www.sciencemag.org/content/355/6323/389/suppl/DC1

Supplementary Materials for

Gender stereotypes about intellectual ability emerge early and influence children's interests

Lin Bian,* Sarah-Jane Leslie, Andrei Cimpian*

*Corresponding author. Email: linbian2@illinois.edu (L.B.); andrei.cimpian@nyu.edu (A.C.)

Published 27 January 2017, *Science* **355**, 389 (2017)
DOI: 10.1126/science.aah6524

This PDF file includes:

Materials and Methods

Participants. All participants were recruited in a small city in the Midwestern United States. Half of them were boys, and half were girls. Study 1 involved 96 children aged 5, 6, and 7 (48 boys, 48 girls; mean age = 5.55y, 6.50y, and 7.44y). Study 2 involved 144 children aged 5, 6, and 7 (72 boys, 72 girls; mean age = 5.50y, 6.48y, and 7.45y). Study 3 involved 64 children aged 6 and 7 (32 boys, 32 girls; mean age = 6.52y and 7.50y). Study 4 involved 96 children aged 5 and 6 (48 boys, 48 girls; mean age = 5.40y and 6.52y).

Demographic information was available for 75% of the families. The racial/ethnic composition of the sample mirrored that of the community in which this research was conducted: 78% of the children were European American, 7% Asian American, 5% African American, 3% Latino or Hispanic, and 7% multi-racial. The median household income was $90,000. Eight-two percent of the parents in the sample had at least a bachelor's degree.

Schooling status information was available for 34% of the children. Of the children for whom we had this information, 5% were in preschool, 21% in kindergarten, 11% in kindergarten or first grade (some schools in our sample combined these grades), 42% in first grade, and 20% in second grade or higher. (The children in the combined kindergarten/first grade program were not included in the analyses testing whether boys' and girls' ages differed in first grade.)

Materials and Procedures. Written informed consent was obtained from each child's parents prior to the test session. Children were tested individually in a quiet room in the lab or at their school. Across all four studies, the experimenter videotaped the sessions and recorded the children's responses on an answer form. At the end of the sessions, children were thanked for their participation and praised for their responses.

Study 1. The study began with a set of 12 screener questions designed to gauge whether children understand the meaning of the key terms "smart" (6 questions) and "nice" (6 questions). The "smart" and "nice" questions were presented to children as separate blocks whose order was counterbalanced. For each of these questions, the experimenter described the behavior of an unfamiliar child (e.g., "This child learns things fast") and then asked participants whether the relevant trait term could be applied to this child (e.g., "Is this child smart, not smart, or are you not sure?"). The questions were accompanied by pictures of individual boys and girls, which were placed behind a cardboard tent (i.e., out of participants' view) in order to avoid interfering with the subsequent tasks, which measured gender stereotypes. Children were corrected if they gave the wrong answer. We used an a priori exclusion criterion of 4/6 correct for each trait; 19 additional children were tested but excluded from the sample because they did not pass this threshold. In addition, 3 children were excluded for refusing to finish the study and 1 for being more than 2.5 standard deviations (*SD*s) away from the stereotype mean. We used a uniform 2.5 *SD* outlier exclusion criterion across studies. The results were robust to these exclusions: If the children who failed the screener questions are added to the final sample, the *P* values for gender differences in own-gender brilliance scores are .97 for the 5-year-olds (vs. .89 in the main text) and .007 for 6- and 7-year-olds (vs. .004 in the main text). Similarly, if the outlier is added to the final sample, the *P* values for gender differences in own-gender brilliance scores are .89 for the 5-year-olds (vs. .89 in the main text) and .002 for 6- and 7-year-olds (vs. .004 in the main text).

After the screener questions, the experimenter administered 3 stereotype tasks in random order. Task (i) consisted of 2 stories, each of which described an unfamiliar person whose gender was purposely left unspecified (see Table S1 for full text). One story was about a "really, really smart" person, and the other was about a "really, really nice" person. After telling the story, the experimenter laid out 4 pictures in a line (2 females and 2 males, randomly interspersed) and asked the child to guess which one of the 4 people might be the person in the story. If children chose a person of the same gender as themselves (e.g., if a girl picked a woman), they were assigned a score of 1 for that trial; otherwise, they received a 0.

In task (ii), children were shown 6 pictures one by one; each picture depicted two individuals. The first 2 trials served as practice trials, and the individuals depicted were all of the same gender as the participant. For the next 4 trials, the pictures consisted of a man and a woman. Children were told that one of the two people was "really, really smart" (on 3 of 6 trials) or "really, really nice" (on the other 3 trials), and they were asked to guess which of the two had the relevant trait. The order of the pictures was counterbalanced. The pictures used in this task and the next were of white men and women, normed for attractiveness ("How attractive does this person look?") and professional dress ("How professionally is this person dressed?") in a sample of 29 adults recruited via Amazon's Mechanical Turk. Race/ethnicity might interact with whether males are assumed to be more brilliant than females, so in future work it will be important to extend this work to targets from other race/ethnicity groups. Similar to task (i), children's responses were scored as a 1 on a trial if they chose the person of the same gender as themselves, and 0 otherwise.

In task (iii), children were asked to complete 3 puzzles. Each puzzle consisted of 2 rows × 4 columns, with the top row consisting of pictures of 2 men and 2 women arranged in a random order. Different pictures of men and women were used for each puzzle. However, the four pieces the child was asked to place in the bottom row of the puzzle remained the same: one piece had the word "smart" on it, one piece had the word "nice," one piece had a picture of a high-heel shoe (stereotypically feminine), and the fourth had a picture of a hammer (stereotypically masculine). Children were given the pieces one by one (and were told the word on them, for the "smart" and "nice" puzzle pieces), in random order, and asked to put these pieces in one of the empty slots on the bottom of the puzzle so as to "match" the pictures of the men and women at the top. Again, children's answers were scored as a 1 if they matched "smart" or "nice" with someone of the same gender as themselves, and 0 otherwise.

Study 2. The procedure of Study 2 was identical to that of Study 1, with three exceptions. First, each task was separated into two blocks: one with pictures of men and women (identical to those in Study 1), and the other with pictures of boys and girls. The order of the two blocks was counterbalanced. The pictures of boys and girls used in this study were normed for attractiveness ("How attractive does this child look?") and age ("How old do you think this child is?") in the same sample of 29 Mechanical Turk adults that rated the pictures used in Study 1. Second, because the addition of the child targets increased the length of the sessions, we omitted the puzzle task (i.e., task (iii)) to avoid taxing children's attention spans. Third, at the end of the sessions, we assessed children's perceptions of boys' and girls' school achievement. Children first saw 4 pictures of unfamiliar children (2 boys and 2 girls) and were asked, "If you had to make a guess, who do you think gets the best grades in school?" With another set of 4 pictures, they were then asked, "If you had to make a guess, who do you think is first in their class?" Finally, participants were asked the same 2 questions again, except this time they had to choose between 2 verbally-presented options: "boys or girls?" Responses across these 4 items were coded as in the stereotype tasks (same-gender choice = 1; other-gender choice = 0) and averaged.

Thirty-one additional children were tested but excluded from the final sample because they failed the initial screener questions (*n* = 28), refused to complete the study (*n* = 2), or were more than 2.5 *SD*s away from the stereotype mean (*n* = 1). The results were again robust to these exclusions: If the children who failed the screener questions are added to the final sample, the *P* values for gender differences in own-gender brilliance scores are .43 for the 5-year-olds (vs. .94 in the main text) and .003 for 6- and 7-year-olds (vs. .001 in the main text). If the outlier is added to the final sample, the *P* values for gender differences in own-gender brilliance scores are .60 for the 5-year-olds (vs. .94 in the main text) and .002 for 6- and 7-year-olds (vs. .001 in the main text).

Study 3. Children were introduced to two novel games ("zarky" and "impok") in counterbalanced order. For each game, the experimenter showed children a picture of it and briefly described its "rules" (see Table S4). Crucially, one game was said to be for "children who are really, really smart," and the other was said to be for "children who try really, really hard." Each game was presented in "smart" format for half of the participants and in "try hard" format for the other half. To ensure that children encoded the crucial ability information about each game, the experimenter asked them to recall it before proceeding to the main set of questions and corrected them if necessary. Next, the experimenter asked 4 questions designed to gauge children's interest in the game (e.g., "Would you want to play the zarky/impok game, or would you not want to play it?"; see Table S5). The order of the questions was randomized. After the first two questions, children were provided with a reminder of the relevant ability information (i.e., that the game is for children who are "really, really smart" vs. who "try really, really hard"). Responses to the four questions were standardized (so that they are on the same scale) and then averaged.

After an abbreviated, simplified set of screener questions (which all children passed), we assessed children's brilliance stereotypes with task (i) from Studies 1 and 2 (the gender-neutral story). However, before children selected the protagonist of the story from among the 4 pictures provided (as in Studies 1 and 2), we also asked them to repeat the story and then coded the gender of the pronouns they used (*20*). The final own-gender brilliance score in this study was an average of these two items (the pronouns they used + the pictures they selected; standardized before averaging).

At the end of the sessions, children received a thorough debriefing that was designed to convey that effort and hard work are the key to success (e.g., "If you try really hard and practice a lot, you can be good at *any* game you want").

One additional child was tested but excluded from the final sample because she was more than 2.5 *SD*s from the interest mean. If the outlier is added to the final sample, the *P* values for gender differences in interest are .023 for the brilliance game (vs. .045 in the main text) and .69 for the try-hard game (vs. .47 in the main text).

Study 4. The procedure of Study 4 was identical to that of Study 3, except children were only told about the game "for children who are really, really smart." Half of the children saw the "zarky" game, and half saw the "impok" game.

Two additional children were tested but excluded from the final sample because they were more than 2.5 *SD*s from the interest mean. If the outliers are added to the final sample, the *P* values for gender differences in interest are .99 for the 5-year-olds (vs. .45 in the main text) and .053 for the 6-year-olds (vs. .057 in the main text). With the outliers from Studies 3 and 4 included, the meta-analytic effect size for the gender difference in 6- and 7-year-olds' interest toward the brilliance game is $d = .54$ (vs. .51 in the main text), 95% confidence interval = [.17, .92], $P = .005$.

Analyses. In Study 1, children's stereotype scores (combined across the 3 tasks) were submitted to a multilevel mixed-effects linear model with trait (smart vs. nice; level-1 predictor), gender (boys vs. girls; level-2 predictor), and age (5- vs. 6- vs. 7-year-olds; level-2 predictor), plus all possible interaction terms, as categorical fixed effects and a random intercept for participants. A second model was also calculated in which the 6- and 7-year-olds were treated as a single group. The models were computed with the *mixed* command in Stata 14.1. The crucial three-way interaction among trait, gender, and age (5- vs. 6- and 7-year-olds) was significant, Wald χ^2 = 10.01, $P = .002$. Four follow-up tests compared boys' and girls' stereotype scores about each trait separately for younger (5-year-old) and older (6- and 7-year-old) children. These tests were computed with the *contrast* command in Stata, which outputs a Wald χ^2 statistic. Adjusting the α level to .013 to account for multiple comparisons (with a Bonferroni correction) would leave our conclusions intact, since the *P* values for all crucial differences are below this threshold.

The analyses in Study 2 were similar, except the multilevel models included an extra level-1 fixed effect (plus its interactions with all other fixed effects): the age of the stereotype targets (adults vs. children). Again, the three-way interaction among trait, gender, and age (5- [younger] vs. 6- and 7-year-olds [older]) was significant, Wald χ^2 = 7.51, $P = .006$. In addition, we used *t* tests and a Pearson correlation to analyze the data on perceptions of school achievement and their relationship to brilliance stereotypes.

In Study 3, children's interest scores (combined across the 4 questions) were submitted to a multilevel mixed-effects linear model with game (smart vs. try-hard; level-1 predictor), gender (boys vs. girls; level-2 predictor), and age (6- vs. 7-year-olds; level-2 predictor), plus all possible interaction terms, as categorical fixed effects and a random intercept for participants. The interaction between gender and game was significant, Wald χ^2 = 5.42, $P = .020$. Two follow-up tests comparing boys' and girls' interest in each game were computed with the *contrast* command in Stata. Boys' and girls' stereotype scores were compared with a *t* test. The indirect effect of children's gender on their interest via "brilliance = males" stereotypes was calculated by using the PROCESS macro (release 2.16.1) in SPSS 24 to perform a bootstrapped (10,000 replications) product-of-coefficients mediation test (*34*, *35*). Children's gender (0 = boys, 1 = girls) was the independent variable in this analysis; the own-gender brilliance score was the mediator; and children's interest in the smart vs. the try-hard game (a difference score) was the dependent variable. The mediator and the dependent variable were standardized prior to entering in the mediation analysis.

In Study 4, children's interest scores were submitted to a linear regression with gender (boys vs. girls), age (5- vs. 6-year-olds), and their interaction as categorical predictors. Standard errors were bootstrapped (10,000 replications). The interaction between gender and age was marginally significant, Wald $\chi^2 = 3.12$, $P = .078$. Follow-ups testing gender differences in children's interest at each age level were again computed with *contrast* in Stata 14.1. The random-effects meta-analysis of gender differences in 6- and 7-year-olds' interest in the smart game was computed with the *metan* command in Stata using the relevant sample sizes, means, and *SD*s from Studies 3 and 4.

Data. The data for these studies are available on Open Science Framework: https://osf.io/yund6/?view_only=9a8505d4e87b456a89f255b43e21234e.

Supplemental Analyses

Studies 1 and 2: First-Trial Data. Was the difference in 6- and 7-year-old boys' vs. girls' own-gender brilliance scores apparent on the first question they were asked, or did it only emerge over the course of the session? Consistent with the first possibility, the difference in 6- and 7-year-old boys' vs. girls' own-gender brilliance scores on the first trial was as large as in the aggregate, $M_{boys} = .73$ vs. $M_{girls} = .50$, Wald $\chi^2 = 4.07$, $P = .044$. In addition, this difference was not significant among 5-year-olds, replicating the analyses in the main text, $M_{boys} = .77$ vs. $M_{girls} = .73$, Wald $\chi^2 = 0.10$, $P = .75$.

Analyses Involving Demographic Variables. Several analyses were conducted to investigate whether demographic variables moderated the main results of Studies 1–4.

Studies 1 and 2: Race/Ethnicity. Do children from different racial/ethnic backgrounds have different beliefs about which gender is "really, really smart" and which is "really, really nice"? Because each of the racial/ethnic minority groups in our sample (e.g., African Americans, Latinos) was small, we combined them into a single group for purposes of this analysis. To increase the power to detect differences by race/ethnicity, we also pooled the data from Studies 1 and 2. We then submitted these data to a multilevel mixed-effects linear model with race/ethnicity (white children vs. children of color; level-2 predictor), trait (smart vs. nice; level-1 predictor), gender (boys vs. girls; level-2 predictor), and age (5- [younger] vs. 6- and 7-year-olds [older]; level-2 predictor), plus all possible interaction terms, as categorical fixed effects and a random intercept for participants. Contrary to the idea that the development of the stereotypes investigated here varies by racial/ethnic group, we found that race/ethnicity did not significantly moderate the key three-way interaction among trait, gender, and age, Wald $\chi^2 = 0.15$, $P = .70$. Inspection of the means revealed broadly similar developmental patterns. For example, own-gender brilliance scores decreased with age for both white girls ($M_{younger} = .67$ vs. $M_{older} = .49$) and girls of color ($M_{younger} = .83$ vs. $M_{older} = .60$).

Studies 1 and 2: Socioeconomic Status (SES). Do children from high- vs. low-SES backgrounds have different beliefs about which gender is "really, really smart" and which is "really, really nice"? To examine this question, we first created a composite SES measure by (1) standardizing the average education level of the parent(s) (which had been converted to years of education prior to standardizing) and the total income of the household, and then (2) averaging these two scores (education and income) into a composite SES variable. Next, we performed the same multilevel analysis as above, except that race/ethnicity was replaced by SES. Again, we found that SES did not significantly moderate the key three-way interaction among trait, gender, and age, Wald $\chi^2 = 0.58$, $P = .45$. Inspection of the means revealed similar developmental trends for high- and low-SES children. For example, both high-SES (+1 *SD*) girls and low-SES (−1 *SD*) girls showed age-related drops in their own-gender brilliance scores (high-SES girls: $M_{younger} = .71$ vs. $M_{older} = .49$; low-SES girls: $M_{younger} = .71$ vs. $M_{older} = .53$).

Studies 3 and 4: Race/Ethnicity. To examine if children's race/ethnicity moderates the observed gender differences in interest toward the smart game, we submitted this variable (pooled across Studies 3 and 4) to a linear regression with race/ethnicity (white children vs. children of color), gender (boys vs. girls), age (5- [younger] vs. 6- and 7-year-olds [older]), and all their two- and three-way interactions as categorical predictors. Similar to the analysis of children's stereotypes, children's racial/ethnic backgrounds did not significantly moderate the key age × gender interaction in children's interest toward the smart game, Wald $\chi^2 = 1.48$, $P = .22$. For example, 6- and 7-year-old girls displayed lower interest in the smart game than 6- and 7-year-old boys regardless of their racial/ethnic group (white children: $M_{boys} = .25$ vs. $M_{girls} = -.22$; children of color: $M_{boys} = .19$ vs. $M_{girls} = -.23$).

Studies 3 and 4: SES. Like race/ethnicity, SES did not significantly moderate the age × gender interaction in children's interest toward the smart game, Wald $\chi^2 = 0.01$, $P = .96$. For example, 6- and 7-year-old girls' interest in the smart game was lower than 6- and 7-year-old boys' interest regardless of whether they came from high-SES ($M_{boys} = .08$ vs. $M_{girls} = -.22$) or low-SES ($M_{boys} = .34$ vs. $M_{girls} = -.22$) families.

Table S1

The Gender-Neutral Stories Used to Assess Children's Stereotypes in Studies 1 and 2

	Story about an Adult (Study 1)	Story about a Child (Studies 1 and 2)
Trait: Smart	There are lots of people at the place where I work. But there is one person who is really special. This person is really, really smart. This person figures out how to do things quickly and comes up with answers much faster and better than anyone else. This person is really, really smart.	When I was your age, there were lots of children at the kindergarten where I went. But there was one child who was really special. This child was really, really smart. This child learned things very quickly and could answer even the hardest questions from the teacher. This child was really, really smart.
Trait: Nice	There are lots of people at the place where I work. But there is one person who is really special. This person is really, really nice. This person likes to help others with their problems and is friendly to everyone at the office. This person is really, really nice.	When I was your age, there were lots of children at the kindergarten where I went. But there was one child who was really special. This child was really, really nice. This child shared their toys with everyone else, and really cared about the other kids. This child was really, really nice.

Table S2

Boys' and Girls' Stereotype Scores in Studies 1 and 2 (Standard Deviations in Parentheses)

Age	Gender	Study 1		Study 2	
		Smart	Nice	Smart	Nice
5-year-olds	Boys	0.71 (0.22)	0.66 (0.22)	0.73 (0.24)	0.63 (0.24)
	Girls	0.69 (0.19)	0.61 (0.31)	0.73 (0.23)	0.77 (0.23)
6-year-olds	Boys	0.65 (0.20)	0.40 (0.25)	0.69 (0.27)	0.49 (0.26)
	Girls	0.48 (0.24)	0.67 (0.15)	0.52 (0.21)	0.73 (0.15)
7-year-olds	Boys	0.68 (0.26)	0.43 (0.24)	0.66 (0.25)	0.48 (0.29)
	Girls	0.54 (0.21)	0.62 (0.18)	0.55 (0.23)	0.74 (0.17)

Table S3

Boys' and Girls' Interest Scores in Studies 3 and 4 (Standard Deviations in Parentheses)

Age	Gender	Study 3		Study 4
		Smart	Try-hard	Smart
5-year-olds	Boys	−	−	−0.08 (0.88)
	Girls	−	−	0.11 (0.84)
6-year-olds	Boys	0.20 (0.71)	−0.09 (0.81)	0.17 (0.63)
	Girls	−0.17 (0.77)	0.10 (0.49)	−0.21 (0.73)
7-year-olds	Boys	0.15 (0.69)	−0.03 (0.90)	−
	Girls	−0.21 (0.88)	0.04 (0.58)	−

Table S4

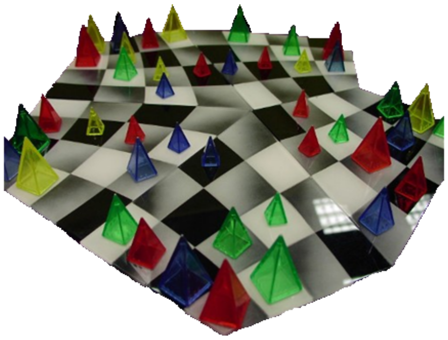
The Games Used to Assess Children's Interest in Studies 3 and 4

Zarky

I want to tell you about this game that I ask children to play sometimes. It's called Zarky, and it's a lot of fun. In this game, what you have to do is to bring the red pieces from this side to this side, one piece at a time, without going in a straight line and without getting them stuck in between the blue pieces. Oh, and here is something else about the Zarky game, and this is important so make sure you're paying attention. This game is not for everyone. It's only for children who are really, really smart [who try really, really hard]. Only smart [hardworking] children can be good at this game.

Impok

I want to tell you about this game that I ask children to play sometimes. It's called Impok, and it's a lot of fun. In this game, what you have to do is to figure out how to get the big pyramids next to each other in the black squares and get the small pyramids next to each other in the white squares in only ten moves and without crossing the grey squares. Oh, and here is something else about the Impok game, and this is important, so make sure you're paying attention. This game is not for everyone. It's only for children who are really, really smart [who try really, really hard]. Only smart [hardworking] children can be good at this game.



Note. In Study 3, each of the games was presented in the "smart" format to half of the children and in the "try-hard" format to the other half. Only the "smart" game versions were used in Study 4.

Table S5

The Four Questions Used to Assess Children's Interest in Studies 3 and 4

1) Imagine I had the Zarky/Impok game right here, in front of you. Would you want to play the Zarky/Impok game, or would you not want to play it?

   [*if "yes"*]
   Would you sort of want to play it (= 3), or really want to play it (= 4)?

   [*if "no"*]
   Would you sort of not want to play it (= 2), or really not want to play it (= 1)?

2) Do you like the Zarky/Impok game, or do you not like it?

   [*if "yes"*]
   Do you sort of like it (= 4), like it (= 5), or really like it (= 6)?

   [*if "no"*]
   Do you sort of not like it (= 3), not like it (= 2), or really not like it (= 1)?

3) Imagine you are playing the Zarky/Impok game. Would playing Zarky/Impok make you feel happy or sad?

   [*if "happy"*]
   Would it make you feel sort of happy (= 4), happy (= 5), or really happy (= 6)?

   [*if "sad"*]
   Would it make you feel sort of sad (= 3), sad (= 2), or really sad (= 1)?

4) If you had a chance to do something tomorrow, would you play the Zarky/Impok game (= 1) or would you do something else (= 0)?

Note. The numerical scoring of each option is indicated in parentheses. Question order was randomized across children. Because the questions used different scales, responses to each were standardized before averaging.

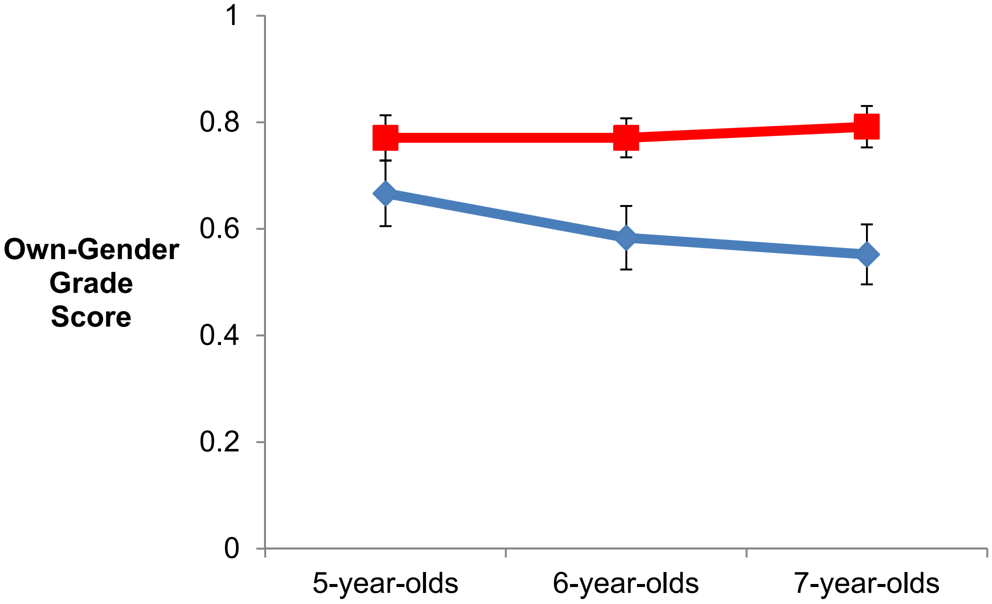


Fig. S1. Boys' (blue) and girls' (red) average proportions of selecting children of the same gender as having top grades in Study 2, by age group (5- vs. 6- vs. 7-year-olds). The error bars represent ± 1 *SE*.

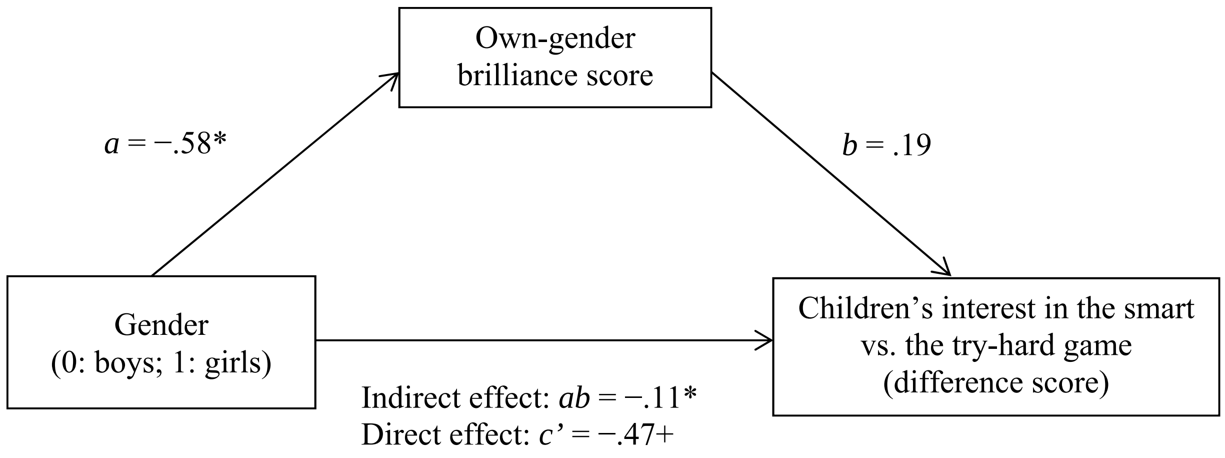

Own-gender brilliance score
a = −.58*
b = .19
Gender (0: boys; 1: girls)
Children's interest in the smart vs. the try-hard game (difference score)
Indirect effect: *ab* = −.11*
Direct effect: *c'* = −.47+

Fig. S2. The difference between boys and girls in their interest toward the smart vs. the try-hard game was mediated by their own-gender brilliance scores. The mediator and the dependent variable were standardized prior to entering in the mediation analysis. * $p < .05$, + $p < .10$

References and Notes

1. W. Wood, A. H. Eagly, Biosocial construction of sex differences and similarities in behavior. *Adv. Exp. Soc. Psychol.* **46**, 55–123 (2012). doi:10.1016/B978-0-12-394281-4.00002-7

2. S. T. Fiske, A. J. C. Cuddy, P. Glick, J. Xu, A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *J. Pers. Soc. Psychol.* **82**, 878–902 (2002). doi:10.1037/0022-3514.82.6.878 Medline

3. D. Cvencek, A. N. Meltzoff, A. G. Greenwald, Math-gender stereotypes in elementary school children. *Child Dev.* **82**, 766–779 (2011). doi:10.1111/j.1467-8624.2010.01529.x Medline

4. S. J. Spencer, C. M. Steele, D. M. Quinn, Stereotype threat and women's math performance. *J. Exp. Soc. Psychol.* **35**, 4–28 (1999). doi:10.1006/jesp.1998.1373

5. S. Galdi, M. Cadinu, C. Tomasetto, The roots of stereotype threat: When automatic associations disrupt girls' math performance. *Child Dev.* **85**, 250–263 (2014). doi:10.1111/cdev.12128 Medline

6. P. G. Davies, S. J. Spencer, D. M. Quinn, R. Gerhardstein, Consuming images: How television commercials that elicit stereotype threat can restrain women academically and professionally. *Pers. Soc. Psychol. Bull.* **28**, 1615–1628 (2002). doi:10.1177/014616702237644

7. M. C. Murphy, C. M. Steele, J. J. Gross, Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychol. Sci.* **18**, 879–885 (2007). doi:10.1111/j.1467-9280.2007.01995.x Medline

8. S. Upson, L. F. Friedman, Where are all the female geniuses? *Sci. Am. Mind* **23**, 63–65 (2012). doi:10.1038/scientificamericanmind1112-63

9. A. Furnham, E. Reeves, S. Budhani, Parents think their sons are brighter than their daughters: Sex differences in parental self-estimations and estimations of their children's multiple intelligences. *J. Genet. Psychol.* **163**, 24–39 (2002). doi:10.1080/00221320209597966 Medline

10. B. Kirkcaldy, P. Noack, A. Furnham, G. Siefen, Parental estimates of their own and their children's intelligence. *Eur. Psychol.* **12**, 173–180 (2007). doi:10.1027/1016-9040.12.3.173

11. A. Lecklider, *Inventing the Egghead: The Battle Over Brainpower in American Culture* (Univ. of Pennsylvania Press, 2013).

12. S.-J. Leslie, A. Cimpian, M. Meyer, E. Freeland, Expectations of brilliance underlie gender distributions across academic disciplines. *Science* **347**, 262–265 (2015). doi:10.1126/science.1261375 Medline

13. A. Cimpian, S.-J. Leslie, Response to Comment on "Expectations of brilliance underlie gender distributions across academic disciplines". *Science* **349**, 391 (2015). doi:10.1126/science.aaa9892 Medline

14. D. Storage, Z. Horne, A. Cimpian, S.-J. Leslie, The frequency of “brilliant” and “genius” in teaching evaluations predicts the representation of women and African Americans across fields. *PLOS ONE* **11**, e0150194 (2016). doi:10.1371/journal.pone.0150194 Medline

15. M. Meyer, A. Cimpian, S.-J. Leslie, Women are underrepresented in fields where success is believed to require brilliance. *Front. Psychol.* **6**, 235 (2015). doi:10.3389/fpsyg.2015.00235 Medline

16. K. T. U. Emerson, M. C. Murphy, A company I can trust? Organizational lay theories moderate stereotype threat for women. *Pers. Soc. Psychol. Bull.* **41**, 295–307 (2015). doi:10.1177/0146167214564969 Medline

17. D. K. Ginther, S. Kahn, Comment on “Expectations of brilliance underlie gender distributions across academic disciplines”. *Science* **349**, 391 (2015). doi:10.1126/science.aaa9632 Medline

18. K. Crowley, M. A. Callanan, H. R. Tenenbaum, E. Allen, Parents explain more often to boys than to girls during shared scientific thinking. *Psychol. Sci.* **12**, 258–261 (2001). doi:10.1111/1467-9280.00347 Medline

19. H. R. Tenenbaum, C. Leaper, Parent-child conversations about science: The socialization of gender inequities? *Dev. Psychol.* **39**, 34–47 (2003). doi:10.1037/0012-1649.39.1.34 Medline

20. N. Ambady, M. Shih, A. Kim, T. L. Pittinsky, Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychol. Sci.* **12**, 385–390 (2001). doi:10.1111/1467-9280.00371 Medline

21. L. S. Liben, R. S. Bigler, H. R. Krogh, Pink and blue collar jobs: Children’s judgments of job status and job aspirations in relation to sex of worker. *J. Exp. Child Psychol.* **79**, 346–363 (2001). doi:10.1006/jecp.2000.2611 Medline

22. Y. Dunham, A. S. Baron, M. R. Banaji, The development of implicit gender attitudes. *Dev. Sci.* **19**, 781–789 (2015). doi:10.1111/desc.12321 Medline

23. D. Cvencek, A. G. Greenwald, A. N. Meltzoff, Implicit measures for preschool children confirm self-esteem’s role in maintaining a balanced identity. *J. Exp. Soc. Psychol.* **62**, 50–57 (2016). doi:10.1016/j.jesp.2015.09.015

24. D. Voyer, S. D. Voyer, Gender differences in scholastic achievement: A meta-analysis. *Psychol. Bull.* **140**, 1174–1204 (2014). doi:10.1037/a0036620 Medline

25. S. L. Beilock, E. A. Gunderson, G. Ramirez, S. C. Levine, Female teachers’ math anxiety affects girls’ math achievement. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 1860–1863 (2010). doi:10.1073/pnas.0910967107 Medline

26. J. P. Robinson-Cimpian, S. T. Lubienski, C. M. Ganley, Y. Copur-Gencturk, Teachers’ perceptions of students’ mathematics proficiency may exacerbate early gender gaps in achievement. *Dev. Psychol.* **50**, 1262–1281 (2014). doi:10.1037/a0035073 Medline

27. G. Cumming, The new statistics: Why and how. *Psychol. Sci.* **25**, 7–29 (2014). doi:10.1177/0956797613504966 Medline

28. F. L. Huang, Investigating the prevalence of academic redshirting using population-level data. *AERA Open* **1**, 1–11 (2015). doi:10.1177/2332858415590800

29. J. Eccles, C. Midgley, T. F. Adler, "Grade-related changes in the school environment: Effects on achievement motivation," in *The Development of Achievement Motivation*, J. G. Nicholls, Ed. (JAI Press, 1984), pp. 282–331.

30. R. Butler, "Competence assessment, competence, and motivation between early and middle childhood," in *Handbook of Competence and Motivation*, A. J. Elliott, C. S. Dweck, Eds. (Guilford Press, 2005), pp. 202–221.

31. L. A. Rudman, Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *J. Pers. Soc. Psychol.* **74**, 629–645 (1998). doi:10.1037/0022-3514.74.3.629 Medline

32. J. L. Smith, M. Huntoon, Women's bragging rights: Overcoming modesty norms to facilitate women's self-promotion. *Psychol. Women Q.* **38**, 447–459 (2013). doi:10.1177/0361684313515840

33. J. Mazei, J. Hüffmeier, P. A. Freund, A. F. Stuhlmacher, L. Bilke, G. Hertel, A meta-analysis on gender differences in negotiation outcomes and their moderators. *Psychol. Bull.* **141**, 85–104 (2015). doi:10.1037/a0038184 Medline

34. A. F. Hayes, M. Scharkow, The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychol. Sci.* **24**, 1918–1927 (2013). doi:10.1177/0956797613480187 Medline

35. A. F. Hayes, *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach* (Guilford Publications, 2013).