The Sex Difference in Mental Rotation Test Scores May Not Reflect a Difference in Mental Rotation Ability

Carole K. Hooven†
Department of Human Evolutionary Biology, Harvard University

Jonathan Wai
Autism and Developmental Medicine Institute, Geisinger Health System
Department of Psychology, Case Western Reserve University

Rogier A. Kievit
MRC-Cognition and Brain Science Unit
Fitzwilliam College, University of Cambridge

Peter T. Ellison
Department of Human Evolutionary Biology, Harvard University

Stephen M. Kosslyn
Minerva Schools at the Keck Graduate Institute

Christopher F. Chabris†*
Autism and Developmental Medicine Institute, Geisinger Health System
Institute for Advanced Study in Toulouse

†These authors contributed equally to the work.

*Address correspondence to:

Christopher F. Chabris
Geisinger Health System
120 Hamm Drive, Suite 2A, MC 60-36
Lewisburg, PA 17837 USA

Email: chabris@gmail.com

Running head: Sex differences on mental rotation tests

The largest reported sex difference in human cognition is found on mental rotation tests, which ask participants to compare pictures of three-dimensional objects and decide whether they depict the same or different objects. When the objects are the same, one can be rotated two- or three-dimensionally to match the other. Across cultures, males score up to one standard deviation higher than females on these tests. We administered two mental rotation tests to 123 participants and found that these higher scores likely do not reflect superiority in the process of mental rotation *per se*, but rather in other aspects of task performance. We found: (1) men are more likely than women to answer correctly when two objects are different, whereas women are more likely to answer incorrectly that they are the same; and (2) individual differences in confidence explain a considerable portion of the male advantage, but differences in spatial encoding ability do not. These results suggest more attention should be paid to individual differences in the various components of spatial ability and task performance, and have implications for evolutionary theories of sex differences in spatial cognition and for efforts to reduce sex differences in spatial ability, especially via training interventions.

The fact that males score higher than females on tests that require rotating objects in mental images is widely accepted in psychology. The most popular mental rotation test, developed by Vandenberg and Kuse[1] (hereafter referred to as "VK") and used in nearly half of all studies of sex differences in mental rotation, consistently yields the largest male advantage: 0.75–1.0 standard deviations according to a meta-analysis[2]. Although it varies by socioeconomic background[3] and sexual orientation[4], a male advantage is found robustly across cultures[5,6] and age ranges[7-9], and it is a critical empirical pillar of theories of human sex differences in spatial abilities, including evolutionary theories[10,11].

The VK test is a paper-and-pencil adaptation of a task developed by Shepard and Metzler[12] (hereafter "SM") that first demonstrated people "mentally rotate" imagined objects to compare them and decide whether they are identical. The consistent behavioral signature of this mental rotation process is a strongly linear increase in response time and error rate as the angular disparity between objects in a stimulus pair increases. The slope of this linear function measures the efficacy of the rotation process; a lower cost (in time and accuracy) for each additional degree of rotation indicates better mental rotation ability. In the VK, however, response time and accuracy are not measured as a function of the angle of disparity on individual trials; instead, the participant attempts to complete a fixed series of trials within a given time limit, and performance is measured simply as the total number of correct answers. **Figure 1** illustrates the SM and VK.

In a previous study[13] we assessed the relationship between performance on a version of the SM task and testosterone level in men. The total score on the task reflects the operation of many distinct cognitive processes. To complete each trial, participants must (a) select two objects to compare, shifting attention to the appropriate "standard" and "target" objects; (b) form

a mental representation of the object to be rotated; (c) mentally rotate the object until its orientation is the same as the standard; (d) compare the two objects; (e) decide whether the objects are the same or different; (f) produce an appropriate response[14,15]. We partitioned performance into two components, one that reflects primarily the mental rotation process itself (i.e., the slope of the "rotation function" relating response speed and accuracy to the angular disparity between the objects, which results from step "c" above) and one that reflects primarily other processes, including visual encoding, preparing for rotation, decision making, and responding (the efficienty of all of which are collectively measured by the intercept of the rotation function). We found that for error rate, testosterone level was related only to the intercept, and only when the two objects were different ("Different" response trials)—not when the objects are the same ("Same" response trials). Importantly, as Shepard and Metzler[12] explained in their original article on the rotation task, the "Different" trials are distractor trials—they are included not to measure the mental rotation process (the "Same" trials do that), but to make the task difficult to perform without mental rotation, and thus ensure that participants must engage in that process.

Based on this within-sex result—and the widely-accepted theory that sex differences in pre- and/or post-natal testosterone levels contribute to sex differences in spatial ability[11]—we hypothesized that the sex difference in performance on the SM task would also be confined to the intercept of Different trials, and that this component of an individual's performance would be the best predictor of that individual's score on the VK test. (That is, differences in non-rotation aspects of performance on the SM would predict differences in scores on the VK.) Accordingly, we administered both tests, in counterbalanced order, to a group of male and female participants, following standard procedures from previous studies.

We also explored whether sex differences in variables that are not directly related to mental rotation or derived from mental rotation tasks might explain the sex differences in mental rotation performance. There is a broad literature on individual differences in confidence for both sexes, ranging from measures of confidence within spatial tasks to real-world performance in behavioural domains[16-18] to sex differences in impulsiveness and risk aversion[19-21]. Additionally, individual differences in the various components[22] of spatial ability and task performance would benefit from an exploratory analytic approach. We thus explored possible mediators of the relationship between sex and mental rotation performance by also administering tests of spatial relations encoding, impulsiveness, and confidence.

Methods

Participants

We tested 123 volunteers (60 male, 63 female; ages 18–60, mean 26 years), who participated for pay after being recruited via advertisements (which did not mention sex differences or spatial ability). Approximately two thirds of the participants were students and one third were local residents. All reported not using drugs or psychoactive medications, no history of psychiatric or neurological illness, and at least a high-school education. Our male and female samples did not differ significantly in age, years of education, handedness, general cognitive ability (measured using a short form of Raven's Advanced Progressive Matrices[23]), vividness of experienced mental imagery (Vividness of Visual Imagery Questionnaire[24]), frequency of imagery use in daily life (Spontaneous Use of Imagery Scale[25]), or length of time awake before testing (all $p > .10$). This research was approved by the Harvard University Committee on the Use of Human Subjects in Research, and written informed consent was obtained. All experiments were performed in accordance with relevant guidelines and regulations.

Materials and Apparatus

The SM was administered using a computerized adaptation of the three-dimensional mental rotation task described by Shepard and Metzler[12]. Stimuli were presented and responses were recorded by an Apple Macintosh computer running OS 9 with a 40.5 cm monitor. Keypresses and response times (in milliseconds) were automatically recorded by PsyScope 1.2.5 software[26]. Each trial consisted of two circles, presented side-by-side, with each containing a block stimulus. We presented a subset of the stimuli used in the original Shepard and Metzler[12] study. As illustrated in **Figure 1a**, each circle (diameter 5.6 cm, or 10.8° of visual angle) contained one two-dimensional representation of a three-dimensional block object (approximately 5.7 cm x 0.64 cm, or 7.5° x 4.2°). An equal number of objects in each pair were presented at angles that differed by 0, 40, 80, 120, or 160 degrees. Half of the stimuli at each angle were Same pairs and half were Different pairs. Accordingly, there were 5 angles x 2 response types x 8 standard objects = 80 total trials. The VK[1] is a paper-and-pencil adaptation of the original SM task. **Figure 1b** shows a sample trial from this test, which consists of four practice trials and 20 test trials, with five on each page.

Procedure

Participants were tested individually in a private room by same-sex investigators (to eliminate the possibility that responses to opposite-sex investigators could affect performance on cognitive tests[27]). The tasks were administered within a larger battery of cognitive and personality measures. Participants were randomly assigned to one of two task orders that determined whether they would complete the SM or VK first; one task was completed about 30 minutes into the study, and the other about two hours later (after unrelated intervening tasks).

Preliminary analysis showed no interactions between sex and order, so we pooled over order in all analyses reported here. Task instructions did not mention spatial ability or sex differences.

For each trial of the SM task, participants decided whether the two stimuli depicted the same object or different objects, and indicated their choice by pressing a corresponding key on the keyboard; 500 ms after participants responded, the next trial began. Participants were told to "respond as quickly and accurately as possible." They completed ten practice trials (with different stimuli from those in the experimental trials), asked any questions they had about the procedure, and then completed the 80 experimental trials, which were presented in a new random sequence for each participant.

For each trial of the VK test, participants selected, by marking an X in the box below them, two of the four shapes on the right that matched the shape on the left (exactly two shapes were correct matches). We administered and scored the test as recommended by its originators[1,28]: Participants were given three minutes for each half of the test, with a one minute break between halves, and a trial was counted as correct if the two correct choices, and only those choices, were selected. The VK score reflects the number of correct trials out of a possible 20 (using alternative scoring methods, such as giving one point per correctly-chosen matching object, did not affect the pattern of results). For correlation and regression analyses, VK scores were converted to error rates.

Data Preparation and Task Validation

For the SM task, we computed mean response times (RTs) for each cell of the design (defined by crossing participant, trial type, and angle) after eliminating error trials and trimming "outlier" trials using an iterative criterion of 2.5 times the mean RT of the remaining trials in that cell (approximately 3% of trials were excluded as RT outliers). Error rates (ERs) represent the

percentage of trials (out of the total in a given cell of the design) on which the participant answered incorrectly. To ensure that our implementation of the SM task validly assessed mental rotation, we performed linear contrasts on the Same trial RTs and ERs, averaged for each participant according to rotation angle. As expected, RT increased linearly with angle, t (122) = 18.04, $p < .0001$, as did ER, t (122) = 14.48, $p < .0001$. (All higher-order contrasts were nonsignificant, $p > .20$). On the basis of these extremely strong linear trends (explaining 73% and 63% of the variance in RT and ER respectively, comparable to the original study of Shepard & Metzler[12]), we obtained slope and intercept measures for each participant by regressing ER and mean RT in each cell on rotation angle (0°, 40°, 80°, 120°, and 160°).

We assessed the split-half reliability of the SM by correlating ERs and RTs derived from the odd- and even-numbered trials, and found that individual differences were largely consistent between odd and even trials: r ranged from .53–.96 for the various measures. (We previously found the measures from a very similar version of this task to have adequate test-retest reliability[13]). To assess the reliability of the VK, we correlated ERs from the first and second halves of the test; the result ($r = .68$) was similar to that reported by other investigators[28] and comparable to the reliabilities of the SM components that we used as predictors of VK in regression analyses ($r = .58–.87$).

Additional Cognitive Tests

Individual differences in spatial encoding were assessed by separate computerized tests involving judgment of *categorical* and *coordinate* spatial relations[29-31]. Stimuli were identical, but appeared in a different pseudo-random order, for each test. On each trial a small dot appeared directly above or below a horizontal bar, at one of 16 discrete distances from the bar, for 150 ms. In the categorical task, the participant decided whether the dot was above or below the bar; in the

coordinate task the participant decided whether the dot was more or less than 8 mm away from the bar (a distance that had been demonstrated earlier on the computer screen). For each task, the 32 possible stimuli appeared once each in the left, central, and right visual fields; participants were instructed to fixate on a central point at all times. ERs for each task were used as indications of spatial encoding abilities. Consistent with previous studies, the coordinate task (14% ER, similar to the SM) was more difficult than the categorical task (3% ER).

Individual differences in confidence were assessed by a paper-and-pencil "trivia quiz" that asked participants to decide whether each of 20 statements was true or false (e.g., "In the year 2000 the population of Brazil exceeded 85 million" [true]). Upon finishing, participants were surprised with a request to estimate how many correct responses they had made. This estimate served as a measure of confidence; to control for individual differences in accuracy, all analyses involving confidence included the number of correct responses as a covariate (a method preferred to the use of difference scores[32]).

Individual differences in impulsiveness were assessed by a computerized delay-discounting (intertemporal choice) task[33] in which the participant makes a series of 27 choices, each between an amount of money "today" and a larger amount delayed some number of days into the future (e.g., "Would you prefer $54 today, or $80 in 30 days?"). Each participant has a 1/6 chance of receiving the outcome of one of his choices, selected at random, in addition to the standard payment for participation. The rate k at which a participant discounts the future value of money is estimated from his choices, and is taken as a measure of impulsiveness. Discount rates correlate with impulsiveness traits from personality tests[33] and with impatient real-world behaviors[34]. The overall mean discount rate (k = 0.015, representing approximately 1.5% per day) was comparable to that found in published studies with the same choice questions. To

correct for the non-normal distribution of discount rates, we used the natural logarithm of *k* in all analyses.

Data Availability

The datasets analyzed in this study are available from the corresponding author on reasonable request.

Results and Discussion

On the VK test, males were more accurate than females (mean score 9.8 vs. 6.9 correct, or $d = 0.66$ standard deviations), $t\ (121) = 3.87$, $p < .001$, replicating the typical finding (see **Table 1** for complete results). **Figure 2** depicts performance as a function of rotation angle on the SM task. Here, males also performed better than females (12.3% vs. 17.4% overall error rate), $t\ (121) = 2.64$, $p < .01$, but the effect was somewhat smaller ($d = 0.46$) than for the VK test. As predicted, males were significantly more accurate than females on the SM on Different, but not Same trials ($d = 0.48$, $p < .01$ for Different, $d = 0.16$, $p = .37$ for Same). Moreover, when considering the components of ER (slopes and intercepts for Same and Different trials, respectively) the only significant male advantage was on the intercept of the rotation function for Different trials, $d = 0.38$, $p < .05$. There were no significant sex differences in any measure involving response time (overall, Same/Different trial type, slope/intercept), $p > .15$ in all cases. Notably, we found the smallest sex difference in the slope for Same trials ($d = 0.04$ for ER, $d = 0.02$ for RT), the measure that best reflects the operation of the rotation process[12]. By re-analyzing data included in a classic published mental rotation article[35], we discovered that the pattern shown in Figure 2 (p. 126) of that older paper (a male advantage largely confined to the error rate intercept of Different trials) has been recorded before—but apparently never noticed or reported.

Next we tested the hypothesis that the male advantage in non-rotation components of mental rotation accounts for differences in scores on the VK test. Because the single VK score does not distinguish among angles or trial types, we used an individual-differences approach and conducted correlation and regression analyses, with VK score as the dependent variable and the components of error rate (Same slope, Same intercept, Different slope, Different intercept) as predictors. We reasoned that if the VK primarily measures mental rotation ability, then one's slope scores on the SM task—the scores that index the efficiency of the mental rotation process—should best predict one's VK score; in contrast, if VK performance is best predicted by some other performance measure on the SM, and if this differs for men and women, then the VK may not be a pure test of mental rotation or of sex differences in spatial ability.

For the entire sample, as shown in **Tables 2 and 3,** the intercept on Different trials (i.e., the measure of all cognitive processes in the task other than mental rotation) was the strongest predictor of VK scores (simple $r = .39$, $p < .0001$) and had the highest weight in the regression ($\beta = 0.39$). The overall regression solution accounted for 24% of the variance in participants' VK performance. Within sex, however, the results differed for males and females. The full regression for the male sample explained 41% of their VK performance variation, compared to only 12% for the female sample. This suggests that the VK test does not measure the exact same cognitive abilities in men and women; in particular, in women the VK may measure something else in addition to mental rotation.

For males, the Different intercept was the most significant predictor ($r = .56$, $p < .0001$), but for the females, the Same slope was the best predictor ($r = .26$, $p < .05$) and the Different intercept was not significantly related to VK performance ($r = .17$). A Fisher transformed test showed that the Different intercept was more strongly related to VK performance in males than

in females (Z = 2.49, $p < .02$). Indeed, the Same slope was comparably predictive for males and females ($r = .34$, $p < .01$ for males, which was not significantly different from females, $Z = 0.48$, $p = .63$). These findings also support the conclusion that scores on the VK do reflect rotation ability in part, but they reflect different aspects of spatial performance in men and women. For men, performance on the VK test is explained primarily by the *non-rotation* components of processing when the objects being compared are different, and secondarily by the efficiency of the rotation process itself when objects are the same. Ironically, although the VK seems to measure mental rotation efficiency for female participants, the *sex difference* in VK performance is driven by primarily non-rotation processes (intercept) on Different trials. This conclusion is based on the fact that the sexes are comparable in the strength with which the Same slope on the SM predicts VK score, but differ in the predictive strength of the Different intercept.

In short, although the VK does reflect some aspects of mental rotation per se (as indicated by the correlations with the Same slope in the SM task), it also reflects other aspects of processing. We suggest that those other aspects, and not a general male superiority in the process of mental rotation *per se*, account for much of the sex difference, a conclusion also reached by researchers who have used somewhat different approaches and samples[36,37]. In light of the relatively low and contrasting correlations between the VK and elements of mental rotation task performance in males and females, we suggest that the use of the VK as a pure measure of sex differences in spatial ability be reconsidered.

If not rotation itself, what cognitive processes differentiate male and female performance on mental rotation tests? Our results show that females are more likely than males to decide mistakenly that different objects are the same, but they do not make increasingly more errors than males as the angular disparity between the objects increases (a pattern also observed by

others[38] for 40–160-degree trials, but not 0-degree trials, which require no mental rotation and may be solved using different strategies). Bias in the choice of Same or Different, and thus the proportion of errors, is affected by the decision strategy a participant uses. We have found, consistent with other results[39], that participants with high error rates are more likely to choose Same when objects are actually different, and that these participants are disproportionately female.

These findings are consistent with the hypothesis that females are more reluctant, or require greater certainty or confidence, to take the risk of deviating from a default response. For both sexes, Same responses were more frequent than Different, even though the task contained equal numbers of the two stimulus types, and thus can be regarded as a default; the Same-bias was greater for females than for males. Indeed, females have been shown to be more risk-averse in a variety of behavioral domains[19-21], and are less confident in competitive test-taking settings[17]. Men tend to be more *over*confident than women[18] and more confident in believing they have found correct answers on mental rotation tests[16]. Exposure to obviously difficult tests of "spatial ability," or to time-restricted tests, may induce greater caution among females. Indeed, when participants are presented with familiar objects to rotate mentally, instead of novel three-dimensional block objects, or when reference to the spatial nature of the VK is removed, the sex difference diminishes[40-43].

An alternative to our "cautious female" hypothesis is that superior male performance is driven not by more accurate decision-making, but by an advantage in visually encoding and maintaining mental representations of the block stimuli. This is hard to reconcile with the fact that the sex difference appears only when the two objects are different: random errors in encoding should produce object representations that falsely appear to be different more often

than they falsely appear to be the same, resulting in a bias toward Different, not Same responses. Thus, the greater Same bias that females exhibit is difficult to explain in terms of a sex difference in spatial encoding. Nonetheless, we added to our analysis the four control tasks described earlier (Categorical spatial relations encoding, Coordinate spatial relations encoding, Impulsive choice, and Confidence) to address this alternative account directly and to pit it against the claim that females are more cautious than males during mental rotation tasks.

Figure 3 shows the sex differences we observed for these four tasks, plotted on a scale of *d* and compared to the VK and SM measures discussed earlier. There was no male advantage on either spatial encoding task ($d < 0$ in each case), but the greater impulsiveness and confidence of males than females were statistically significant ($p < .05$ for Impulsiveness, $p < .01$ for Confidence) and comparable in magnitude to the male advantages in overall mental rotation test scores and Different Intercept components (*d*-values 0.37–0.66).

If our "cautious female" hypothesis is correct and the "superior male spatial encoding" alternative is not, then individual differences in Confidence and/or Impulsiveness, but not Categorical or Coordinate encoding, should add to our ability to predict scores on the VK test. We conducted a new regression analysis to test this, with VK as the dependent variable and the four SM components already analyzed, plus the four new measures, as independent variables. In addition to the previous predictors (Different Intercept, Same Slope, and Different Slope; see **Tables 2 and 3**), only Confidence was an independently significant predictor of VK error rate ($\beta = –0.20$, $t = –2.46$, $p < .02$; for all other new measures $p > .20$).

Finally, we asked whether the direct effect of sex on VK score was *mediated* by any of our four non-rotation measures. An effect of an independent variable X on a dependent variable Y is mediated when X also affects a third variable M (the mediator), M affects Y, and the effect

of X on Y is reduced when M is controlled for in a regression[44]. A finding of significant mediation lends support to the hypothesis that the influence of X on Y is transmitted, at least in part, through an effect of X on the process or trait represented by M. We found that only Confidence mediated the sex difference in VK scores ($p < .02$ according to standard methods[44]); Categorical and Coordinate encoding did not show significant sex differences and did not significantly predict VK, and while Impulsiveness was greater in males, it was not related to VK. Confidence mediated only 8% of the sex difference in VK score, but taken together with the failure to find any influence of spatial encoding abilities on VK performance, and the incremental predictive power of confidence for individual differences, this result bolsters our argument that a difference in decision strategy—and not in the process of mental rotation itself—best explains the consistently different performance of males and females on the VK test. This adds to the literature supporting the possibility that different decision strategies are used by males and females in spatial tasks[45,46]. Prior research examining confidence in mental rotation tests defined it as the extent to which one believes they had the correct answers *on the rotation test*[16], so our findings add to and extend these findings by showing that individual differences in a general trait of confidence—more broadly construed and measured—may be important to consider and explore in more detail.

Conclusions

Our results have pragmatic implications for spatial skill training interventions, which some researchers have suggested might narrow male-female differences in science, technology, engineering, and mathematics (STEM) achievement[47-49] given the link between spatial abilities—measured as a general construct—and later STEM achievement[50]. Our finding that sex differences might well primarily lie in the non-rotation aspects of the mental rotation test

suggests that decision strategies, confidence, and individual differences in traits and abilities not directly tapped by spatial tasks would be worthwhile to pursue in future studies of sex differences and of training regimens. It is possible that investigating individual differences in spatial ability components might lead to improvements in such interventions, although solid links between such interventions and "far transfer"[51,52] to untrained tasks, as well as transfer to the long-term outcomes predicted by spatial ability (e.g., educational and career achievement in science and engineering) remain to be made.

Finally, our results have implications for the "hunter-gatherer" theory of sex differences in spatial cognition, which posits that a present-day male advantage in mental rotation ability is a consequence of the sexual division of labor in human evolutionary history[10]. According to this theory, because males hunted, they navigated larger ranges than females, and were subject to selection for the cognitive skills that facilitate tracking prey and returning from sojourns in unfamiliar territory. This idea is undermined by findings that spatial cognition is not identical at large scales (navigation) and small scales (mental rotation)[53,54], and that spatial sex differences also covary with day range in nonhuman species (e.g., voles[55,56]). Moreover, it is not consistent with our findings that: (1) males are better than females only in the non-rotation component of one half of the trials (different objects) in mental rotation tasks; (2) this aspect of performance is the best predictor of male performance on paper-and-pencil rotation tests, but not of female performance; (3) testosterone exerts its influence primarily on this same component of rotation test performance[13]; and (4) differences in confidence, but not spatial encoding abilities, contribute to the sex difference in mental rotation performance. Our results suggest that although men may perform better than women on putative tests of "mental rotation," their mental rotation

processes are not more efficient, and therefore the sex difference in mental rotation tests is not necessarily a difference in mental rotation ability.

Acknowledgements

We thank Steven Gangestad, Steven Pinker, William Thompson, and Richard Wrangham for helpful discussions, and Aerfen Whittle, Kirill Babikov, Jacob Sattelmair, Carrie Morris, Lee Chung, Thomas Jerde, and Jonathon Schuldt for research assistance. C.F.C. was supported by a DCI Postdoctoral Fellowship; this work was supported by NIH grant R01-MH60734 and NSF grant REC-0106760 to S.M.K.

Author Contributions Statement

C.K.H., P.T.E., S.M.K., and C.F.C. designed the research. C.K.H. and C.F.C. conducted the study and analysed the data. C.K.H., C.F.C., J.W., and R.A.K. drafted the manuscript. All authors critically edited the manuscript.

Competing Financial Interests Statement

The author(s) declare no competing financial interests.

References

1. Vandenberg, S. G., & Kuse, A. R. Mental rotations, a group test of three-dimensional spatial visualization. *Percept. Mot. Skills*. **47**, 599–604 (1978).

2. Voyer, D., Voyer, S., & Bryden, M. P. Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychol. Bull.* **117**, 250–270 (1995).

3. Levine, S. C., Vasilyeva, M., Lourenco, S. F., Newcombe, N. S., & Huttenlocher, J. Socioeconomic status modifies the sex difference in spatial skill. *Psychol. Sci.* **16**, 841–845 (2005).

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

5. Silverman, I., Choi, J., & Peters, M. The hunter-gatherer theory of sex differences in spatial abilities: Data from 40 countries. *Arch. Sex. Behav.* **36**, 261–268 (2007).

6. Silverman, I., Phillips, K., & Silverman, L. K. Homogeneity of effect sizes for sex across spatial tests and cultures: Implications for hormonal theories. *Brain Cogn.* **31**, 90–94 (1996).

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

8. Maylor, E. A., Reimers, S., Choi, J., Collaer, M. L., Peters, M., & Silverman, I. Gender and sexual orientation differences in cognition across adulthood: Age is kinder to women than men regardless of sexual orientation. *Arch. Sex. Behav.* **36**, 235–249 (2007).

9. Quinn, P. C., & Liben, L. S. A sex difference in mental rotation in young infants. *Psychol. Sci.* **19**, 1067-1070 (2008).

10. Silverman, I., & Eals, M. Sex differences in spatial abilities: Evolutionary theory and data in *The adapted mind: Evolutionary psychology and the generation of culture* (eds. Barkow, J. H. & Cosmides, L.) 533–549 (Oxford University Press, 1992).

11. Kimura, D. *Sex and cognition.* (MIT Press, 1999).

12. Shepard, R. N., & Metzler, J. Mental rotation of three-dimensional objects. *Science.* **171**, 701–703 (1971).

13. Hooven, C. K., Chabris, C. F., Ellison, P. T., & Kosslyn, S. M. The relationship of testosterone to components of mental rotation. *Neuropsychologia.* **42**, 782–790 (2004).

14. Kosslyn, S. M. *Image and mind*. (Harvard University Press, 1980).

15. Karadi, K., Kallai, J., & Kovacs, B. (2001). Cognitive subprocesses of mental rotation: Why is a good rotator better than a poor one? *Percept. Mot. Skills.* **93**, 333-337 (2001).

16. Cooke-Simpson, A., & Voyer, D. Confidence and gender differences on the mental rotation test. *Learn. Individ. Differ.* **17**, 181-186 (2007).

17. Lenney, E., & Gold, J. A. Sex differences in self-confidence: The effects of task completion and of comparison to competent others. *Pers. Soc. Psychol. Bull.* **8**, 74–80 (1982).

18. Pallier, G. Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex Roles.* **48**, 265–276 (2003).

19. Eckel, C. C., & Grossman, P. J. Men, women, and risk aversion: Experimental evidence in *Handbook of experimental economics results* (Vol. 1) (eds. Plott, C., & Smith, V.) (Elsevier, 2008).

20. Hay, D. F., & Lockwood, R. Girls' and boys' success and strategies on a computer-generated hunting task. *Brit. J. Dev. Psychol.* **7**, 17–27 (1989).

21. Wilson, M., & Daly, M. Competitiveness, risk taking, and violence: The young male syndrome. *Ethol. Sociobiol.* **6**, 59–73 (1985).

22. Lohman, D. F. (1994). Spatial ability in *Encyclopedia of intelligence* (Vol. 2) (ed., Sternberg, R. J.) 1000-1007 (Macmillan, 1994).

23. Bors, D. A., & Stokes, T. L. Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educ. Psychol. Meas.* **58**, 382–398 (1998).

24. Marks, D. F. (1972). Individual differences in the vividness of visual imagery and their effect on function in *The function and nature of imagery* (ed., Sheehan, W. P.) 83-108 (Academic Press, 1972).

25. Reisberg, D., Pearson, D. G., & Kosslyn, S. M. Intuitions and introspections about imagery: The role of imagery experience in shaping an investigator's theoretical views. *Appl. Cogn. Psychol.* **17**, 147–160 (2003).

26. Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behav. Res. Methods Instrum. Comput.* **25**, 257–271 (1993).

27. Roney, J. R., Mahler, S. V., & Maestripieri, D. Behavioral and hormonal responses of men to brief interactions with women. *Evol. Hum. Behav.* **24**, 365–375 (2003).

28. Vandenberg, S. G., Kuse, A. R., & Vogler, G. P. Searching for correlates of spatial ability. *Percept. Mot. Skills*. **60**, 343–350 (1985).

29. Hellige, J. B., & Michimata, C. Categorization versus distance: Hemispheric differences for processing spatial information. *Mem. Cognit.* **17**, 770–776 (1989).

30. Kosslyn, S. M. Seeing and imagining in the cerebral hemispheres: A computational approach. *Psychol. Rev.* **94**, 148–175 (1987).

31. Laeng, B., Chabris, C. F., & Kosslyn, S. M. Asymmetries in encoding spatial relations in *The asymmetrical brain* (eds., Hugdahl, K., & Davidson, R. J.) 303–339 (MIT Press, 2003).

32. Cohen, J., & Cohen, P. *Applied multiple regression/correlation analysis for the behavioral sciences.* (Erlbaum, 1983).

33. Kirby, K. N., Petry, N. M., & Bickel, W. K. Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. *J. Exp. Psychol. -Gen.* **128**, 78–87 (1999).

34. Chabris, C. F., Laibson, D. I., Morris, C. L., Schuldt, J. P., & Taubinsky, D. Individual laboratory-measured discount rates predict field behavior. *J. Risk Uncertain.* **37**, 237-269 (2008).

35. Tapley, S., & Bryden, M. An investigation of sex differences in spatial ability: Mental rotation of three-dimensional objects. *Can. J. Psychol.* **31**, 122–130 (1977).

36. Boone, A. P., & Hegarty, M. Sex differences in the mental rotation tasks: Not just in the mental rotation process! *J. Exp. Psychol. -Learn. Mem. Cogn.* Advance online publication (2017).

37. Brosnan, M., Daggar, R., & Collomosse, J. The relationship between systematizing and mental rotation and the implications for the extreme male brain theory. *J. Autism Dev. Disord.* **40**, 1-7 (2010).

38. Voyer, D., Butler, T., Cordero, J., Brake, B., Silbersweig, D., Stern, E., & Imperato-McGinley, J. The relation between computerized and paper-and-pencil mental rotation tasks: A validation study. *J. Clin. Exp. Neuropsychol.* **28**, 928–939 (2006).

39. Kerkman, D. D., Wise, J. C., & Harwood, E. A. Impossible mental rotation problems: A mismeasure of women's spatial abilities? *Learn. Individ. Differ.* **12**, 253–269 (2000).

40. Alexander, G. M., & Evardone, M. Blocks and bodies: Sex differences in a novel version of the mental rotations test. *Horm. Behav.* **53**, 177-184 (2008).

41. Goldstein, D. G., Haldane, D., & Mitchell, C. Sex differences in visual-spatial ability: The role of performance factors. *Mem. Cognit.* **18**, 546–550 (1990).

42. Scali, R. M., Brownlow, S., & Hicks, J. L. Gender differences in spatial task performance as a function of speed or accuracy orientation. *Sex Roles.* **43**, 359–376 (2000).

43. Sharps, M. J., Price, J. L., & Williams, J. K. Spatial cognition and gender: Instructional and stimulus influences on mental image rotation performance. *Psychol. Women Q.* **18**, 413–425 (1994).

44. MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. A comparison of methods to test mediation and other intervening variable effects. *Psychol. Methods.* **7**, 83–104 (2002).

45. Doyle, R. A., Voyer, D., & Lesmana, M. Item type, occlusion, and gender differences in mental rotation. *Q. J. Exp. Psychol.* **69**, 1530-1544 (2016).

46. Heil, M., & Jansen-Osmann, P. Sex differences in mental rotation with polygons of different complexity: Do men utilize holistic processes whereas women prefer piecemeal ones? *The Q. J. Exp. Psychol.* **61**, 683-689 (2008).

47. Sorby, S. A. A course in spatial visualization and its impact on the retention of female engineering students. *J. Women Minor. Sci. Eng.* **7**, 153-172 (2001).

48. Sorby, S. A., & Baartmans, B. J. The development and assessment of a course for enhancing the 3-D spatial visualization skills of first year engineering students. *J. Eng. Educ.* **89**, 301-307 (2000).

49. Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. The malleability of spatial skills: A meta-analysis of training studies. *Psychol. Bull*. **139**, 352-402 (2013).

50. Wai, J., Lubinski, D., & Benbow, C. P. Spatial ability for STEM domains: Aligning over fifty years of cumulative psychological knowledge solidifies its importance. *J. Educ. Psychol.* **101**, 817-835 (2009).

51. Shipstead, Z., Redick, T. S., & Engle, R. W. Is working memory training effective? *Psychol. Bull*. **138**, 628-654 (2012).

52. Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. L. Do “brain training” programs work? *Psychol. Sci. Public Interest.* **17**, 103-186 (2016).

53. Kozhevnikov, M., & Hegarty, M. A dissociation between object manipulation spatial ability and spatial orientation ability. *Mem. Cognit.* **29**, 745–756 (2001).

54. Zacks, J. M., Vettel, J. M, & Michelson, J. (2003). Imagined viewer and object rotations dissociated with event-related fMRI. *J. Cognitive Neurosci.* **15**, 1002–1018 (2003).

55. Gaulin, S. J. C., & FitzGerald, R. W. Sex differences in spatial ability: An evolutionary hypothesis and test. *Am. Nat.* **127**, 74–88 (1986).

56. Jones, C. M., Braithwaite, V. A., & Healy, S. D. The evolution of sex differences in spatial ability. *Behav. Neurosci.* **117**, 403–411 (2003).

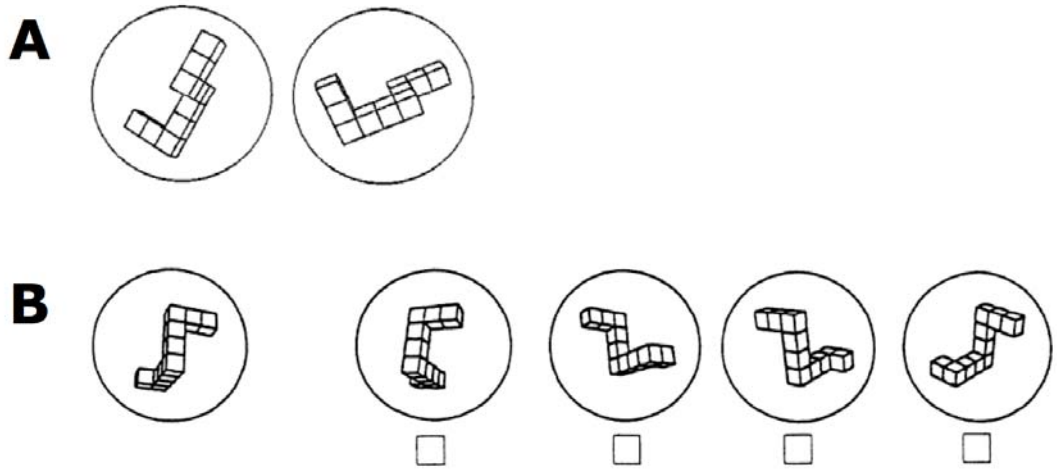
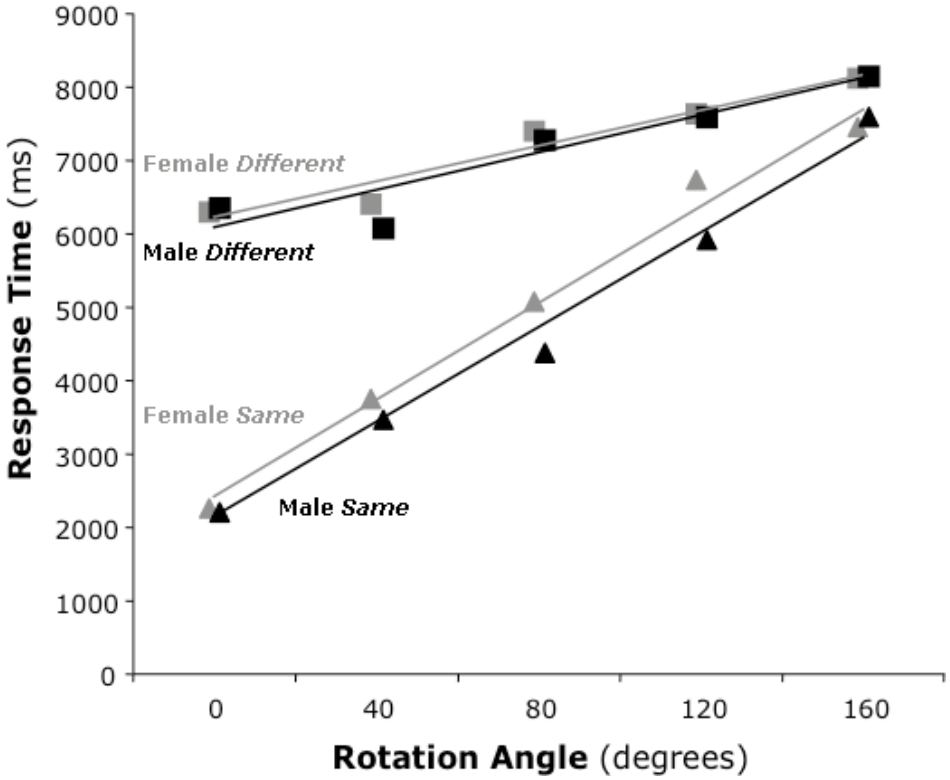


Figure 1. Two tasks used to measure mental rotation ability. *(A)* Sample trial from the Shepard and Metzler[12] task ("SM"). Two images remain in view together while the participant decides whether they depict the same object or two different objects, and response time and error rate are recorded for each trial. *(B)* A sample trial from the Vandenberg and Kuse[1] pencil-and-paper adaptation ("VK"). The VK presents pictures of the same objects as those in the SM, but the format is different. Each trial of the VK requires participants to determine which two of four "comparison" objects are identical to a "standard" object, regardless of differences in orientation. The comparison objects are usually (but not always) presented at different orientations from the standard.


9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Response Time (ms)
Female Different
Male Different
Female Same
Male Same
0
40
80
120
160
Rotation Angle (degrees)

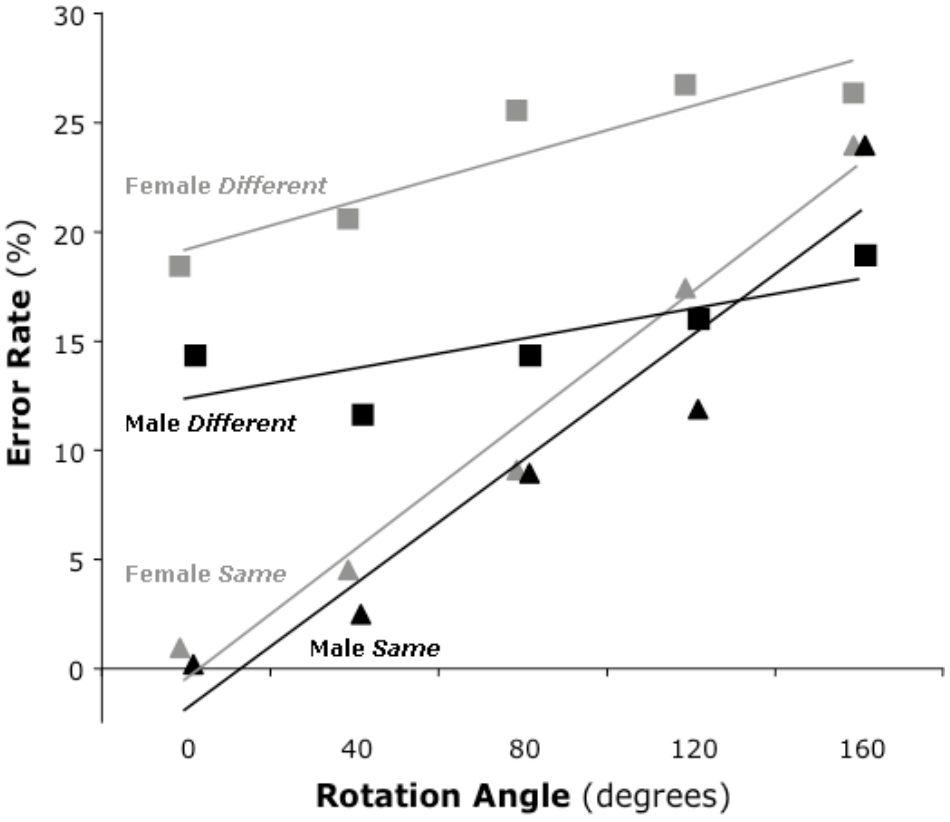

30
25
20
15
10
5
0
Error Rate (%)
Female Different
Male Different
Female Same
Male Same
0
40
80
120
160
Rotation Angle (degrees)

Figure 2. Performance on mental rotation tasks. *(Top)* Response time as a function of rotation angle (i.e., the misalignment between the two objects), plotted for male (N = 60; black symbols) and female (N = 63; gray symbols) participants and Same (⑩) and Different (⑤) response trials. There are no sex differences in the slopes or intercepts of linear fits to the data (shown as black and gray lines for males and females). *(Bottom)* Error rate as a function of rotation angle, plotted in the same way. There is no sex difference for Same trials in either slope or intercept, but for Different trials, female participants have a higher intercept than male participants. *Note:* Our participants seem to mentally rotate three-dimensional objects at slower rates than reported in some previous studies; this may be due to factors such as our lengthy testing session, our inclusion of older participants in addition to college students, and other researchers' use of highly-practiced subjects who receive many hundreds of trials.

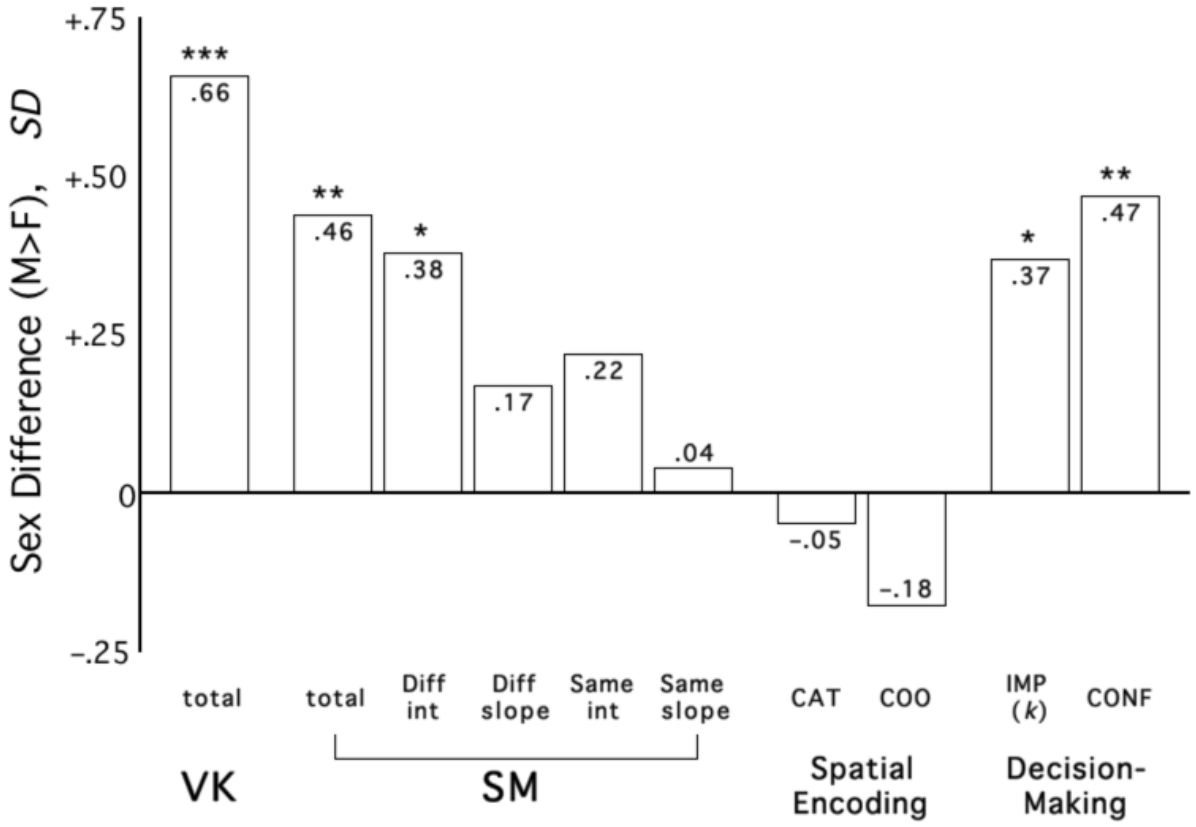


Figure 3. Comparative magnitude of sex differences observed on mental rotation and other cognitive tasks (* $p < .05$, ** $p < .01$, *** $p < .001$). There are significant differences (*d*, units of standard deviation) between male and female performance for the VK, SM, SM Different intercept, impulsiveness (IMP), and confidence (CONF) measures, with males scoring higher in all cases, but not for the categorical (CAT) and coordinate (COO) spatial relations encoding measures, or for the Different slope, Same intercept, or Same slope measures from the SM.

Table 1. Means, standard errors (SE), sex differences (*t* statistic and Cohen's *d*—difference in means divided by the pooled standard deviation), on the SM and VK tests.

Measure	All (n=123)		Male (n=60)		Female (n=63)		*t* (121)	*D*
	Mean	SE	Mean	SE	Mean	SE		
SM Error Rate (%)								
Overall	14.91	0.99	12.29	1.32	17.40	1.41	2.64**	0.46
Different trials	19.43	1.59	15.08	2.08	23.57	2.28	2.74**	0.48
Same trials	10.39	0.97	9.50	1.34	11.23	1.41	0.89	0.16
Different intercept	15.85	1.60	12.38	1.99	19.17	2.42	2.16*	0.38
Different slope	0.04	0.01	0.03	0.01	0.06	0.02	0.95	0.17
Same intercept	–1.20	0.53	–1.88	0.67	–0.56	0.82	1.24	0.22
Same slope	0.14	0.01	0.14	0.02	0.15	0.02	0.22	0.04
SM Response Time (ms)								
Overall	6009	281	5893	295	6119	473	0.40	0.07
Different trials	7140	338	7094	358	7184	568	0.13	0.02
Same trials	4877	246	4692	257	5054	414	0.74	0.13
Different intercept	6141	304	6071	369	6207	481	0.22	0.04
Different slope	12	2	13	2	12	3	0.16	–0.03
Same intercept	2239	120	2077	168	2393	171	1.32	0.24
Same slope	33	2	33	2	33	4	0.13	0.02
VK Score (out of 20)	8.29	0.39	9.77	0.59	6.89	0.46	3.87***	0.66

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 2. Correlations (Pearson's *r*) between predictor variables and VK error rates for all participants and for males and females separately. *Top:* SM task components (error rates) as predictors. *Bottom:* Measures of spatial encoding (error rates), impulsiveness, and confidence as predictors. (Note that the negative Confidence-VK correlation means that greater confidence is associated with *fewer errors* on the VK.)

SM Variable	All (n=123)	Male (n=60)	Female (n=63)
Overall performance	.48****	.62****	.26*
Different trials	.42****	.52****	.22
Same trials	.30***	.42***	.16
Different intercept	.39****	.56****	.17
Different slope	.05	–.03	.06
Same intercept	.04	.13	–.12
Same slope	.30***	.34**	.26*

Measure	All (n=123)	Male (n=60)	Female (n=63)
Categorical encoding	–.03	–.04	.00
Coordinate encoding	.06	.13	.04
Discount rate (impulsiveness)	–.05	.05	–.02
Confidence (*pr*)†	–.19*	–.12	–.11

* p < .05, ** p < .01, *** p < .001, **** p < .0001, † number of correct answers partialled out

Table 3. Summary of multiple regression analyses for prediction of VK score (re-expressed as error rate). *Top:* Model with only SM task components as independent variables. *Bottom:* Model with SM components and additional cognitive measures. (In each analysis, all predictors were entered together in a single block.)

	All (R^2=.24, n=123)			Males (R^2=.41, n=60)			Females (R^2=.12, n=63)		
	ß	b	*t*	ß	b	*t*	ß	b	*t*
Different intercept	.39	0.48	4.57****	.50	0.73	4.66***	.20	0.19	1.43
Different slope	.19	33.37	2.26*	.07	19.79	0.70	.16	19.20	1.17
Same intercept	.07	0.26	0.86	.22	0.94	1.93	–.10	–0.28	–0.79
Same slope	.25	42.67	3.00**	.33	54.94	2.53*	.23	35.09	1.83

	All (R^2=.29, n=123)			Males (R^2=.44, n=60)			Females (R^2=.14, n=63)		
	ß	b	*t*	ß	b	*t*	ß	b	*t*
Different intercept	.39	0.48	4.61****	.53	0.78	4.68***	.22	0.21	1.55
Different slope	.20	35.08	2.38*	.06	16.16	0.56	.18	22.43	1.31
Same intercept	.09	0.34	1.11	.19	0.84	1.67	–.05	–0.14	–0.34
Same slope	.26	44.96	3.07**	.30	50.03	2.51*	.26	39.06	1.90
Categorical encoding	–.10	–0.62	–1.19	.02	0.13	0.20	–.08	–0.48	–0.48
Coordinate encoding	.04	0.07	0.46	.12	0.21	1.13	.01	0.01	0.04
Discount rate (impulsiveness)	–.05	–0.82	–0.59	–.06	–1.15	–0.55	–.01	–0.19	–0.09
Confidence†	–.20	–1.18	–2.46*	–.10	–0.64	–0.91	–.15	–0.76	–1.11

* p < .05, ** p < .01, *** p < .001, **** p < .0001, † number of correct answers partialled out