Personality and Individual Differences 239 (2025) 113111

Contents lists available at ScienceDirect

Personality and Individual Differences

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

Boys are smart (and really dull and pretty average): Testing replication and validity of the Brilliance Stereotype

Yue Li [a,*], Timothy C. Bates [b]

[a] School of Social Sciences, University of Westminster, W1W 6UW London, UK
[b] Department of Psychology, University of Edinburgh, EH8 9JZ Edinburgh, UK

ARTICLE INFO

Keywords:
Gender stereotype
Career interests
Brilliance stereotype
Young children
Cross-culture

ABSTRACT

A Brilliance Stereotype associating high intellectual ability with men and not women with possible downstream impacts on interests or work has been reported. Here, we report five replications and extensions testing this finding (total $N = 737$). Studies 1 and 2 were direct replications and found no support for the male brilliance stereotype: Instead, 10-year-old boys and girls both chose own-gender targets as smartest. Study 3 tested stereotyping of the opposite of brilliance – being very dull. Contrary to the brilliance stereotype model, males were stereotyped as dull by both girls and boys ($OR = 0.22$, $p < .001$). Study 4 added additional validity checks, but no difference in brilliance stereotype was found between boys and girls ($p = .517$). We also tested the causal claim that brilliance stereotypes impact career interests. Large gender differences were found for occupational interests (e.g. nursing ($\beta = 0.73$ CI_{95} [0.48, 0.98], $t = 5.68$, $p < .001$, scientist/engineer ($\beta = -0.61$ CI_{95} [−0.88, −0.35], $t = -4.60$, $p < .001$). Despite this, the brilliance stereotype showed no relationship to any occupational interests (p-values 0.523 to 0.999). Brilliance stereotype, and effects of brilliance stereotype lack internal coherence and predictive validity.

1. Introduction

Stereotypes are group representations (Hilton & von Hippel, 1996) which are often accurate (Jussim et al., 2016), and may influence treatment and choices of individuals (Ellemers, 2018). A recent report suggested that girls as young as 6-years-old show a brilliance stereotype – treating innate ability as a male, not a female trait. Evidence for this included matching photos to vignettes describing a "*really, really smart*" person: Older boys and girls were more likely to choose a male photo (Bian et al., 2017). As stereotype theory is scientifically important and predictions impact STEM and workplace policy (Leslie et al., 2015), replication ensuring the method produces coherent and valid results is crucial (Eronen & Bringmann, 2021; Open Science Collaboration, 2015). We therefore tested replication of the brilliance stereotype reported in study-1, task (i) of Bian et al. (2017) – hereafter "study1i" – collecting data in five studies across two countries as well as using additional conditions to test the coherence and interpretability of stereotyping in vignette responses. We first briefly summarise the target study.

Bian et al. (2017) conducted four studies on gender stereotypes about brilliance (hereon termed brilliance stereotypes) and interest in brilliance-related activities in children from 5 to 7 years old. They found that by age-6-years, girls endorsed a brilliance stereotype towards males and, compared to boys, had less interest in games requiring innate ability (children were told that the game was not for everyone but only for those who are really smart, and only smart children could be good at the game). One method used to test brilliance stereotypes was vignette matching. Specifically, in study-1i Bian et al. (2017) tested children aged 5, 6 or 7 years (16 boys and girls at each age). Children saw vignettes describing a "*really, really smart*" person, and a "*really, really nice*" person and chose an image from 2 male and 2 female targets most likely to be the person in the vignette. At age 5, both boys and girls showed overwhelming (61–71 %) own-gender bias – associating brilliance and niceness with their own gender (Bian et al., 2017). This own-gender effect has been reported previously (e.g. Cvencek et al., 2011, 2016; Dunham et al., 2016). In older groups, however, girls were unbiased, associating brilliance at random to female and male targets (e.g., at age 6, 48 % of girls chose a same-gender target as very brilliant). At the same age, boys retained the own-gender bias shown by five-year olds, choosing a male 65 % of the time (Wald $\chi^2 = 8.10$, $p = .004$). Interestingly, niceness stereotypes changed significantly in the older children, with both girls and boys choosing a female target as nice. The conclusion drawn was that environmental influences lead boys and girls to

* Corresponding author.
E-mail addresses: y.li1@westminster.ac.uk (Y. Li), tim.bates@ed.ac.uk (T.C. Bates).

https://doi.org/10.1016/j.paid.2025.113111
Received 6 December 2024; Received in revised form 12 February 2025; Accepted 14 February 2025
Available online 20 February 2025

stereotype females as nicer and males as more intellectually able than females, and that these then manifest in interests and career choices.

These results are potentially significant (Leslie et al., 2015; Schuster & Martiny, 2016; Shapiro & Williams, 2011). However, despite ~1684 citations, to our knowledge no independent replication has been reported. Replication and validation are essential (Simmons et al., 2011), as is testing generalization to different cultures (Flore et al., 2019; Flore & Wicherts, 2015). Therefore, we conducted five studies testing if these results replicate in different cultures (China, then UK, which has the same concern over STEM recruitment) then, in part due to failures to replicate in these first two studies, we examine the validity and coherence of the vignette method itself.

2. Study 1

It is important for replication that subjects do not differ in ways predicted to alter results. Bian et al. (2017) ruled out several factors as irrelevant to the brilliance stereotype. Vignettes of adult and child stimuli yielded similar results, and effects were constant across race, ethnicity, and socioeconomic status (SES; Bian et al., 2017, p. S6). These, then, are not predicted to impact replication. However, whether the effects are cross-culturally constant is unclear. Gender stereotyping may differ between cultures, as it varies over time within a culture (Madon et al., 2001). In Study 1, therefore, we used vignette matching to test 10-year-old children in China.

Gender stereotypes in China reflect traditional gender roles rooted in Confucian norms of obedience to males: until married, females need to obey their fathers; and after marrying, they need to obey their husbands; if husbands pass away, females should obey their sons (Bowen et al., 2007). Although traditions have moderated gender stereotypes of females as inferior remain. For instance, Tsui (2007) found that most teachers believed males are superior on mathematics/science, attributing this to superior male spatial ability. Based on these norms, we replicated and extended Bian's study in a different culture and an older age group, i.e., in a sample of 10-year-old Chinese students. As stereotypes should increase with time (Bian et al., 2017), we expected to find effects at least as large or larger than those reported in the target study.

As a control, we tested if the brilliance-stereotype was reflected in a stereotype of girls as more likely to have *low*-ability. We therefore added a vignette describing a person with "*very low*" intellectual ability to examine if low ability would be stereotyped as female, as predicted by a gender stereotype regarding basic ability. This was added at the end of the test items thus not interfering with the presentation conditions for the earlier vignettes. If stereotyping for brilliance is in the "male" direction, logically, a vignette about a person with low ability should be stereotyped by children as more likely female.

We also tested replication of the niceness effect. This provides a useful internal control for the replication. Unlike cognitive ability, where data show negligible mean gender differences (Deary et al., 2007), studies of traits linked to niceness reveal replicable differences favouring females in sensitivity, warmth and agreeableness (Del Giudice et al., 2012; Lippa, 2010; Weisberg et al., 2011). We hypothesised both based on Bian et al. (2017) and wider literature on gender differences in personality, that boys and girls would pick a female target as most likely to be very nice.

Our hypotheses in study 1 were as follows.

1. Girls would be less likely to choose a same-gender target as a match for the narrative of a really smart person than boys would be to choose a same-gender target, replicating the finding of Bian et al. (2017).
2. Boys and girls would both be more likely to choose a female target as a match a really nice person compared to male targets, replicating niceness stereotyped as female (Bian et al., 2017).
3. Female targets would be chosen as most likely to be slow-minded, testing the matching stereotype at the low end of brilliance is consistent with stereotyping of lower female ability.

2.1. Method

2.1.1. Participants

A total of 227 children (118 boys, 109 girls) were recruited from a public primary school in China during the spring term of 2017. Participants were aged from 7 years 8 months to 11 years 4 months old (mean = 10.10, SD = 0.52). The socioeconomic status of the region in which we recruited participants is low (e.g., 21 % below the Chinese national average income; National Bureau of Statistics of the People's Republic of China, 2017).

2.1.2. Materials

Our experimental materials and procedure closely followed study-1i (Bian et al., 2017), with differences noted below (see also Table 1, tabulating points of similarity and difference between Bian et al. (2017) and studies 1–4 in the present paper). The vignettes in study-1i in Bian et al. (2017) describing a "*really, really smart*" person and a "*really, really nice*" person were translated in Chinese. A back-translation into English was made by a bilingual PhD student, and iterated until the vignette was back-translated without loss of fidelity. The novel third vignette, describing a "*really, really slow-minded*" person was modified from the smart vignette. Vignettes are in the Appendix.

In choosing the target images, we matched the characteristics of the stimuli reported in study-1i. These were not described in detail, but a closely related study (task ii study 1) described the images used as being of "*white men and women, normed for attractiveness… and professional dress in a sample of 29 adults recruited via Amazon's Mechanical Turk*" (p. S3). We followed these criteria in searching for images for the present study. Bian et al. (2017) did not specify the exact image search: we used Google images to search. Matching our sample, we sourced images of Chinese men and women, normed for attractiveness and in professional dress. We used the keywords "Chinese", "professional dress ID photos", "blue background colour" and "face front" to decrease image background confounds. The experimenter (YL) selected twenty female male images ranked highly in the search and four female Chinese PhD students were asked to each select two images for each gender which they thought matched in attractiveness and professional appearance. Two male and two female images matching most often were chosen as stimuli.

Our procedure differed from that of Bian et al. in three respects. First, in Bian et al. study-1i, different sets of four images were used for each vignette (Bian, personal communication 11/05/2018). We were unaware of this, interpreted the method as implying images were recycled across vignettes. We therefore used only one set of images. Second, Bian et al. (2017) presented the smart and nice conditions in a random order. To reduce effects of previous conditions on the critical brilliance stereotype effect, we presented the "*really, really smart*" condition first in all cases. For this reason, the images were novel to children at the point of answering the brilliance stereotype question, and thus the differences in methods cannot affect the data for the key task. Finally, children answered a paper-based questionnaire in their classrooms, rather than individually in a laboratory or classroom. Thus, the vignettes were printed with the four target images printed in a row at the bottom of the page.

2.1.3. Procedure

Children were tested in their classrooms. The vignettes were presented in the same session as a brief questionnaire for a separate study. All studies were approved by the Research Ethics Committee of University of Edinburgh (reference number 172–1819/7, approved on 3rd Nov 2016). After gaining consent, the experimenter asked the children

Table 1
Matches and differences between Bian et al. (2017) study-1i and the present studies 1, 2, 3 and 4.

	Bian et al. (2017)	The present paper			
	Study 1 task (i)	Study 1	Study 2	Study 3	Study 4
Participants	$N = 32$ (Equal numbers of boys and girls) Two additional groups of n = 32 tested at ages 5 and 6.	$N = 227$ (118 boys, 109 girls)	$N = 100$ (52 boys, 48 girls)	N = 200 (111 boys, 89 girls)	$N = 210$ (119 boys, 91 girls)
Age	Mean age = 7.44y, 6.50y, and 5.55y for each age group respectively.	Mean age = 10.10y.	Mean age = 7.96y.	Mean age = 10.04y.	Mean age = 8.03y.
Ethnicity	78 % European American, 7 % Asian American, 5 % African American, 3 % Latino or Hispanic, and 7 % multi-racial.	100 % Chinese	100 % from England	85 % from England, 4.5 % from Wales, 2 % Northern Ireland, 8 % Scotland and 0.5 % did not provide their nationality.	100 % from England
Test situation	Individually tested in a quiet room in the lab or at their school.	Children were asked to answer the questionnaire individually in their classroom.	Children were asked to do an online questionnaire.	Children were asked to do an online questionnaire.	Children were asked to do an online questionnaire.
Vignette stimuli	Two vignettes: One about a "really, really smart" person, one about a "really, really nice" person.	Ditto	Ditto	Ditto	Ditto
Dull vignette	Not used	Vignette about a "really, really slow-minded" person.	Ditto	Ditto	Ditto
Photo stimuli	Photos of adult males (n = 2) and female (n = 2) **Different photo-sets for nice and smart vignettes**	Photos of adult males (n = 2) and female ($n = 2$) targets drawn from a search engine and matched for attractiveness and professional status. **Photos reused between the three vignettes**	Photos of adult males (n = 2) and female (n = 2) **Different photo-sets for nice and smart vignettes**	Photos of adult males (n = 2) and female (n = 2) **Different photo-sets for nice and smart vignettes**	Photos of adult males (n = 2) and female (n = 2) **Different photo-sets for nice and smart vignettes**
DV	A choice of a person of the same gender as the participant was scored 1, otherwise, 0.	Ditto	Ditto	Ditto	Ditto
Exclusions	19 children excluded for matching fewer than 4/6 pre-screens regarding meaning of smart and nice. 3 children for refusal to finish. 1 for stereotype >2.5 SDs from the mean.	No exclusions.	No exclusions.	No exclusions.	No exclusions.
Order of testing	Smart and nice vignettes **presented in randomised order**.	Smart vignette **presented first**, "nice" vignette given second, followed by dull vignette.	Smart, nice and dull vignettes **presented in randomised order**.	Smart, nice and dull vignettes **presented in randomised order**.	Smart, nice and dull vignettes **presented in randomised order**.
School grades	NA	Self-reported by children	Reported by parents	NA	Reported by parents

to fill out a demographic questionnaire (including age, gender, and school grades). Children then read the first vignette (a "*really, really smart*" person), and chose an image from the four in front of them that was most likely to be the person described in the vignette. They then read vignette 2 (a "*really, really nice*" person), and asked to pick the most likely target image (the children were told that each image could be chosen more than once). Finally, they read the third vignette (a "*really, really slow-minded*" person) and identified the image most likely to match the vignette. At the end of the session oral praise and thanks were given to participants.

2.2. Results

All data and analysis code are open-access and available at https://osf.io/n8j7h/. All hypotheses with binary outcomes were tested by logistic regression using R's *glm* function with the binomial family, i.e., "logit" link function. Participant gender was a predictor. Age was covaried, but in each case made no substantive difference to results. The DV was the sex of chosen target (male vs. female). Differences in own-gender stereotype effects were conducted by recoding choices as same- or opposite-sex.

We first tested the prediction that girls would choose an opposite gender target while boys would select a same-gender target as a match to a vignette about a really smart person. This was rejected. Instead, 64 % of girls and 67 % of boys chose a target of their own gender, a highly significant bias towards stereotyping their own-gender as "*really, really smart*" vignette with an odds-ratio (*OR*) of 3.80 (CI_{95} [2.17, 6.79], $z = 4.61$, $p < .001$). There was no significant difference between boys and girls in this own-gender stereotype ($OR = 0.94$, CI_{95} [0.53, 1.65], $z = -0.23$, $p = .817$). Thus, contrary to prediction, girls (like boys) had strong stereotypes favouring their own, not the male gender for brilliance.

We next examined stereotypes at the other end of the spectrum: stereotypes about being very mentally slow. Again, the result was opposite to that predicted, but this time both girls and boys stereotyped a male target as the "*really, really slow-minded*" person (84 % of girls and 64 % of boys), a highly significant bias against males ($OR = 0.33$ CI_{95} [0.17, 0.63], $z = -3.29$, $p < .001$). This bias against boys as very dull, though strong in both genders, was significantly stronger in girls ($OR = 0.11$, CI_{95} [0.06, 0.20], $z = -6.71$, $p < .001$).

Finally, we tested the hypothesis that boys and girls would stereotype females as nicer. This time, the prediction was supported with 69 % of boys and 79 % of girls stereotyping a female target as nice, namely, boys were more likely to choose the opposite gender as nicest and girls were more likely to choose their own gender as nicest ($OR = 8.13$, CI_{95} [4.45, 15.33], $z = 6.65$, $p < .001$). The difference between boys and girls on stereotyping females as nicer was not significant ($OR = 1.56$, CI_{95} [0.84, 2.93], $z = 1.39$, $p = .164$).

2.3. Discussion

Three findings were of interest in study 1. First, boys and girls failed to show a brilliance stereotype – instead showing an own gender bias. Second, boys and girls both stereotyped niceness as a female trait. And

third, both genders stereotyped males as extremely dull. Despite replicating the bias towards females for niceness, then, we were unable to replicate the bias towards males for brilliance reported by Bian et al. (2017), and instead found evidence for a stereotype against males when it came to stereotypes about extreme low ability, despite a large sample ($n = 200$) and subjects at age (10 years) when these effects are predicted to be even stronger.

A candidate account of this failure to replicate might run to cultural differences. This is contra-prediction given the stronger male orientation of the cultural milieu of the children in our study 1 sample, but clearly replicating in a culture more similar to the US sample of Bian et al. (2017) would be informative. Additionally, we wished to address some minor procedural items in a second replication, introduced below.

3. Study 2

We wished to enhance the power of the design to explore children's gender stereotyping beliefs by making five changes. The first change was to conduct the study in a Western country (England) more culturally like the U.S. We also adopted the exact images from study-1i from Bian et al. (2017) for the smart and nice vignette and images for our active control testing gender stereotypes for low ability used images from task (iii) Bian et al. (2017). Thus, our images for the brilliance, nice and slow-minded vignettes were identical to those in the original study. To control presentation order effects, vignette order was randomised. Importantly, study 2 recruited participants of the age as the group showing the largest effect reported by Bian et al. (2017), i.e., 7-years-old. If brilliance stereotypes have the policy implications ascribed to them, this age should not be relevant, but given the initial failure to replicate the effect, we wished to make study 2 a direct replication. The final change made was to add a career interests scale at the end of the study allowing us to test if brilliance stereotype correlated with gendered career interests, which are well established (e.g., Ellis, 2011), but, to our knowledge, little studied at this young age in contemporary samples.

Study 2 hypotheses were:

1. Girls would be less likely to choose a same-gender target as a match for the vignette of a smart person compared to boys.
2. Both boys and girls would be more likely to choose a female target as a nice person.
3. Both boys and girls would be more likely to choose a female target as most likely to be slow-minded.
4. Statistically significant gender differences in occupations would emerge in the sample of 7-year-old children.
5. Brilliance stereotypes would predict interest in careers perceived to require high level of innate ability (e.g., scientist and engineer).

3.1. Method

3.1.1. Participants

In total, 100 parent-child dyads resident in England were recruited from an online volunteer pool (Prolific – an online research platform that allows researchers to recruit participants from diverse backgrounds including ethnicity, gender and more) during the spring term of 2019. Our child participants (52 were boys and 48 were girls) were aged from 6 years 3 months to 9 years 2 months old (mean = 7.96, SD = 0.86). Among the adult participants, 53 % of participants were mothers. Participants' family socioeconomic status (SES) was measured by asking the highest educational qualification attained by a parent in the family and the highest-earning parent's employment type. There were 54 % of family had a parent achieved a bachelor's or a higher degree and 64 % of families had a parent working as professionals, administrators, or officials.

3.1.2. Materials

Gender stereotype stimuli: The three vignettes used to test whether children have gender stereotypes about smartness, niceness and dullness were identical to those used in our study 1, with a few differences: First, because participants in our study 2 were English, we adopted the smart and nice vignettes and images directly from study-1i (Bian et al., 2017) without translation and back-translation. For the "*really, really slow-minded*" vignette we adopted a set of four images that were used in their task (iii) study 1.

Career interests were measured with six items assessing different career interests based on items in the table developed by Ellis (2011, p. 556). The first five items asked if children would be interested in a given career when they grew up (i.e., "*I would like to work as an executive manager/in law enforcement/as a nurse/as a scientist or engineer/as a university professor*"). Children were indicated their interest on a 5-point Likert scale: "Not at all interested" to "Very interested". Item six asked if children would like to spend more time at work or with parenting and childcare (work-life balance). Children marked their preference on a 7-point slider with 1 being long work hours and 7 being more time with parenting and childcare.

3.1.3. Procedure

Before starting to answer the survey, parents read the information sheet and asked children for their oral consent. Once parents and children both consented, parents clicked the accept button on the survey page to provide their and their children's formal consent.

After consenting, a welcome message and study instructions were shown to parents. These introduced the survey as including two sections, that they needed to finish the first section and then ask their child to come to the screen to answer the second section. The first section consisted demographic questions including parents' Prolific ID, gender and age, educational and employment status, child year and month of birth, and gender and 2018 key stage 1 national test score.

After filling out demographic information, a message reminded parents needed to ask their child to come to the screen to answer the following questions. Once the child sat in front of the screen, parents were asked to click the continue button on the screen to start the children's section. Children firstly were instructed that they would see three brief vignettes about a person and for each vignette, they need to choose an image that they thought to be the person described. Following the instruction, the first vignette about a "*really, really smart*" person with a set of four images were shown to children on the screen. Once children chose an image and clicked the arrow to the next page, they saw the second vignette about a "*really, really nice*" person with a second set of four images. Once children chose an image and clicked the arrow to the next page, they saw the third vignette about a "*really, really slow-minded*" person and a third set of four images. Finally, a new page with the 6-item career interests scale was shown to children. They then completed the interest items. Participants were then debriefed with a message showing that they had finished all the items, and they would receive a £1.50 as compensation once the experimenter verified their answers.

3.2. Results

Hypotheses were tested using logistic regression with children's gender predicting the selected target, controlling for age. As in study 1, for brilliance, the prediction that girls would show a significantly lower same-gender preference compared to boys was not supported ($OR = 0.83$ CI_{95} [0.36, 1.92], $z = -0.43$, $p = .670$). Instead, both boys (62 %) and girls (60 %) were significantly biased towards their own gender as brilliant ($OR = 2.40$ CI_{95} [1.08, 5.49], $z = 2.12$, $p = .034$).

At the opposite end of the brilliance dimension – a really slow-minded person – again contrary to prediction but in line with study 1, 69 % of boys and 79 % of girls chose a male target as the "*really, really slow-minded*" person: A significant bias ($OR = 0.11$, CI_{95} [0.04, 0.28], $z = -4.59$, $p < .001$) not differing significantly in magnitude between

boys and girls ($OR = 0.56$ CI_{95} [0.22, 1.40], $z = -1.22$, $p = .222$).

Stereotyping of females as nice was replicated: 83 % of boys and 81 % of girls choosing a female target as really nice: a significant bias ($OR = 20.36$ CI_{95} [7.66, 60.58], $z = 5.75$, $p < .001$), not differing in magnitude between boys and girls ($OR = 0.95$ CI_{95} [0.33, 2.71], $z = -0.10$, $p = .922$).

We next tested for gender differences in career interest at age 7 using six linear regression models with career interests as the dependent variables, children's gender as the independent variable and age as a covariate. Boys were coded as 1, girls coded as 2, so positive differences indicate a larger preference among girls than boys. Large gender differences were found for interest in working in law enforcement ($\beta = -0.48$ CI_{95} [−0.86, −0.09], $t = -2.47$, $p = .015$) and working as a nurse ($\beta = 0.96$ CI_{95} [0.61, 1.32], $t = 5.38$, $p < .001$). No other career interests showed significant differences (executive manager ($p = .764$, scientist or engineer: $p = .300$, university professor: $p = .539$, preferred work-life balance: $p = .484$). We tested if children's brilliance stereotypes predicted career interests, adding brilliance target and its interaction with participant gender to the regression. In no case was the predicted stereotype × interest interaction significant (p values 0.603 for "executive manager", 0.667 for "law enforcement", 0.282 for "nurse", 0.700 for "scientist or engineer", 1.000 for "university professor", and 0.412 for spending more time at work versus parenting activities).

3.3. Discussion

Study 2 showed similar results to study 1. The core hypothesis of a brilliance stereotype failed to replicate in this sample of 100 seven-year-old English children. We again found that both boys and girls stereotyped males as slow-minded. We also replicated the bias against boys as less nice. Children showed gender-linked differences in career interest in only two of the five career interests, but no evidence was found for any effect of brilliance stereotype on career interests.

Thus, in a similar culture and identical age, and with a sample more than three times the size (100 versus 32 seven year olds), and based on the small telescopes logic proposed by Simonsohn (2015), the result constitutes a valid failure to replicate with power to disconfirm the research hypothesis. Findings in Studies 1 and 2 reduce the likelihood that either culture or age affected the replicability of the brilliance stereotype, and decrease confidence in the finding, especially alongside significant negative stereotypes against males in the same study. We wished to conduct a third study to increase confidence in the nature of the finding and begin to understand better what information the method is conveying.

4. Study 3

Study 3 replicated study 2 with two modifications. First, we wished to test the findings in a sample the same age (10 years) as in study 1 where stereotype effects should be larger still. Second, we wished to test the findings in a larger sample gathered UK-wide. The hypotheses in study 3 were identical to those in study 2.

4.1. Method

4.1.1. Participants

In total 200 parent-child dyads were recruited from a UK-wide volunteer panel during the spring term of 2019. Our participants (111 boys and 89 girls) were aged from 6 years 3 months to 13 years 2 months old (mean = 10.04, SD = 2.20). Regarding SES, 51 % of parents achieved a bachelor's or a higher degree, 65 % of parents work as professionals, administrators, or officials, which is not atypical of modern UK.

4.1.2. Materials

The materials used in study 3 were identical to those used in study 2, except two questions asking parents' age and gender were deleted from the demographic survey.

4.1.3. Procedure

All experimental procedures were identical to those of study 2. Participants were compensated with 50p once their answers were verified by the experimenter.

4.2. Results

We first tested if girls were less likely than boys to choose a same-gender target as a match for the vignette of a really smart person. A significant difference was found ($OR = 0.33$, CI_{95} [0.18, 0.59], $z = -3.70$, $p < .001$) with 68 % of boys choosing a same-gender target compared to 42 % of girls. Thus, contrary to studies 1 and 2, girls were less likely to choose same-gender target as smart.

Regarding stereotypes for slow-mindedness, we again found a significant bias against males ($OR = 0.22$ CI_{95} [0.12, 0.39], $z = -4.91$, $p < .001$) with 63 % of boys and 73 % of girls choosing a male target as "*really, really slow-minded*". The bias of boys and girls did not differ ($OR = 0.63$ CI_{95} [0.34, 1.15], $z = -1.48$, $p = .138$).

Stereotypes for niceness replicated the perception of females as more likely to be nice seen in study 2 ($OR = 9.15$ CI_{95} [4.83, 18.07], $z = 6.60$, $p < .001$ with both girls (79 %) and boys (71 %) stereotyping females as nice and no significant gender difference in this bias ($OR = 1.59$ CI_{95} [0.82, 3.13], $z = 1.35$, $p = .176$).

Finally, we tested gender differences in career interests using linear regressions, controlling for age. Gender differences were found in interest in working as an executive manager ($\beta = -0.30$ CI_{95} [−0.58, −0.03], $t = -2.16$, $p = .032$), nurse ($\beta = 0.68$ CI_{95} [0.42, 0.94], $t = 5.18$, $p < .001$), and scientist/engineer ($\beta = -0.57$ CI_{95} [−0.84, −0.29], $t = -4.11$, $p < .001$). No gender difference was found in interests of working in law enforcement, university professor, nor in preferred work-life balance (p values 0.265, 0.712, and 0.194 respectively). Brilliance stereotypes showed no association with career interests when tested as in study 2, with non-significant interaction of participants' gender with the sex of the brilliance target ($p = .235$ for executive manager, 0.781 for law enforcement, 0.363 for nurse, 0.056 for scientist or engineer, 0.184 for university professor, and 0.605 for work-life balance).

4.3. Discussion

As in studies 1 and 2, significant stereotypes against boys as less nice and duller emerged and with equal strength in boys and girls. Unlike these studies 1 & 2, girls exhibited a bias to choose a male as the smart target. This, and the discrepancy with their significant bias in the same study to choose a male target as being very dull lead us to wish to further test the robustness of the results, but also to test the coherence of the vignette-matching method itself. Regarding gender differences in children's career interests, we found significant bias towards males in executive manager and scientist or engineer, and bias towards females in working as a nurse. These gender differences in career interests were compatible with relevant research showing large gender differences in interests (e.g., Master et al., 2021; Su & Rounds, 2015), which further replicated most strongly in countries with the lowest stereotype pressure (Stoet & Geary, 2018).

5. Study 4

Study 4 differed in four ways from study 3. The vignette method is vulnerable to confounds in the images chosen – imagine a set in which the women all wore lab coats. To manipulate and thus better understand this potential limitation, in study 4, the images for the smart vignette were swapped with those used the nice vignette. We also used a fixed M/F order for the images with the left-most image female with 50 % probability. Third, we presented the vignette and image probe on separate "pages" to match the presentation in Bian et al. (2017), and

added a manipulation check, asking participants to indicate the kinds of person described in the vignettes by typing a word that described them, to verify they understood the vignettes. These were inserted immediately prior to the career interest scale. Hypotheses for study 4 were identical to those in study 3.

5.1. Method

5.1.1. Participants

In total, 210 English parent-child dyads were recruited from a volunteer pool (Prolific; 119 boys and 91 girls, aged 6 years 4 months to 9 years 3 months old, mean = 8.03, SD = 0.88) during the spring term of 2019. For the family SES, 56 % parents achieved a bachelor's or a higher degree, and 65 % parents work as professionals, administrators, or officials.

5.1.2. Materials

The materials used in study 4 were identical to those used in study 2 and 3, with the exception that the images used for the nice and smart vignettes were swapped, images were presented in alternating order and appeared separately from the vignettes.

5.1.3. Procedure

The experimental procedure was identical to those used in study 3. Participants were compensated with 50p once their answers were verified by the experimenter.

5.2. Results

We first tested if girls were statistically significantly less likely than boys to choose their own gender as a smart person. Effect of own-gender on sex of the chosen target was statistically at chance, with 55 % of boys and 48 % of girls choosing an own-gender target ($OR = 0.76$ CI_{95} [0.44, 1.31], $z = -0.99$, $p = .320$). A glm predicting which gender was chosen as smartest (boy or girl) based on the gender of the respondent, and controlling for age showed no evidence for a gender difference ($OR = 1.20$ CI_{95} [0.69, 2.10], $z = 0.65$, $p = .517$).

By contrast, 75 % of boys and 76 % of girls chose a male target as very dull, replicating again this very large and highly significant stereotyping of boys as very dull ($OR = 0.11$ CI_{95} [0.05, 0.20], $z = -6.87$, $p < .001$). No gender difference was found in this stereotype ($OR = 0.93$ CI_{95} [0.49, 1.75], $z = -0.22$, $p = .827$).

Next, we tested whether children had a gender stereotype towards female gender for niceness, again, by using the same logistic regression, with same independent variable, dependent variable and covariate as in study 2 and 3. A small majority of both boys (53 %) of boys and girls (60 %) chose a female target as nicest ($OR = 1.71$ CI_{95} [0.99, 3.00], $z = 1.91$, $p = .056$, with no significant gender difference for this bias ($OR = 1.36$ CI_{95} [0.78, 2.37], $z = 1.08$, $p = .279$).

We tested gender differences in children's career interests, with differences emerging in nursing ($\beta = 0.73$ CI_{95} [0.48, 0.98], $t = 5.68$, $p < .001$) and working as a scientist or engineer ($\beta = -0.61$ CI_{95} [−0.88, −0.35], $t = -4.60$, $p < .001$). No gender difference was found in interests of working as an executive manager ($p = .796$), law enforcement ($p = .135$), university professor ($p = .726$), or for work-life balance ($p = .157$). Importantly, stereotypes assessed by picture matching about brilliance showed no hint of a stereotype × gender interaction effect (p values 0.910 for executive manager, 0.999 for law enforcement, 0.903 for nurse, 0.826 for scientist or engineer, 0.523 for university professor, and 0.815 for work-life balance).

5.3. Discussion

As in study 1, 2 and 3, both boys and girls had a gender stereotype towards female gender for niceness, and male gender for dullness. For brilliance stereotypes, a similar result as in study 3 was found, that most of both boys and girls chose a male target as a match for the really smart person. Thus, in over 400 British children (across studies 3 and 4), both boys and girls believed that boys were more likely to be highly intelligent, but also more likely to have very low intelligence. These findings might suggest that a stereotype of males as having higher variance in ability compared to females – a finding in some empirical studies (Deary et al., 2007) – but were not predicted and seem hard to reconcile with stereotype theory.

In summary, after four close replications, two studies have rejected the original finding that girls stereotype males as very smart, and two studies confirmed the finding. We reliably found support for an outcome at the low end of brilliance which contradicts the brilliance stereotype – namely both boys and girls stereotyped males, not females, as very dull. Moreover, we found no support for any connection of these stereotypes to career interests: the main driver of interest proposed for this suggested phenomenon. The large aggregate sample size (over 700 children in total across studies 1 to 4) makes it unlikely that substantive effects would be missed by low power. These contradictory results thus raise questions regarding the vignette matching method itself: Not only its reliability, but also the problematic finding that the method is typically deployed testing only a positive stereotype, but that, as found here, when a mirror-image negative stereotype is tested (dullness), it yields results incompatible with the positive stereotype result. This raises a potential confound: the matching task is also compatible with simple bias to pick a person of a certain gender whenever there is no strong evidence one way or the other, irrespective of what question is being posed. We wished to test this prediction and thus we conducted a follow-up experiment.

6. Study 5

Study 5 used similar methods to the previous studies. Children would read a vignette about a person with a certain ability level, then choose an image from a set of four as a match to the description. To clarify the results from studies 2 and 4, we wished to rule out the possibility that, when they have no other strong guide, boys and girls may choose male targets irrespective of the question asked. Or else if children were more likely to stereotype a female target as very average. Thus, we re-tested subject in studies 2 and 4, adding a vignette of a "very average" person, predicted that that male targets would be more likely to be chosen as a match to someone who is very average. This profile of choices is incompatible with stereotyping and is compatible with a default bias emerging in older children (especially girls) to attend to and choose male targets.

6.1. Method

6.1.1. Participants

A total of 268 participants (223 boys and 45 girls, mean age = 8.03, SD = 0.87) were recruited from studies 2 ($N = 84$) and 4 ($N = 184$) during the spring term of 2019.

6.1.2. Materials

Study 5 consisted of a vignette describing a person of average ability: "*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 average in how smart they are. They are right in the middle. This person figures out how to do things about as quickly and comes up with answers about as fast and about as good as most other people. With regards to being smart, this person is right in the middle*". This was followed by four images from task (iii) study 1 of Bian et al. (2017). A validity-check item asked children to indicate what kind of person the vignette was about, and children were asked to choose one answer from four options: "*very bright*", "*very dull*", "*averagely bright*", "*very unhappy*".

6.1.3. Procedure

After consenting, a welcome message and the study instruction were shown to children. Then children saw a vignette about a person who has average ability on the screen. Children were asked to read the vignette and a set of four images was shown on the next page. Children were asked to choose one image from four options that they thought to be the person described in the vignette. This was followed by the validity check item, presented on the following screen. Participants were compensated once their answers were verified by the experimenter.

6.2. Results

Responding to the average ability vignette, 51 % of boys and 53 % of girls chose a boy as very average. We tested for differences between boys and girls in their choice of an own-gender target, and minimal evidence a gender bias was found ($OR = 0.85$ CI_{95} [0.44, 1.61], $z = -0.51$, $p = .609$) and boys and girls did not differ significantly in rates of choosing targets matching their own-gender as average ability ($OR = 0.90$ CI_{95} [0.47, 1.71], $z = -0.32$, $p = .750$).

While the subjects and the materials used in study 2 and 4 were similar, children in study 2 selected their own gender as a match to smartness whereas in study 4 both males and females selected males as smartest on average. To try and understand this difference in result, we tested how these children would perform on their choices of the gender for average ability. As the participants in study 4 overwhelmingly stereotyped extreme dullness as a male trait, it does not seem these subjects were straightforwardly biased against females for brilliance. But perhaps, we wondered, these children would be more likely to stereotype a female target as very average compared to participants in study 2. However, no support for this idea was found: A logistic regression with target sex as the dependent variable and participant gender and study as independent variables, along with age as a covariate indicated no difference in the children's choices between study 2 and 4 ($z = 0.44$, $p = .661$).

6.3. Discussion

Study five tested whether children would be more likely to select a female target as having average ability than select a male target. Contrary to prediction, boys and girls exhibited no significant preference for a male target as a match to the average ability vignette. The results from our first five studies support two reliable and robust biases against boys – stereotyping them as very dull and less likely to be nice – along with a highly variable bias regarding brilliance. Before discussing the overall findings of the present research and implications for brilliance stereotypes, we wished to synthesise the results of studies 1–4 meta-analytically. These results are presented next.

7. Study 6

Across four studies, we tested whether children would be more likely to stereotype males as very smart, females as very nice and, in a novel extension, males as very dull. While we found consistent evidence that female targets were more likely to be chosen as very nice and male targets as very dull, we found varying results regarding the brilliance stereotype. Specifically, in two studies, children selected their own-gender targets as very smart while in two an overall bias to male targets as very smart emerged. Such inconsistent findings are expected in modestly powered research, and while our samples exceeded those used in other studies, the power of each study is likely not high, given that participants answer only one question on each kind of stereotype. To increase power, we combined the results of studies 1 through 4 using meta-analysis to bring to bear the total sample in the four studies. As this was relatively large (i.e., over 700 children), the resulting meta-analytic effects will permit a well-powered test of the effect of gender stereotypes towards smartness, niceness, and dullness. Using the *WebPower* package (Zhang & Yuan, 2018) to estimate power for a logistic regression in a single study with $n = 350$ subjects in each of the male and female groups (total $N = 700$) would have power of 99.999 % to detect an effect comparable to Bian's estimates in their study 1, i.e., where 75 % of males choose own gender compared to 50 % for females. Power remained high for detection of a reduced differences, e.g., 97 % power to detect a smaller 65 % vs. 50 % difference.

7.1. Method

The results in our studies 1, 2, 3 and 4 were combined for meta-analysis. The effect size measure used in these meta-analyses was the log odds ratio of boys who chose their own-gender as the smartest/nicest/dullest compared to girls who chose their own-gender. These analyses were done using the *metafor* package (Viechtbauer, 2010) in R.

7.2. Results

The sample consisted of all 737 participants from studies 1–4 (400 boys, 337 girls). A random effect meta-analysis was conducted for testing the overall effect size of the smart, nice and dull vignettes respectively. For smartness, a negative overall effect was found ($g = -0.42$, CI_{95} [−0.90, 0.06]; see Fig. 1 for the forest plot) but this effect was not significant ($p = .088$). No significant evidence of heterogeneity among the studies was found ($\tau^2 = 0.14$, $Q(3) = 7.32$, CI_{95} [0.00, 3.00], $p = .063$. Heterogeneity accounted for 58.83 % (I^2) of total variability). Thus, overall, there was no significant preference among boys and girls on choosing a male or female target as the match of a smart person.

For niceness, a significant overall effect was found ($g = 1.92$, CI_{95} [0.91, 2.93], $p = .0002$; see Fig. 2). Evidence for significant heterogeneity was apparent ($\tau^2 = 0.93$, $Q(3) = 27.05$, CI_{95} [0.21, 14.74], $p < .001$), accounting for 88.60 % (I^2) of total variability. Overall, however, both boys and girls had a significant stereotype of female targets being nice.

Finally, for dullness, the overall effect was significant ($g = -2.01$, CI_{95} [−2.40, −1.63], $p < .0001$; see Fig. 3). No support for study heterogeneity was found ($\tau^2 = 0.03$, $Q(3) = 3.41$, CI_{95} [0.00, 1.53], $p = .332$). Both boys and girls had a significant dullness stereotype for males being the match of a dull person.

7.3. Discussion

Our meta-analytic finding in study 6 suggested that boys and girls did not differ significantly in their likelihood of selecting their own or an opposite gender target as the smartest one (see Fig. 1). Turning to children's gender stereotypes towards niceness and dullness, the meta-analytic results suggested that children's own genders had strong and significant effects on their selection of a target for the niceness and dullness vignettes, with large preferences for both boys and girls to stereotype a girl as most likely to be nicest (see Fig. 2) and a boy as most likely to be very dull (see Fig. 3).

8. General discussion

In a five close replications and extensions of study-1i of Bian et al. (2017), totalling 737 children across different ages and cultures, we tested 1) Whether children aged from 7 to 10-years-old stereotype brilliance as a male trait and females as very nice; 2) Whether these effects (if found) would be present in children across different cultures; 3) Whether brilliance stereotypes predicted career interests; and 4) Whether the vignette matching method yields internally coherent results. While we consistently found that both boys and girls picked females as likely to be very-nice, support for stereotypes about brilliance were mixed at best. In studies 1 and 2, both boys and girls chose their own gender as the smartest (i.e., the opposite to the predicted effect for girls of this age). In studies 3 and 4 (UK samples totalling over 400

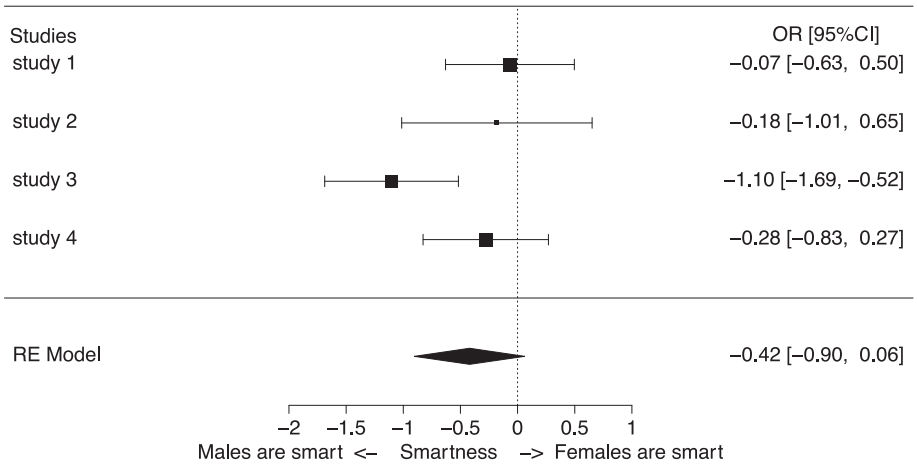

Fig. 1. Meta-analytic effect of gender on choosing one's own gender as smart (studies 1–4).

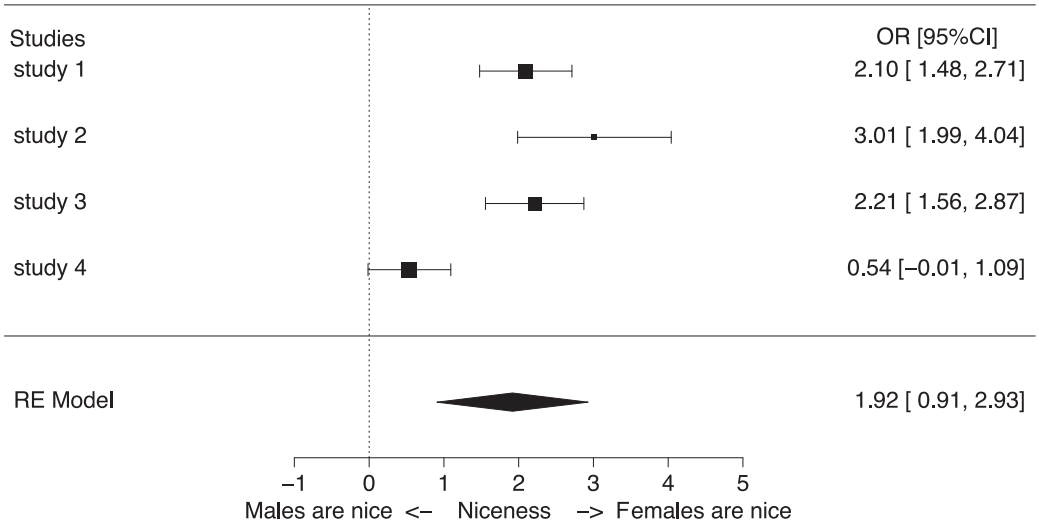

Fig. 2. Meta-analytic effect of gender on choosing one's own gender as nice (studies 1–4).

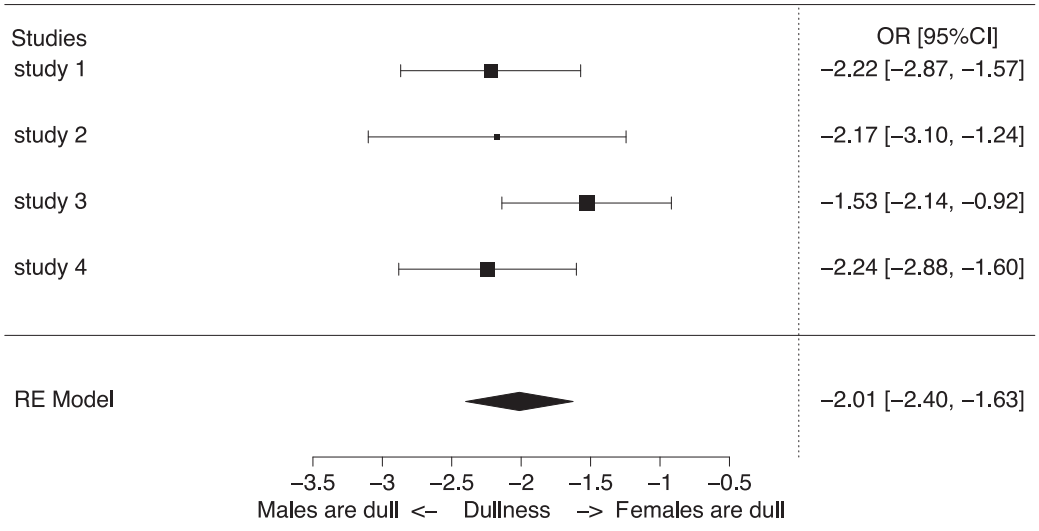

Fig. 3. Meta-analytic effect of gender on choosing one's own gender as dull (studies 1–4).

Table 2
The genders children selected for a match of the smart, nice and slow-minded person in Bian et al. (2017) and the present studies 1, 2, 3 and 4.

	Bian et al. (2017)	Present paper			
	Study 1	Study 1	Study 2	Study 3	Study 4
Smart	5 yrs: Own gender 6 yrs: Male 7 yrs: Male	Own gender	Own gender	Male gender	Male gender
Nice	5 yrs: Own gender 6 yrs: Female 7 yrs: Female	Female gender	Female gender	Female gender	Female gender
Slow-minded	NA	Male gender	Male gender	Male gender	Male gender

children), girls appeared to stereotype boys as smart (also see Table 2). These mixed findings and contradictions about brilliance make support for replication weak.

Studies 2 through 4 which investigated gender differences in children's career interests and the possible effects of brilliance stereotypes on children's career interests, we found significant and consistent gender differences among 7-to-10-year children in interest in working as a nurse (favoured by females). We also found gender differences in interests for other careers such as working as executive manager or working as a scientist or engineer, but these results were not consistent across the three studies. Importantly, across three studies, while we found at least one gender-linked difference in career interest (working as a nurse), we found no significant support for any effect of children's brilliance stereotypes on their career interests. This pattern of results, while compatible with sex differences in career interests (Halpern et al., 2007; Su & Rounds, 2015), is incompatible with stereotypes being a cause of these differences in career preference.

We replicated that finding that both genders stereotype males as less likely to be nice. Given considerable data supporting a genuine difference favouring females in traits such as agreeableness, sensitivity etc. (e.g., Del Giudice et al., 2012), it is possible that this effect, which both Bian et al. (2017) and the present studies were able to demonstrate, reflects a true mean difference, internalised as an accurate stereotype (Jussim et al., 2016; Löckenhoff et al., 2014) shared by both genders.

We next turn to the core hypothesis that females would show a bias to stereotype brilliance as a male trait. We tested this in children of different ages (7 and 10 years old), across different cultures (China and Britain) over two years (from 2017 to 2019) to establish reliable findings. We verified this gender stereotype in two of our British samples where males were more likely to be chosen as the smartest person by both genders. However, we failed to replicate this effect in the Chinese sample and one of the British samples – in both of these, both genders stereotyped brilliance as an own-gender trait. Thus, two of our close replication studies yielded a result opposite the original report by Bian et al. (2017) for children of this age, while two studies were compatible with brilliance stereotype but in the presence of an incompatible dullness stereotype and not evidence for effects on career interests. In addition, it seems that changes in culture over time are unlikely to be relevant to children's gender stereotypes towards brilliance. These divergent results lead us to explore the vignette matching method itself in more detail.

To validly index gender-linked stereotypes, the vignette method requires that picking a male target to match a brilliance vignette indicates a coherent stereotype. If males are stereotyped as brilliant, they should not only match a brilliance vignette, but also be *less likely* to be match a low brilliance vignette, i.e., they should also be less likely to be chosen as matching vignettes of a very dull, or even average target. The method as implemented originally, taps only one level of this range: brilliance. If, for instance, children were simply biased to pick opposite-sex images over same-sex, then any adjective would be matched to that sex: yielding incoherent stereotypes of the other sex as really smart, and really average, and really dull. We tested this assumption by adding a new "slow-minded" condition and by conducting a follow-up study using an "average person" vignette. For the "slow-minded" condition, we found that both genders more often picked a male as most likely to be "*really, really slow-minded*", this stereotype of males as more likely to be low in ability is, like the niceness stereotype, consistent with a genuine sex difference. Males are more likely to attain very low scores on measures of cognitive ability (Deary et al., 2007). However, if the outcome was driven by an intuitive awareness of this finding, we would have expected to also find a reliable bias against females at the "*really, really brilliant*" end of the distribution, which failed to emerge reliably *and* a bias for females being stereotyped as more likely to be average (i.e., falling closer to the mean). Study 5 tested this, yielding no significant difference for male versus female targets in a "very average" vignette, and with a *male* target again slightly preferred in this study by both boys and girls.

The results of these studies, then, are twofold. First, to be informative about bias, studies must measure both ends of the proposed stereotype. Second, having done this, the results of the present studies are incompatible with models of brilliance stereotypes as a simple bias. They are also incompatible with a stereotype of differing variances in a dullness to brilliance dimension in the sexes. Instead, the results indicate the vignette matching method can generate incoherent results such as males being smart, dull and average. Despite the samples typically being larger, outcomes fluctuated across samples, undermining the reliability of the method. Importantly, the brilliance stereotype failed to predict any of its proposed downstream consequences in the form of STEM linked career preferences, further undermining the idea that any stereotype effect is causal of material outcomes.

8.1. Limitations and suggestions for future studies

Besides gender stereotypes, other factors such as gender roles (Rudman & Phelan, 2010), motivational beliefs (Wang, 2012), social media and culture constraints (Jaoul-Grammare, 2023) may also play a role in children's career choices as they grow up. Future studies could consider these when examining relationships between gender stereotypes and career choices. In addition, a reviewer noted that when testing the effects of stereotypes, participants' self-held implicit biases (automatic associations based on previous experiences e.g., associating males and mathematics ability, females and verbal ability), rather than explicit stereotypes (conscious beliefs), significantly impacted children's performance on spatial ability tasks (Guizzo et al., 2019). Future studies could test if this is the case for children's career interests, and, in general, also explore what does influence women's and men's career choices (Ceci et al., 2014). Moreover, researchers (Stevenson & Stigler, 1994) suggested that Eastern cultures tend to view cognitive ability as a more malleable trait than Western cultures. Thus, when considering cultural differences in brilliance stereotypes, it would be worth testing whether children in the UK and China perceive the word "smart" differently and how this may influence their choices in the vignette matching method. Finally, it would be of value in future replication attempts to include large samples, preferably in a pre-registered and collaborative model (Clark & Tetlock, 2023) and including independently validated methods, rather than the vignette matching approach.

9. Conclusion

Without controls for low and average levels of a stereotyped trait, results of the method emerged as unstable but also may not be validly interpreted. When these controls are included, the method yielded incoherent results. Our results suggest the vignette matching method can generate incoherent and uninterpretable results, and, thus, is not a valid or reliable for testing gender stereotypes about innate ability. This

places results obtained using this method in doubt. Additional tests are needed, bringing to bear multiple competing explanations for differences in life choices, e.g., stereotypes, interests, talents, personality, and socio-cultural moderators (Bian et al., 2017; Deary et al., 2007; Del Giudice et al., 2012; Stoet & Geary, 2018).

CRediT authorship contribution statement

Yue Li: Writing – review & editing, Writing – original draft, Resources, Investigation, Data curation. **Timothy C. Bates:** Writing – review & editing, Supervision, Formal analysis.

Declaration of competing interest

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

Appendix

Vignette 1: Brilliance

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.

Vignette 2: Niceness

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.

Vignette 3: Dullness

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 slow-minded. This person figures out how to do things slowly and comes up with answers much slower and worse than anyone else. This person is really, really slow-minded.

Data availability

I have shared the link of our data in the manuscript.

References

Bian, L., Leslie, S. J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science, 355*(6323), 389–391. https://doi.org/10.1126/science.aah6524

Bowen, C., Wu, Y., Hwang, C., & Scherer, R. F. (2007). Holding up half of the sky? Attitudes toward women as managers in the People's Republic of China. *The International Journal of Human Resource Management, 18*(2), 268–283. https://doi.org/10.1080/09585190601102455

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological Science in the Public Interest, 15*(3), 75–141. https://doi.org/10.1177/1529100614541236

Clark, C. J., & Tetlock, P. E. (2023). Adversarial collaboration: The next science reform. In *Ideological and political bias in psychology: Nature, scope, and solutions* (pp. 905–927). Springer.

Cvencek, D., Greenwald, A. G., & Meltzoff, A. N. (2011). Measuring implicit attitudes of 4-year-olds: The preschool implicit association test. *Journal of Experimental Child Psychology, 109*(2), 187–200. https://doi.org/10.1016/j.jecp.2010.11.002

Cvencek, D., Greenwald, A. G., & Meltzoff, A. N. (2016). Implicit measures for preschool children confirm self-esteem's role in maintaining a balanced identity. *Journal of Experimental Social Psychology, 62*, 50–57. https://doi.org/10.1016/j.jesp.2015.09.015

Deary, I. J., Irwing, P., Der, G., & Bates, T. C. (2007). Brother-sister differences in the g factor in intelligence: Analysis of full, opposite-sex siblings from the NLSY 1979. *Intelligence, 35*(5), 451–456. ://000249214800004.

Del Giudice, M., Booth, T., & Irwing, P. (2012). The distance between Mars and Venus: measuring global sex differences in personality. *PLoS One, 7*(1), Article e29265. https://doi.org/10.1371/journal.pone.0029265

Dunham, Y., Baron, A. S., & Banaji, M. R. (2016). The development of implicit gender attitudes. *Developmental Science, 19*(5), 781–789. https://doi.org/10.1111/desc.12321

Ellemers, N. (2018). Gender stereotypes. *Annual Review of Psychology, 69*, 275–298. https://doi.org/10.1146/annurev-psych-122216-011719

Ellis, L. (2011). Identifying and explaining apparent universal sex differences in cognition and behavior. *Personality and Individual Differences, 51*(5), 552–561. https://doi.org/10.1016/j.paid.2011.04.004

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science, 16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Flore, P. C., Mulder, J., & Wicherts, J. M. (2019). The influence of gender stereotype threat on mathematics test scores of Dutch high school students: A registered report. *Comprehensive Results in Social Psychology, 3*(2), 140–174. https://doi.org/10.1080/23743603.2018.1559647

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology, 53*(1), 25–44. https://doi.org/10.1016/j.jsp.2014.10.002

Guizzo, F., Moe, A., Cadinu, M., & Bertolli, C. (2019). The role of implicit gender spatial stereotyping in mental rotation performance. *Acta Psychological (Amsterdam), 194*, 63–68. https://doi.org/10.1016/j.actpsy.2019.01.013

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8*(1), 1–51. https://doi.org/10.1111/j.1529-1006.2007.00032.x

Hilton, J. L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology, 47*, 237–271. https://doi.org/10.1146/annurev.psych.47.1.237

Jaoul-Grammare, M. (2023). Gendered professions, prestigious professions: When stereotypes condition career choices. *European Journal of Education, 59*(2). https://doi.org/10.1111/ejed.12603

Jussim, L., Crawford, J. T., Anglin, S. M., Chambers, J. R., Stevens, S. T., & Cohen, F. (2016). Stereotype accuracy: One of the largest and most replicable effects in all of social psychology. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 31–63). Psychology Press.

Leslie, S. J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science, 347*(6219), 262–265. https://doi.org/10.1126/science.1261375

Lippa, R. A. (2010). Sex differences in personality traits and gender-related occupational preferences across 53 nations: Testing evolutionary and social-environmental theories. *Archives of Sexual Behavior, 39*(3), 619–636.

Löckenhoff, C. E., Chan, W., McCrae, R. R., De Fruyt, F., Jussim, L., De Bolle, M., … Terracciano, A. (2014). Gender stereotypes of personality: Universal and accurate? *Journal of Cross-Cultural Psychology, 45*(5), 675–694. https://doi.org/10.1177/0022022113520075

Madon, S., Guyll, M., Aboufadel, K., Montiel, E., Smith, A., Palumbo, P., & Jussim, L. (2001). Ethnic and national stereotypes: The Princeton trilogy revisited and revised. *Personality and Social Psychology Bulletin, 27*(8), 996–1010. https://doi.org/10.1177/0146167201278007

Master, A., Meltzoff, A. N., & Cheryan, S. (2021). Gender stereotypes about interests start early and cause gender disparities in computer science and engineering. *Proceedings of the National Academy of Sciences, 118*(48). https://doi.org/10.1073/pnas.2100030118

National Bureau of Statistics of the People's Republic of China. (2017). China statistical year book. http://www.stats.gov.cn/tjsj/ndsj/2016/html/0411EN.jpg.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Rudman, L. A., & Phelan, J. E. (2010). The effect of priming gender roles on women's implicit gender beliefs and career aspirations. *Social Psychology, 41*(3), 192–202. https://doi.org/10.1027/1864-9335/a000027

Schuster, C., & Martiny, S. E. (2016). Not feeling good in STEM: Effects of stereotype activation and anticipated affect on women's career aspirations. *Sex Roles, 76*(1–2), 40–55. https://doi.org/10.1007/s11199-016-0665-3

Shapiro, J. R., & Williams, A. M. (2011). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles, 66*(3–4), 175–183. https://doi.org/10.1007/s11199-011-0051-0

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U. (2015). Small telescopes: detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Stevenson, H., & Stigler, J. W. (1994). *The learning gap: Why our schools are failing and what we can learn from Japanese and Chinese education*. Simon and Schuster.

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science, 29*(4), 581–593. https://doi.org/10.1177/0956797617741719

Su, R., & Rounds, J. (2015). All STEM fields are not created equal: People and things interests explain gender disparities across STEM fields. *Frontiers in Psychology, 6*, 189. https://doi.org/10.3389/fpsyg.2015.00189

Tsui, M. (2007). Gender and mathematics achievement in China and the United States. *Gender Issues, 24*(3), 1–11. https://doi.org/10.1007/s12147-007-9044-2

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3). https://doi.org/10.18637/jss.v036.i03

Wang, M. T. (2012). Educational and career interests in math: A longitudinal examination of the links between classroom environment, motivational beliefs, and

interests. *Developmental Psychology, 48*(6), 1643–1657. https://doi.org/10.1037/a0027247

Weisberg, Y. J., Deyoung, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the Big Five. *Frontiers in Psychology, 2*, 178–188. https://doi.org/10.3176/tr.2009.1.01

Zhang, Z., & Yuan, K.-H. (2018). *Practical statistical power analysis using Webpower and R*. ISDSA Press.