Gender Differences in the Mental Rotations Test (MRT) Are Not Due to Task Complexity

Corinna Titze, Martin Heil, and Petra Jansen

Institute of Experimental Psychology, University of Düsseldorf, Germany

Abstract. Gender differences are one of the main topics in mental rotation research. This paper focuses on the influence of the performance factor task complexity by using two versions of the Mental Rotations Test (MRT). Some 300 participants completed the test without time constraints, either in the regular version or with a complexity reducing template creating successive two-alternative forced-choice tasks. Results showed that the complexity manipulation did not affect the gender differences at all. These results were supported by a sufficient power to detect medium effects. Although performance factors seem to play a role in solving mental rotation problems, we conclude that the variation of task complexity as realized in the present study did not.

Keywords: gender differences, visual-spatial cognition, mental rotation, task complexity

Whereas females outperform males on, for example, measures of verbal fluency, males outperform females on certain tests of spatial ability (e.g., Halpern, 1992). This male advantage is largest on tasks involving mental rotation (Voyer, Voyer, & Bryden, 1995), and within these, it is largest on the MRT (Mental Rotations Tests; Vandenberg & Kuse, 1978; Peters, Laeng, Latham, Jackson, Zaiyouna, & Richardson, 1995). The MRT consists of 24 items, each presenting a 3-D target block figure (taken from Shepard & Metzler, 1971) and four choice figures. Two of these are identical to the target but are rotated in depth, while two can not be matched regardless of how they are rotated. Usually (see Voyer et al., 1995), a single point is given if and only if both correct matches are identified. The test is usually presented in two sets of 12 items each, with a 3-min time constraint for each set.

Despite an enormous amount of research, the cause(s) for the gender difference, however, are still far from being understood. The explanations being offered include psychosocial ones (such as stereotype threat, Shih, Pittinsky, & Ambady, 1999; sex role identification, Signorella & Jamison, 1986; or differential experience and socialization, Baenninger & Newcombe, 1989) and biological-neuronal ones (such as sex hormone level, Imperato-McGinley, Pichardo, Gautier, Voyer, & Bryden, 1991; rate of maturation, Sanders & Soares, 1986; or cerebral lateralization, McGlone, 1980) and, for all of them, a certain amount of empirical support exists.

At the same time, however, these explanations run the risk of overgeneralization and, thus, it might help to identify in detail the empirical facts that have to be explained, especially to identify which task factors do and do not affect the size of the gender difference. For our present study, two task factors are especially relevant, i.e., time constraints and task complexity. Goldstein, Haldane, and Mitchell (1990) reported findings that the gender difference on the MRT disappears when subjects were allowed sufficient time to attempt all items or when the scoring procedure was controlled for the number of items attempted. In contrast to Goldstein et al. (1990), however, and in line with the majority of the published data (see, e.g., Delgado & Prieto, 1996; Resnick, 1993), Masters (1998) showed that the gender difference was *not* affected by time constraints. Neither the scoring method nor the time limits used modified the size of the gender difference, a result which was also obtained by Peters (2005). Peters found evidence that although females attempted fewer items than males under the standard timing condition, the magnitude of the gender difference did not change when subjects did the MRT with double the usual time allowed for the test. Thus, one can conclude that time constraints are not critical for gender differences in the MRT.

Additionally, the Peters study presented one very interesting and, hitherto, not sufficiently considered finding that might be crucial for understanding the causes for the gender difference in the MRT: In Experiment 3 of Peters (2005), a selected sample of subjects dealt with the MRT twice. These subjects were drawn from a larger sample pretested on the MRT and were selected on the basis of scoring within one standard deviation of their gender mean. As a consequence, gender differences in both the initial and the second MRT administration resulted in an effect size of Cohen's (1977) *d* greater than 1. In between the two MRT administrations, subjects solved a chronometric version of

DOI 10.1027/1614-0001.29.3.130

1.a

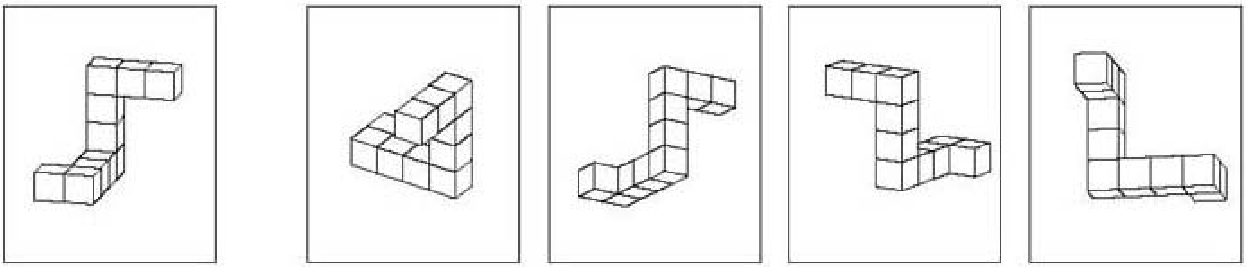

Figure 1. An example of an item used in the Mental Rotations Test. The target item is shown on the left and the four sample stimuli are presented aside. Always two of these are identical to the target item but are rotated in depth.

the mental rotation task with pairs of cube stimuli of the kind encountered in the MRT. In chronometric studies of mental rotation, two stimuli are presented with varying angular disparity, and response times and error rates are measured when participants decide whether these two match when mentally aligned. Interestingly, in this sample preselected to establish substantial and reliable gender differences in the MRT, no gender difference at all was observed for the chronometric version of mental rotation (see also Jansen-Osmann & Heil, 2007, for corroborative results). This finding raises the (important) question of the task factors that differ between the MRT and the chronometric version of mental rotation and that might be responsible for the (non)existence of gender differences in performance. Thus, identifying these task factors might be crucial to understanding the causes of the gender difference.

Obviously, the MRT and the chronometric version of mental rotation differ in a number of aspects. The MRT is a paper and pencil test that differs from a computer-administered task. On the one hand, the computer task forces participants to respond to all items whereas with the paper-pencil test, participants have greater control how they manage their time and effort. For example, they could decide to spend more time on difficult items, to skip these but return to them later, double-check the answers at the end, etc. On the other hand, only one item has to be compared with another one in the computer task, whereas in the MRT participants have to compare one item with four items. This task is much more complex than the two-alternative forced-choice task because all five items are presented at the same time, look very similar, and might give the impression of a very complex and difficult task. This impression might evoke more strongly the stereotype of being unable to solve spatial problems in women (compare Shih et al., 1999).

To investigate this phenomenon of task complexity we compared two groups of participants with one solving the MRT in the standard version but without any time limit. The second group solved the MRT with the aid of a template that, per item, breaks down the two-out-of-four alternatives choice into four consecutive two-alternative forced-choices, similar to the tasks in the computer tests.

Methods

Participants

In this study 300 subjects (150 women) participated. Their age ranged from 18 to 35 years (M = 24). This homogeneous group of participants was recruited on campus and needed to have the school-leaving examination to be allowed to take part in this experiment. Half of the subjects were assigned randomly to the template group; the other half to the nontemplate group.

Material and Procedure

We used the paper-pencil MRT, (Version A) redrawn by Peters et al. (1995), which was originally developed by Vandenberg and Kuse (1978) with figures created by Shepard and Metzler (1971). This test consists of two sets with 12 items that contain respectively a target item on the left side and four sample stimuli on the right. Participants had to identify those two out of four sample stimuli that show the target item in a rotated version. Figure 1 shows an example of the items used.

In the original test of Peters et al. (1995), the items were presented to the participants on four DIN A-4 sheets, with six items per sheet and a 3-min deadline to solve a set of 12 items (6 min for the entire test). Instructions were given in written form, followed by three training items so that participants became familiar with the task. The correct solutions of these training items were shown at the end of the page. We used the original test but abandoned any time limit, i.e., participants were allowed as much time as needed to solve all 24 items. Moreover, participants were explicitly instructed to attempt a solution to all items.

In the template group, participants solved the MRT with the help of a 50 × 30 cm black board template with a 5 × 20 cm horizontal whole in its middle. The template allowed participants to watch exactly one item at a time. Additionally we used two more black board templates measuring 5 × 7 cm and 5 × 4 cm, respectively. These templates were constructed to mask three out of the four sample stimuli to ensure that subjects always compared just one sample stimulus with the target item. In this way, the two-out-of-four alternatives choice usually inherent in the MRT was broken down into four consecutive two-alternative forced-choices that subsequently were checked to fulfill the criterion of two positive and two negative choices.

The participants were tested individually. An investigator was present in each test session in both experimental groups. First, participants read the instructions and solved the three training items. Participants in the template group were not allowed to inspect all four sample stimuli before solving an item. They just saw the target item on the left side all the time and only one by one of the sample stimuli. By using the template, three of the four sample stimuli were masked by the investigator so that the participant had to compare each sample stimulus individually with the target item. When the participant had made his or her decision, this stimulus was masked and the next one was uncovered until all four sample stimuli were decided. When one item had been finished, participants checked with the help of the investigator if exactly two answers were marked as correct. If there were correct answers missing or if more than two sample stimuli were marked correct, the participant was asked to correct his or her choices but was now allowed to view all four sample stimuli at the same time, so that they had to fulfill again a two-alternative forced-choice task. To be sure that participants could not see the test as a whole the covering templates were operated by the investigator.

Statistical Analysis

The standard scoring method proposed by Peters et al. (1995) was used: One point was given if and only if both correct sample stimuli of a target item were marked correctly. Thus, participants could obtain 24 points maximum.

The design of the study involved two factors: Gender (male, female) and Experimental group (with or without template). The dependent variable was the number of correctly answered items in the MRT-A. Given a total sample size of $N = 300$ and a desired level of significance $\alpha = .05$, effects of size $d = 0.5$ (that is, medium effects as defined by Cohen, 1977) could be detected with a probability $1\text{-}\beta = .95$[1].

Results and Discussion

An ANOVA revealed a significant main effect of Gender with $F(1, 296) = 13.26, p < .001$; men gaining in mean two points more than women. The effect size of this gender difference amounted to $d = 0.42$. No main effect of factor Experimental group, $F(1, 296) = .32$, *ns*, was found. There also was no significant interaction between both factors, $F(1, 296) = .07$, *ns*. The effect sizes did not differ significantly depending upon whether participants did or did not use the complexity reducing template ($d = 0.45$ vs. $d = 0.40$). The means and their standard errors are presented in Table 1.

Table 1. Means and standard errors in the MRT as a function of subjects' gender and experimental group

Gender	Complexity of stimuli	Mean correct	Standard error
Male ($N = 150$)	with pattern ($N = 75$)	21.21	.54
	without pattern ($N = 75$)	20.76	.50
	total	20.99	.37
Female ($N = 150$)	with pattern ($N = 75$)	19.01	.56
	without pattern ($N = 75$)	18.91	.59
	total	18.99	.40
Total ($N = 300$)	with pattern ($N = 150$)	20.14	.40
	without pattern ($N = 149$)	19.83	.39
	total	19.98	.28

Discussion

The present study aimed to identify relevant task factors, such as task complexity, which might affect the size of the gender difference in mental rotation. Recently, Peters (2005) observed no gender differences at all for the chronometric version of mental rotation with pairs of cube stimuli similar to the ones used in the MRT in a sample preselected to establish substantial and reliable gender differences in exactly this MRT. The absence of gender differences in a same/different judgment with comparable stimuli in a sample that reliably produced a gender-effect size of about one standard deviation in the MRT is an intriguing finding that deserves to be explained. Moreover, the reason for this discrepancy might also turn out to be a prime candidate for explaining gender differences in spatial cognition in general.

Obviously, the MRT and the chronometric version used by Peters (2005) differ in a number of aspects, and substantial research efforts are needed to identify the relevant task factors. In this paper, we concentrated on one important aspect, i.e., the complexity of the task, especially the decision required. Whereas the MRT asks for two-out-of-four alternative choices, the chronometric versions usually use a two-alternative forced-choice task. Our aim was to realize a test situation comparable to the standard MRT as much as possible. Therefore, we used the original paper-and-pencil MRT und reduced the complexity of the experimental group from the two-out-of-four alternative choices to four consecutive two-alternative choices, which subsequently had to be checked to fulfill the criterion of two positive and two negative ones. This manipulation capitalized on the findings of, e.g., Masters (1998) and Peters (2005) that the gender differ-

[1] All power calculations reported in this article were conducted using the G·Power program (Erdfelder, Faul, & Buchner, 1996).

ence in the MRT was not affected by time constraints. In line with these results, we obtained a reliable gender difference of medium effect size according to the definition of Cohen (1977) although (1) subjects were tested individually, (2) no time constraint was realized at all, and (3) all subjects attempted to solve all items. The complete removal of time constraints, in fact, directly asks for an individual testing situation. As a consequence, the average number of 20 problems solved correctly was substantially larger than what is usually observed with time constraints in a group-testing situation (e.g., about 12 problems in Peters, 2005). Nevertheless, although all subjects in both the control and the experimental group attempted to solve all items, men, on average correctly, solved two more problems than women did.

The manipulation of task complexity, however, did not affect the size of the gender difference at all. This conclusion is supported by the power of 96% to detect an effect of complexity. Whatever the relevant task factor is to explain the intriguing finding of Peters (2005), task complexity does not seem to be critical. More research is needed to solve this puzzle. We started with the aim of eliminating the gender difference in the MRT by introducing as few modifications to the MRT as possible, therefore, we did not split the MRT in separate two-alternative forced-choice tasks on a computer screen. However, an alternative approach might be to start from the chronometric version of the mental rotation task to identify what modifications are needed to evoke the gender difference that is absent in the chronometric version but is reliable in the MRT. With this paper we made a first contribution to the research efforts that are needed – and as a result, we have to exclude task complexity as realized here as a factor responsible for the gender difference in the MRT.

Acknowledgments

This study was supported by the German Research Foundation. We thank Michael Peters for his friendly permission to use the Mental Rotations Test (MRT) in our study. We thank our student assistants for help during data acquisition.

References

Baenninger, M., & Newcombe, N. (1989). The role of experience in spatial test performance: A meta-analysis. *Sex Roles*, *20*, 327–344.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Delgado, A.R., & Prieto, G. (1996). Sex differences in visuospatial ability: Do performance factors play such an important role? *Memory and Cognition, 24*, 504–510.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, and Computers, 28*(1), 1–11.

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

Halpern, D.F. (1992). *Sex differences in cognitive abilities*. Hillsdale, NJ: Erlbaum.

Imperato-McGinley, J., Pichardo, M., Gautier, T., Voyer, D., & Bryden, M.P. (1991). Cognitive abilities in androgen insensitive subjects – Comparison with control males and females from the same kindred. *Clinical Endocrinology, 34,* 341–347.

Jansen-Osmann, P., & Heil, M. (2007). Suitable stimuli to obtain (no) gender differences in the speed of cognitive processes involved in mental rotation. *Brain and Cognition*, *64*, 217–227.

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

McGlone, J. (1980). Sex differences in human brain asymmetry: A critical survey. *Behavioral and Brain Sciences, 3*(2), 215–263.

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

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

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

Sanders, B., & Soares, M.P. (1986). Sexual maturation and spatial ability in college students. *Developmental Psychology*, *22*, 199–203.

Shih, M., Pittinsky, T.L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science*, *10*, 80–83.

Shepard, R.N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*(3972), 701–703.

Signorella, M.L., & Jamison, W. (1986). Masculinity, femininity, androgyny, and cognitive performance: A meta-analysis. *Psychological Bulletin*, *100*, 207–228.

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

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.

Corinna Titze

Institute of Experimental Psychology
Heinrich-Heine-University
Universitätsstr. 1
D-40225 Düsseldorf
Germany
Tel +49 211 811-2134
Fax +49 211 811-5037
E-mail corinna.titze@uni-duesseldorf.de

