Do Gender-Related Stereotypes Affect Spatial Performance? Exploring When, How and to Whom Using a Chronometric Two-Choice Mental Rotation Task

Carla Sanchis-Segura*, Naiara Aguirre, Álvaro J. Cruz-Gómez, Noemí Solozano and Cristina Forn

Departamento de Psicología Básica, Clínica y Psicobiología, Facultad de Ciencias de la Salud, Universitat Jaume I, Castellón, Spain

It is a common belief that males have superior visuospatial abilities and that differences in this and other cognitive domains (e.g., math) contribute to the reduced interest and low representation of girls and women in STEM education and professions. However, previous studies show that gender-related implicit associations and explicit beliefs, as well as situational variables, might affect cognitive performance in those gender-stereotyped domains and produce between-gender spurious differences. Therefore, the present study aimed to provide information on when, how and who might be affected by the situational reactivation of stereotypic gender-science beliefs/associations while performing a 3D mental rotation chronometric task (3DMRT). More specifically, we assessed the explicit beliefs and implicit associations (by the Implicit Association Test) held by female and male students of humanities and STEM majors and compared their performance in a 3DMRT after receiving stereotype- congruent, incongruent and nullifying experimental instructions. Our results show that implicit stereotypic gender-science associations correlate with 3DMRT performance in both females and males, but that inter-gender differences emerge only under stereotype-reactivating conditions. We also found that changes in self-confidence mediate these instructions' effects and that academic specialization moderates them, hence promoting 3DMRT performance differences between male and female humanities, but not STEM, students. Taken together, these observations suggest that the common statement "males have superior mental rotation abilities" simplifies a much more complex reality and might promote stereotypes which, in turn, might induce artefactual performance differences between females and males in such tasks.

Keywords: gender stereotypes, stereotype threat, mental rotation, implicit association test, STEM

INTRODUCTION

Although in elementary, middle, and high school, girls and boys take math and science courses in roughly equal numbers, only around 20 percent of STEM graduates are women, a number that declines even further in the workplace (Hill et al., 2010; European Commission, 2016). Because STEM related careers are expected to grow faster than the average rate for all occupations

(National Science Board, 2010) and are among the best paid jobs (National Association of Colleges and Employers, 2009; European Commission, 2016), the underrepresentation of women in STEM studies severely increases the risk of exclusion and precarization in their future incorporation into the labor market. Yet, this is not only a problem for women. The absence of women from STEM education and careers is a waste of talent for those fields (European Commission, 2014; Norland et al., 2016) and also an economic cost for society as a whole. Indeed, it has been estimated that closing the gender gap in the STEM field would increase the EU GDP per capita by 0.7–0.9% in 2030 and by 2.2–3.0% in 2050 (Maceira, 2017). Accordingly, the gender segregation that characterizes the STEM field at the educational and professional level is seen with increasing social and institutional concern (Hill et al., 2010; European Commission, 2016).

The underrepresentation of women in STEM studies and professions has been traditionally considered a consequence of an innate higher proficiency of males in math and visuospatial abilities (Benbow et al., 2000; Baron-Cohen, 2004). Popularized through expressions such as "math is hard for girls" (Barbie, 1994, see Ben-Zeev et al., 2005) or "women cannot read maps" (Pease and Pease, 2004), the notion that males excel over females in these cognitive domains has become a widely shared social belief. However, scientific evidence does not support these claims and presents a much more complex reality (Ceci et al., 2009; Wang and Degol, 2013).

Thus, although older studies regularly identified a males' advantage in math performance (Glennon and Callahan, 1968; Maccoby and Jacklin, 1974; Benbow and Stanley, 1980, 1983), more recent large-sample studies and metaanalyses have revealed that gender differences in mathematics achievement tend to be inconsistent and small (d = 0.05, Lindberg et al., 2010; d = 0.06, Voyer and Voyer, 2014). Moreover, the size and even the direction of average gender differences in math performance widely varies among countries (ds ranging −0.42 to 0.40, Else-Quest et al., 2010; OECD, 2016) and they are correlated to national gender equity indexes (Reilly, 2012). Similarly, the proportion of females and males scoring at the 95th or 99th percentiles also differ among countries (Guiso et al., 2008; Machin and Pekkarinen, 2008) and they are highly correlated to national gender equality indexes, (Guiso et al., 2008; Hyde and Mertz, 2009). Finally, theoretical models demonstrate that the number of women in STEM studies and professions is substantially lower than that predicted from their math performance (Hyde and Mertz, 2009). Taken together, these and other data (reviewed in Spelke, 2005; Halpern et al., 2007; Ceci et al., 2009; Wang and Degol, 2013) strongly argue against the notion that males have innate or "hard-wired" superior math abilities that could account for the underrepresentation of women in STEM studies and occupations.

On the other hand, spatial ability is the cognitive domain in which differences between males and females are most commonly replicated and reported (Voyer et al., 1995; Hyde, 2014). Among the tasks in which such differences are observed, mental rotation tasks (MRT) produce the largest effects (Linn and Petersen, 1985; Halpern, 2013), which meta-analyses and

large-sample studies have estimated as being medium to large (Linn and Petersen, 1985; Voyer et al., 1995; Peters et al., 2007; Hyde, 2014). Conversely to what has been observed for math, gender differences in MRT are observed in all countries (Silverman et al., 2007) and their size do not seem to have declined over time (Masters and Sanders, 1993).

Given their high replicability, males-females' differences in MRT performance have been traditionally regarded as "sex differences" in visuospatial competence that arise from brain specializations imposed by the organizing actions of testosterone during prenatal development (Grimshaw et al., 1995; Baron-Cohen, 2004; Kempel et al., 2005; Peters et al., 2007; Vuoksimaa et al., 2010) and/or from the sexual division of labor in human early evolutionary history (Silverman and Eals, 1992; Silverman et al., 2007). However, despite its popularity both inside and outside the scientific realm, the empirical evidence that supports these views is far from conclusive (Fausto-Sterling, 2003; Jones et al., 2003; Jordan-Young, 2010). Indeed, there is a poor correlation between visuospatial abilities and the indirect indices of prenatal testosterone exposure (Puts et al., 2008) and the "sex differences" regularly observed in this cognitive domain are moderated by subjects' age (Geiser et al., 2008; Titze et al., 2010), experience and training (Uttal et al., 2012) as well as by task-related factors [e.g., time constrains (Voyer, 2011; Maeda and Yoon, 2013; kinds of stimuli (Alexander and Evardone, 2008; Ruthsatz et al., 2017)]. Furthermore, the biological and socio-cultural factors traditionally assigned to sex and gender are irremediably entangled and, in practice, it is not possible to separate their relative contribution to males and females' behaviors as they form a complex set of intertwined influences, referred to as sex/gender (Fausto-Sterling, 2003; Kaiser et al., 2009; Springer et al., 2012). Accordingly, the study of behavioral and cognitive similarities, and the differences between females and males, require more complex and integrative formulations than those provided by traditional categorical divides (e.g., male vs. female; biological vs. social, etc.), and should incorporate the interactions among predisposing, experiential and situational variables (Jordan-Young and Rumiati, 2012; Springer et al., 2012; Rippon et al., 2014).

In line with this, accumulated evidence indicates that factors traditionally assigned to "gender" might boost the differences in MRT performance ordinarily attributed to "sex." Indeed, it is well known that stereotypic beliefs about cognitive female-male differences can exert long-term effects on the acquisition of both interests and skills (Eccles, 1987; Bussey and Bandura, 1999), but may also have more immediate effects by affecting performance when situationally activated. Thus ever since childhood, self- or others' endorsement of commonly held stereotypic beliefs and implicit associations about genders (e.g., "science-male"; Nosek et al., 2002) reduce female performance in cognitive domains culturally viewed as "masculine" (e.g., math; Ambady et al., 2001; Beilock et al., 2010; Cvencek et al., 2011), and dwindle their interest in pursuing STEM-related studies and professions (Schmader et al., 2004; Kiefer and Sekaquaptewa, 2007a; Watt et al., 2012; Wang and Degol, 2013; Ertl et al., 2017).

Cognitive performance may also be affected by mere awareness of, rather than belief in, stereotypes of the different

abilities of targeted groups of persons. Thus when situational variables implicitly or explicitly activate stereotypes, they might induce a so-called 'stereotype threat' in the negatively stereotyped group members, and promote a reduction in their confidence and cognitive performance in those tasks perceived as being relevant to the activated stereotype (Steele and Aronson, 1995; Maass and Cadinu, 2003; Pennington et al., 2016). Accordingly, several studies have shown that the situational cues (e.g., received task instructions) that explicitly state or implicitly activate gender-related stereotypes reduce females' performance in experimental tasks and tests measuring visuospatial abilities (McGlone and Aronson, 2006; Moè and Pazzaglia, 2006; Campbell and Collaer, 2009; Hausmann et al., 2009; Heil et al., 2012; Neuburger et al., 2015). However by encouraging downward social comparisons with a denigrated outgroup, the same situational conditions to promote stereotype reactivation might boost self-confidence and performance in non-negatively stereotyped groups (Blanton et al., 1999; Walton and Cohen, 2003). Accordingly, the explicit or implicit activation of stereotypes on the allegedly different visuospatial abilities of males and females also results in increased male performance in MRT (Moè and Pazzaglia, 2006; Campbell and Collaer, 2009; Hausmann et al., 2009), and in other cognitive domains ordinarily perceived as "masculine" (e.g., math; Kiefer and Sekaquaptewa, 2007b).

Although these and other studies clearly establish that endorsement, implicit interiorization or situational activation of gender-related stereotypes might promote opposite effects in males and females' performance in math and visuospatial tasks, less is known about the individual variables that can moderate these effects (Maass and Cadinu, 2003). This is partly due to the generalized experimental treatment of females and males as being two distinct, but internally, homogenous groups and is also owing to focalization on average-based comparisons. Therefore, in the present study, we decided to compare subgroups of females and males with presumably different degrees of visuospatial abilities (STEM-Males ≥ STEM-Females > HUM-Males ≥ HUM-Females) and stereotypic gender-science beliefs/associations (STEM-Males = HUM-Females > HUM-Males > STEM-Females; see Nosek and Smyth, 2011) in a mixed design that allowed us to establish statistical relationships within, between and across groups.

More specifically, we assessed the relationship between the implicit and explicit gender-science biases held by a single cohort of female and male students of STEM and humanities' majors and their MRT performance after receiving stereotype-congruent ("males will do better"), stereotype-incongruent ("females will do better") or stereotype-nullifying ("no gender differences are expected") experimental instructions. After taking into account the results of previous studies, we hypothesized that 3DMRT performance should relate to the interactive effects between the academic trajectory (STEM vs. humanities) and situational variables (received instructions) rather than their raw categorization as females or males. In this way, by reactivating preexisting gender-related explicit beliefs/implicit associations, the received instructions should differentially modify 3DMRT performance in each group and promote specific constellations of between-group differences in each experimental condition. These differences were expected to be larger after receiving stereotype-congruent instructions, when task difficulty increased and among participants endorsing stereotypic views of females and males (a more specific hypotheses' formulation is provided in the different subheadings of the Results section). Moreover, correlational and linear-regression analyses were used to specifically explore whether the influence of gender-science biases on the participants' 3DMRT performance was: (1) similar in females and males; (2) similar in STEM and humanities students; (3) similar across the different experimental conditions. Finally, mediation analyses were used to test the a priori hypothesis that these gender-related biases influence 3DMRT performance by decreasing/ increasing the participants' self-confidence.

MATERIALS AND METHODS

This study was carried out in accordance with the recommendations of the ethical standards of the American Psychological Association. The protocol was approved by the Ethics Standards Committees of the Universitat Jaume I. In accordance with the Declaration of Helsinki all subjects gave written informed consent prior to participating.

Participants

Participants were university students at the Universitat Jaume I (Spain) who self-volunteered in response to an invitational email. To be included in the study, participants had to meet the following inclusion criteria: (1) to be in their first university year; (2) to maintain a consistent academic specialization in STEM or humanities since the last two high school years. The initial sample comprised 110 subjects, but five subjects were excluded from the statistical analysis due to incomplete reports of relevant demographic data or to violations of the inclusion criteria. Thus, 105 participants were included in this study (see **Table 1** for the sample details), which were subdivided into four groups according to their self-reported gender and college major. Two of these groups, STEM males (STEM-M; $N = 30$) and Humanities females (HUM-F; $N = 25$), had stereotypic gender-major combinations and the other two, STEM females (STEM-F; $N = 28$) and Humanities males (HUM-M; $N = 22$), had non-stereotypic gender-major combinations. All the participants signed informed consent and their collaboration was awarded with €20.

Measures

All the experimental tasks were programmed and presented in individual personal computers using the Millisecond Inquisit software package 4.0 (Millisecond©). The experimental tasks completed by all the participants included in presentation order: a demographic data form (on which participants reported their gender, age and university major), a mental rotation task, the Gender-Science implicit association test (IAT) and a single-item question to assess explicit beliefs on the suitability of females and males for scientific studies/professions.

TABLE 1 | The sample's demographic and academic characteristics.

	Males	Females
Computer sciences	10	8
Engineering	20	20
Total STEM	30	28
Journalism	8	8
Education	5	15
Other humanities studies	9	2
Total HUMANITIES	22	25
Total participants	52	53
Age	19.10 ± 1.20	18.96 ± 1.30

This table presents the number of participants of each gender and major. Male and female participants' ages (mean ± SD) are in the bottom row.

3D Mental Rotation Task (3DMRT)

To construct our 3DMRT task, we used the stimuli set developed and validated by Ganis and Kievit (2015). As in the classical paper-and-pencil mental rotation task designed by Shepard and Metzler (1971), each stimulus displays two abstract figures (a baseline object and a target object) composed of 7–11 cubes, arranged on four arms and connected end-to-end in a sequence. Ganis and Kievit (2015) provided eight different stimuli variations, grouped into two main categories: four "same" stimuli (those at which the baseline and target objects can be made to coincide with each other through a 0°, 50°, 100°, or 150° rotation on the vertical axis) and four "different" stimuli [whenever this is not possible, one figure arm (or more) is flipped]. Thus, by using the different rotation angles of a single figure, this set of stimuli allows the parametric manipulation of task difficulty. Furthermore, since the number of cubes and other characteristics of figures are identical in "different stimuli" and "same stimuli," the task cannot be carried out merely by taking into account the number of cubes in the objects or any other spurious cue.

Our 3DMRT comprised three phases, which correspond to three experimental conditions, each preceded by a different set of instructions (see Procedure). In all these experimental phases, we used six versions (2 categories × 3 rotation angles, 50°, 100°, and 150°) of eight different stimuli across 48 time-restricted trials (duration: 7.5 s; ITI: 0.5 s). These time parameters were the same as those used by Ganis and Kievit (2015) when validating the current stimuli set. Their inclusion was a necessary control to ensure a similar task performance pace for all the participants, which allows administering the necessary instructions before each experimental phase. In each trial, the computer screens displayed a baseline (left) and a target figure (right). The target figure could be a "same" or a "different" rotated (50, 100, or 150°) version of the baseline figure, but both figures had the same number of cubes and arms arrangement in all cases. The participants were asked to respond by pressing the "b" key (masked with a green tag) on their computer keyboard if they decided that the objects in a pair were the same, or by pressing the "n" key (masked with a red tag) if they decided that the two objects differed. Accuracy (number of correct responses) and latency to respond were automatically measured and, at the end of each phase, subjects were asked to provide (by means of a sliding bar of 10 discrete steps) an estimation of the percentage of correct responses achieved. This additional requirement provided an overall measure confidence in task execution, similar to that used by Estes and Felker (2012).

Implicit Association Test

The Implicit Association Test (Greenwald et al., 1998) is commonly used to assess implicit stereotypic associations, such as those which differentially link males to sciences and females to humanities (Nosek et al., 2002; Smyth and Nosek, 2015). For this study, the Gender-Science IAT script provided at http://www.millisecond.com/download/library/ (the Milisecond Test Library) was adapted to and translated into Spanish for this study (see Supplementary Table 1). This provided script implements the standard IAT procedure, which consists of 7 phases.

Phase 1 (Target category sorting training; 20 trials)

Participants are asked to discriminate and classify the target stimuli (male/female names) that appear at the center on the screen into one of the two categories (female/male) displayed in top corners by pressing the left ("E") or the right ("I") key on the computer's keyboard.

Phase 2 (Attribute sorting training; 20 trials)

Participants are asked to similarly classify attribute stimuli (majors) into one of the two categories (humanities/ STEM) displayed in the top corners of the computer's screen using the same keys than in the previous phase.

Phase 3 (Test block. Stereotype consistent target-attribute pairing; 20 trials)

Participants are asked to perform a combined categorization task by responding with the "E" key to both target and attribute stimuli belonging to the categories (female/humanities) placed on the left top corner and with the "I" key to both target and attribute stimuli belonging to the categories (male/ STEM) displayed on the right top corner of the computer screen.

Phase 4 (Test block; Stereotype consistent target-attribute pairing)

This phase is identical to the previous one but consists of 40 trials.

Phase 5 (Target category sorting training; 20 trials)

This phase is identical to phase 1 but the target sides are switched, so participants must classify male names by pressing the "I" key and the female names by pressing the "E" key. Twenty trials.

Phase 6 (Test block. Stereotype inconsistent target-attribute pairing; 20 trials)

This phase is identical to phase 3, but the category-attribute pairs are reversed. Thus, female names and STEM majors share the same response key ("E") whereas the male names and humanities majors are classified by pressing the "I" key.

Phase 7 (Test block. Stereotype inconsistent target-attribute pairing)

This phase is identical to the previous one but consists of 40 trials.

The provided script automatically counterbalances the order presentation of phases 3–4 and 5–6, so half of the participants

perform first the test blocks containing stereotype consistent trials and the other half the stereotype inconsistent test blocks. This script also automatically calculates the so-called *d*-scores (Greenwald et al., 2003). *d*-scores are standardized deviation scores that range between +2 and −2, whose interpretation is similar to that of Cohen's *d* statistics. Following the general convention, the IAT protocol used herein were arranged to provide positive *d* values for stereotype-consistent associations (e.g., "science = male/humanities = female") and negative d values for stereotype-inconsistent associations.

Explicit Beliefs

Participants were asked to explicitly declare and quantify their beliefs as to whether males and females differ in their suitability for "scientific tasks." We literally posed this question as "Who is better suited for science?" and the participants provided answers by a sliding bar of 10 discrete steps. Thus, setting the bar at 1 and 10 indicated that males/females were maximally suited for science, respectively (while setting it at 5 indicated no differences in this respect). Individual scores were computed as 5, minus the provided answer. In this way, and similarly to the IAT *d*-scores, positive (1–4) values quantified the presumed differences to favor males and negative (−1 to −4) values quantified the presumed differences to favor females.

Procedure

The experiment was carried out during six different experimental sessions, and each session involved 15–20 participants. As group composition might create a threatening environment for negatively stereotyped groups (Inzlicht and Ben-Zeev, 2000), we matched the participants in each session for gender and academic specialization into four similarly sized groups (see Supplementary Table 2). At the beginning of each session, three female experimenters greeted the participants in the laboratory, and they randomly assigned them to an individual desk equipped with a personal computer. After giving their informed consent, the experimenters asked the participants to fill in the *Demographic data* form and to wait for further instructions.

After all the participants had completed this first step, a senior researcher introduced the 3DMRT task, and informed them that it comprised three successive phases that should be initiated after her explicit instructions. Before starting each phase, and with the help of a video projection system, the researcher explained the task generalities (goal, response keys, etc.) and provided the specific instructions for each experimental condition. Phases were labeled and presented to the participants as "optimized for women," "optimized for men" and "neutral." The experimenter also emphasized that the selection of the stimuli of each phase was in accordance with previous studies in which they proved to be differently processed and resulted in enhanced performance for females or males, or had led to similar results between genders, respectively. These explanations came along with faked figures of brain scans and bar graphs, which displayed such differential results, which also appeared in the written instructions that the participants had to individually read on their computer screens before starting each phase. In order to increase distinguishability between

conditions, the stimuli of the "optimized for women," "optimized for men" and "neutral" conditions appeared on a pink, a blue and a white background, respectively (see Supplementary Figure 1). The order of these three experimental conditions was randomized across the six experimental sessions as a strategy to prevent any practice/learning effect (see Supplementary Table 3).

After finishing the 3DMRT task, the same leading researcher introduced the IAT as a word-sorting task by carefully avoiding any reference to gender or gender-related differences and provided the pertinent instructions for its completion. This cautious introduction to the IAT intended to minimize the chance of any carry-over effects from previous experimental phases. The provided instructions, which emphasized responding quickly, but accurately, also came in writing, shown on the individual screen of each participant's computer before starting the IAT.

Finally, participants were instructed to answer a single explicit question to assess their beliefs as to whether males or females are more capacitated for science (see *Explicit Beliefs* in the Measures section). Once they answered this question, participants were thanked and economically rewarded for their participation.
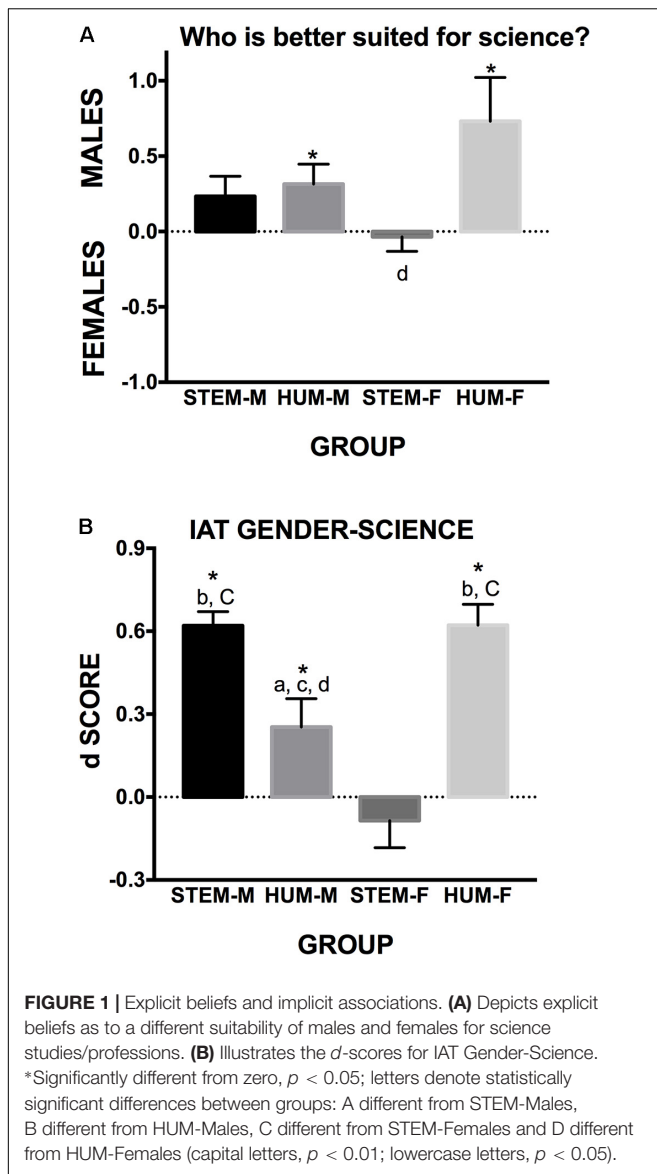
Data and Statistical Analyses

All the data included in the present study are provided as Supplementary Material (Supplementary Data Sheet S1). Data were analyzed using SPSS 23 (IBM Corp.) and PRISM 7.0 (GraphPad Inc.) for Mac OS X. Figures were constructed using PRISM 7.0 GraphPad Inc.).

One-sample Student's *t*-tests were used to evaluate whether or not the explicit beliefs and implicit associations held by each group of participants were significantly different from zero. Between-group differences in these variables as well as in the observed and expected 3DMRT performance were evaluated by design-appropriate ANOVAs, followed by Tuckey HSD *post hoc* comparisons. The relationship between explicit and implicit gender-related biases and observed/expected 3DMRT performance was initially evaluated by means of Pearson's *r* correlation index. However, in a second step, linear-regression-based procedures were used to explore in further detail the relationship between the IAT scores and observed/ expected 3DMRT performance scores. These more fine-grained analyses included: (1) the evaluation of a possible moderating effect of academic specialization on the influence of implicit gender-science associations over the observed and expected 3DMRT performance scores; (2) The evaluation of a possible mediatory role of confidence on the effects of these implicit associations on the observed 3DMRT performance.

RESULTS

Explicit Beliefs and Implicit Associations

H1: The participants, especially those of groups with gender-major stereotypic combinations, will hold explicit beliefs and implicit associations that preferentially link science to males and humanities to females.

FIGURE 1 | Explicit beliefs and implicit associations. **(A)** Depicts explicit beliefs as to a different suitability of males and females for science studies/professions. **(B)** Illustrates the *d*-scores for IAT Gender-Science. *Significantly different from zero, *p* < 0.05; letters denote statistically significant differences between groups: A different from STEM-Males, B different from HUM-Males, C different from STEM-Females and D different from HUM-Females (capital letters, *p* < 0.01; lowercase letters, *p* < 0.05).

To ascertain whether or not the participants held explicit beliefs as to a differential suitability of females and males for science, one-sample Student's *t*-tests were used. As shown in **Figure 1A**, the size of this belief significantly differed from zero in HUM-Males ($t_{21} = 2.309$, $p < 0.05$) and HUM-Females ($t_{24} = 2.520$, $p < 0.05$), and approached statistical significance in the STEM-Males group ($t_{29} = 1.756$, $p = 0.09$). In a second step, we analyzed the between-group differences by means of one-way ANOVA. The group factor reached statistical significance ($F_{3,101} = 2.86$; $p < 0.05$; $\eta_p^2 = 0.091$) which, as revealed by the Tukey HSD *post hoc* comparisons, was driven solely by a difference between the HUM-Females and the STEM-Females groups ($p < 0.05$, Cohen's $d = 0.739$).

On the other hand, one-sample Student's *t*-test revealed that STEM-Males ($t_{21} = 12.29$, $p < 0.001$), HUM-Males ($t_{21} = 2.46$, $p < 0.05$) and HUM-Females ($t_{24} = 8.24$, $p < 0.001$),

but not STEM-Females ($t_{27} = -0.872$, $p = 0.391$), exhibited a significant implicit "male-science/female-humanities" stereotypic association (**Figure 2B**). A one-way ANOVA ($F_{3,101} = 18.12$, $p < 0.001$; $\eta_p^2 = 0.350$) yielded a group effect on the size of this bias. The Tukey HSD *post hoc* comparisons revealed that this bias was larger in groups with gender-major stereotypic combinations than in those with non-stereotypic combinations (STEM-Males > HUM-Males: $p < 0.05$, Cohen's $d = 1.01$; STEM-Males > STEM-Females: $p < 0.01$, Cohen's $d = 1.74$; HUM-Females > HUM-Males: $p < 0.05$; Cohen's $d = 0.89$; HUM-Females > STEM-Females: $p < 0.01$; Cohen's $d = 1.57$). This bias was also larger in HUM-males than in STEM-Females ($p < 0.05$; Cohen's $d = 0.65$).

These results confirmed Hypothesis 1. However, explicit beliefs and implicit associations are two distinct cognitive and poorly correlated ($r = 0.147$, $p = 0.134$) constructs and, only in the second one, the groups with gender-major stereotypic combinations (STEM-Males and HUM-Females) clearly obtained higher bias scores than those with non-stereotypic combinations (STEM-Females and HUM-Males).

3DMRT Observed Performance

H2: The experimental groups will differ in their observed 3DMRT performance (STEM-Males ≥ STEM-Females > HUM-Males ≥ HUM-Females).

H3: The received instructions will differentially modify 3DMRT performance in each group and will hence lead to specific constellations of between-group differences in each experimental phase.
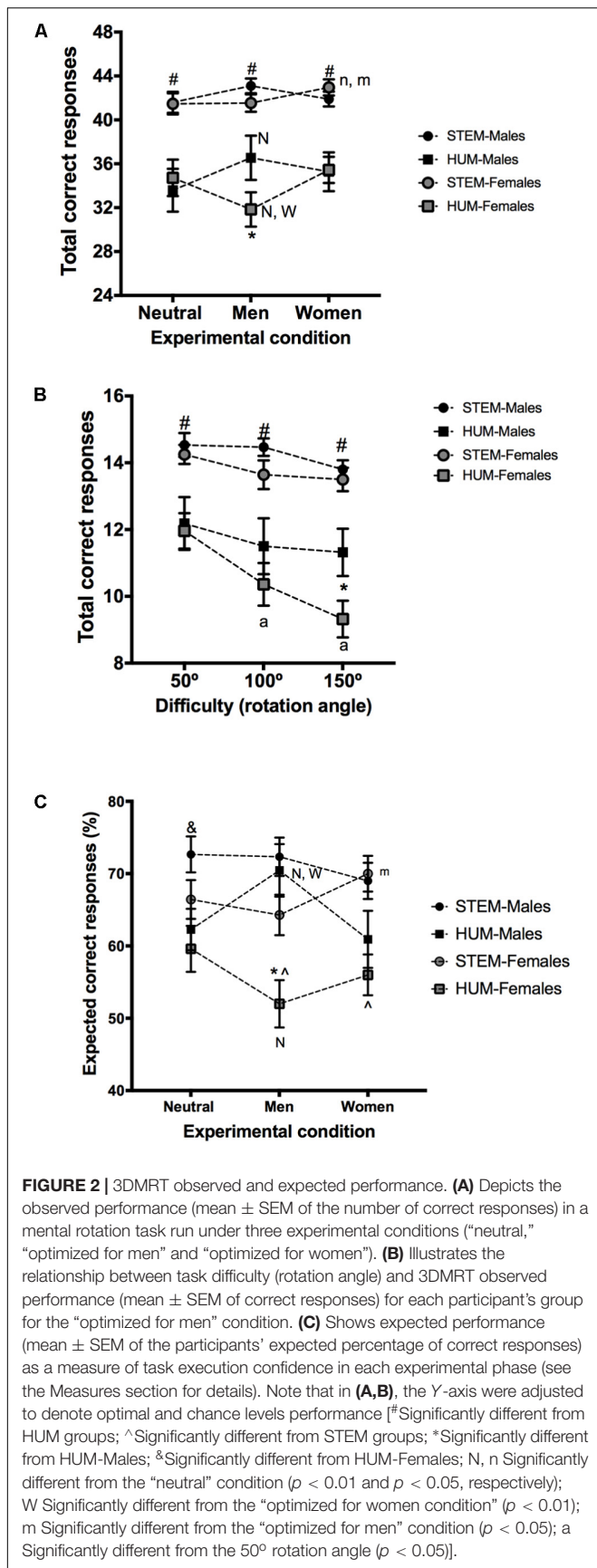
H4: The ability of the experimental instructions to promote gender-related differences in 3DMRT performance will increase with task difficulty.

3DMRT performance was assessed by two main variables: latency to respond and the number of correct responses. As latencies to respond did not differ between groups for any experimental condition (Supplementary Table 4), we do not discuss them further.

Regarding the number of correct responses (**Figure 2A**), a two-way repeated measures ANOVA revealed a significant effect for the group factor ($F_{3,101} = 17.16$, $p < 0.001$; $\eta_p^2 = 0.338$), but not for the condition factor ($F_{2,202} = 3.08$, $p = 0.18$), although the interaction between both factors was significant ($F_{6,202} = 2.98$, $p < 0.001$; $\eta_p^2 = 0.160$). This significant group × condition interaction allowed us to explore how the performance of each group varied across the three experimental conditions (within group comparisons) as well as the between group differences for each one.

Effects of the Received Instructions in Each Experimental Phase (Within Group Comparisons)

The Tukey HSD-based comparisons showed that the performance of the two STEM groups remained largely stable across the three experimental conditions. However, the less conservative Student's *t*-tests for related samples revealed a slight enhancement of STEM-Females' performance for the

FIGURE 2 | 3DMRT observed and expected performance. (A) Depicts the observed performance (mean ± SEM of the number of correct responses) in a mental rotation task run under three experimental conditions ("neutral," "optimized for men" and "optimized for women"). (B) Illustrates the relationship between task difficulty (rotation angle) and 3DMRT observed performance (mean ± SEM of correct responses) for each participant's group for the "optimized for men" condition. (C) Shows expected performance (mean ± SEM of the participants' expected percentage of correct responses) as a measure of task execution confidence in each experimental phase (see the Measures section for details). Note that in (A,B), the Y-axis were adjusted to denote optimal and chance levels performance [#Significantly different from HUM groups; ^Significantly different from STEM groups; *Significantly different from HUM-Males; &Significantly different from HUM-Females; N, n Significantly different from the "neutral" condition ($p < 0.01$ and $p < 0.05$, respectively); W Significantly different from the "optimized for women condition" ($p < 0.01$); m Significantly different from the "optimized for men" condition ($p < 0.05$); a Significantly different from the 50° rotation angle ($p < 0.05$)].

"optimized for women" condition compared to the other two ("neutral": $t_{27} = 2.075$, $p = 0.048$; Cohen's $d = 0.332$; "optimized for males": $t_{27} = 2.655$, $p = 0.013$; Cohen's $d = 0.352$). The same t-test based analysis did not reveal any significant variation in the STEM-Males group.

The experimental phase had more pronounced effects on the HUM groups. The intra-group Tukey HSD-based comparisons revealed that HUM-Females' performance dropped for the "optimized for men" condition to become lower than under the "neutral" ($p < 0.05$, Cohen's $d = -0.364$) and the "optimized for women" ($p < 0.05$ Cohen's $d = -0.528$) conditions. Conversely, HUM-Males displayed increased performance under the "optimized for men" condition, which became significantly higher ($p < 0.05$, Cohen's $d = 0.318$) than for the "neutral" condition.

Between-Group Differences in Each Experimental Phase

Under all the experimental conditions STEM-Females and STEM-Males outperformed HUM-Females and HUM-Males (Tukey HSD $p < 0.01$; Cohen's d, ranging from 0.98 to 1.83), but no differences between the two STEM groups were observed. HUM-Males outperformed HUM-Females for the "optimized for men" condition ($p < 0.05$, Cohen's $d = 0.557$), but not for any other experimental condition.

Taken together, these results confirmed Hypotheses 2 and 3 by showing that the 3DMRT performance of STEM-Males, STEM-Females, HUM-Males and HUM-Females differed, and that some of their differences (remarkably those between genders) only arose when receiving gender-loaded task instructions.

Task Difficulty and Gender-Related Differences in 3DMRT Observed Performance

Figure 2B depicts the relationship between task difficulty (rotation angle) and the observed 3DMRT performance for each participants group for the "optimized for men" condition (the only one at which we observed gender-related differences). A two-way repeated measures ANOVA yielded significant group ($F_{3,101} = 17.16$, $p < 0.001$; $\eta_p^2 = 0.338$), rotation angle ($F_{2,202} = 14.90$, $p < 0.001$; $\eta_p^2 = 0.128$) and interaction ($F_{6,202} = 2.29$, $p < 0.05$; $\eta_p^2 = 0.064$) effects.

All the groups showed rotation-related decreases in performance but, as revealed by the Tukey HSD post hoc comparisons, this effect was statistically significant only in HUM-Females (50° vs. 100° $p < 0.01$, Cohen's $d = 0.557$; 50° vs. 150° $p < 0.001$, Cohen's $d = 0.977$). The between-group comparisons revealed that the two STEM groups outperformed both HUMs groups, regardless of the rotation angle ($p < 0.01$ in all cases). Moreover, when difficulty was maximal (150°) HUM-Females gave fewer correct responses than HUM-Males ($p < 0.05$, Cohen's $d = 0.656$), which hence confirms Hypothesis 4.

Participants' Expected 3DMRT Performance

H5: The received instructions will differentially modify the self-reported expected performance (confidence) in each

group, which will then result in specific patterns of between-group differences in each experimental phase.

Figure 2C depicts the participants' expected percentage of correct responses for each experimental condition. A two-way repeated measures ANOVA yielded significant effects for the group factor ($F_{3,101} = 6.94$, $p < 0.001$; $\eta_p^2 = 0.171$) and for the group × experimental condition interaction ($F_{6,202} = 5.36$, $p < 0.01$; $\eta_p^2 = 0.137$). This significant group x condition interaction allowed us to explore how this self-reported index of the participants confidence varied within each group across the three experimental conditions as well as the between-group differences in this variable under each experimental condition.

Effects of the Received Instructions in Participants' Expected Performance in Each Experimental Phase (Within-Group Comparisons)

STEM males showed stable levels of expected performance across all the experimental phases. Conversely, all the other groups exhibited significant variations of expected performance depending on the received instructions. Thus HUM-Females' expected performance dropped under the "optimized for men condition" and hence became significantly lower than for the "neutral" condition; Tukey HSD $p < 0.05$, Cohen's $d = 0.471$). The opposite effect appeared for HUM-Males, with enhanced expected performance under the "optimized for men" condition (Tukey HSD $p < 0.01$, Cohen's $d = 0.566$ and Tukey HSD $p < 0.001$, Cohen's $d = 0.657$ compared to the "neutral" and the "optimized for women" conditions, respectively). Finally, the Student's t-tests for related samples, but not the Tukey HSD-based comparisons, revealed a selective increase in STEM-Females' expected performance under the "optimized for women" condition ($t_{27} = 2.741$, $p = 0.01$; Cohen's $d = 0.261$ and $t_{27} = 1.780$, $p = 0.08$ compared to the "optimized for men" and "neutral" condition, respectively).

Between-Group Differences in Each Experimental Phase

The Tukey HSD *post hoc* comparisons revealed that HUM-Females had the lowest expected performance (confidence) scores in all the experimental phases. Thus, for the "neutral" condition, only the HUM-Females and the STEM-Males groups significantly differed ($p < 0.01$, Cohen's $d = 0.881$). For the "optimized for women condition," HUM-Females reported lower expected performance scores than STEM-Males ($p < 0.01$, Cohen's $d = 0.932$), and also than STEM-Females ($p < 0.01$, Cohen's $d = 1.028$). Finally, under the "optimized for men" condition, the HUM-Females group differed from all the other groups: STEM-Males ($p < 0.001$, Cohen's $d = 1.314$), STEM-Females ($p < 0.02$, Cohen's $d = 0.789$) and HUM-Males ($p < 0.001$, Cohen's $d = 1.138$).

Taken together, the results of Sections "Effects of the Received Instructions in Participants' Expected Performance in Each Experimental Phase (Within-Group Comparisons)" and "Between-Group Differences in Each Experimental Phase" confirmed hypothesis 5.

Relationships Between Variables

H6: Observed and expected 3DMRT performance will be directly related between them, and will also show gender-dependent correlations with explicit beliefs and implicit associations preferentially linking males and science.

Observed and expected 3DMRT performance directly correlated with one another: ("neutral" condition $r = 0.536$, $p < 0.000$; "optimized for the men" condition $r = 0.596$, $p < 0.000$; "optimized for the women" condition $r = 0.468$, $p < 0.000$). Moreover, these performance-related variables correlated in a gender-dependent manner with the explicit and, more notably, the implicit "gender-science" biases (**Table 2**).

The **Table 2** results confirm Hypothesis 6 and also show that the implicit "male-science/female humanities" associations are more closely related to 3DMRT performance than explicit beliefs. These results also indicate that the same "male-science/female-humanities" association might have opposite functional consequences on female and male 3DMRT performance.

In order to confirm this last observation, we calculated an IAT-derived "influence" index. More specifically, females IAT scores were multiplied by −1, and those of males by 1. This transformation does not change the strength of the implicit Gender-Science associations revealed by the IAT, but slightly modifies the interpretation of the performed correlations, which now provide an index of the expectable "influence" of these implicit gender-related associations on 3DMRT performance rather than a plain measure of their co-variation. As expected, this IAT-derived "influence" index correlated directly with the observed 3DMRT performance ("neutral" condition: $r = 0.246$, $p < 0.02$; "optimized for males" condition: $r = 0.425$, $p < 0.000$; "optimized for females" condition: $r = 0.319$, $p < 0.001$). Similar correlations were found for expected 3DMRT performance ("neutral" condition: $r = 0.204$, $p < 0.04$; "optimized for males" condition: $r = 0.400$, $p < 0.000$; "optimized for females" condition: $r = 0.277$, $p < 0.005$).

By means of this IAT-derived "influence" index, we sought to investigate three additional research questions:

(Q1) Does the implicit "male-science/female humanities" association equally affect 3DMRT observed and expected performance in STEM and humanities students?

To answer this question, we used the regression-based moderation testing procedure proposed by Baron and Kenny (1986). Because the highest correlations between IAT-derived "influence" scores and performance measures were observed for the "optimized for men," we focused on this condition. Regarding observed performance (**Figure 3A**), the slope of the regression line for the HUM group significantly differed from zero ($F_{1,45} = 4.47$, $p = 0.04$), unlike that calculated for the STEM group ($F_{1,56} = 1.01$, $p = 0.31$). These slopes showed a clear trend toward being significantly different between them ($Z = 1.48$, $p = 0.06$). Similarly, as shown in **Figure 3B**, the slope of the regression line for the expected performance of the HUM ($F_{1,45} = 16.25$, $p < 0.001$), but not that of the STEM groups ($F_{1,56} = 1.24$, $p = 0.26$), significantly differed from zero, and yielded a significant inter-groups difference in this case ($Z = 2.19$,

TABLE 2 | Correlations by gender.

	Explicit		Implicit	
	Females	**Males**	**Females**	**Males**
Correct responses "neutral" condition	**r = −0.277**	r = 0.053	**r = −0.299**	**r = 0.330**
	p = 0.044	p = 0.710	**p = 0.030**	**p = 0.017**
Correct responses "optimized for men" condition	**r = −0.366**	r = −0.030	**r = −0.433**	**r = 0.304**
	p = 0.007	p = 0.834	**p = 0.001**	**p = 0.029**
50°	**r = −0.311**	r = −0.042	**r = −0.352**	r = 0.260
	p = 0.023	p = 0.765	**p = 0.010**	p = 0.060
100°	**r = −0.280**	r = −0.146	**r = −0.356**	r = 0.124
	p = 0.042	p = 0.302	**p = 0.009**	p = 0.383
150°	**r = −0.327**	r = −0.215	**r = −0.452**	r = 0.239
	p = 0.017	p = 0.125	**p = 0.001**	p = 0.087
Correct responses "optimized for women" condition	**r = −0.276**	r = −0.153	**r = −0.470**	**r = 0.345**
	p = 0.045	p = 0.280	**p < 0.000**	**p = 0.012**
Expected correct responses "neutral" condition	r = −0.139	r = 0.005	r = −0.018	**r = 0.303**
	p = 0.322	p = 0.996	p = 0.898	**p = 0.029**
Expected correct responses "optimized for men" condition	r = −0.253	r = −0.022	r = −0.260	r = 0.199
	p = 0.067	p = 0.878	p = 0.060	p = 0.157
Expected correct responses "optimized for women" condition	**r = −0.318**	r = −0.014	**r = −0.310**	**r = 0.337**
	p = 0.020	p = 0.920	**p = 0.024**	**p = 0.014**

Pearson's r index was used to quantify the correlation between Gender-Science explicit beliefs and implicit associations (IAT d scores) and the different indexes of 3DMRT performance. Statistically significant correlations (p < 0.05) are highlighted in bold.

$p = 0.01$). These results confirmed that academic specialization is a significant moderator of the IAT Gender-Science "influence" on the observed and expected 3DMRT performance, and revealed that this bias exclusively affected HUM students.

Confirming these results, we found significant correlations between the IAT "influence" scores and the observed and expected performance in HUM, but not in STEM, students (**Table 3**). The same correlational analysis revealed that the "influence" of the implicit "male-science/female-humanities" association on observed and expected 3DMRT performance varied for the different experimental conditions (see below).

(Q2) When do implicit biases affect expected and observed 3DMRT performance?

Several results of the present study were suggestive of a specific effect of the implicit "male-science/female humanities" association on the observed and expected 3DMRT performance of HUM, but not of STEM, students for the "optimized for men" condition. In order to confirm these effects and to explore their specificity, we ran a series of regression analyses.
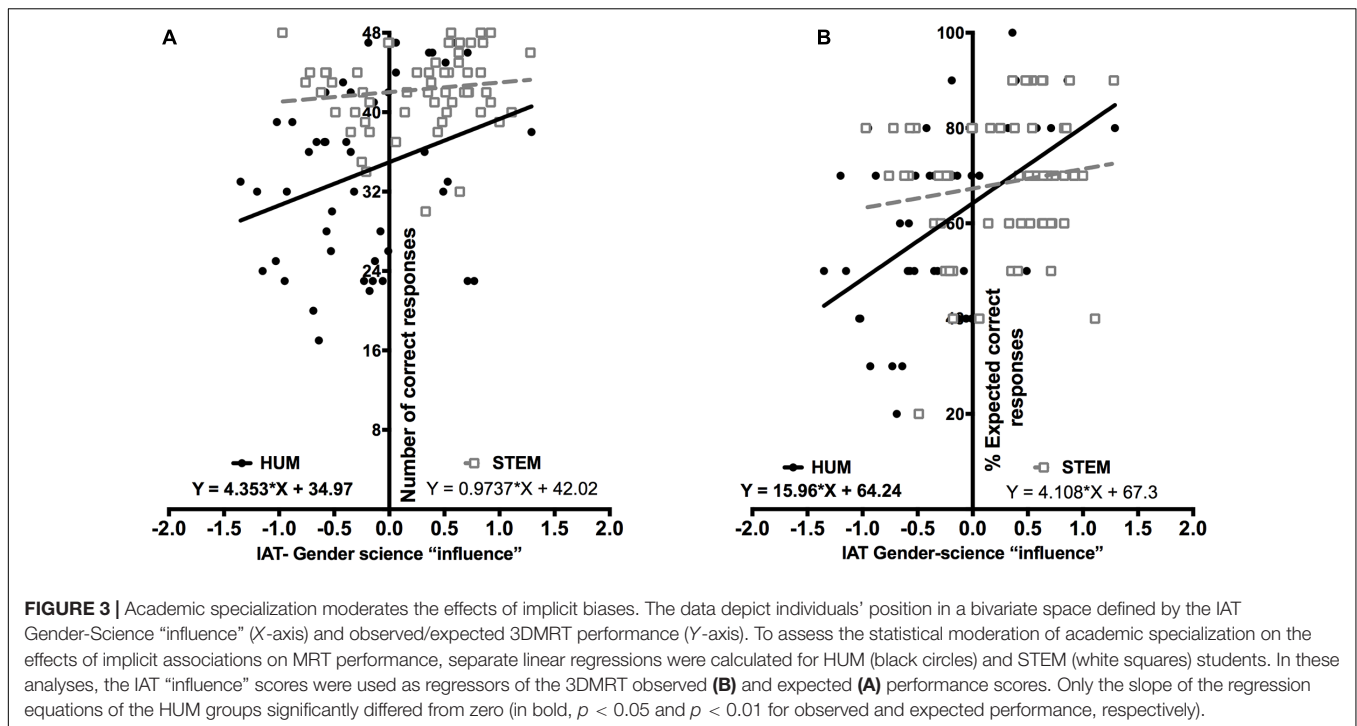
As shown in **Table 4A**, the IAT "influence" scores (but not gender, age, university major, or explicit beliefs) achieved statistical significance as predictors of the observed 3DMRT performance of HUM students for the "optimized for men" condition. Similarly, the IAT "influence" was the only significant predictor of the expected performance of HUM students under this experimental condition (**Table 4B**). Conversely, neither the IAT "influence," nor gender, age, university major or explicit beliefs achieved statistical significance as predictors of 3DMRT observed or expected performance of HUM students

for the "neutral" or the "optimized for women" conditions, nor as predictors of STEM students' performance. Therefore, a specific effect of the implicit "male-science/female humanities" association on HUM students' 3DMRT performance was confirmed.

(Q3) Does expected performance (confidence) mediate the effects of implicit "male-science/female-humanities" association on 3DMRT observed performance?

Previous studies (Steele, 1997; Walton and Cohen, 2003; Estes and Felker, 2012) have suggested that, by reducing confidence, gender stereotypes promote decrease female performance in "male cognitive domains," such as math or mental rotation. Therefore, we sought to explore whether our measure of confidence (expected performance) would mediate the "influence" of implicit gender-science associations on the 3DMRT observed performance. Taking into account all the previous results of our own study, we should solely observe this effect in HUM students (the only ones who displayed gender-related differences) and under the "optimized for men" condition (the only one at which we observed these differences). We tested this *a priori* hypothesis following the regression method for simple mediation described by Baron and Kenny (1986).

As shown in **Figure 4**, when the IAT "influence" and expected performance scores were simultaneously included in a single regression equation, only the second remained a strong predictor ($\beta = 0.674$, $p < 0.000$) of observed performance, while the predictive value of IAT "influence" scores' came very close to zero ($\beta = −0.047$, $p = 0.727$. That is, when the effect of confidence was taken into account, the influence of the

FIGURE 3 | Academic specialization moderates the effects of implicit biases. The data depict individuals' position in a bivariate space defined by the IAT Gender-Science "influence" (X-axis) and observed/expected 3DMRT performance (Y-axis). To assess the statistical moderation of academic specialization on the effects of implicit associations on MRT performance, separate linear regressions were calculated for HUM (black circles) and STEM (white squares) students. In these analyses, the IAT "influence" scores were used as regressors of the 3DMRT observed **(B)** and expected **(A)** performance scores. Only the slope of the regression equations of the HUM groups significantly differed from zero (in bold, $p < 0.05$ and $p < 0.01$ for observed and expected performance, respectively).

implicit "male-science/female-humanities" association in HUM students' 3DMRT performance was entirely eliminated. The specificity of this mediatory effect was ratified by testing several alternative models with the same regression-based procedure. These additional tests included assessing: (1) the same model in STEM students under the "optimized for men" condition; (2) the same model in HUM students under the "neutral" and "optimized for women" conditions; (3) the reverse model (observed performance mediates expected performance of HUM students under the "optimized for men" condition). As expected, the results of all these tests were negative (for details, see the figures and text included in the Supplementary Materials Image 1 file).

DISCUSSION

Our main results can be summarized as follows: (1) university students hold explicit beliefs and implicit associations that preferentially link science to males and humanities to females; (2) participants' science-related beliefs and associations vary according to an academic specialization (STEM *vs.* humanities) *per* gender interaction; (3) under experimental conditions specifically aimed to nullify or counteract these participants' stereotypic beliefs and associations, academic specialization was the only relevant predictor of 3DMRT performance; (4) when the received experimental instructions reactivated participants' stereotypes on gender-visuospatial abilities, explicit beliefs and, more significantly, gender-science implicit associations, were able to affect 3DMRT performance; (5) changes in confidence mediated these effects and academic specialization moderated them.

Explicit and Implicit Gender-Science Biases

The stereotypical notion of males being more suited for science was explicitly endorsed by the HUM-Males and, to a larger extent, by the HUM-Females groups (**Figure 1A**). As expected from previous studies (Greenwald and Banaji, 1995; Nosek et al., 2002), this explicit belief did not significantly correlate with the implicit Gender-Science associations revealed by the IAT and correlated solely with 3DMRT performance in females, but not in males (**Table 2**). This observation, together with the results of our linear regression-based analyses (see **Table 4**, but also Supplementary Tables 5–7) and those of some previous studies (Hyde et al., 1990; Schmader et al., 2004; Nosek et al., 2009), suggest that explicit gender-science beliefs are less accurate predictors and/or less powerful influencers of cognitive performance than implicit attitudes.

The participants also exhibited an implicit "science-male/humanities-female" association that correlated significantly with the 3DMRT performance in both females and males (**Table 2**). This bias was larger among the gender-major stereotypic combination groups (STEM-Males = HUM-Females; **Figure 1B**) than in those with non-stereotypic combinations (HUM-Males > STEM-Females). This observation is in agreement with cognitive-consistency principles (Nosek et al., 2002), with the results of a massive online survey conducted with college-educated people (Smyth and Nosek, 2015), and also with studies which show that STEM-majoring females hold weaker implicit gender-math stereotypes than both males from the same field and female and male humanities students (Nosek and Smyth, 2011; Smeding, 2012). Taken together, these studies suggest that the implicit "science-

TABLE 3 | Correlations between the IAT "influence" scores and 3DMRT performance in HUM and STEM students.

	HUM	STEM
Correct responses "neutral" condition	$r = -0.044$	$r = 0.171$
	$p = 0.771$	$p = 0.200$
Correct responses "optimized for men" condition	**$r = 0.300$**	$r = 0.121$
	$p = 0.04$	$p = 0.367$
50°	$r = 0.157$	$r = 0.123$
	$p = 0.802$	$p = 0.357$
100°	$r = 0.107$	$r = 0.123$
	$p = 0.473$	$p = 0.341$
150°	**$r = 0.287$**	$r = 0.127$
	$p = 0.05$	$p = 0.341$
Correct responses "optimized for women" condition	$r = 0.072$	$r = 0.123$
	$p = 0.630$	$p = 0.357$
Expected correct responses "neutral" condition	$r = 0.038$	$r = 0.005$
	$p = 0.802$	$p = 0.971$
Expected correct responses "optimized for men" condition	**$r = 0.515$**	$r = 0.092$
	$p < 0.000$	$p = 0.494$
Expected correct responses "optimized for women" condition	$r = 0.233$	$r = 0.056$
	$p = 0.114$	$p = 0.676$

Pearson's r correlation indices and associated p-values are provided. Statistically significant correlations (p < 0.05) are in bold.

male/humanities-female" association is highly related to academic/ professional career orientation. Moreover, since our study was conducted in freshman students, our results show that this implicit association is acquired before starting university and suggest that it might influence the students' choice of college major, hence contributing to the asymmetrical representation of girls and boys in STEM and humanities studies.

Interaction Between Implicit Associations and "Neutralizing," "Stereotypic" and "Counter-Stereotypic" Instructions and Its Effects on 3DMRT Performance

When interacting with situational cues (received instructions), the implicit "male-science/female-humanities" association was able to influence 3DMRT performance. As expected, the effects of this implicit bias were substantially smaller when arranging situational cues to nullify latent stereotypes ("neutral" condition) than under the experimental conditions which aimed to activate them (see the correlation values in **Tables 2**, **3**). Indeed, under this "stereotypes' neutralizing condition," STEM-students outperformed HUM-students, and no gender-related differences between these high and low performance groups were found (**Figure 2A**). Accordingly, regression analyses revealed that neither gender nor gender-related explicit beliefs or implicit associations were relevant predictors of 3DMRT performance under this experimental condition, which was significantly related only to academic specialization (see Supplementary Table 5). Thus our results confirm those of previous studies (Quinn and Spencer, 2001; Campbell and Collaer, 2009; Marchand and Taasoobshirazi, 2013), which also observed

that stereotype nullification by experimenter-controlled cues suppressed gender-related differences in visuospatial abilities and other cognitive domains for which males' superiority has been traditionally reported. As discussed below, these observations have important theoretical implications in the study and interpretation of "sex-differences" but also practical implications when trying to design educational interventions aimed to increase the representation of girls and women in STEM majors and professions.

The introduction of counter-stereotypic gender-related instructions ("optimized for women condition") did not substantially change the groups' 3DMRT performance. In this atypical situation, STEM-students displayed higher task accuracy than HUM-students but, once again, no gender-related differences were found (**Figure 2A**). Accordingly, linear regression-based analyses revealed that academic specialization, but not participants' gender, gender-related beliefs or implicit associations, became a significant predictor of 3DMRT performance under this experimental condition (see Supplementary Table 7). However, the "optimized for women" and "neutral" conditions were not identical as only the former promoted a slight enhancement of observed and expected performance in STEM-, but not HUM-, females (**Figures 2A,C**). The different reaction of STEM- and HUM-Females to counter-stereotypic instructions could lie in their distinct *a priori* beliefs and implicit associations (**Figure 1**). Thus, lacking any explicit or implicit Gender-Science bias, STEM-Females benefited from females' encouraging instructions, whereas the high and self-demoting biases held by HUM-Females made it impossible for them to benefit from the same positive endorsement. These observations replicate those made in previous studies (Moè and Pazzaglia, 2006; Wraga et al., 2006; Moè, 2009; Heil et al., 2012), which also found that instructions which stressed females'

TABLE 4 | Step-forward linear regression of the **(A)** observed and **(B)** expected performance of HUM students for the "optimized for men" condition.
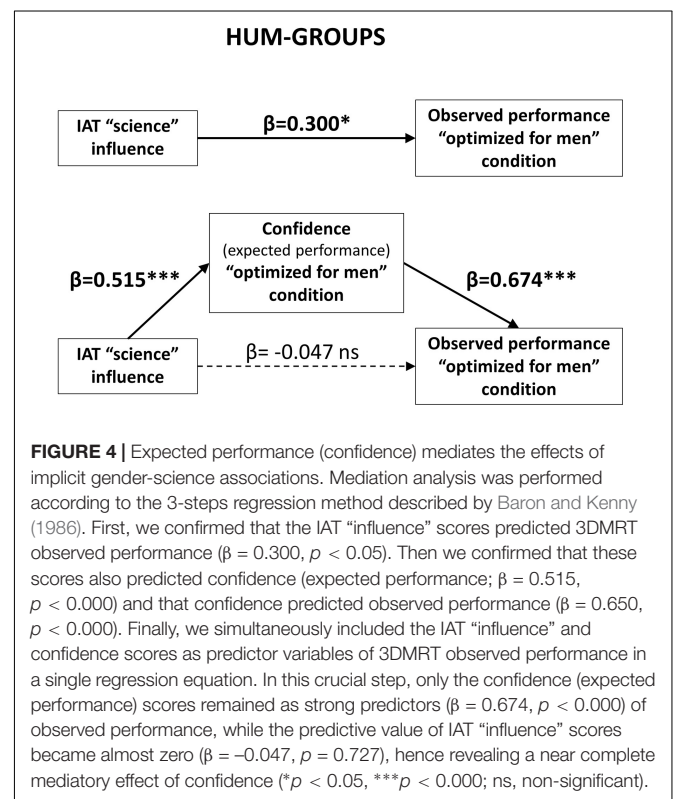
		Beta	t	p-value
(A) Observed performance				
Included in the	Constant	–	26.45	<0.000
model	IAT "influence"	0.300	2.11	0.04
Excluded from	Age	0.194	1.33	0.18
the regression	Gender	−0.107	−0.51	0.60
model	University major	−0.182	−1.24	0.21
	Gender-science explicit belief	−0.232	1.65	0.10
	Model summary	R	Adjusted R^2	p-value
		0.300	0.07	0.04
(B) Expected performance				
Included in the model	Constant	–	25.26	<0.000
	IAT "influence"	0.515	4.02	<0.000
Excluded from	Age	0.197	1.51	0.13
the regression	Gender	−0.27	−1.50	0.14
model	University major	−0.124	−1.24	0.35
	Gender-science explicit belief	−0.178	−1.39	0.16
	Model summary	R	Adjusted R^2	p-value
		0.515	0.249	<0.000

Separate stepwise forward linear regression analyses were conducted in the HUM and STEM groups to compare the predictive power of the IAT "influence" scores and of other possible predictors on the 3DMRT observed and expected performance at each experimental condition. For these analyses, nominal variables were coded as follows: gender (males = 1, females = 2) and university major (computer sciences = 1, engineering = 2, journalism = 3, education = 4, other humanities' studies = 5). Similar regression analyses were conducted for STEM students, but no variable entered in the model).

superiority in mental rotation tasks increased their performance, and that this increase was more marked for those females who did not sustain *a priori* beliefs about males' visuospatial superiority (Moè and Pazzaglia, 2006). However, in line with some (Moè, 2009; Heil et al., 2012), but not with other (Moè and Pazzaglia, 2006; Wraga et al., 2006) preceding studies, counter-stereotypic instructions did not bring about any change in STEM- or HUM-Males task performance. The reasons why these studies found distinct results remain unclear, but they might be indicative of a relatively weaker capacity of counter-stereotypic instructions to induce 3DMRT performance changes, especially if they result in a threat, and/or if subjects subscribe to the stereotypes contradicted by received instructions.

In contrast, stereotype-congruent instructions resulted in significant gender-related changes in 3DMRT performance. More specifically under the "optimized for men" condition, the 3DMRT accuracy of HUM-Females markedly diminished, but substantially increased in HUM-Males, and hence became significantly different between them and from their own performance under the other two experimental conditions (**Figure 2A**). Thus our results agree with those of previous studies, which have shown that experimental instructions which explicitly state females' inferiority in visuospatial abilities reduce females' performance in mental rotation tasks (Martens et al., 2006; Moè and Pazzaglia, 2006; Wraga et al., 2006; Campbell and Collaer, 2009; Hausmann et al., 2009; Moè, 2009; Heil et al., 2012), but increases males' performance (Moè and Pazzaglia, 2006; Campbell and Collaer, 2009; Hausmann et al., 2009).

In line with this, it has been proposed that confidence might underlie between gender differences in 3DMRT performance (Estes and Felker, 2012) as well as instructions-driven

HUM-GROUPS

IAT "science" influence — β=0.300* → Observed performance "optimized for men" condition

Confidence (expected performance) "optimized for men" condition

β=0.515*** β=0.674***

IAT "science" influence — β= -0.047 ns → Observed performance "optimized for men" condition

FIGURE 4 | Expected performance (confidence) mediates the effects of implicit gender-science associations. Mediation analysis was performed according to the 3-steps regression method described by Baron and Kenny (1986). First, we confirmed that the IAT "influence" scores predicted 3DMRT observed performance (β = 0.300, p < 0.05). Then we confirmed that these scores also predicted confidence (expected performance; β = 0.515, p < 0.000) and that confidence predicted observed performance (β = 0.650, p < 0.000). Finally, we simultaneously included the IAT "influence" and confidence scores as predictor variables of 3DMRT observed performance in a single regression equation. In this crucial step, only the confidence (expected performance) scores remained as strong predictors (β = 0.674, p < 0.000) of observed performance, while the predictive value of IAT "influence" scores became almost zero (β = −0.047, p = 0.727), hence revealing a near complete mediatory effect of confidence (*p < 0.05, ***p < 0.000; ns, non-significant).

performance changes in gender-stereotyped cognitive domains (Steele, 1997; Walton and Cohen, 2003). More specifically, it has been suggested that stereotype reactivation might induce

a self-confidence threat that disrupts task performance in the negatively stereotyped group (Schmader et al., 2008), but may induce a self-confidence boost that increases performance in the non-negatively stereotyped group (Blanton et al., 1999; Walton and Cohen, 2003). In agreement with this proposal, we observed that (probably by re-activating previously held stereotypic associations; **Figure 1B** and **Table 3**) the stereotype-congruent instructions of the "optimized for men" condition promoted disparate changes not only in the 3DMRT performance of the HUM-Females and HUM-Males groups (**Figure 2A**), but also in their confidence (**Figure 2C**), and that confidence mediates the influence of implicit associations on 3DMRT observed performance (**Figure 4**).

However, stereotype-congruent instructions do not uniformly affect females or males' performance as academic specialization moderates their effects (**Figure 3**). Accordingly, gender as a binary category did not come over as a significant predictor of 3DMRT performance for the "optimized for men" condition, which was instead mainly predicted from participants' academic specialization (Supplementary Table 6). Moreover, although the IAT "influence" scores were also significant predictors of 3DMRT performance under this experimental condition (Supplementary Table 6), their effects were restricted to HUM students (**Table 4**). Thus, despite having very different implicit Gender-Science associations (**Figure 1B**), the 3DMRT performance of STEM-Females and STEM-Males under the "optimized for men" condition was high, similarly to that observed for the "neutral" and "optimized for women" conditions and was indistinguishable between them (**Figure 2A**). These results, together with those of **Table 3** and Supplementary Figure 2, suggest that academic training or related academic experiences that result in a high level of task performance and/or confidence are able to suppress the influence of the gender-related implicit associations triggered by stereotypic experimental instructions. Our results and conclusions agree with those of a previous study (Hausmann, 2014), which showed that female arts, but not female STEM or male, students, reduced their 3DMRT performance after the reactivation of gender stereotypes. Similarly, gender stereotypes reactivation promotes a reduction of math performance of female psychology, but not of female engineering, students (Crisp et al., 2009).

Limitations and Implications

Under the different experimental conditions of the present study, academic specialization, but not the participants' gender, was the most relevant variable to predict 3DMRT performance. Our results also reveal that the within-gender differences that derived from academic specialization (STEM vs. HUM) are larger than those observed between genders. Indeed, we only observed between-gender differences in 3DMRT performance in HUM, but not STEM, students, and these differences solely emerged in response to stereotype-reactivating experimental instructions. These findings contrast with the common belief that males have better spatial abilities than females (Devlin, 2001; Blanton et al., 2002) and with the ordinarily reported higher performance of males in mental rotation tasks in studies that specifically aim to identify "sex differences" (Linn and Petersen, 1985; Silverman and

Eals, 1992; Grimshaw et al., 1995; Kempel et al., 2005; Peters et al., 2007; Silverman et al., 2007; Vuoksimaa et al., 2010; Halpern, 2013; Hyde, 2014; National Science Foundation, 2015).

At this respect, it should be noted that while we used a chronometric two-choice task, most research into sex differences in mental rotation use the pen-and-paper Mental Rotations Test (MRT) developed by Vandenberg and Kuse (1978). The MRT tends to produce larger sex differences (average $d = 1$) than chronometric tasks (average $d = 0.3$) and many studies using this second kind of procedures did not observe between genders differences (Voyer, 2011). Therefore, it might be argued that we did not observe the regularly reported gender differences because we did not use the "right" task for this. However, mental rotation chronometric tasks are as valid as psychometric tests (Voyer et al., 2006) and the MRT should not be considered as a benchmark when assessing and comparing the mental rotation abilities of males and females. In fact, the MRT does not seem to provide a pure measure of mental rotation abilities, and its singular ability to detect between gender differences might be related to the specific aspects of this test rather than to the responders' visuospatial abilities (Kerkman et al., 2000; Voyer and Hou, 2006; Hooven et al., 2008; Bors and Vigneau, 2011). Thus, while the results obtained with either chronometric or psychometric MRTs may differ and have a limited generalizability between each other, the use of a chronometric task does not limit the validity of the results observed in the present study.

Yet, it might be argued that, because gender differences observed in mental rotation chronometric tests are small (average $d = 0.3$), our study may lack the necessary statistical power to detect them. Therefore, the results of the present study should be interpreted with caution and replicated in a larger sample of participants. However, it should be noted that, although some small effects might have failed to reach statistical significance, these power limitations did not preclude by identifying the effects of academic specialization and stereotype-reactivating experimental instructions. This hence reveals that 3DMRT performance (at least as measured in our chronometric task) is much more dependent on these factors than on the participants' gender. Moreover, it should be also noted that the present study was not primarily intended to assess overall gender differences in visuospatial abilities but to identify a possible relationship between gender-science stereotypes and the participants performance in a specific 3DMRT task and that our study has power enough to detect even small to moderate correlations ($\approx\rho = 0.26$ if involving all participants and $\approx\rho = 0.32$ for any two subgroups of participants).

In this regard, it should also be emphasized that our study did not fail to identify between-gender differences in MRT performance but showed that these differences seem to emerge under particular testing conditions and involve some, but not all, male and female participants. Yet, precisely because gender differences in 3DMRT performance depend on task and respondents' characteristics (Sharps et al., 1994; Levine et al., 2005; Jansen-Osmann and Heil, 2007; Alexander and Evardone, 2008; Lippa et al., 2010), it might be concluded that the "sex differences" in mental rotation abilities do not arise from "sex" *per se,* but from its interaction with biographical

(e.g., academic specialization) and situational variables (e.g., received instructions). In this way, our results also argue against the attempt to explain the scarce representation of women in STEM studies and professions as a result of "hardly-wired" sex differences in visuospatial and math abilities. On the contrary, our results suggest that gender socialization and stereotypes might have a larger impact in situational performance in these cognitive domains and, thereby, in shaping the perceived competence and motivation to pursuit STEM careers. These conclusions fall in line with those of other studies that have indicated an important role of females and males' differential preferences, experiences and activities in the development of their visuospatial abilities (Flaherty, 2005; Feng et al., 2007; Sander et al., 2010; Nazareth et al., 2013; Moè, 2016). Moreover, the results and conclusions of our study also align with recent proposals which have suggested that in brain and behavior-related studies, sex and gender or, more properly, their composite resultant (sex/gender), should be considered a source of differential interactive effects with other variables rather than a binary-independent factor (Springer et al., 2012; Rippon et al., 2014; Joel and Fausto-sterling, 2016).

CONCLUSION

We observed that experimental instructions might reactivate implicit biases and promote increased/decreased 3DMRT performance, but training and/or other experiences related to academic specialization moderate these effects. In this way, the present study provides evidence about when (after receiving stereotype-congruent, but not stereotype-incongruent or stereotype-nullifying instructions), how (by increasing or reducing confidence) and who (HUM, but not STEM students) might be influenced by implicit gender-science associations while performing a chronometric mental rotation task. Our results also highlight that within-gender differences might be as large as, or even bigger than, those observed between genders and, therefore, that males and females are not two uniform populations (neither in their mental rotation abilities, nor in their

reaction to gender-stereotypes reactivation). Therefore, stating that "males have higher visuospatial abilities than females" is a misleading simplification that might contribute to perpetuate stereotypes. Those stereotypes and their detrimental impact on individual performance might progressively undermine the confidence and self-perceived competence of girls in cognitive domains ordinarily labeled as "masculine," hence reducing their interest in pursuing STEM-related academic and professional careers. However, as also suggested by some results of the present study (STEM/HUM females comparison), training and positive academic experiences in those cognitive domains promote resilience against pervasive gender-science stereotypes and provide a promising avenue when trying to enhance the number of women enrolled in STEM majors.

AUTHOR CONTRIBUTIONS

FUNDING

SUPPLEMENTARY MATERIAL

REFERENCES

Alexander, G. M., and Evardone, M. (2008). Blocks and bodies: sex differences in a novel version of the Mental Rotations Test. *Horm. Behav.* 53, 177–184. doi: 10.1016/j.yhbeh.2007.09.014

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

Barbie. (1994). "Responses of threat vs. challenge mediated arousal to stereotypes alleging intellectual inferiority," in *Gender Differences in Mathematics*, eds A. Gallagher and J. C. Kaufman (Cambridge, MA: Cambridge University Press).

Baron, R. M., and Kenny, D. A. (1986). The moderator-mediator variable distinction in social the moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers.* 51, 1173–1182. doi: 10.1037/0022-3514.51.6.1173

Baron-Cohen, S. (2004). *The Essential Difference*. London: Penguin Books.

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

Benbow, C. P., Lubinski, D., Shea, D. L., and Eftekhari-Sanjani, H. (2000). Sex differences in mathematical reasoning ability at age 13: their status 20 years later. *Psychol. Sci.* 11, 474–480. doi: 10.1111/1467-9280.00291

Benbow, C. P., and Stanley, J. C. (1980). Sex differences in mathematical ability: fact or artifact? *Science* 210, 1262–1264. doi: 10.1126/science.7434028

Benbow, C. P., and Stanley, J. C. (1983). Sex differences in mathematical reasoning ability: more facts. *Science* 222, 1029–1031. doi: 10.1126/science.6648516

Ben-Zeev, T., Carrasquillo, C. M., Ching, A. M. L., Kliengklom, T. J., McDonald, K. L., Newhall, D. C., et al. (2005). "Math is hard!" (Barbie, 1994). Responses of threat vs. challenge-mediated arousal to stereotypes alleging intellectual inferiority," in *Gender Differences in Mathematics*, eds A. M. Gallagher and J. C. Kaufman (New York, NY: Cambridge University Press), 48–72.

Blanton, H., Buunk, B. P., Gibbons, F. X., and Kuyper, H. (1999). When better-than-others compare upward: choice of comparison and comparative evaluation as independent predictors of academic performance. *J. Pers. Soc. Psychol.* 76, 420–430. doi: 10.1037/0022-3514.76.3.420

Blanton, H., Christie, C., and Dye, M. (2002). Social identity versus reference frame comparisons: the moderating role of stereotype endorsement 1. *J. Exp. Soc. Psychol.* 38, 253–267. doi: 10.1006/jesp.2001.1510

Bors, D. A., and Vigneau, F. (2011). Sex differences on the mental rotation test: an analysis of item types. *Learn. Individ. Dif.* 21, 129–132. doi: 10.1016/j.lindif.2010.09.014

Bussey, K., and Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychol. Rev.* 106, 676–713. doi: 10.1037/0033-295X.106.4.676

Campbell, S. M., and Collaer, M. L. (2009). Stereotype threat and gender differences in performance on a novel visuospatial task. *Psychol. Women Q.* 33, 437–444. doi: 10.1111/j.1471-6402.2009.01521.x

Ceci, S. J., Williams, W. M., and Barnett, S. M. (2009). Women's underrepresentation in science: sociocultural and biological considerations. *Psychol. Bull.* 135, 218–261. doi: 10.1037/a0014412

Crisp, R. J., Bache, L. M., and Maitner, A. T. (2009). Dynamics of social comparison in counter-stereotypic domains: stereotype boost, not stereotype threat, for women engineering majors. *Soc. Influ.* 4, 171–184. doi: 10.1080/15534510802607953

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

Devlin, A. S. (2001). *Mind and Maze. Spatial Cognition and Environmental Behavior.* Westport, CT: Greenwood Publishing Group, 328. doi: 10.1002/acp.938

Eccles, J. S. (1987). Gender roles and women's achievement-related decisions. *Psychol. Women Q.* 11, 135–172. doi: 10.1111/j.1471-6402.1987.tb00781.x

Else-Quest, N. M., Hyde, J. S., and Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychol. Bulletin* 136, 103–127. doi: 10.1037/a0018053

Ertl, B., Luttenberger, S., and Paechter, M. (2017). The impact of gender stereotypes on the self-concept of female students in STEM subjects with an under-representation of females. *Front. Psychol.* 8:703. doi: 10.3389/fpsyg.2017.00703

Estes, Z., and Felker, S. (2012). Confidence mediates the sex difference in mental rotation performance. *Arch. Sex. Behav.* 41, 557–570. doi: 10.1007/s10508-011-9875-5

European Commission (2014). *Cape Town: Carlos Moedas. Reaching Gender Equality in Science, Technology, Engineering and Mathematics.* Available at: https://ec.europa.eu/commission/2014-2019/moedas/announcements/reaching-gender-equality-science-technology-engineering-and-mathematics_en

European Comission (2016). *SHE Figures 2015.* Available at: https://ec.europa.eu/research/swafs/pdf/pub_gender_equality/she_figures_2015-final.pdf

Fausto-Sterling, A. (2003). "The problem with sex/gender and nature/nurture," in *Debating Biology,* eds S. J. Williams, L. Birke, and G. Bendelow (London: Routledge), 123–132. doi: 10.4324/9780203987681

Feng, J., Spence, I., and Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychol. Sci.* 18, 850–855. doi: 10.1111/j.1467-9280.2007.01990.x

Flaherty, M. (2005). Gender differences in mental rotation ability in three cultures: Ireland, Ecuador and Japan. *Psychologia* 48, 31–38. doi: 10.2117/psysoc.2005.31

Ganis, G., and Kievit, R. (2015). A new set of three-dimensional shapes for investigating mental rotation processes: validation data and stimulus set. *J. Open Psychol. Data* 3:e3. doi: 10.5334/jopd.ai

Geiser, C., Lehmann, W., and Eid, M. (2008). A note on sex differences in mental rotation in different age groups. *Intelligence* 36, 556–563. doi: 10.1016/j.intell.2007.12.003

Glennon, V. J., and Callahan, L. G. (1968). *Elementary School Mathematics: A Guide to Current Research.* Washington, DC: Association for supervision and Curriculum Development.

Greenwald, A., and Banaji, M. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychol. Rev.* 102, 4–27. doi: 10.1037/0033-295X.102.1.4

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *J. Pers. Soc. Psychol.* 74, 1464–1480. doi: 10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., and Banaji, M. R. (2003). Understanding and using the implicit association test: an improved scoring algorithm. *J. Pers. Soc. Psychol.* 85, 197–216. doi: 10.1037/0022-3514.85.2.197

Grimshaw, G. M., Sitarenios, G., and Finegan, J. A. (1995). Mental rotation at 7 years: relations with prenatal testosterone levels and spatial play experiences. *Brain Cogn.* 29, 85–100. doi: 10.1006/brcg.1995.1269

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). Culture, gender, and math. *Science* 320, 1164–1165. doi: 10.1126/science.1154094

Halpern, D. F. (2013). *Sex Differences in Cognitive Abilities,* 4th Edn. Hove: Psychology Press. doi: 10.4324/9781410605290

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

Hausmann, M. (2014). Arts versus science - Academic background implicitly activates gender stereotypes on cognitive abilities with threat raising men's (but lowering women's) performance. *Intelligence* 46, 235–245. doi: 10.1016/j.intell.2014.07.004

Hausmann, M., Schoofs, D., Rosenthal, H. E. S., and Jordan, K. (2009). Interactive effects of sex hormones and gender stereotypes on cognitive sex differences- A psychobiosocial approach. *Psychoneuroendocrinology* 34, 389–401. doi: 10.1016/j.psyneuen.2008.09.019

Heil, M., Jansen, P., Quaiser-Pohl, C., and Neuburger, S. (2012). Gender-specific effects of artificially induced gender beliefs in mental rotation. *Learn. Individ. Dif.* 22, 350–353. doi: 10.1016/j.lindif.2012.01.004

Hill, C., Corbett, C., and St Rose, A. (2010). *Why so Few? Women in Science, Technology, Engineering, and Mathematics title.* Washington, DC: American Association of University Women.

Hooven, C. K., Chabris, C. F., Ellison, P. T., Kievit, R. A., and Kosslyn, S. M. (2008). *The Sex Difference on Mental Rotation Tests Is Not Necessarily a Difference in Mental Rotation Ability.* New York, NY: Christopher F. Chabris.

Hyde, J. S. (2014). Gender similarities and differences. *Annu. Rev. Psychol.* 65, 373–398. doi: 10.1146/annurev-psych-010213-115057

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., and Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect: a Meta-Analysis. *Psychol. Women Q.* 14, 299–324. doi: 10.1111/j.1471-6402.1990.tb00022.x

Hyde, J. S., and Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proc. Natl. Acad. Sci. U.S.A.* 106, 8801–8807. doi: 10.1073/pnas.0901265106

Inzlicht, M., and Ben-Zeev, T. (2000). A threatening intellectual environment: why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychol. Sci.* 11, 365–371. doi: 10.1111/1467-9280.00272

Jansen-Osmann, P., and Heil, M. (2007). Suitable stimuli to obtain (no) gender differences in the speed of cognitive processes involved in mental rotation. *Brain Cogn.* 64, 217–227. doi: 10.1016/j.bandc.2007.03.002

Joel, D., and Fausto-sterling, A. (2016). Beyond sex differences: new approaches for thinking about variation in brain structure and function. *Philos. Trans. R. Soc. B* 371:20150451. doi: 10.1098/rstb.2015.0451

Jones, C. M., Braithwaite, V. A., and Healy, S. D. (2003). The evolution of sex differences in spatial ability. *Behav. Neurosci.* 117, 403–411. doi: 10.1037/0735-7044.117.3.403

Jordan-Young, R., and Rumiati, R. I. (2012). Hardwired for sexism? Approaches to sex/gender in neuroscience. *Neuroethics* 5, 305–315. doi: 10.1007/s12152-011-9134-4

Jordan-Young, R. M. (2010). *Brain Storm The Flaws in the Science of Sex Differences.* London: Harvard University Press.

Kaiser, A., Haller, S., Schmitz, S., and Nitsch, C. (2009). On sex/gender related similarities and differences in fMRI language research. *Brain Res. Rev.* 61, 49–59. doi: 10.1016/j.brainresrev.2009.03.005

Kempel, P., Gohlke, B., Klempau, J., Zinsberger, P., Reuter, M., and Hennig, J. (2005). Second-to-fourth digit length, testosterone and spatial ability. *Intelligence* 33, 215–230. doi: 10.1016/j.intell.2004.11.004

Kerkman, D. D., Wise, J. C., and Harwood, E. A. (2000). Impossible "mental rotation" problems. *Learn. Individ. Dif.* 12, 253–269. doi: 10.1016/S1041-6080(01)00039-5

Kiefer, A. K., and Sekaquaptewa, D. (2007a). Implicit stereotypes and women's math performance: how implicit gender-math stereotypes influence women's susceptibility to stereotype threat. *J. Exp. Soc. Psychol.* 43, 825–832. doi: 10.1016/j.jesp.2006.08.004

Kiefer, A. K., and Sekaquaptewa, D. (2007b). Implicit stereotypes, gender identification, and math-related outcomes: a prospective study of female college students: research report. *Psychol. Sci.* 18, 13–18. doi: 10.1111/j.1467-9280.2007.01841.x

Levine, S. C., Vasilyeva, M., Lourenco, S. F., Newcombe, N. S., and Huttenlocher, J. (2005). Socioeconomic status modifies the sex difference in spatial skill. *Psychol. Sci.* 16, 841–845. doi: 10.1111/j.1467-9280.2005.01623.x

Lindberg, S. M., Hyde, J. S., Petersen, J. L., and Linn, M. C. (2010). New trends in gender and mathematics performance: a meta-analysis. *Psychol. Bull.* 136, 1123–1135. doi: 10.1037/a0021276

Linn, M. C., and Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: a meta-analysis. *Child Dev.* 56, 1479–1498. doi: 10.2307/1130467

Lippa, R. A., Collaer, M. L., and Peters, M. (2010). Sex differences in mental rotation and line angle judgments are positively associated with gender equality and economic development across 53 nations. *Arch. Sex. Behav.* 39, 990–997. doi: 10.1007/s10508-008-9460-8

Maass, A., and Cadinu, M. (2003). Stereotype threat: when minority members underperform. *Eur. Rev. Soc. Psychol.* 14, 243–275. doi: 10.1080/10463280340000072

Maccoby, E. E., and Jacklin, C. N. (1974). *The Psychology of Sex Differences.* Stanford, CA: Stanford University Press.

Maceira, H. M. (2017). Economic benefits of gender equality in the EU. *Intereconomics* 52, 178–183. doi: 10.1007/s10272-017-0669-4

Machin, S., and Pekkarinen, T. (2008). Global sex differences in test score variability. *Science* 322, 1331–1332. doi: 10.1126/science.1162573

Maeda, Y., and Yoon, S. Y. (2013). A meta-analysis on gender differences in mental rotation ability measured by the Purdue spatial visualization tests: visualization of rotations (PSVT:R). *Educ. Psychol. Rev.* 25, 69–94. doi: 10.1007/s10648-012-9215-x

Marchand, G. C., and Taasoobshirazi, G. (2013). Stereotype threat and women's performance in physics. *Int. J. Sci. Educ.* 35, 3050–3061. doi: 10.1080/09500693.2012.683461

Martens, A., Johns, M., Greenberg, J., and Schimel, J. (2006). Combating stereotype threat: the effect of self-affirmation on women's intellectual performance. *J. Exp. Soc. Psychol.* 42, 236–243. doi: 10.1016/j.jesp.2005.04.010

Masters, M. S., and Sanders, B. (1993). Is the gender difference in mental rotation disappearing? *Behav. Genet.* 23, 337–341. doi: 10.1007/BF01067434

McGlone, M. S., and Aronson, J. (2006). Stereotype threat, identity salience, and spatial reasoning. *J. Appl. Dev. Psychol.* 27, 486–493. doi: 10.1016/j.appdev.2006.06.003

Moè, A. (2009). Are males always better than females in mental rotation? Exploring a gender belief explanation. *Learn. Individ. Dif.* 19, 21–27. doi: 10.1016/j.lindif.2008.02.002

Moè, A. (2016). Does experience with spatial school subjects favour girls' mental rotation performance? *Learn. Individ. Dif.* 47, 11–16. doi: 10.1016/j.lindif.2015.12.007

Moè, A., and Pazzaglia, F. (2006). Following the instructions!. Effects of gender beliefs in mental rotation. *Learn. Individ. Dif.* 16, 369–377. doi: 10.1016/j.lindif.2007.01.002

National Association of Colleges and Employers (2009). *Salary Survey.* Houston, TX: NACE.

National Science Board (2010). *Science and Engineering Indicators 2010 (NSB 10-01).* Arlington, VA: National Science Foundation.

National Science Foundation (2015). *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2015.* Arlington, VA: Special Report NSF, 17–310.

Nazareth, A., Herrera, A., and Pruden, S. M. (2013). Explaining sex differences in mental rotation: role of spatial activity experience. *Cogn. Process.* 14, 201–204. doi: 10.1007/s10339-013-0542-8

Neuburger, S., Ruthsatz, V., Jansen, P., and Quaiser-Pohl, C. (2015). Can girls think spatially? Influence of implicit gender stereotype activation and rotational axis on fourth graders' mental-rotation performance. *Learn. Individ. Dif.* 37, 169–175. doi: 10.1016/j.lindif.2014.09.003

Norland, M., Mora, T., and Kostschwar, B. (2016). *Is Gender Diversity Profitable? Evidence from a Global Survey.* Available at: https://ssrn.com/abstract=2729348

Nosek, B. A., Banaji, M., and Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dyn.* 6, 101–115. doi: 10.1037/1089-2699.6.1.101

Nosek, B. A., and Smyth, F. L. (2011). Implicit social cognitions predict sex differences in math engagement and achievement. *Am. Educ. Res. J.* 48, 1125–1156. doi: 10.3102/0002831211410683

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., et al. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10593–10597. doi: 10.1073/pnas.0809921106

OECD (2016). *PISA 2015 Results: Excellence and Equity in Education, PISA*, Vol. I. Paris: OECD Publishing. Available at: https://www.mecd.gob.es/inee/dam/jcr:54fd088e-f421-49c7-8ee2-852aff57682f/pisa2015-results-eng-vol1.pdf

Pease, A., and Pease, B. (2004). *Why Men Don't Listen and Women Can't Read Maps: How We're Different and What to Do About It.* Emmaus, PA: Rodale, INC.

Pennington, C. R., Heim, D., Levy, A. R., and Larkin, D. T. (2016). Twenty years of stereotype threat research: a review of psychological mediators. *PLoS One* 11:e0146487. doi: 10.1371/journal.pone.0146487

Peters, M., Manning, J. T., and Reimers, S. (2007). The effects of sex, sexual orientation, and digit ratio (2D:4D) on mental rotation performance. *Arch. Sex. Behav.* 36, 251–260. doi: 10.1007/s10508-006-9166-8

Puts, D. A., McDaniel, M. A., Jordan, C. L., and Breedlove, S. M. (2008). Spatial ability and prenatal androgens: meta-analyses of congenital adrenal hyperplasia and digit ratio (2D:4D) studies. *Arch. Sex. Behav.* 37, 100–111. doi: 10.1007/s10508-007-9271-3

Quinn, D. M., and Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *J. Soc. Issues* 57, 55–71. doi: 10.1111/0022-4537.00201

Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *PLoS One* 7:e39904. doi: 10.1371/journal.pone.0039904

Rippon, G., Jordan-Young, R., Kaiser, A., and Fine, C. (2014). Recommendations for sex/gender neuroimaging research: key principles and implications for research design, analysis, and interpretation. *Front. Hum. Neurosci.* 8:650. doi: 10.3389/fnhum.2014.00650

Ruthsatz, V., Neuburger, S., Rahe, M., Jansen, P., and Quaiser-Pohl, C. (2017). The gender effect in 3D-Mental-rotation performance with familiar and gender-stereotyped objects – a study with elementary school children. *J. Cogn. Psychol.* 29, 717–730. doi: 10.1080/20445911.2017.1312689

Sander, E., Quaiser-Pohl, C., and Stigler, C. (2010). Factors influencing the development of mental-rotation ability the role of socio-cultural background. *Int. J. Dev. Sci.* 4, 18–30. doi: 10.3233/DEV-2010-4102

Schmader, T., Johns, M., and Barquissau, M. (2004). The costs of accepting gender differences: the role of stereotype endorsement in women's experience in the math domain. *Sex Roles* 50, 835–850. doi: 10.1023/B:SERS.0000029101.74557.a0

Schmader, T., Johns, M., and Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychol. Rev.* 115, 336–356. doi: 10.1037/0033-295X.115.2.336.An

Sharps, M. J., Price, J. L., and Williams, J. K. (1994). Spatial cognition and gender: instructional and stimulus influences on mental image rotation performance. *Psychol. Women Q.* 18, 413–425. doi: 10.1111/j.1471-6402.1994.tb00464.x

Shepard, R. N., and Metzler, B. (1971). Mental rotation of three-dimensional objects. *Science* 171, 702–703. doi: 10.1126/science.171.3972.701

Silverman, I., Choi, J., and Peters, M. (2007). The Hunter-gatherer theory of sex differences in spatial abilities: data from 40 countries. *Arch. Sex. Behav.* 36, 261–268. doi: 10.1007/s10508-006-9168-6

Silverman, I., and Eals, M. (1992). "Sex differences in spatial abilities: evolutionary theory and data," in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, eds J. Barkow and L. Cosmides (New York, NY: Oxford University Press), 533–549.

Smeding, A. (2012). Women in science, technology, engineering, and mathematics (STEM): an investigation of their implicit gender stereotypes and stereotypes' connectedness to math performance. *Sex Roles* 67, 617–629. doi: 10.1007/s11199-012-0209-4

Smyth, F. L., and Nosek, B. A. (2015). On the gender-science stereotypes held by Scientists: explicit accord with gender-ratios, implicit accord with scientific identity. *Front. Psychol.* 6:415. doi: 10.3389/fpsyg.2015.00415

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science?: a critical review. *Am. Psychol.* 60, 950–958. doi: 10.1037/0003-066X.60.9.950

Springer, K. W., Mager Stellman, J., and Jordan-Young, R. M. (2012). Beyond a catalogue of differences: a theoretical frame and good practice guidelines for researching sex/gender in human health. *Soc. Sci. Med.* 74, 1817–1824. doi: 10.1016/j.socscimed.2011.05.033

Steele, C. M. (1997). A threat in the air. How stereotypes shape intellectual identity and performance. *Am. Psychol.* 52, 613–629. doi: 10.1037/0003-066X.52.6.613

Steele, C. M., and Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *J. Pers. Soc. Psychol.* 69, 797–811. doi: 10.1037/0022-3514.69.5.797

Titze, C., Jansen, P., and Heil, M. (2010). Mental rotation performance and the effect of gender in fourth graders and adults. *Eur. J. Dev. Psychol.* 7, 432–444. doi: 10.1080/17405620802548214

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., et al. (2012). The malleability of spatial skills: a meta-analysis of training studies. *Psychol. Bull.* 139, 352–402. doi: 10.1037/a0028446

Vandenberg, S. G., and Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Percept. Mot. Skills* 47, 599–604. doi: 10.2466/pms.1978.47.2.599

Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: a meta-analysis. *Psychon. Bull. Rev.* 18, 267–277. doi: 10.3758/s13423-010-0042-0

Voyer, D., Butler, T., Cordero, J., Brake, B., Silbersweig, D., Stern, E., et al. (2006). The relation between computerized and paper-and-pencil mental rotation tasks: a validation study. *J. Clin. Exp. Neuropsychol.* 28, 928–939. doi: 10.1080/13803390591004310

Voyer, D., and Hou, J. (2006). Type of items and the magnitude of gender differences on the Mental Rotations Test. *Can. J. Exp. Psychol.* 60, 91–100. doi: 10.1037/cjep2006010

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

Voyer, D., Voyer, S. S., and Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. *Psychol. Bull.* 117, 250–270. doi: 10.1037/0033-2909.117.2.250

Vuoksimaa, E., Kaprio, J., Kremen, W. S., Hokkanen, L., Viken, R. J., Tuulio-Henriksson, A., et al. (2010). Having a male co-twin masculinizes mental rotation performance in females. *Psychol. Sci.* 21, 1069–1071. doi: 10.1177/0956797610376075

Walton, G. M., and Cohen, G. L. (2003). Stereotype lift. *J. Exp. Soc. Psychol.* 39, 456–467. doi: 10.1016/S0022-1031(03)00019-2

Wang, M. T., and Degol, J. (2013). Motivational pathways to STEM career choices: using expectancy-value perspective to understand individual and gender differences in STEM fields. *Dev. Rev.* 33, 304–340. doi: 10.1016/j.dr.2013.08.001

Watt, H. M., Shapka, J. D., Morris, Z. A., Durik, A. M., Keating, D. P., and Eccles, J. S. (2012). Gendered motivational processes affecting high school mathematics participation, educational aspirations, and career plans: a comparison of samples from Australia, Canada, and the United States. *Dev. Psychol.* 48, 1594–1611. doi: 10.1037/a0027838

Wraga, M., Duncan, L., Jacobs, E. C., Helt, M., and Church, J. (2006). Stereotype susceptibility narrows the gender gap in imagined self-rotation performance. *Psychon. Bull. Rev.* 13, 813–819. doi: 10.3758/BF03194048