UNIVERSITY OF CALIFORNIA

Santa Barbara

Sex Differences in Mental Rotation Tasks

A Thesis submitted in partial satisfaction of the

requirements for the degree Master of Arts

in Psychology

by

Alexander Paul Boone

Committee in charge:

Professor Mary Hegarty, Chair

Professor Richard Mayer

Professor Russell Revlin

June 2015

The thesis of Alexander Paul Boone is approved.

Russell Revlin

Richard Mayer

Mary Hegarty, Committee Chair

June 2015

Sex Differences in Mental Rotation Tasks

ABSTRACT

Sex Differences in Mental Rotation Tasks

by

Alexander Paul Boone

The Mental Rotations Test (Vandenberg & Kuse, 1978) consistently produces large sex differences favoring males (Voyer, Voyer, & Bryden, 1995). This test requires participants to select two of four answer choices that are rotations of a probe stimulus. Incorrect choices (i.e., foils) are either mirror reflections of the probe or structurally different. In contrast, in the original mental rotation task (Shepard & Metzler, 1971) participants judge whether two stimuli are the same but rotated or different by mirror reflection. It was hypothesized that the large sex difference in the Mental Rotations Test emerges as a result of males noticing and capitalizing on the orientation independent features of structurally different foils. Two experiments indicated greater accuracy and faster reaction times for structurally different compared to mirror trials for both sexes. A significant male advantage in accuracy was found for both trial types. Males and females did not differ in reaction time for either trial type. Although no evidence was found to suggest that differences in capitalizing on an orientation independent strategy accounts for the large sex difference, results suggest that the mental rotation process is not the only source of the sex difference in these tasks.

The process of mental rotation allows one to simulate the outcome of rotating an object before carrying out a high-cost physical task such as lifting a heavy object to place it in a tight space. Across many studies, several tasks of spatial ability have shown sex differences typically favoring males, and the largest effect among those is for mental rotation (Voyer, Voyer, & Bryden, 1995). Mental rotation is seen as an important aspect of spatial thinking (National Research Council, 2006) and has garnered much interest in the domains of Science, Technology, Engineering, and Mathematics (STEM) education generally (Wai, Lubinkski, & Benbow, 2009) and specifically in chemistry (Steiff, 2007) and training (Uttal et al., 2013) as well as in evolutionary psychology (Ecuyer-Dab & Robert, 2004; Hooven, Chabris, Ellison, & Kosslyn, 2004). It is interesting to note, however, that the largest sex difference in mental rotation tasks comes primarily from a specific paper and pencil mental rotation task developed by Vandenberg and Kuse (1978; VK MRT). Other mental rotation paradigms yield small to medium sex differences (Voyer et al., 1995). This suggests that VK MRT may also be measuring something other than mental rotation per se such as ability to identify the usefulness of a specific strategy.

Since the development of the VK MRT task, many studies have sought to identify the source of the sex difference in spatial ability in mental rotation tasks focusing on the process of mental rotation itself (Kail, Carter, & Pellegrino, 1979; Mumaw, Pellegrino, Kail, & Carter, 1984), task-specific details such as timing (Resnick, 1993; Masters, 1998; Peters, 2005), task and figure complexity (Titze, Heil, & Jansen, 2010; Vanrie, Willems, & Wageman, 2001; Yuille & Steiger, 1982), and strategy usage (Hegarty, 2010). Although the VK MRT is the format through which the largest sex difference is seen, it is not the only task used to assess mental rotation or spatial ability generally.

The earliest work on 3-D object mental rotation comes from the chronometric studies of Shepard & Metzler (1971; SM). In this paradigm, two block figures are presented side by side. Participants were asked to make a same/different judgment concerning each pairwise presentation by pulling a corresponding lever concluding the trial. In “same” trials, the two block figures were the same block figure but rotated through ten levels of angular disparity between zero and 180 degrees at 20-degree intervals (i.e., 0, 20, 40, 60…160, 180). In “different” trials, the two diagrams were always mirror reflections of one another. Eight hundred trials in this task are “same” trials while the other 800 are “different.” Four hundred same trials are rotations in the depth plane, while the other 400 are rotations in the picture plane. In the original study, a linear relationship was found between angular disparity of the block figures and reaction time such that the greater disparity between the two block figures constituted longer reaction times. This finding suggested that participants were mentally rotating the objects into congruence and then making the “same” decision. Reaction times for different trials were not analyzed.

In the VK MRT paradigm, a single Shepard & Metzler block figure is presented to the left of four alternative choice response items (Figure 1). Participants are instructed to select the two alternative answer choices that are the same as the probe but rotated. Twenty total items are presented in two sets of 10. Participants are given three minutes to complete each set of 10 items. The two incorrect choices in each trial (henceforth, foils) are either mirror images of the probe or structurally different from the probe (Figure 1). Half of the items in this task show mirror foils while the other half show structurally different foils. It is this task that typically exhibits the large male advantage in performance, up to one standard deviation (Voyer et al., 1995).

As can be seen, there are a number of areas in which these two tasks deviate in the nature of the cognitive processes involved. First, in the SM task, a participant makes pairwise comparisons between two block figures (Figure 1). In contrast, participants must make as few as three and as many as five comparisons to ensure a correct response for VK MRT. The number of comparisons would necessarily increase the amount of time per item, as would any amount of rotation for those comparisons. Further, the timing factor between the tasks is different. For VK MRT a participant is allowed only three minutes for each set of 10 items (i.e., six minutes total). For SM, timing for each trial is self-paced although some researchers set arbitrary upper limits.

Most of the literature concerning the sex difference in mental rotation has been based on the VK MRT rather than the SM paradigm. Given that these two tasks vary in a number of dimensions, it is questionable that the sex difference is actually a difference in the process of mental rotation itself. Voyer et al. (1995) showed that for VK MRT simply modifying the strictness in scoring technique reduces the sex difference effect, although the effect is still large (.94 for strict scoring versus .70 for liberal scoring).

Less of the research on sex differences in mental rotation has focused on the SM paradigm. Shepard and Metzler (1971) did not comment on differences between the sexes in their tasks; however, it is likely that this difference would have been small in that particular study as these participants were selected for high spatial ability and were highly practiced at the task. Tapley and Bryden (1977) used the SM paradigm to assess sex differences in mental rotation over the course of two testing sessions. They found that males were significantly more accurate and produced significantly faster rates of rotation than females. Titze et al. (2010) sought to evaluate the sex differences in a pairwise presentation of mental rotation figures to investigate whether the reduced complexity of the SM could drive the effects seen in VK MRT. In this study, participants were presented pairwise comparisons singly on a piece of paper and participants marked either “same” or “different.” These researchers found a large effect favoring males on par with that found in VK MRT; however, it is unclear how closely this task approximates the SM paradigm, given methodological differences. Similarly, Peters (2005) sought to establish if the sex differences found in VK MRT were similar in an SM paradigm where timing was the factor of importance. In this study, participants compared pairwise Shepard and Metzler figures presented on a computer. Participants were tested on 24 trials for four days. Males significantly outperformed females, but neither sex showed a reaction time benefit. Results from this study are interpreted with caution as subjects were selected based on their previous mental rotation study performance (all subjects were within one standard deviation of their gender means) and well as the power associated with using only 24 trials in the modified Shepard and Metzler task. Additionally, in this case, it is unclear whether the trials used by Peters (2005) represent new comparisons each day or if they were the same 24 trials across all four days. Increasing the number of trials in a SM task would bring the field closer in understanding of the nature of the sex differences in mental rotation tasks.

Research into the neural processes involved in mental rotation and the associated sex difference has found medium effects (Butler et al., 2006; Voyer et al., 2006) but modified the paradigm by constraining each trial to seven seconds regardless of a response which is a short reaction time for large angular disparity trials in a typical SM task (e.g., 160 degrees). Secondly, trial block runs were constructed of five total trials, such that three were same trials and two were different trials. Although the difference between the number of same and different trials within a block run is small, the effect of bias cannot be ruled out.

The most salient difference between the VK MRT and SM tasks concerns trials in which the to-be-compared block figures are different from the probe (or each other in the case of SM). In a single VK MRT trial, the two foils can be either mirror reflections of the probe *or* structurally different in configuration whereas in the SM task the "different" trial stimuli are only mirror reflections of one another. Certain characteristics of the structurally different probes in VK MRT potentially allow participants to discover the correct answer without rotating any of the answer choices. This key difference could give rise to the large gender difference found in VK MRT. Interestingly, Shepard and Metzler (1971) explicitly mention *not* using structurally different foils in case participants should notice these differences and not rotate at all.

In a previous study (Hegarty, 2010), participants were asked to report the strategies they used to solve a mental rotation task in a post experiment questionnaire. Many participants reported using mental imagery strategies such as mentally rotating the figures (e.g., "I imagined one or more of the objects turning in my mind") or changing one's perspective (e.g., "I imagined looking at the object from a different viewpoint or perspective.") in order to make a judgment. Non-rotation strategies included counting blocks (pure analytic) and general test taking strategies such as skipping difficult items.

Most important for the current research is the spatial analytic strategy. The spatial analytic strategy involves noting the relative directions of sections of the block diagrams or noting the position of the tails of the answer choices to deduce if they were parallel or perpendicular with respect to the probe (e.g., "I figured out whether the two end arms of the target were parallel or perpendicular to each other and eliminated answer choices based on this feature."). This strategy is orientation independent inasmuch as the rotation factor between compared block figures is irrelevant. Noticing the position of key features such as the tail direction (Figure 2) and then comparing the standard to the alternatives is moderately correlated with performance (Hegarty, 2010). Indeed, this strategy would represent a case in which only visual perception (and not mental rotation) is necessary to see that the stimuli would constitute a “different” response (i.e., orientation independent; Takano, 1989; Jolicoeur, 1990). Other empirical work has suggested that participants can pick up on this type of perceptual difference that exists between mirror and structurally different foils. Geiser, Lehmann, and Eid (2006) identified a group of participants, labelled “non-rotators” that showed generally poor performance on the VK MRT but based on good performance on structurally different foils items, suggested that they did not rotate the block figures but rather were able to make use of feature-based differences (e.g., tail directions) when available.

Research that has focused on the question of whether the type of foil in VK MRT gives rise the sex differences in VK MRT has revealed conflicting results (Voyer & Hou, 2006; Voyer & Doyle, 2010; Bors & Vigneau, 2011). Voyer and Hou (2006) administered a mental rotation task similar to VK MRT (Peters, Laeng, Latham, Jackson, Zaiyouna, & Richardson, 1995) to a large sample of respondents. In their experiment, the researchers elected to use unlimited time, as most participants do not get to the final few items under the traditional timing of this task. The ubiquitous sex difference in overall performance was found, but there was no effect of foil type or an interaction. To these researchers, the lack of a significant interaction effect suggested that the gender difference in VK MRT is not due to the structural foils because males do better in both foil types. In later work, Voyer and Doyle (2010) found an effect of foil type. This finding was downplayed as it was qualified by an effect of occlusion in the block figures. Finally, Bors and Vigneau (2011) found evidence that mirror items were more difficult and additionally found a sex by foil type interaction indicating male superiority in mirror foil trials rather than structurally different foil trials.

It was the goal of the present research to further explore the potential that structurally different foils in a mental rotation task could give rise to the sex difference in VK MRT, specifically a strategy that identifies the foil type present in each item. From self-reports of strategy usage favoring this orientation-independent strategy and empirical evidence that males sometimes show this foil type advantage, conflicting evidence warrants a closer look at the effect of foil type as a potential source of the differences between sexes in the VK MRT. In two experiments, we sought to explore the relationship between foil type and the performance of each sex using two mental rotation paradigms. In Experiment 1, VK MRT with 40 total items (20 mirror, 20 structurally different) was used where foil type order was counterbalanced. In Experiment 2, participants responded to 160 same/different Shepard & Metzler paradigm trials where a different trial was either mirror or structurally different.

Within the context of two experiments, I tested the hypothesis that the large gender difference, favoring males in the VK MRT, arises in part from strategic differences between males and females. The strategy difference in question makes use of the incorrect answers (i.e., foils) present in a given trial. Males may spontaneously notice structural differences between the probe and the foils allowing for quick, strategic elimination of incorrect choices and selection of correct alternatives. Females, on the other hand, may not notice this strategy and instead rotate all answer choices making them less accurate and potentially slower than males. An interaction between sex and "different" foil type would be expected if males do benefit over females from the presence of structurally different foils. The effect size for the difference between males and females for structurally different trials should be relatively large while the effect size difference between the genders for mirror trials should relatively small. Secondarily, if males show performance benefits from the structurally different trials and this strategy exists consciously, then we could expect faster reaction times for males than for females.

An alternative hypothesis in these experiments is that the sex difference is accounted for by something other than capitalizing on structurally different foils. Other sources of variance between males and females could exist in any of the other necessary processes of this task: encoding the stimuli, the rotation process, or even the judgments. The two experiments reported herein focus primarily on whether structurally different foils could lead to the large sex differences. Experiment 2 investigates other potential sources of this sex difference through means of spatial visualization measures.

Experiment 1

In the first experiment, the relationship between foil type and accuracy in MRT items was investigated using the Vandenberg and Kuse (1978) task. The traditional timing constraint of three minutes per set of 10 items was used in order to adhere to the original task. Of particular interest were sex differences between foil types. Relative to mirror trials, better performance is expected for both males and females for structurally different trials due to the visually salient structural differences (i.e., parallel versus perpendicular tails), which obviate the need for mental rotation. Second, it was hypothesized that the difference between foil types will be qualified by a male advantage on the structurally different version shown in an interaction between sex and trial type. To investigate these hypotheses, two versions of the original VK task were created to maximize the number of trials. In one version, all 20 trials were mirror object foils and in another version all 20 trials were structurally different object foils. If males are better at the VK MRT simply because they are better at noticing the structurally different foils and use them to identify correct answers, then a larger gender difference in structurally different trials relative to mirror trials is expected.

Method

Participants

Participants consisted of 84 (42 female) University of California, Santa Barbara undergraduates who participated in return for course credit. Subjects were discarded from analyses for skipping pages within the task (n = 2) or recently participating in a similar mental rotation study (n = 2). Forty males and 40 females were included in the final analyses.

Stimuli

Two versions of the Vandenberg and Kuse (1978) Mental Rotations Test were constructed using the original format. In these new versions, 20 new items were created for a total of 40 items. To construct two unique versions, structurally different foils were removed from items 3, 4, 7, 8, 11, 12, 15, 16, 19 and 20 and replaced with mirror foils. The same procedure was carried out for items 1, 2, 5, 6, 9, 10, 13, 14, 17, and 18 replacing the mirror foils with structurally different foils. In the structurally different foils version, if the tails were parallel to one another in the probe figure, the tails of the foils were perpendicular to one another and vice versa (see Figure 2) in order to make all structurally different trials orientation independent. This gives a total of 40 items (20 mirror, 20 structurally different) and preserves the order of the original task as well as the correct answers. The first two pages of the original VK task were used to preserve the instructions and practice items.

Information about strategy usage was collected via a questionnaire in which seven often reported strategies used to solve mental rotation tasks were listed. Participants marked each strategy they used and were allowed to mark as many as they wished (see Appendix A for wording).

Experimental Design

A 2 (foil type) x 2 (sex) x 2 (task order) mixed design was used. All participants saw each version of the VK mental rotation task (within subjects). Order of the two versions was between subjects with 20 males and females receiving either version first.

Procedure

Participants gave informed consent before participation. Participants were then given a test booklet and asked to read the instructions only. When completed, the participants performed three practice items by marking an "x" by the item that was the same as the standard probe. They were oriented to their incorrect choices if any were chosen. As in the original task, three minutes was allotted per set of 10 items. The two foil type versions of the MRT were split into two sets of 10 items each. Each participant completed the two mirror sets then the two structurally different sets, or vice versa. After completing the final set, participants indicated their task strategies via the questionnaire.

Results

Accuracy

First, structurally different foil trials are predicted to be more accurate than mirror trials as it obviates the need to rotate all stimuli in a trial. Moreover, if males are better in VK MRT because of this structurally different foil type, then the difference between the sexes for structural foil trials should be larger than it is for mirror trials. Consistent with the first hypothesis, a significant difference in performance between the two new versions of the MRT was found such that performance on structurally different foil trials (structural M = 11.61, SD = 4.75) was better than performance on mirror foil trials (mirror M = 8.44, SD = 4.82, d = .66). A 2 (task version: mirror and structurally different) x 2 (sex) repeated measures ANOVA was conducted on accuracy data. Table 1 presents descriptive statistics by sex for both different trial types. This analysis indicated that the structurally different foil type version produced significantly higher scores than the mirror foil version, $F(1,78) = 40.81$, $p < .001$, as can be seen in Figure 3. A main effect of sex was also found, $F(1,78) = 12.91$, $p = .001$, such that males performed significantly better at both versions. The task version by sex interaction was not significant, $F(1,78) = 0.82$, $p = .37$, *ns*, suggesting that males and females benefitted equally from structurally different foil trials. In terms of performance, this result provides no evidence that males perform VK MRT better than females due purely to one type of foil within the VK MRT.

To determine if the order of the versions influenced the results above, a repeated measures 2 (version type) x 2 (version order) ANOVA was conducted. A main effect of order was not significant, $F(1,78) = 1.61$, $p = .21$, although a significant interaction between order and version was revealed, $F(1,78) = 11.39$, $p = .001$, due to the fact that the mirror version did not show accuracy changes between orders.

Number of Items Attempted and Proportional Scoring

As has been argued previously, males could show a large advantage in VK MRT for two other reasons: males may be faster, so that they attempt more items or get more of the items they attempt correct. Table 2 presents descriptive statistics for number of trials attempted and proportion correct for each version type by sex. A 2 (foil type) by 2 (sex) repeated measures ANOVA revealed that more items were attempted in the structural version, $F(1,78) = 4.28$, $p = .04$, though males did not attempt more items than females overall, $F(1,78) = 2.86$, $p = .10$, *ns*. Further, the interaction is not significant, $F(1,78) = 1.98$, $p = .16$, *ns*.

The proportion of attempted items that were correct was entered into a 2 (foil type) x

2 (sex) repeated measures ANOVA and revealed that the structurally different foil version showed a greater proportion of items correct, $F(1, 78) = 64.16$, $p < .001$, and males were better proportionally, $F(1,78) = 10.71$, $p = .002$. Again, no interaction was found between foil version and sex, $F(1, 78) = .01$, $p = .93$, suggesting that males do not benefit from the structurally different foils more than females. These results indicate that it is not the case that the number of attempted items or scoring methodology produces evidence that structurally different foils benefit males more than females. These results suggest that rather than the pure number of attempt items, or speed, it is the accuracy for the attempted items that is important.

Strategy Usage

No differences were seen between males (males $M = 5.05$, $SD = 1.18$) and females ($M = 5.15$, $SD = 1.35$) in the overall number of strategies reported, $t(78) = .35$, $p > .05$, *ns* Proportion of participant endorsement is presented in Appendix A alongside the wording for each strategy. One question in the strategy questionnaire asked about the use of the structurally different foil strategy. Males and females did not differ in reports of this strategy, χ^2 (1, $N = 80$) = .72, $p = .40$, ns. Further, chi square tests of independence revealed no significant differences between the sexes on any of the eight self-report strategy questions (including the write-in option), all χ^2 (1, $N = 80$) < 1.00 $p > .05$, *ns*. This finding could be due to the fact that participants were presented with eight equally valid strategies. Participants were able to select all strategies that were used rather than the one or two that they used the most. It is possible that single strategy selection or rank order responses could show differences between sexes in strategy.

Discussion

In the results of this experiment, no evidence is found to support the hypothesis that males capitalize on orientation-independent features of the structurally different foils more than females in the Mental Rotations Test. Participants were more accurate for structurally different foil trials than mirror foil trials overall. Males were more accurate than females for both mirror and structurally different trials. The absence of an interaction between foil type and sex indicates that males did not benefit more from the presence of the structurally different foils than females. Additionally, males and females report using the structurally different foil strategy equally. These lines of evidence provide no evidence that the sex difference in mental rotation arises because males are better at a specific foil type strategy. However, in these data, we also find no evidence that males are inordinately better at mirror trials, which suggests that another process may be involved when considering the sex difference in this task, such as encoding the stimuli, the rotation process itself, or in making or confirming the judgment. These data prompted a second study to investigate whether the sex difference in mental rotation tasks could arise because a) males are faster at structurally different foil trials and/or b) overall time limits mask the effects of the foil strategy.

Experiment 2

In a second experiment, the Shepard and Metzler (1978) chronometric paradigm was used to investigate the effects of various foil types during a mental rotation task. In the original version of this task, only mirror reflections were used for "different" trials in order to eliminate the possibility that participants would use special features to forgo rotation. In order to fully explore the effects of foil type during mental rotation tasks, structurally different foil trials were added. Generally, this paradigm allows for analysis of accuracy *and* reaction times to draw out the effect of each type of trial: same, mirror different, and structurally different.

As suggested by Experiment 1, structurally different foil trials should exhibit better performance than mirror trials. If structurally different trials lead to better performance as a function of a perceptual process (i.e., noticing the orientation independence of the structurally different foils) then reaction time for structurally different foil trials should be faster than mirror trials. Further, if males are better able to notice this difference, then an interaction between trial type accuracy and sex is expected. If, however, this particular trial type leads to better performance simply because the structurally different foils increase the visual differences between the to-be-compared block figures (cf. Cooper & Podgorney, 1976) allowing those answer choices to be easily dropped as potential correct answers by all subjects, then we would expect to see no interaction between "different" trial types (i.e., mirror and structurally different) and sex in terms of accuracy *or* reaction time.

If the sex difference between males and females in mental rotation tasks is due to the process of mental rotation itself, differences could manifest between either a) accuracy and/or reaction time for mirror trials whereby males perform better and/or faster than females and/or b) for same trials where males could show more accurate and/or faster responses to items with greater angular disparity between block figures. Slope and intercept analysis for same trials will help tease apart whether the sex difference arises through the mental rotation process (different slopes) or perhaps some other process such as encoding the stimuli (different intercepts).

Although the focus of this work concerns the nature of sex differences in mental rotation tasks, it is also important to investigate the role of spatial ability in mental rotation tasks as little work has focused on this type of comparison. Relating the sex differences to other factors that could influence any of the processes of the task such as spatial visualization abilities or other factors wholly separate from spatial abilities such as stereotype threat or self-efficacy is also beneficial. Mental rotation tasks typically load on a factor that encompasses speeded rotation, defined by Lohman (1988) as spatial relations. Relating the ability to mentally rotate with a factor that is separable from spatial relations can help us understand the cause sex difference effects seen in mental rotation tasks. One candidate factor that is somewhat separable from spatial relations, although shares common elements with it, is the spatial visualization factor (Hegarty & Waller, 2005). The tasks that load on this factor tend to show mixed findings concerning the magnitude and direction of the sex differences (Voyer et al., 1995) making it in an interesting comparison. Tasks that load on this factor include the Paper Folding Test, Cube Comparison Test, and the Surface Development Test (Ekstrom, French, Harman, & Derman, 1976). The cognitive process that is carried out when comparing structurally different block objects is not necessarily a process that can be measured by a task that loads on spatial relations factor, but rather one of spatial visualization.

Independent of sex differences, it is also fruitful to understand how those participants high in spatial visualization abilities and those participants low in spatial visualization ability differ on a mental rotation task. Of course high and low spatial visualization abilities should indicate rather large differences, the key question is the nature of the differences: are the ability difference similar to the pattern of sex differences? How much larger are the differences? Therefore, the same predictions and hypotheses concerning the sex difference can be applied to high and low spatial visualization ability participants as a between subjects variable. Specifically, high spatial visualization ability participants may realize the usefulness of the structurally different foil types more than low spatial visualization ability participants. An alternative hypothesis is that the orientation independence of structurally different foils affords low spatial visualization ability participants a better chance at solving problems as suggested by Geiser et al. (2006) allowing these participants to focus on feature-based differences rather than mental rotation.

High spatial visualization ability participants also may be faster at solving the various MRT trial types leading to differences between ability levels. The investigation of high and low spatial visualization ability participants may inform questions concerning whether the large sex difference seen in MRT is potentially more aptly described as a spatial visualization ability difference. Slope and intercept analysis of accuracy and reaction time data for high and low spatial visualization ability participants will be pivotal in understanding the nature of the sex differences and differences due to spatial visualization ability.

Method

Participants

Subjects in this experiment were 134 University of California, Santa Barbara undergraduates who participated for course credit. Five participants were dropped from later analyses due to outlier reaction times that were either very short (310 ms) or very long (greater than 20,000 ms) and producing more than 12.5% of trials above or below fast (500 ms) and slow (20,000 ms) response time cutoffs leaving a total of 129 participants (65 females) in the final analyses. Additionally, reaction time outliers for a given trial that was below 500ms and or above 20,000ms were removed from analysis and represented 1.2% of the sample. Participants were run in groups of up to three in the same testing room.

Stimuli

The original drawings of the Shepard and Metzler (1971) objects were presented on a Samsung SyncMaster 2223 monitor with a horizontal visual angle of 23.62 degrees and a vertical visual angle of 12.47 degrees.

In “same” trials, block figures were structurally the same, differing only by rotation at 10 levels of angular disparity between 0 to 180 degrees. Different trials came in two types. Mirror different trials presented two block figures that were mirror opposites. Structurally different trials showed two block figures that were structurally different such that the tails of one block figure were parallel to each other while the tails of other figure were perpendicular to one another. Figure 4 shows examples of each trial type.

Forty trials were created for each trial type. Same trials were repeated in order to balance the number of same and different responses, for a total of 160 trials. Mirror items and structural items were pseudo-matched to the ten levels of angular disparity.

Experimental Design

A 3 (trial type) x 2 (sex) x 10 (angular disparity) mixed design was used. All participants saw all trial types and angular disparity between block figures was a within subjects factor for same trials only. The mental rotation paradigm was always completed first and trials were presented in random order. Three spatial visualization tasks taken from ETS' Kit of Factor-Referenced Tests (Ekstrom et al., 1976) were used to measure the spatial ability of our participants. A 3x3 Latin square was used to order these three tasks to eliminate order effects.

For the Surface Development Test, participants were asked to construct a flat piece of paper presented on the left side of the page into the three dimensional structure on the right by matching the corresponding numbers and letters. Participants were given six minutes for both sets of six items for a total of 12 minutes. Correct responses were awarded a full point while a fifth of a point was subtracted for incorrect responses.

For each item in the Cube Comparison Test two cubes were presented with a letter drawn on each visible face. The participant's task was to decide whether the cubes were the same, but rotated, or different based on the letters that are visible. Three minutes were given for both sets of 21 items for a total of six minutes. One point was given for a correct answer and zero points were awarded for incorrect answers.

The Paper Folding Test requires participants to view a folded piece of paper with a hole punched through its thickness, mentally unfold the paper, and select one of four answer choices that match where the holes should be located. Three minutes were given for each set of 10 item sets for a total of six minutes. One point was given for a correct answer and a fourth of a point was subtracted for incorrect answers.

Procedure

Participants gave informed consent prior to participation. Instructions and trials were performed at a computer using E-prime experiment software (Schneider, Eschman, & Zuccolotto, 2012). During instructions each item type was explained and participants performed six practice trials representing each trial type twice. Participants were then explicitly instructed only to use the left hand for same judgments by pressing a key marked with an "S" and only to use the right hand for different judgments by pressing a key marked with "D." After all instructions and practice trials, participants started the experimental trials. This portion of the experiment was self-paced. After each trial, a central fixation cross was displayed for 500ms. In total, participants saw 160 trials (80 same, 40 mirror, 40 structurally different).

After completing all mental rotation trials, the participants performed the spatial abilities measures. Each group of participants performed the same order of spatial abilities measures so that different timing conditions did not disrupt other participants. Participants read the instructions for each paper test, the experimenter then reiterated the instructions, and they were told to start. See Appendix B for specific wording of the instructions for each spatial ability measure. When the participants had finished all spatial ability measures they were debriefed and given credit for their participation.

Results

Sex Differences

According to our predictions, if the large sex difference in mental rotation is due to the discovery of the structurally different foils, then we would expect a greater difference between males and females for structurally different foil trials relative to mirror different trials. Descriptive statistics and t tests for all trial types are presented in Table 3. Accuracy data for “different” trials were analyzed via a 2 (different trial type: mirror and structural) x 2 (sex) repeated measures ANOVA and revealed that mirror trials showed the lowest accuracy performance, $F(1, 127) = 130.08$, $p < .001$, $\eta_p^2 = .51$, $d = .86$, and that males performed significantly better on both different trial types, $F(1, 127) = 14.19$, $p < .001$, $\eta_p^2 = .10$, $d =$ 61. Contrary to our prediction, the interaction effect was not significant, $F(1, 127) = .72$, $p =$ .40, *ns*.

The sex difference may arise due to faster reaction times by males for different foil type trials relative to females. For this prediction to be correct, then a main effect of sex as well as an interaction between foil type and sex would be present. A 2 (different foil type: mirror and structural) x 2 (sex) repeated measures ANOVA was conducted on median reaction time data. This analysis revealed that mirror trials took significantly longer ($M =$ 5099.10, $SD = 1800.33$) than structurally different trials ($M = 4754.32$, $SD = 1750.53$), $F(1, 127) = 14.70$, $p < .001$, $\eta_p^2 = .10$, $d = .19$, but no other significant main effects or interactions were found, indicating that males and females did not significantly differ in the amount of time to solve either different trial type, $F(1, 127) < 1$, *ns*. Figure 5 shows reaction time by sex for each trial type.

It is possible that the male advantage in VK MRT is because males have more accurate performance at high levels of disparity than females, which would reflect an advantage of mental rotation rather than strategic differences between the sexes. Accuracy and median reaction time data were subjected to 10 (disparity) x 2 (sex) repeated measures ANOVAs. Mauchly's test indicted a violation in the assumption of sphericity ($\chi^2(44) = .16$, $p < .000$) and Greenhouse-Geisser estimates were used as a correction for degrees of freedom ($\varepsilon = .76$). As seen in Figure 6a, accuracy starts to decline significantly after 60 degrees of angular disparity between block figures, $F(6.87, 127) = 58.32$, $p < .001$, $\eta_p^2 = .32$, and males significantly outperform females, $F(1, 127) = 7.80$, $p < .01$, $\eta_p^2 = .06$, d = .26. As suggested by a significant interaction effect, males perform better at higher disparities, $F(6.87, 127) = 2.36$, $p = .02$, $\eta_p^2 = .02$.

Given that males outperform females in accuracy for higher disparity trials, males may perform faster specifically at higher levels of disparity. Another 10 (disparity) x 2 (sex) repeated measures ANOVA was conducted using median reaction time data. Mauchly's test indicted a violation in the assumption of sphericity ($\chi^2(44) = .05$, $p < .000$) and Greenhouse-Geisser estimates were used as a correction for degrees of freedom ($\varepsilon = .63$). As seen in Figure 6b, larger disparities take significantly longer, $F(5.69, 127) = 124.49$, $p < .001$, $\eta_p^2 = .50$, however males and females do not differ significantly in the amount of time taken to solve items nor are males faster at higher disparities, $F(1, 127) < 1.40$, *ns*.

Slope Intercept Analysis. Finally, if the difference between males and females is actually in the mental rotation process, a significant difference should be seen for the slope of the line defining the time to mentally rotate the two objects into congruence. If a significant difference is seen between the sexes in the intercept term, this would point to something other than mental rotation giving rise to the sex difference such as time to encode the stimuli. Mean and standard deviation including Cohen's d for slope and intercept can be found in Table 4. One-way ANOVA analyses were carried out on both accuracy and reaction time data for same trials. Neither slope nor intercept was significantly different between the sexes for accuracy although the effect size was moderate, $F(1,127) < 3.41$, $p = .07$, d > .30, while intercept was significantly different between the sexes in reaction time data, $F(1,127) = 3.95$, $p = .05$, d = .35, such that females showed a larger intercept than males suggesting that the sex difference in mental rotation arises due to some process other than mental rotation.

Spatial Visualization Ability Median Split

Means and standard deviation statistics for all spatial visualization abilities measures can be found in Table 5. Males performed significantly better on all three measures, all $t(127) > 2.70$, $p < .01$. All measures were significantly correlated with each other as well as with overall accuracy in the mental rotation task (all $r > .37$, $p < .01$). See Table 6 for the correlation matrix. Scores on these measures were standardized via z-score transformation and averaged together to produce a composite score of spatial visualization for each participant. From this composite score, a median split was used to separate high and low spatial visualization ability participants. Forty males and 24 females were present in the high spatial ability group while 24 males and 41 females were present in the low ability group. This variable was used as a between subjects variable in the subsequent analyses.

As stated above, high spatial visualization participants are predicted to be more accurate than low spatial visualization participants. An interaction between trial type and spatial visualization ability would indicate one group receiving benefit from either "different" trial type. Descriptive and *t* test statistics for each trial type are presented in Table 7. A 2 (different trial type: mirror and structural) x 2 (spatial ability) repeated measures

ANOVA was conducted on accuracy data. As predicted, high spatial visualization ability participants performed better on each of the different trial types, $F(1, 127) = 9.64$, $p = .002$, $\eta_p^2 = .52$, $d = 1.19$, and lower accuracy was seen for mirror foils for both groups of participants, $F(1, 127) = 138.76$, $p < .001$, $\eta_p^2 = .31$, $d = .86$. Interestingly, a significant interaction effect was opposite of the expected pattern suggesting that low spatial ability participants show greater accuracy gains in structurally different trials than high spatial ability participants, $F(1, 127) = 56.03$, $p < .001$, $\eta_p^2 = .07$. This runs counter to the suggestion that only high spatial ability participants make use of the benefit afforded by the structurally different figures.

Although low spatial ability participants made gains in accuracy for structurally different trials relative to mirror foil trials, it is possible that high spatial ability participants are faster at solving mental rotation trials especially structurally different trials. A 2 (foil type: mirror and structural) x 2 (spatial ability) repeated measures ANOVA was conducted on median reaction time data. As can be seen in Figure 7, structurally different trials were performed faster than mirror trials, $F(1, 127) = 14.81$, $p < .001$, $\eta_p^2 = .10$, $d = .19$, and the two spatial visualization ability levels did not significantly differ in reaction time nor was the interaction significant, $F(1, 127) < 1.27$, $p > .05$, *ns*.

As was seen with the sex variable, if the difference arises as a function of the mental rotation process, then the accuracy of high and low spatial visualization ability participants should start to deviate from one another as the angular disparity increases across the trials. A 10 (disparity) x 2 (spatial visualization ability) repeated measures ANOVA was conducted with accuracy data. Mauchly's test indicted a violation in the assumption of sphericity ($\chi^2(44) = .16$, $p < .001$) and Greenhouse-Geisser estimates were used as a correction for degrees of freedom ($\varepsilon = .78$). As seen in Figure 8a, performance at higher levels of disparity was less accurate than at low levels of disparity, $F(6.98, 127) = 59.23$, $p < .001$, $\eta_p^2 = .32$, and high spatial visualization ability participants are more accurate than low spatial visualization ability participants overall, $F(1, 127) = 30.40$, $p < .001$, $\eta_p^2 = .19$, $d = .61$. As indicated by the significant interaction, high spatial visualization ability participants were more accurate at greater angular disparity than low visualization spatial ability participants, $F(6.98, 127) = 4.51$, $p < .001$, $\eta_p^2 = .03$.

Finally, a 10 (disparity) x 2 (spatial visualization ability) repeated measures ANOVA was conducted using median reaction time. Mauchly's test indicted a violation in the assumption of sphericity ($\chi^2(44) = .05$, p < .000) and Greenhouse-Geisser estimates were used as a correction for degrees of freedom ($\varepsilon = .63$). Shown in Figure 8b, reaction time increased with angular disparity, $F(5.70, 127) = 125.26$, $p < .001$, $\eta_p^2 = .50$; however, spatial visualization ability levels did not significantly differ in reaction time, $F(1, 127) = .82$, $p > .05$, $\eta_p^2 = .01$, $d = .13$, nor was the interaction effect significant, $F(5.70, 127) = 1.65$, $p = .14$, $\eta_p^2 = .01$. This suggests that, although more accurate for trials at greater angular disparity levels, high spatial visualization ability participants are not faster at higher levels of disparity than low spatial visualization ability participants.

Slope Intercept Analysis. One-way ANOVA analyses were carried out for both slope and intercept for accuracy and reaction time data for high and low spatial visualization participants for same trials. Means and standard deviations for slope and intercept can be found in Table 8. For accuracy data, both slope, $F(1, 127) = 6.72$, $p = .01$, d = .44, and intercept $F(1, 127) = 20.25$, $p < .001$, d = .82, were significant such that high spatial visualization participants showed lower slopes than low spatial visualization participants. High spatial visualization participants also showed higher intercepts than low spatial visualization participants. This is expected as better performance should reflected values closer to one, which would indicate perfect performance. For reaction time, only intercept was significantly different, $F(1, 127) = 8.42$, $p = .004$, d = .51, such that high spatial visualization ability participants showed lower intercepts than low spatial visualization ability participants. As with the sex variable, this result suggests that between high and low spatial visualization ability participants, another process beyond mental rotation such as encoding or search is the key factor in the differences between the groups.

Does Spatial Visualization Ability Account for the Sex Difference?

Given that the effects of spatial visualization ability largely mirrored the effects found for sex in accuracy for each trial type and that males showed better performance for all trial types, it could be the case that the sex difference in mental rotation is, in part, due to a spatial visualization ability difference. Males and females were compared on each trial type using the spatial composite score as a covariate in a univariate ANOVA. The sex difference for the same, F(1,127) = 1.56, p = .21, $\eta_p^2 = .01$, and mirror trials, F(1,127) = 1.98, p = .16, $\eta_p^2 = .02$, were eliminated after controlling for spatial ability while for structural trials, the sex difference favoring males was still highly significant, F(1,127) = 6.22, p = .01, $\eta_p^2 = .05$. This suggests that at least for mirror and same trials, the difference between males and females in mental rotation can be explained fully by spatial visualization ability while the structural trials indicate that another factor beyond spatial visualization ability may be contributing to the sex difference.

Discussion

As with Experiment 1, the results of Experiment 2 indicate that participants provided more accurate and faster responses for structurally different trials relative to mirror trials. Males were more accurate, although not faster, for both structurally and mirror different trials relative to females; however, the male advantage in mental rotation tasks was not due to the hypothesized orientation independent strategy of noticing structural differences in block figures. That is, males did not show disproportionate benefit from structurally different trials in accuracy or reaction time over females relative to mirror trials. For same trials, males did show an advantage in accuracy, specifically for angular disparities between 60 and 160 degrees. This benefit did not extend to reaction time. Further, an analysis of slope and intercept between the sexes revealed no differences in accuracy but a significantly different intercept for reaction time data.

When participants were grouped by high and low spatial visualization ability, the results largely followed those for sex in that structurally different trials showed faster and more accurate responses than mirror trials, but were more extreme. High spatial ability participants were more accurate for both mirror and structurally different trials than low spatial ability participants. Interestingly, however, the difference between mirror and structurally different trials was greater for low spatial ability participants than high spatial ability participants. This is in line with evidence from Geiser et al. (2006) that suggests that some low spatial participants who cannot perform the mental rotation instead use an orientation independent strategy. As such, these participants benefit from the orientation independence more than high spatial participants. This result suggests that low spatial ability participants use orientation independent features in foils to identify correct answer choices rather than performing the mental rotation. Reaction time for neither mirror nor structurally different trials was affected by spatial ability level. For same trials, high spatial visualization ability participants were more accurate across the ten levels of angular disparity and significantly more accurate for higher order disparity levels than low spatial visualization ability participants. Although high and low spatial visualization participants were found to be significantly different in slope and intercept for accuracy data, similar results to sex were found for reaction time. This suggests that the sex differences in the mental rotation task is less of a mental rotation difference but rather a difference is another process such as time to encode or discriminate for structurally different foils.

The difference in accuracy between high and low spatial visualization ability participants (*d* = 1.09) was greater than the difference between males and females (*d* = .70). This suggests that the sex differences found in VK MRT may be due to something other than mental rotation ability such as spatial visualization ability differences, and this effect is obscured by the fact that more males typically fall in the high spatial ability ranges. Here, some males fell in the low spatial ability group, and the effects for accuracy between high and low spatial ability groups increased relative to the difference between males and females. However, although spatial ability differences seem to be a large component in the differences between males and females for mental rotation tasks, spatial visualization ability or other factors of spatial ability are not likely to account for the whole story. Other factors such as stereotype threat and general self-efficacy may play a larger role than is often assumed. Taken together, these results suggest that, although mental rotation and spatial visualization ability are linked, the VK MRT also measures other factors that seem to increase the difference between the sexes.

One important result from this experiment that speaks to the importance of not only investigating sex differences but also ability differences was that the sex differences for same and mirror trials were eliminated when controlling for spatial visualization ability, but not for structurally different trials. This result provides at least some credibility to the idea that for structural trials, something beyond spatial visualization ability leads to differences in performance between the sexes.

General Discussion

The focus of the current research was to investigate the large sex difference found in mental rotation tasks. Half of the trials within the VK MRT present structurally different foils while the other half show mirror foils. The structurally different trials can be solved using an orientation independent strategy in which mental rotation is unnecessary to determine that an object is different from a target object (Takano, 1989; Just & Carpenter, 1985). It was hypothesized that this type of foil contributes to the large sex differences in that it aids male more so than female performance. In both experiments presented here, supporting evidence was found that the structurally different trials produced more accurate performance and faster reaction times than mirror trials across all participants. However, males and females made similar gains in performance for structurally different foils relative to mirror foils. This indicates that males are not the sole beneficiaries of the hypothesized orientation independent advantage afforded by the structurally different trials.

Empirical Findings

The results reported here fit well with Voyer and Hou's (2006) finding that males do not benefit over females due to one foil type, although males performed better across foil types. Additionally, the current work is at odds with work that indicated an interaction between gender and foil type such that males benefited more than females from mirror trials (Bors & Vigneau, 2011). No evidence was found here to suggest that males were better due to either foil type.

Additionally, sex differences in mental rotation tasks do not seem to be differences in processing speed of the block figures. In the second experiment, males were not faster than females for mirror or structurally different trials. Similarly, in Experiment 1 males were not found to have completed more items than females in a paper-based version of the MRT. This evidence indicates that the large gender difference is not due to faster processing of the foils by males. This result opposes the finding of Tapley and Bryden (1977) that males performed rotations faster, but is inline with Peters (2005).

Analysis of “same” trials in Experiment 2 provides some insights into the gender differences in performance for VK MRT. It was revealed that accuracy between the sexes began to deviate significantly when the two comparison block figures were presented at greater than 60 degrees of angular disparity, although no reaction time differences were found. This result suggests that females may have a different decision criterion than males for labeling stimuli as “different” since a greater number of same trials for females were erroneously labeled as “different.” This is potentially a consequence of encoding and/or rotation errors. Thus, the gender difference in VK MRT may lie in perceptual differences between males and females. Males may also be benefiting from an ability to hold the pieces of the mental image together long enough to make a “same” judgment at larger disparities. In contrast, females may either a) inaccurately encode the image and/or b) lose parts of the mental image during the rotation process, each of which would lead to inaccuracies.

Finally, slope and intercept analysis for same trials found that reaction time slopes were similar between sexes but the intercept was significantly different. Similarly, when evaluating slope and intercept for reaction time with high and low spatial visualization participants, again, intercept was the significantly different but slope was not. This supports the view that the sex and ability differences are also wrapped up in other processes besides that of mental rotation.

Theoretical Implications

Two theoretical implications are important to emphasize. One implication of the current work concerns the nature of the gender difference in VK MRT as it relates to using a particular strategy focusing on block figure tail directions to solve mental rotation trials. Although the difference between the sexes may not exist as a function of this particular strategy but possibly as a function of other processes within the task such as time to encode the spatial nature of the stimuli or in the decision making process. Beyond that, other factors abound as potential components of laboratory tasks that may influence this sex differences such as stereotype threat and self-efficacy in general.

Another implication of the current work concerns individual differences in other tests of spatial ability. When controlling for spatial visualization ability on performance in Experiment 2, the gender differences were eliminated in same and mirror trials, but not structurally different trials.

Practical Implications

These empirical and theoretical implications then lead directly to practical implications about how best to assess mental rotation ability. First, elimination of structurally different trials in VK MRT could help in reducing the sex differences. The inclusion of such trials increases the likelihood that something other than mental rotation is being assessed such as the ability to notice structural difference. Mental rotation ability, as is currently assessed by VK MRT, may be overestimating mental rotation ability, as the process is unnecessary on half the trials. To that point, the sex difference for same and mirror trials can be eliminated when controlling for spatial visualization ability but not structurally different trials.

Finally, another practical implication of the current work is the relative differences in what the VK MRT and SM paradigms afford the researcher. In the former case, many factors can overestimate the measure of mental rotation. In the SM paradigm, there are many outcome measures allowing for a deeper understanding of the processes at play. For instance, as reported here, SM allows for measurement of reaction time, slope, and intercept analysis for each trial perhaps allowing for more informative conclusions to be drawn rather than just a sweeping assessment that may point to a larger difference than exists.

Limitations and Future Directions

One possible explanation of the finding that sexes differ in reaction time intercept is that this represents time to encode the 3-D nature of the 2-D objects on each trial. Indeed, previous research in this area has noted that occlusion is an important aspect concerning the nature of the sex differences in VK MRT (Voyer & Hou, 2006; Voyer & Doyle, 2010). To understand the contribution of perception in mental rotation tasks, our lab is developing three-dimensional stimuli with salient depth cues to aid participants in identifying the anterior and posterior features of the block objects. This could reduce encoding inaccuracies potentially leading to a reduction in the gender difference, putting males and females on an even field.

We found no evidence that orientation independence benefitted males more than females in accuracy or reaction time. A possible limitation of Experiment 2 is that the stimuli were presented randomly rather than in a predictable order as they are in the VK MRT (e.g., mirror, mirror, structural, structural). An experimental procedure of blocking each type of different trial, as was used in Experiment 1, could help make sense of when participants decide to use a strategy that focuses on orientation independent features. For instance, participants may only employ this strategy near the end of a mental rotation task after assuring themselves that it works on some trials and subsequently use it for the last few items. Collection of reaction time data in this case will indicate if participants ever explicitly notice this difference and if they do, *when* they decide to employ it. The reaction time differences between different foil trials should be stark if participants decide to switch from using mental rotation to the orientation independent strategy. An organization of the data focusing on reaction time as a function of trial order will be particularly illuminating.

Another limitation for VK MRT concerns the angular disparities between block figures within a given trial. Angular disparity between the probe and correct answers, *and* between correct answers within or across items on the task was not explicitly manipulated or controlled in the VK MRT or subsequent versions such as Peters et al. (1995). Given that males and females deviate in performance as angular disparity increases, as seen in Experiment 2, controlling for angular disparity could be important. Future work focusing on systematically varying the angular disparity for each correct answer within an item will help determine what role angular disparity plays in the gender differences found in paper-based versions of the MRT.

Conclusion

Given the large gender difference found using the Mental Rotations Test and its subsequent use to promote theories of education and human evolution, it is important to understand the nature of performance differences in this task and eliminate task specific variables that could influence the nature of the sex difference. Overall, the work considered here provides strong evidence that a specific foil/trial type is not the locus of the large sex difference in VK MRT. Structurally different foil trials provide some advantage to all participants not just males or those with high spatial visualization ability. Further, this work emphasizes that the mental rotation process is not the only source of sex differences in mental rotation tasks.

References

Bors, D. A., & Vigneau, F. (2011). Sex differences on the mental rotation test: An analysis of item types. *Learning and Individual Differences*, *21*(1), 129-132.

Butler, T., Imperato-McGinley, J., Pan, H., Voyer, D., Cordero, J., Zhu, Y. S., ... & Silbersweig, D. (2006). Sex differences in mental rotation: Top–down versus bottom–up processing. *Neuroimage*, *32*(1), 445-456.

Cooper, L. A., & Podgorny, P. (1976). Mental transformations and visual comparison processes: effects of complexity and similarity. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 503-514.

Ecuyer-Dab, I., & Robert, M. (2004). Spatial ability and home-range size: examining the relationship in Western men and women (Homo sapiens). *Journal of Comparative Psychology*, *118*(2), 217-231.

Ekstrom, R. B., French, J. W., Harman, H. H., & Derman, D. (1976). Manual for kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Geiser, C., Lehmann, W., & Eid, M. (2006). Separating "rotators" from "nonrotators" in the mental rotations test: A multigroup latent class analysis. *Multivariate Behavioral Research*, *41*(3), 261-293.

Hegarty, M. (2010). Components of spatial intelligence. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 52, pp. 265–297). San Diego, CA: Academic Press.

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

Jolicoeur, P. (1990). Identification of Disoriented Objects: A Dual- systems Theory. *Mind & Language*, *5*(4), 387-410.

Kail, R., Carter, P., & Pellegrino, J. (1979). The locus of sex differences in spatial ability. *Perception & Psychophysics*, *26*(3), 182-186.

Masters, M. S. (1998). The gender difference on the Mental Rotations Test is not due to performance factors. *Memory and Cognition, 26,* 444-448.

Mumaw, R. J., Pellegrino, J. W., Kail, R. V., & Carter, P. (1984). Different slopes for different folks: Process analysis of spatial aptitude. *Memory & Cognition*, *12*(5), 515-521.

National Research Council. (2006). Learning to think spatially. Washington DC: National Academies Press.

Peters, M. (2005). Sex differences and the factor of time in solving Vandenberg and Kuse mental rotation problems. *Brain & Cognition, 57,* 176-184.

Peters, M., Laeng, B., Latham, K., Johnson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse Mental Rotations Test: Different versions and factors that affect performance. *Brain & Cognition*, 28, 39−58.

Resnick, S. M. (1993). Sex differences in mental rotations: An effect of time limits? *Brain & Cognition*, 21, 71–79.

Schneider, W., Eschman, A., and Zuccolotto, A. (2012). E-Prime User's Guide. Pittsburgh: Psychology Software Tools, Inc.

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

Stieff, M. (2007). Mental rotation and diagrammatic reasoning in science. Learning and Instruction, 17, 219–234.

Takano Y. (1989). Perception of rotated forms: A theory of information types. *Cognitive Psychology*, 21, 1 - 59

Tapley, S. M., & Bryden, M. P. (1977). An investigation of sex differences in spatial ability: Mental rotation of three-dimensional objects. *Canadian Journal of Psychology*, 31, 122−130.

Titze, C., Heil, M., & Jansen, P. (2010). Pairwise presentation of cube figures does not reduce gender differences in mental rotation performance. *Journal of Individual Differences*, *31*(2), 101.

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

Vandenberg, S. G., & Kuse, A. R. (1978). Mental Rotation, a group test of three-dimensional spatial visualization. *Perceptual & Motor Skills*, 47, 599−604.

Vanrie, J., Willems, B., & Wageman, J. (2001). Multiple routes to object matching from different viewpoints: Mental rotation versus invariant features. *Perception*, 30, 1047-1056.

Voyer, D., & Doyle, R. A. (2010). Item type and gender differences on the Mental Rotations Test. *Learning and Individual Differences*, *20*(5), 469-472.

Voyer, D., & Hou, J. (2006). Type of items and the magnitude of sex differences on the Mental Rotation Test. *Canadian Journal of Experimental Psychology*, 60, 91−100.

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

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

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, *101*(4), 817–835.

Yuille, J. C., & Steiger, J. H. (1982). Nonholistic processing in mental rotation: Some suggestive evidence. *Perception & Psychophysics*, *31*(3), 201-209.

Table 1.

Descriptive statistics for each version type by sex in Experiment 1.

	Males		Females		
	M	SD	M	SD	Cohen's d
Mirror	9.8	5.07	7.08	4.19	0.58
Structural	13.43	4.51	9.8	4.32	0.82

Note. M = Mean. SD = Standard Deviation.

Table 2

Average number of items attempted by MRT version type and sex with proportional scoring descriptive statistics in Experiment 1.

	Number Attempted		Proportion Correct	
	M	*SD*	*M*	*SD*
Male Mirror	14.65	3.99	0.66	0.24
Male Structural	16.1	3.91	0.84	0.17
Female Mirror	13.93	3.7	0.52	0.24
Female Structural	14.2	4.13	0.69	0.21

Note. M = Mean. SD = Standard Deviation.

Table 3

Accuracy for each MRT trial type by sex in Experiment 2.

	Males		Females			
	M	*SD*	*M*	*SD*	*t*	Cohen's d
Overall	0.85	0.09	0.78	0.11	3.97***	0.7
Same	0.87	0.09	0.82	0.11	2.80**	0.5
Mirror	0.77	0.19	0.67	0.18	3.11**	0.54
Structural	0.9	0.1	0.82	0.13	4.04***	0.69

** $p < .01$; *** $p < .001$.

Note. M = Mean. SD = Standard Deviation.

Table 4

Mean and standard deviations for slope and intercept analysis of the sex variable in Experiment 2.

	Male		Female		
	M	*SD*	*M*	*SD*	Cohen's *d*
ACC Slope	-0.0011	0.0009	-0.0014	0.0008	.35
ACC Intercept	0.97	0.06	0.94	0.09	.39
RT Slope	17.31	9.68	17.29	11.10	.00
RT Intercept	2140.56	766.89	2426.63	864.42	.35

Note. M = Mean. SD = Standard Deviation.

Table 5

Descriptive statistics for each spatial ability measure by sex in Experiment 2.

	Male			Female		
	M	*SD*	*SEM*	*M*	*SD*	*SEM*
Paper folding	13.04	3.95	0.49	10.47	4.42	0.55
Surface development	38.27	14.32	1.79	29.04	16.77	2.08
Cube comparison	21.23	9.34	1.17	16.8	9.2	1.14

Note. M = Mean. SD = Standard Deviation. SEM = Standard Error of the Mean. Males performed significantly better on all three measures ($p < .01$).

Table 6

Correlations between spatial ability measures and overall mental rotation accuracy in Experiment 2.

Measure	1	2	3
1. Paper folding	--		
2. Surface development	.62	--	
3. Cube comparison	.47	.50	--
4. Overall MRT ACC	.37	.50	.55

Note. All correlations $p < .01$.

Table 7

Spatial ability median split accuracy by trial type in Experiment 2.

	High		Low			
	M	SD	M	SD	*t*	Cohen's d
Same	0.89	0.07	0.80	0.11	5.51***	0.98
Mirror	0.82	0.13	0.62	0.2	6.74***	1.19
Structural	0.92	0.07	0.80	0.13	6.73***	1.14

*** $p < .001$

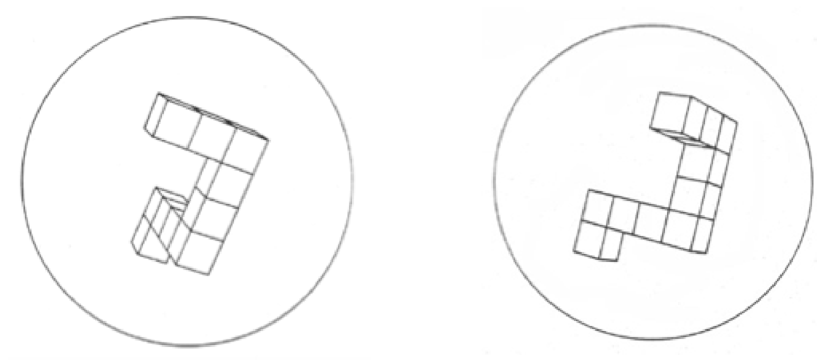
Note. M = mean. SD = Standard Deviation.

Table 8

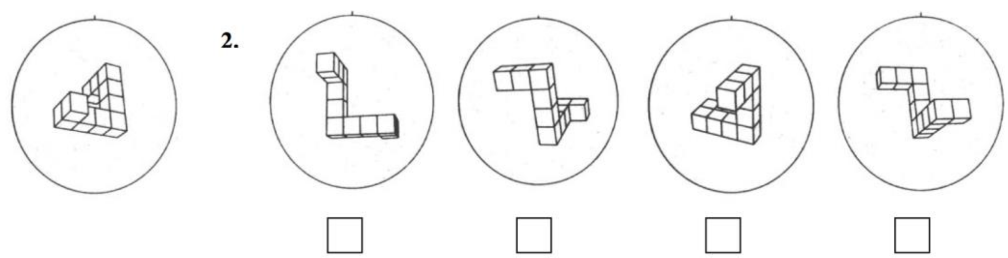
Mean and standard deviations for slope and intercept analysis of the spatial visualization median split variable in Experiment 2.

	High		Low		
	M	SD	M	SD	Cohen's d
ACC Slope	-0.0010	0.0009	-0.0014	0.0009	.44
ACC Intercept	0.99	0.05	0.93	0.09	.82
RT Slope	18.28	9.21	16.33	11.40	.19
RT Intercept	2077.70	639.51	2488.52	937.96	.51

Note. M = mean. SD = Standard Deviation.



a)



b)

Figure 1. Mental rotation item presentation between a) Shepard and Metzler (1971) and b) Vandenberg and Kuse (1978). Note that in SM one trial is presented at a time while in VK MRT several items are presented on a page.

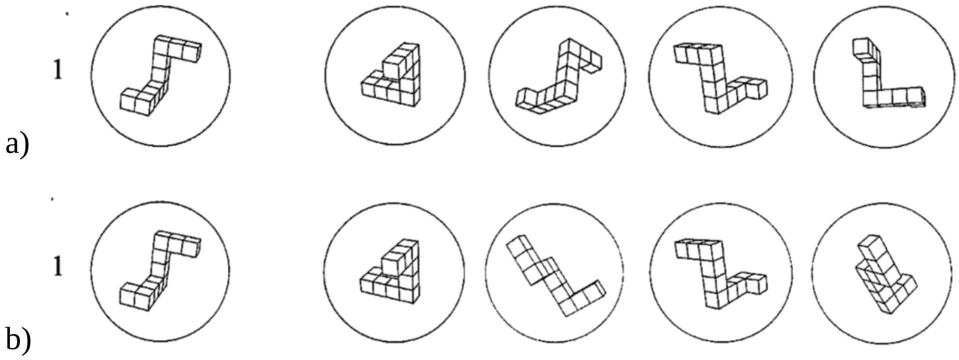


Figure 2. Mirror (a) versus structurally different (b) trials used in Experiment 1. Note that in (b) the tails of the probe on the left are parallel while the tails of the foils on the right are perpendicular.

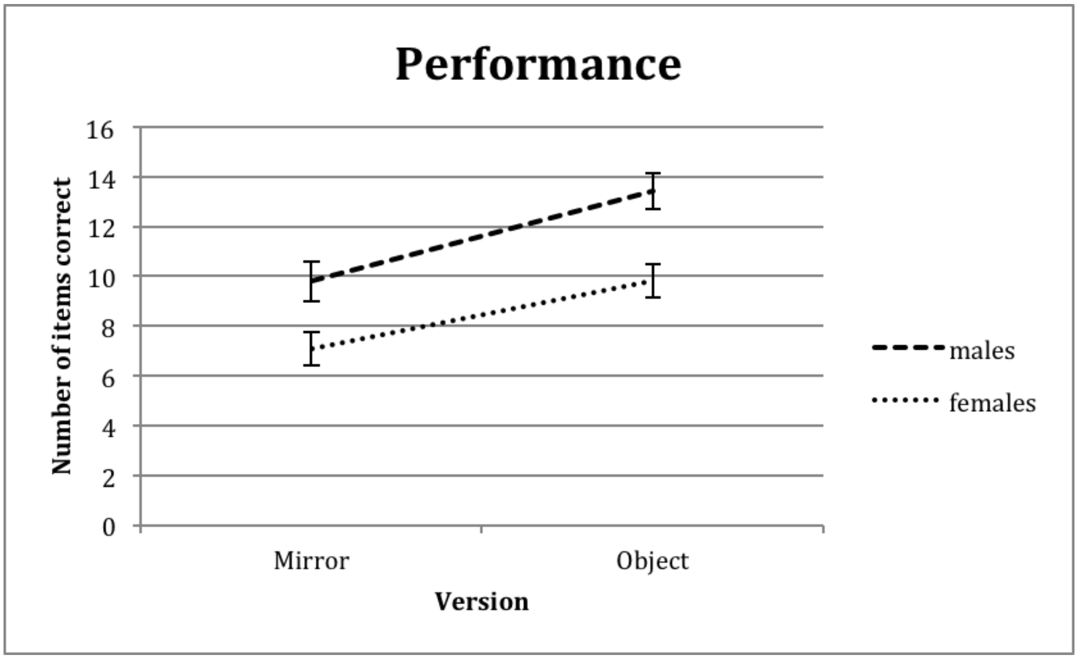


Figure 3. Performance by sex for both foil-type specific versions of the mental rotations task used in Experiment 1. Error bars indicate standard error of the mean.

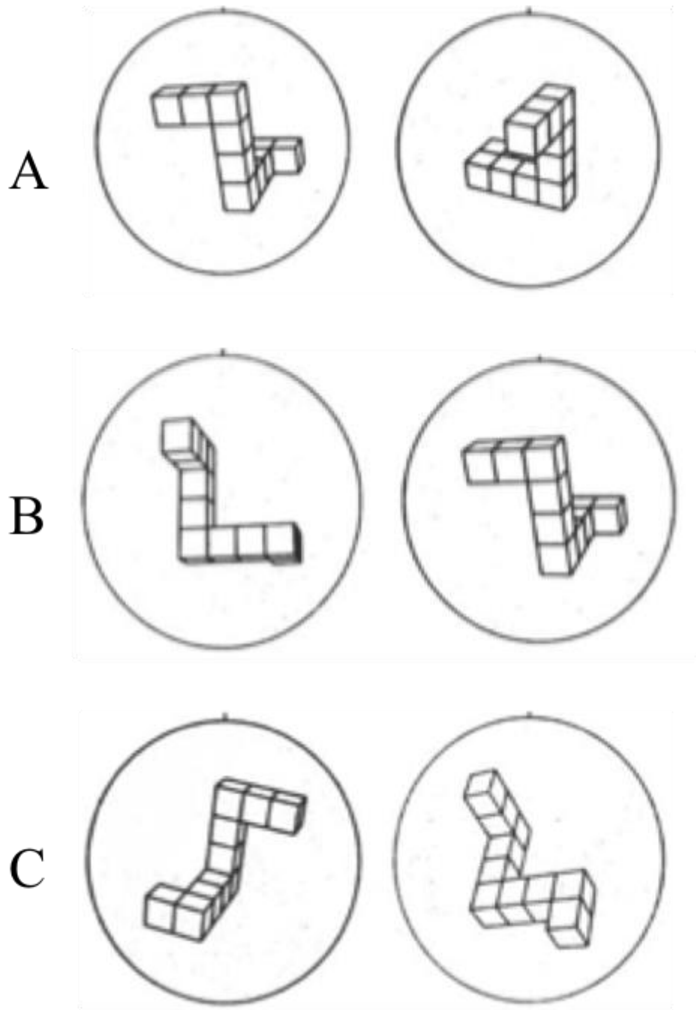


Figure 4. Shepard & Metzler paradigm trial types used in Experiment 2: Same (a), mirror different (b), and structurally different with parallel and perpendicular tails (c).

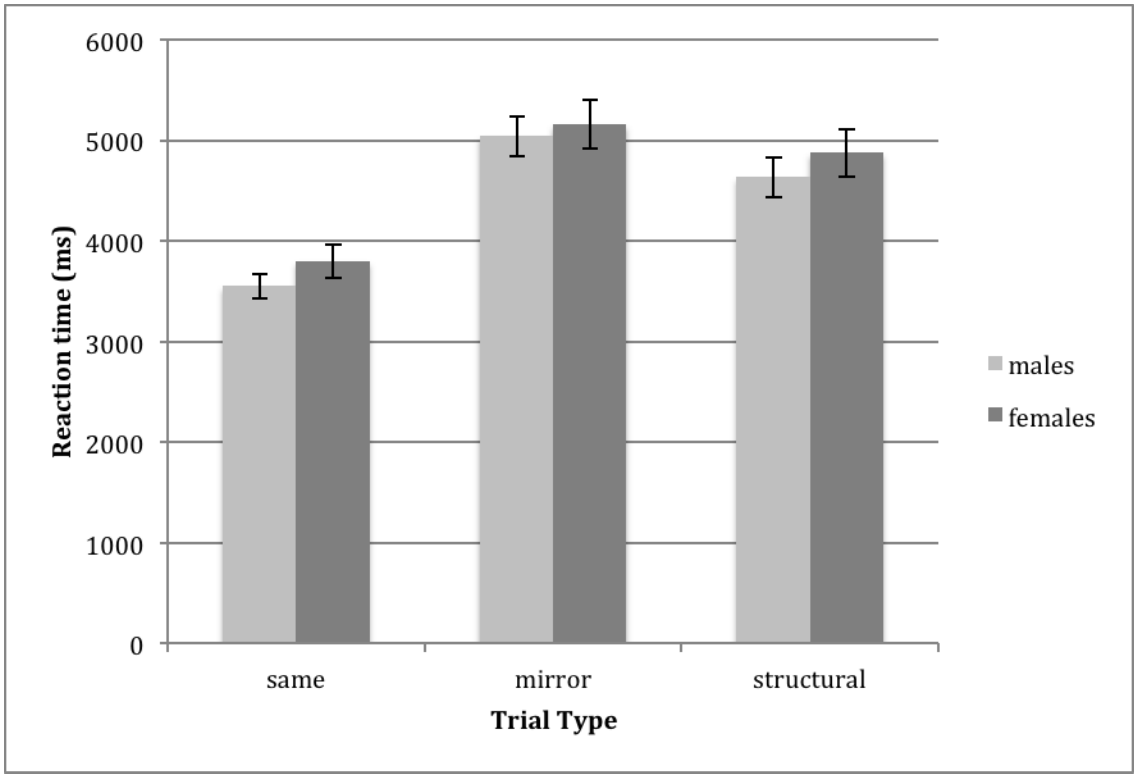
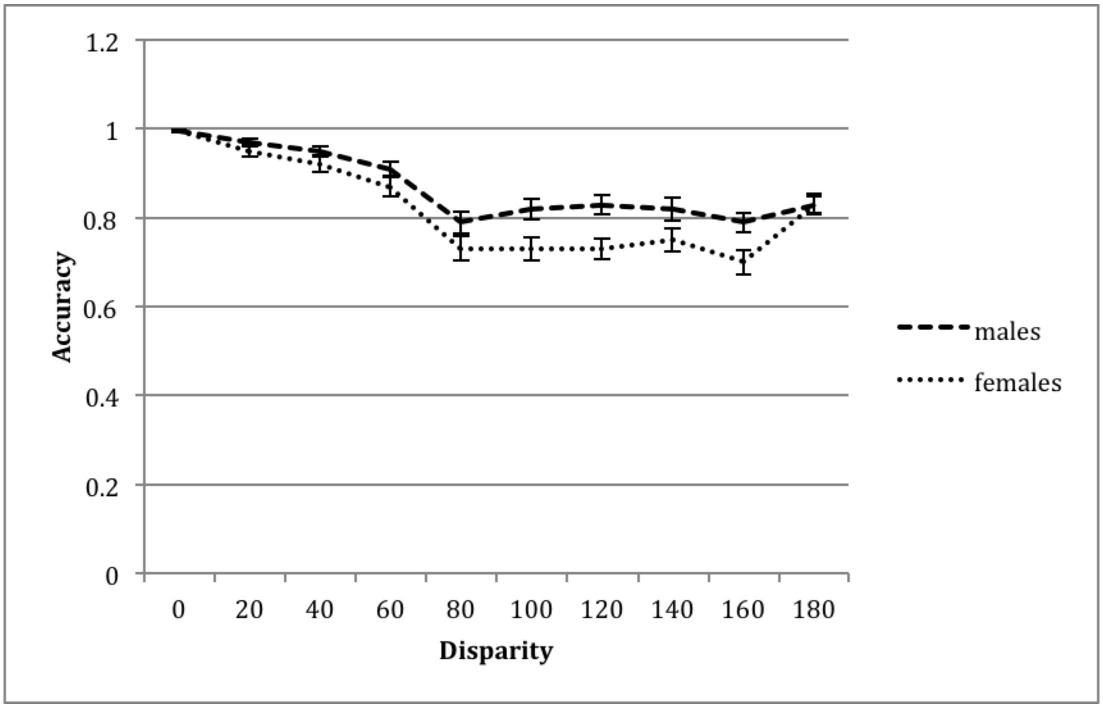


Figure 5. Reaction time in milliseconds for each trial type in Experiment 2 for by sex. Error bars indicate standard error of the mean.

a)



b)

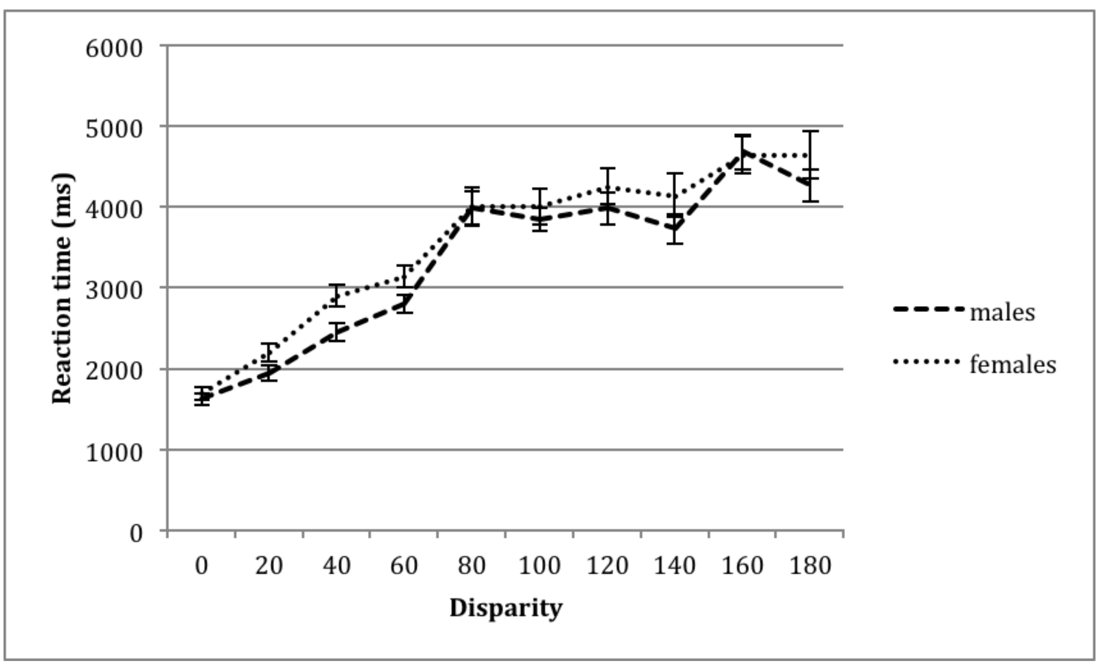


Figure 6. Accuracy (a) and reaction time in milliseconds (b) plots by disparity level for males and females in Experiment 2. Error bars indicate standard error of the mean.

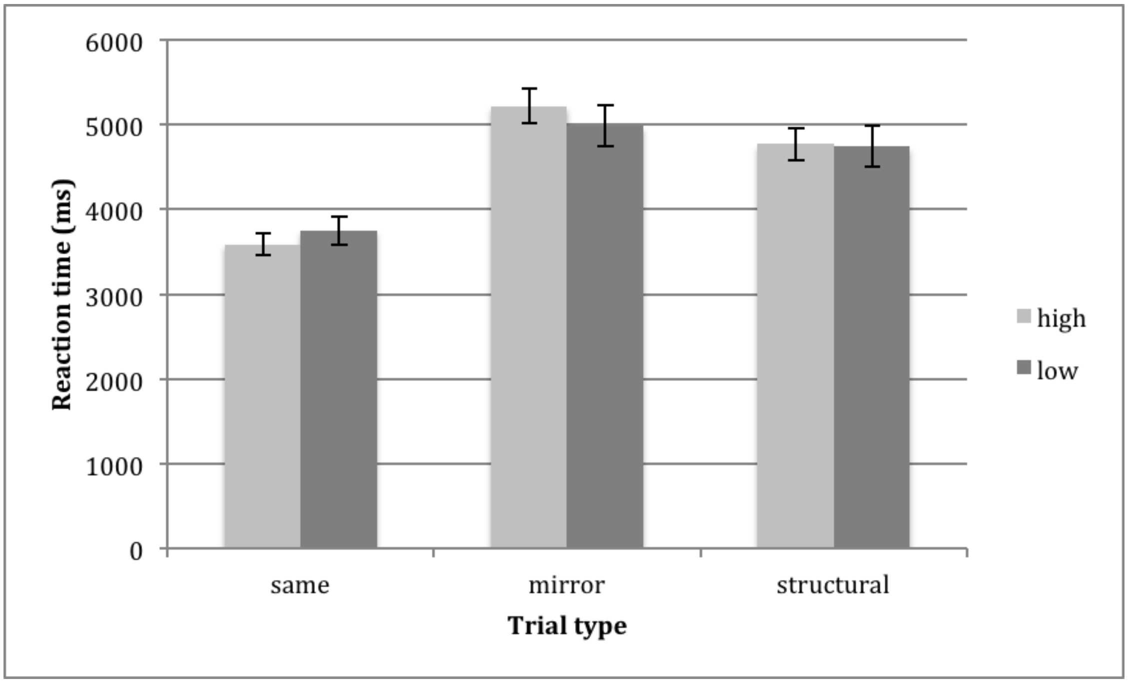
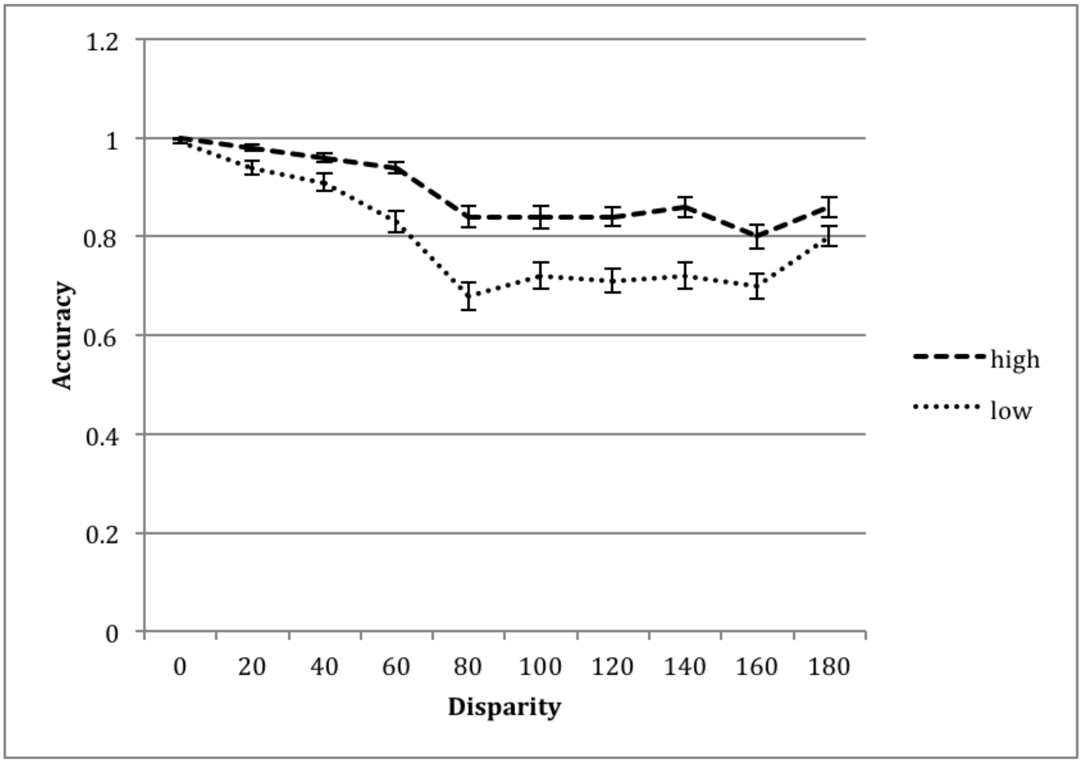
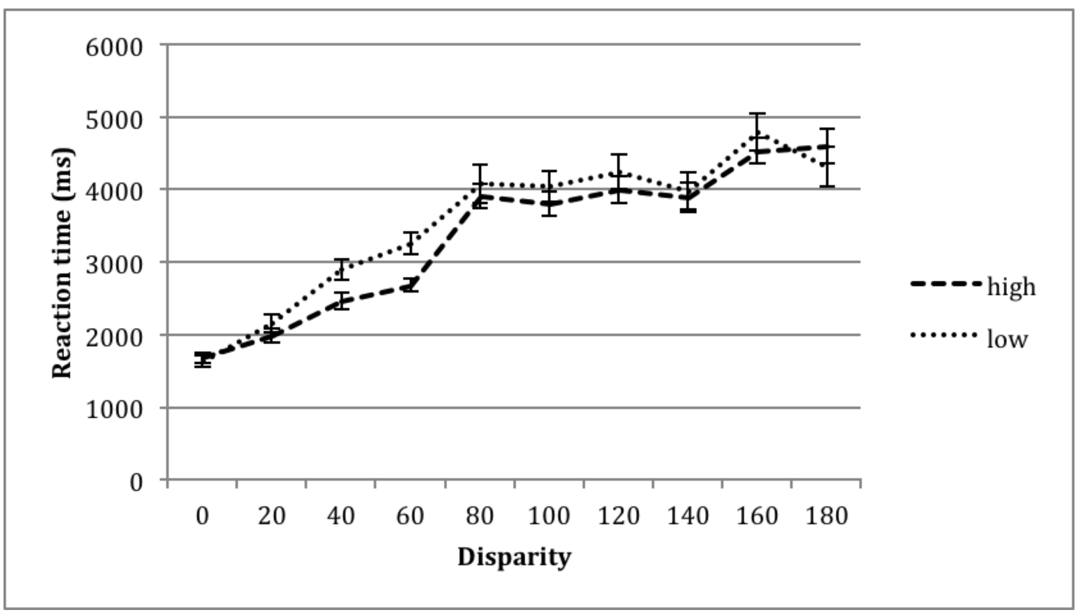


Figure 7. Reaction time in milliseconds by trial type for high and low spatial ability participants in Experiment 2. Error bars indicate standard error of the mean.



a)



b)

Figure 8. Accuracy (a) and reaction time in milliseconds (b) plots for each level of disparity for high and low spatial ability in Experiment 2. Error bars indicate standard error of the mean.

Appendix A

Mental rotation strategy questionnaire used after participation in Experiment 1. Parentheses after each strategy reflect the proportion of males and females participants from Experiment 1 endorsing that strategy.

Mental Rotation:

1. Please indicate which of the following strategies you used to work out the answer to the rotation items that you just completed. Tick as many as apply to your strategies.

__ I imagined the target turning in my mind. (67.5% males, 77.5% females)

__ I imagined the answer choices turning in my mind. (80% males, 82.5% females)

__ I imagined looking at the object from a different viewpoint or perspective. (75% males, 75% females)

__ I counted the number of cubes in different arms of the target and checked whether this matched in the different answer choices. (80% males, 77.5% females)

__ I noted the directions of the different arms of the target with respect to each other (up, to the right, down etc.) and checked whether theses directions matched the answer choices. (90% males, 95% females)

__ If an answer choice was hard to see, I skipped over it and tried to respond without considering that answer choice in detail. (22.5% males, 30% females)

__ I figured out whether the two end arms of the target were parallel or perpendicular to each other and eliminated answer choices based on this feature. (For example if the two end arms were parallel in the target, I eliminated any answer choices in which they were not parallel). (72.5% males, 62.5% females)

__ I used the following strategy not listed above (20% males, 15% females)

Please describe your strategy here:

Appendix B

Experimenter script for each spatial ability measure used in Experiment 2

Paper folding test:

"In this test, you will see drawings of a square sheet of paper that has two to three folds. In the final drawing of the folded paper, a hole is punched through all of the thickness of the paper at that point. Your job is to select one of the five drawings that shows how the paper would look if it was fully reopened. Your score on the test is based on both the number of questions that you get correct and incorrect. So it is not advantageous to guess unless you are pretty sure of the answer. You will have three minutes to complete each set of ten. Do not go beyond those ten items."

Cube comparison test:

"In this test, you will see two drawings of a cube. Assuming that no cube can have two faces alike, you should indicate whether the two cubes can be of the same cube or if they are different cubes. Be sure to consider the orientation of the letters on a given face. No numbers, letters, or symbols appear on more than one face for a given cube. Your score on the test is based on both the number of questions that you get correct and incorrect. So it is not advantageous to guess unless you are pretty sure of the answer. You will have three minutes to complete 21 items. Do not go beyond those 21 items."

Surface development test:

"In this test, you will be asked to visualize how a piece of paper can be folded to form an object. You will see two drawings – a drawing of a piece of paper on the left with dotted lines indicating where it can be folded, and on the right is the paper completely folded into an object. Your job is to figure out which of the lettered edges on the object are the same as the numbered edges on the drawing. Make note of how the side of the drawing marked with the X will always be the same as the side of the object marked with the X. The drawing must always be folded so that the X will be on the outside. Your score on the test is based on both the number of questions that you get correct and incorrect. So it is not advantageous to guess unless you are pretty sure of the answer. You will have six minutes to complete six items. Do not go beyond the first six."