	19	42	131	159
Girl	0.509	0.501	0	1	159
ToM pass rates	0.509	0.501	0	1	159
Number of ToM tests	1.591	0.713	1	3	159
Performance on memory task	5.019	2.971	0	12	157
Household income	26,255	20,300	2,500	82,000	100

Notes: Experimental Sample includes all children who participated in the experimental study. Non-experimental sample only includes children who did not participate in the experimental study. Number of ToM tests refers to the number of times that a ToM test was administered for a child. We use the number of correctly-recalled animals to measure performance on the memory task. Household income is self-reported data.

We have information about household income for 601 of the children in our sample (226 missing values). Figure C.2 shows the household income in our sample. For comparison, we also plotted the household income distribution for the US population based on 2014 census data.[17] It is clear that low-income families are overrepresented in our sample. The median income in our sample is $20,500, against $57,065 for the US. Less than 10 percent of families have an income above

[17] US Census Bureau. http://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-06.html

$75,000 in our sample, against 36 percent in the US. Figure C.3 shows the income distribution by ethnicity. While there is some variation in income for both white and non-white families, households with an income in the highest category (over $75,000) are mostly white families.

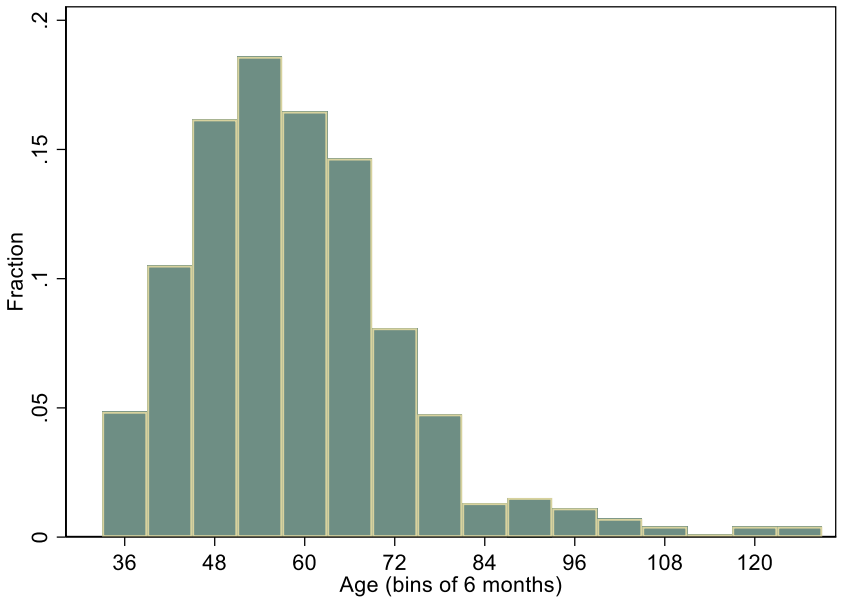

Figure C.1: Distribution of age (*N*=990).

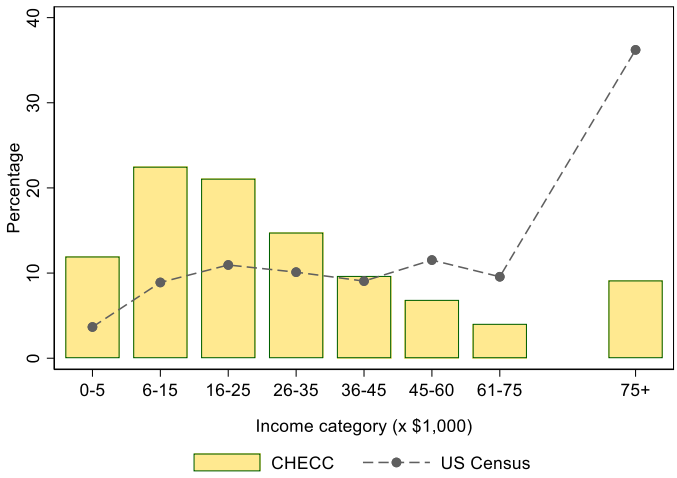

Figure C.2: Distribution of Household Income (percentages). CHECC is self-reported household income by families in our sample. US Census data are from the US Census Bureau (2014). N = 601 (missing observations for 226 children).

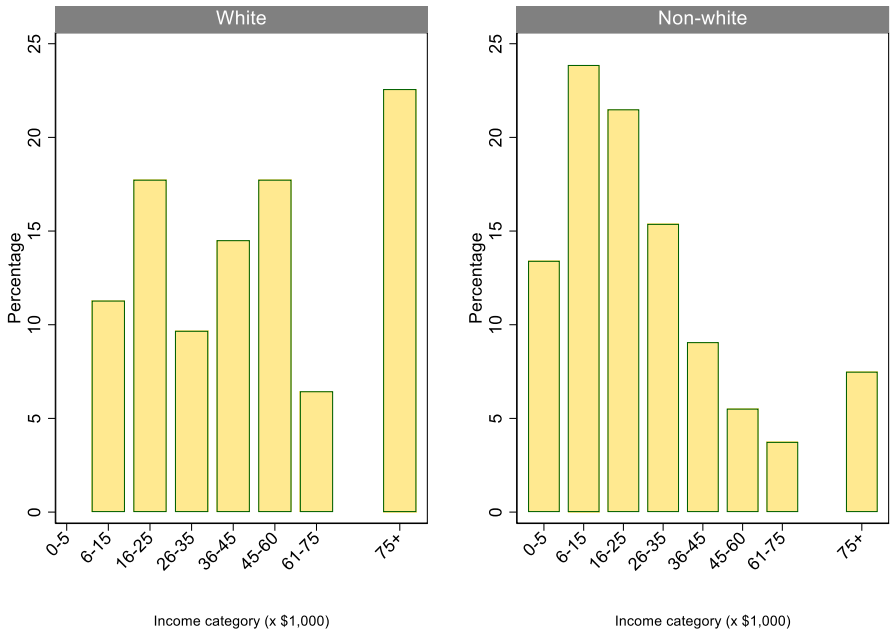

Figure C.3: Income Distribution by Ethnicity.

We have information about ethnicity for 778 of the children in our sample (49 missing values). Figure C.4 shows the distribution of ethnicity. The vast majority (89 percent) in our sample is Black or Hispanic. 10 percent of children are White.

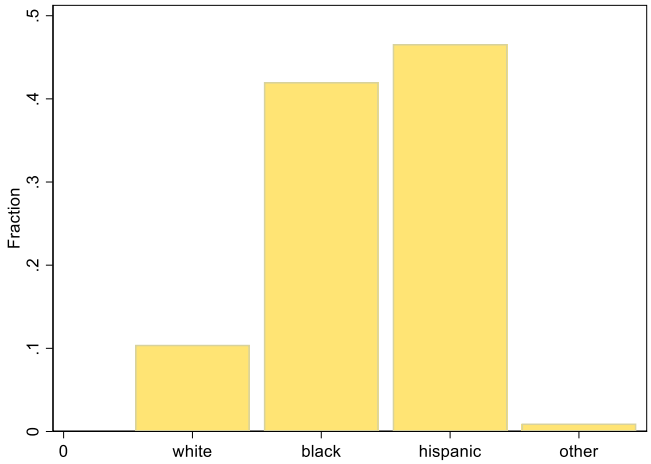

Figure C.4: Distribution of Ethnicities. $N = 778$.

Figure C.5 shows the distribution of education levels. We categorize the highest achieved education level into college degree, high school diploma, and less than high school diploma. For fathers we have missing information for 326 children. Of fathers for whom we have education status, 54 percent has a high school diploma and 16 percent has a college degree. For mothers we have 235 missing observations. 48 percent of mothers has a high school diploma and 29 percent has a college degree.

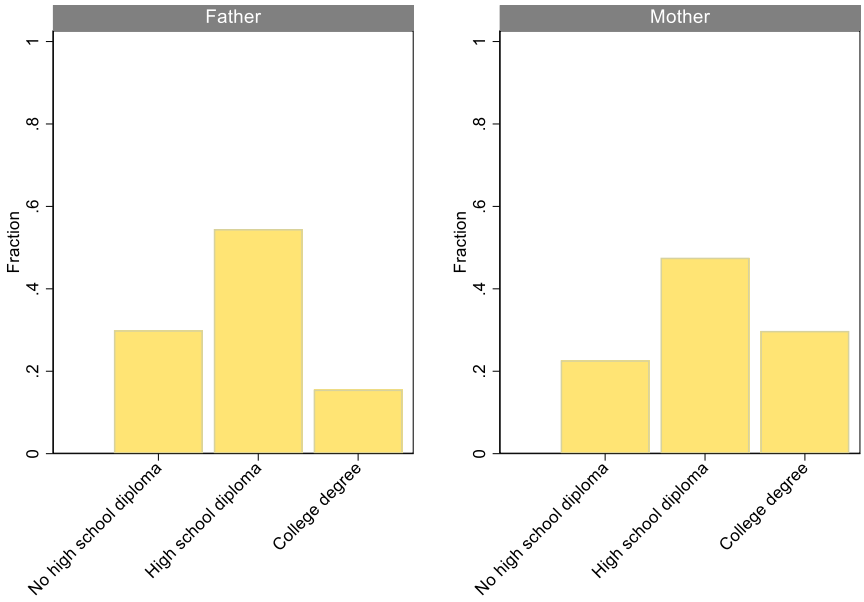

Figure C.5: Highest achieved education level. $N = 501$ (left panel), $N = 592$ (right panel).

Figure C.6 shows the distribution of parents' employment status. We distinguish between a full-time job, part-time job, and no paid job. For fathers we have 335 missing observations. 12 percent has a part-time job and 68 percent has a full-time job. For mothers we have 246 missing observations. 14 percent has a part-time job and 35 percent has a full-time job.

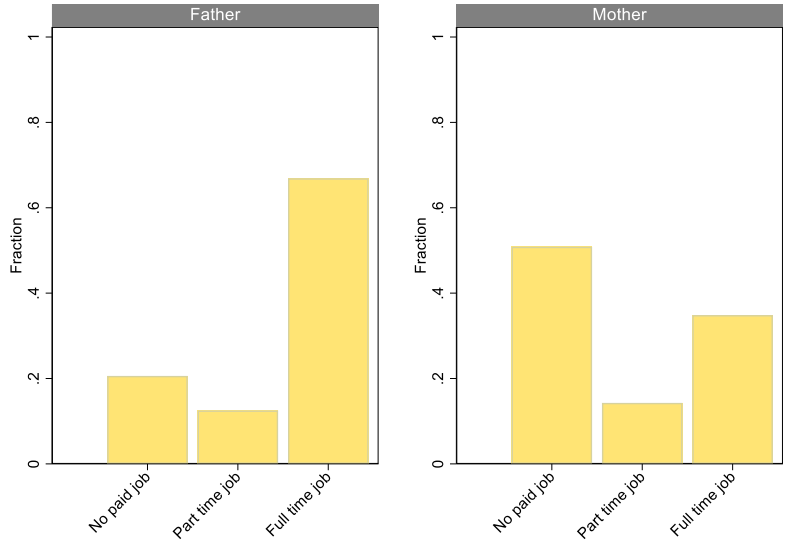

Figure C.6: Employment status. $N = 492$ (left panel), $N = 581$ (right panel).

Table C.2 shows some correlations between income, education level, employment status, and ethnicity. Most correlation coefficients are highly significant and in the expected direction.

Table C.2: Correlation Matrix

	Household income	Father college degree	Mother college degree	Father works full-time	Mother works full-time	Non-white
Household Income	1.000					
Father College degree	0.376***	1.000				
Mother College degree	0.464***	0.349***	1.000			
Father works full-time	0.305***	0.170***	0.126***	1.000		
Mother works full-time	0.402***	0.160***	0.255***	0.030	1.000	
Non-white	-0.243***	-0.126***	-0.114***	-0.099**	-0.035	1.000

Cognitive and non-cognitive scores are available for 772 children (55 missing values). Figure C.7 shows the distributions of test scores. Table C.3 shows the correlations between the different measures. The scores on the math and vocabulary parts are strongly correlated. The correlation with the memory part is also significant but much weaker. The cognitive and non-cognitive test scores are not very strongly correlated. Table C.3 also reports the correlation between the test scores and performance on the main task in the experiment. Not surprisingly, given that the task was a memory game, performance on the experimental task correlates most strongly with the memory component of the cognitive score.

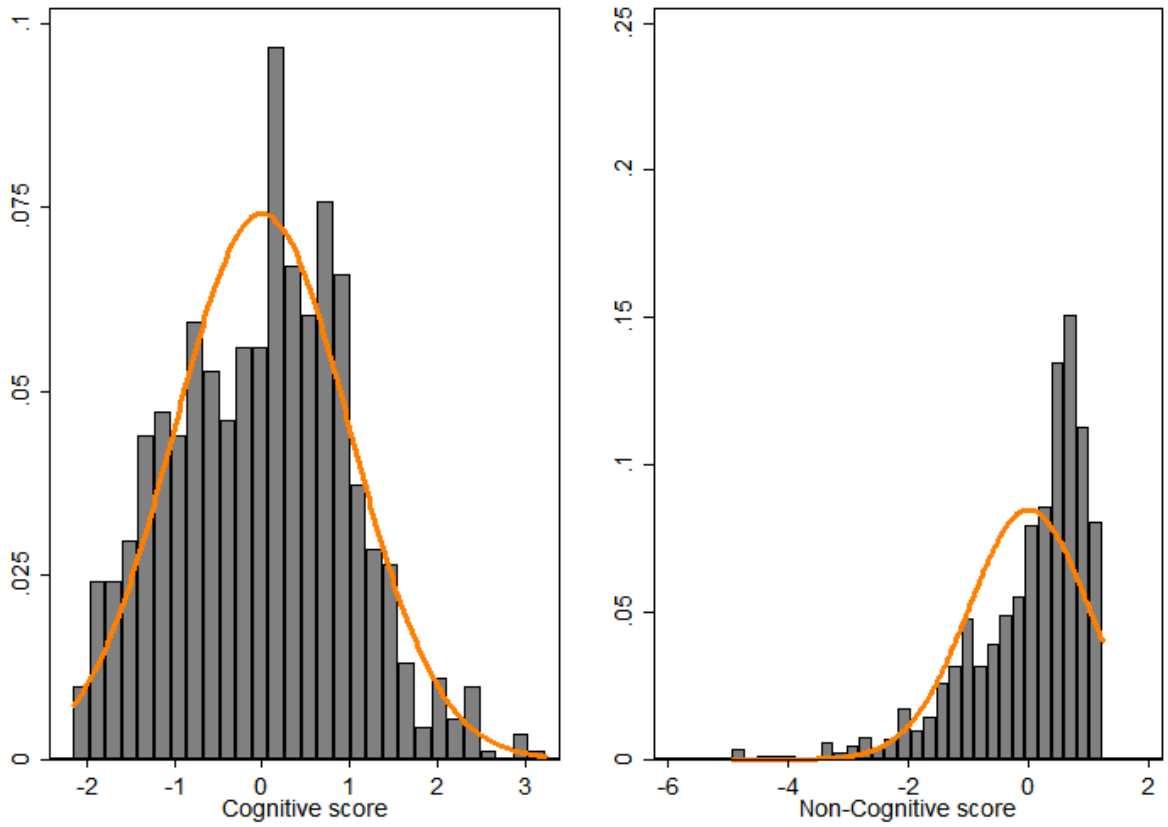

Figure C.7: Distributions of cognitive and non-cognitive test scores. Solid lines are fitted Normal distribution plots. $N = 772$.

Table C.3: Correlation Matrix

	Cog math	Cog vocabulary	Cog memory	Non-cognitive	Memory game
Cog math	1.000				
Cog vocabulary	0.813***	1.000			
Cog memory	0.548***	0.426***	1.000		
Non-cognitive	0.627**	0.501***	0.534***	1.000	
Memory game	0.395**	0.238**	0.475***	0.406***	1.000

Notes: Memory game is the performance on the main task of the experiment.

Appendix D: Balance of Covariates

To verify that the randomization of children over treatments was successful in terms of observables, Table D.1 shows the mean of several covariates by treatment. In most cases, the means are comparable across treatments. Children in the Contest treatment appear to be somewhat older. We cannot reject for any of the variables that they come from the same distribution (Kruskal-Wallis test).

Table D.1: Descriptive statistics – Covariates by treatment

Treatment	Baseline	Social	Contest	Test equality (*p*-value)
Cognitive score	0.49	0.33	0.40	0.930
Non-cognitive score	0.24	0.01	0.25	0.488
Performance main task	5.11	4.88	5.06	0.852
Female	0.5	0.48	0.55	0.827
Age (months)	78.5	76.52	85.57	0.120
Non-white	0.95	0.91	0.91	0.946
Household income (in $1,000)	23,10	28,12	27,57	0.517
Father college degree	0.15	0.15	0.18	0.979
Mother college degree	0.33	0.37	0.32	0.954
Father works full time	0.35	0.67	0.68	0.155
Mother works full time	0.48	0.35	0.36	0.684

Notes: *p*-values are based on a Kruskal-Wallis test

Appendix E: Additional analysis and Robustness checks

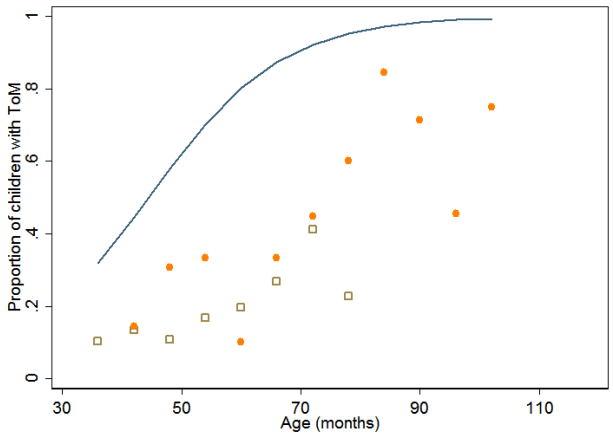

Figure E.1: Proportion of children with ToM. Mean rates of ToM by age category (grouped in bins of 6 months). Open squares are from the non-experimental studies. Solid circles are from the experimental study. The solid blue line is the mean proportion of children with ToM in other studies (based on the meta-study by Wellman et al., 2001). All children older than 100 months are grouped into a single bin (bin 102). Only bins with at least 5 observations are included (972 observations used, 18 observations dropped).

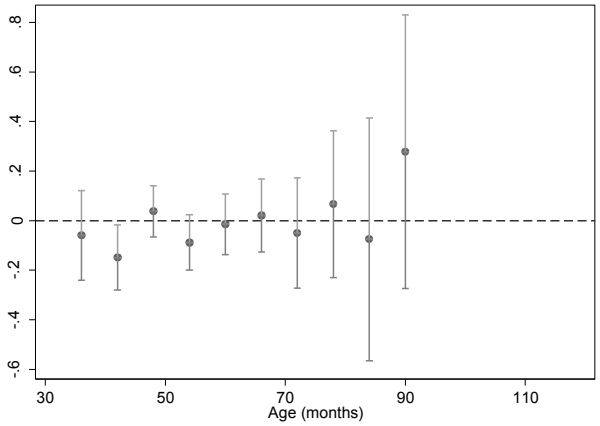

Figure E.2: Gender differences in ToM by age bin. Mean ToM rate males – females. Error bars indicate the 95 percent CI. Only bins with at least five observations are included (967 observations used, 23 observations dropped.).

Table E.1: ToM and Ethnicity

	(1)	(2)
Age (in months)	1.044***	1.045***
	(0.009)	(0.009)
Non-white	0.499***	
	(0.134)	
Black		0.517**
		(0.147)
Hispanic		0.495**
		(0.141)
Other		0.255
		(0.227)
Constant	0.033***	0.031***
	(0.020)	(0.019)
Controls	Yes	Yes
Observations	942	940

Notes: Logistic regressions reporting odds ratios. Controls: Female, Took test before, Experimental sample. Robust standard errors (clustered at the child level) in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table E.2:ToM and test scores

	(1)	(2)	(2)
Sample	Test within 90 days	All	All
Test Scores			
Math	1.305	1.117	1.158
	(0.371)	(0.194)	(0.253)
Vocabulary	0.923	1.042	0.903
	(0.204)	(0.151)	(0.182)
Memory	0.816	0.895	0.877
	(0.144)	(0.105)	(0.140)
Self-regulation	0.874	0.917	1.153
	(0.189)	(0.120)	(0.217)
Age at test data ToM	1.044**		
	(0.022)		
Age at test date cognitive scores		1.037***	1.034***
		(0.006)	(0.010)
Controls	Yes	No	Yes
Observations	450	910	515

Notes: Logistic regressions reporting odds ratios. Column (1): cases where ToM test and cognitive scores are administered within 90 days of each other. Columns (2) and (3): Entire sample. Controls: Took test before, Experimental sample. Test scores are standardized (mean 0, standard deviation 1). Robust standard errors (clustered at the child level) in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table E.3 Influences on Theory of Mind

	(1)	(2)	(3)	(4)
	Pre-K/Kindergarten		Grade 1+	
Social	0.955	1.047	3.692*	4.999**
	(0.517)	(0.582)	(2.557)	(3.954)
Contest	1.938	1.932	3.077*	6.188**
	(1.076)	(1.119)	(1.902)	(4.932)
Performance memory game		0.916		1.277*
		(0.077)		(0.167)
Female		1.237		2.303
		(0.582)		(1.462)
Took ToM test before		1.874		4.872**
		(0.878)		(3.276)
Constant	0.476*	0.404	1.083	0.076*
	(0.183)	(0.246)	(0.437)	(0.100)
Observations	88	87	71	70

Notes: Logistic regression model, reporting odds ratios. Robust standard errors (clustered at the child level) in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Computations of the out-of-sample predictions.

In section 3.3 we made out-of-sample predictions. These were derived as follows. We first estimated specification (4) of Table 4. Based on the estimated coefficients, we predicted for each age the predicted ToM based on the following values: female = 0.5, non-white = 0, all test scores are 0.5 (half a s.d. above the mean), household income = $57,065, father college = 0.35, mother college=0.35, father full time employment=0.9, mother full time employment = 0.5. The values of the socio-economic indicators were chosen to approximately reflect the median values in the US population. The predicted ToM rate was obtained based on the actual age distribution of children in our sample.