APPLYING SDT TO A SPEEDED PAPER AND PENCIL TEST

John S. Monahan and Maureen A. Harke
Department of Psychology, Central Michigan University, Mt. Pleasant, MI 48859
john.monahan@cmich.edu

Abstract

Forty-eight women and 48 men completed the Revised Mental Rotations Test (Peters,1995) (RMRT) twice, three weeks apart. Between tests half of them completed six practice tests made from figures similar to but different from those on the RMRT. RMRT has 24 items; each has a target figure and four alternative figures, two rotations of the target plus mirror images or structurally different figures. Items are scored correct if both rotations are marked, or otherwise, incorrect. Standard RMRT scores, sensitivity (d'), and bias (β) were analyzed. It appears that practice aids women more than men and does so by improving rotated pattern recognition and by speeding up the recognition process for some types of figures thus producing higher d' on some types of items and reducing β on all types.

Measured gender differences in cognitive abilities appear to have diminished over time (Wraga, Duncan, Jacobs, Helt, & Church, 2006), but mental rotation, continues to yield large statistically significant differences favoring men. Linn and Peterson (1985) found that mental rotation, continued to yield a large effect size favoring men. A more recent meta-analysis (Voyer, Voyer, & Bryden, 1995) found a linear increase in effect size with increasing age across categories suggesting the possibility that gender differences in spatial ability may be affected by experience, sexual differentiation, or both. The largest effect was found with the Mental Rotations Test (MRT) (Vandenberg & Kuse, 1978).

Sexual differentiation alone fails to take into account how experience influences ability and brain function (Voyer, 1995). In addition to the biological differences between women and men, there are experiential differences between genders (Terlecki & Newcombe, 2005), such as early childhood toy preference and play (Voyer et al., 2000). For example, puzzles are completed by mentally or physically rotating pieces to make comparisons. Dolls and board games do not emphasize spatial relationships and may be less likely to influence spatial ability. Voyer et al. found that men and women who reported childhood preference for spatial toys performed better on the MRT than those who preferred non-spatial toys. Never-the-less, men's performance was greater overall.

If long-term involvement in spatially oriented activities affects spatial ability, short-term experiences may also. The effects of practice are debated among researchers. One idea is that participants retrieve stored stimuli from memory and therefore answer more questions faster and more accurately on repeated tests without improvement on novel stimuli (Heil, Rosler, Link, & Bajric, 1998; Tarr & Pinker, 1989; Wiedenbauer, Schmid, & Jansen-Osman, 2007). Feng, Spence, and Pratt (2007) found that 10 hours practice on an action video game reduced gender difference in the MRT. Another is that practice sessions as brief as one hour may improve participants' performance independent of the orientation of the practiced stimuli indicating an improvement in spatial ability (Murray, Jolicoeur, McMullen, & Ingleton, 1993).

For the MRT and its successor, Peters's (1995) Revised MRT (RMRT), standard scoring procedure is to count as correct each item for which both correct alternatives were marked. This scoring method ignores some of the additional information available: items for which only one correct alternative is marked and items for which foils are marked incorrectly. One way to assess that additional information is via signal detection theory (SDT) analysis.

Method

Participants

There were 96 college students recruited from introductory psychology classes: 48 women and 48 men, 24 of each in the practice and control groups. The practice group took the RMRT, practiced for 6 sessions over three weeks using similar, but different items, and then took the RMRT again. The control group took the RMRT, did not practice and after three weeks took the RMRT again. Practice group participants received course credit plus $25. Control groups participants received course credit plus $5. Practice and control group participants were tested in separate semesters.

Materials

RMRT. The RMRT consists of 24 items, each with a target figure and four alternatives. Two alternatives are rotations of the target; the other two alternatives are foils. Voyer and Hou (2006) classify items as occluded if one of the figures, target or foil, has a section not visible because of a rotation of the figure, or non-occluded if each figure has all sections visible. Eight non-occluded items have two figures, structurally different from the target, as foils; two structurally different foil items have occluded figures. Another eight non-occluded items have rotated mirror images of the target as foils; four mirror image foil items have occluded figures. Two items have mixed foils; one has non-occluded figures; the other has an occluded figure.

Practice Items. Three sets of 24 practice items, five figures per item, were devised using a CAD program. Each figure consisted of 10 cubic sections with at least three right angles projected onto a two dimensional surface. None of the practice figures were identical to any of the figures in the RMRT. Items consisted of a target figure and four alternative figures: two were images of the target rotated on different axes, and two were distracter alternatives: a mirror image of the target and a structurally different figure. Each set of practice items was reordered by randomization, and a reordered set of figures was devised from each item. The original target figure became a rotation of the new target figure, and the order of alternatives was randomized.

Procedure

Participants were administered the initial RMRT. Practice participants were administered two practice sets per week for three weeks beginning the following week. The practice sets for each week consisted of an original set of 24 items and a reordered set of those items with figures within items, including the target, reordered. These were administered on separate days. At the end of each practice session, participants were shown a list of the correct answers to each item. Control group participants did not receive any practice. After the three week period, participants were again administered the RMRT. Correct answers to the RMRT were never shown.

Administration of the tests and practice sets was conducted using the instructions for the RMRT (Peters, 1995), including allowing 3 min for each of the two 12 item sections with a 2 min break between sections. Participants were subsequently debriefed and paid for their service.

Results

Figure 1 shows items correct, *d'*, and β for tests 1 and 2. Tests were scored as the number of items correct, i.e. the number of items in which both correct alternatives were marked, yielding a maximum score of 24. SDT measures were scored using each of the alternatives and foils of each item, thus there were 48 signal present items and 48 signal absent items.

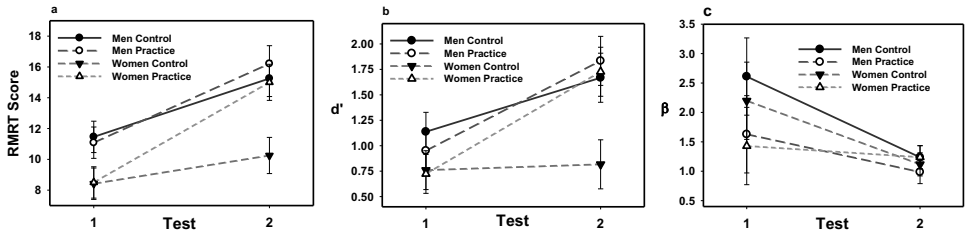

Figure1: Overall results for tests 1 and 2 as a) items correct, b) d', and c) β. Error bars represent SEM.

Similar scoring occurred for each of the three item types, structurally different, mirror image, and occluded alternative, except that proportion rather than number of correct items was used. Initial analyses were conducted using partially repeated measures ANOVA. Condition means were compared using α=.005.

Items Correct. Test, $F(1,92) = 113.75$, p<.0001, sex, $F(1,92) = 8.93$, p=.002, and the interaction between test and practice, $F(1,92) = 4.24$, $p = .042$, had significant effects on items correct. Practice group men had higher scores than women on test 1, but not test 2. Control group men had higher scores than women on both tests. Both women and men attempted about four more items on the second than the first test.

d'. Test, $F(1,92) = 47.86$, $p < .001$, and the interaction between test and practice, $F(1,92) = 13.17$, $p < .001$, had significant effects on *d'*. Sex had a marginal effect, $p = .052$. Practice group men's mean *d'* was not significantly greater than women's on test 1 or 2. Control group men's *d'* was greater than control group women's on test 2. Practice group men and women improved *d'* from test 1 to test 2.

Table 1. *Proportion Correct by Item Type, Test, and Gender for Control and Practice Group Participants*

	Control Group			Practice Group		
Test	Men	Women	d	Men	Women	d
			Structurally Different Foils, Eight Items			
1st	0.54 (0.29)*	0.43 (0.23)	0.44	0.54 (0.26)*	0.36 (0.17)	0.77
2nd	0.70 (0.25)*	0.59 (0.22)	0.45	0.78 (0.21)*	0.69 (0.21)**	0.42
			Mirror Image Foils, Eight Items			
1st	0.54 (0.29)*	0.38 (0.28)	0.58	0.49 (0.29)	0.43 (0.26)	0.23
2nd	0.68 (0.31)*	0.40 (0.32)	0.89	0.66 (0.33)	0.67 (0.31)**	0.03
			Occluded Alternatives, Seven Items			
1st	0.36 (0.19)	0.28 (0.23)	0.37	0.37 (0.25)	0.29 (0.18)	0.35
2nd	0.54 (0.30)*	0.28 (0.26)	0.91	0.60 (0.28)	0.54 (0.26)**	0.25

Note: Standard deviations in parenthesis. d is Cohen's d and reflects gender difference effect size for participants in the same group for each test.

*Men's proportion correct greater than women's from the same group, $p < .005$.

**Practice group proportion correct greater than control group for the same gender, $p < .005$

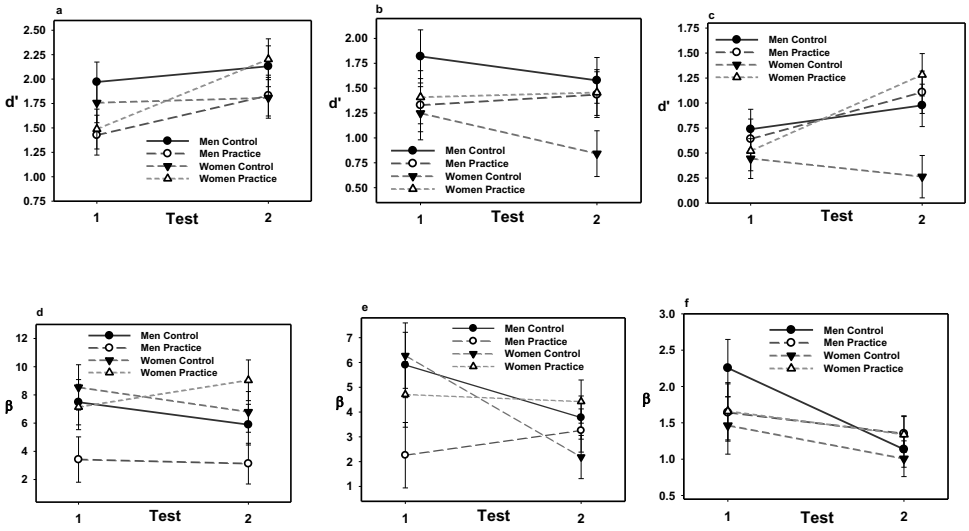

Figure 2: a), b), and c) are *d'* and d), e), and f) are β for structurally different foils, mirror image foils, and occluded alternative items for practice and control group women and men.

β. Only test had a significant effect on β, $F(1,92) = 8.25$, $p = .005$. Overall β on test 1 was greater than on test 2. Thus except for the control group women, Items correct and d' improved from test 1 to test 2, and for everyone, criterion became stricter.

Proportion correct on the three subtypes of items are shown in Table 1. Figure 2 shows *d'* and β for items with structurally different foils, mirror image foils, and occluded alternatives.

Structurally Different Foils. Proportion correct on non-occluded, structurally different foil items were affected by test, $F(1,92) = 92.85$, $p < .0001$, sex, $F(1,92) = 8.38$, $p = .005$, and the interaction of test and practice, $F(1,92) = 6.85$, $p = .01$. With structurally different foil items, although women and men in both groups improved their proportion correct from test 1 to test 2, men scored higher than women in the same group on both tests. The result was different with *d'*; only test had a significant effect on *d'*, $F(1,92) = 6.68$, $p = .011$. Practice group women and men improved their *d'* from test 1 to test 2, and women's d' was greater than men's on test 2. On the other hand only sex had a significant effect on β, $F(1,92) = 5.97$, $p = .016$. Practice group men had a stricter criterion than practice group women on both tests.

Mirror Image Foils. Proportion correct on non-occluded, mirror image foils was affected by test, $F(1,92) = 31.6$, $p < .0001$, sex, $F(1,92) = 4.94$, $p = .029$, and the interaction between test and practice, $F(1,92) = 5.95$, $p = .017$. Practice group women and men and control group men improved proportion correct from test 1 to test 2, and control group men out scored control group women on both tests. None of the three factors or their interactions had an effect on *d'*, nevertheless, control group men had higher *d'* than control group women on both tests. Test, $F(1,92) = 4.39$, $p = .039$, and the interaction between test and practice, $F(1,92) = 6.95$, $p = .01$, had significant effects on β. Control women reduced β from test 1 to test 2. Practice group men had a stricter criterion than women on test 1, but control group women had a stricter criterion than men on test 2.

Occluded Alternatives. Proportion correct on occluded alternative items was affected by test, $F(1,92) = 49.32$, $p < .0001$, sex, $F(1,92) = 7.03$, $p = .009$, the interaction between test and practice, $F(1,92) = 10.19$, $p = .002$, and the interaction among test, sex, and practice,

$F(1,92) = 4.17$, $p = .044$. Practice group women and men and control group men increased their proportion correct from test 1 to test 2, and control group men scored higher than women on test 2. Both test, $F(1,92) = 8.82$, $p < .004$, and the interaction of test and practice, $F(1,92) = 7.31$, $p < .008$, had significant effects on *d'*. Practice group women improved their *d'* from test 1 to test 2, and control group men had a greater *d'* than did women. Only test had a significant effect on β, $F(1,92) = 6.72$, $p < .011$. Control group women's and men's criteria became stricter from test 1 to test 2.

Discussion

Figure 1 shows that with adequate practice, women can score well on the RMRT, but that previous test taking without practice has little effect on women's performance. This result is confirmed by *d'* data, again women with practice improved their sensitivity to a level similar to men's, but women without practice did not. Interestingly, everyone's criterion became stricter by test 2, regardless of sex or practice. Possibly, all had a better idea of what to look for. Thus, although *d'* includes more of the available data in its calculations, the overall picture it presents is little different from the usual number of correct items.

For specific item types, the results were somewhat different. For structurally different foil items, proportion correct results were different from the overall results: men in both groups scored higher on both tests. These were the easiest items according to proportion correct scores. The *d'* scores for these items showed a different pattern: practice group women and men improvement in *d'* from test 1 to test 2, and practice group women had higher *d'* than men on test 2. Practice group men had a stricter criterion than women on both tests.

For mirror image foil items, proportion correct was higher for control, but not practice, group men than women on both tests. Practice group men and women and control group men improved their proportion correct from test 1 to test 2. Control group women did not. The *d'* scores did not increase for men or women in either group. There was a reduction in β from test 1 to test 2 for control group women. Practice group men had a lower β than women on test 1, but control group women had a lower β than men on test 2.

For occluded alternative items, all but control group women increased their proportion correct from test 1 to test 2. Practice group men scored higher than women on test 2. Practice group men scored higher than women on test 2. Practice group women increased their *d'*, and control group men had higher *d'* then women on test 2. Control group women and men made their criterion stricter from test 1 to test 2.

Both men and women in the practice group increased their overall *d'*, indicating that practice improves overall sensitivity rotated figures. However, for two of the item types, structurally different foils and occluded alternatives, almost all improved d'. Structurally different and mirror image foil items were in the practice items, thus the improvement in practice group women's *d'* on structurally different foil items may be due to practice. On the other hand, there were no occluded figures on the items in the practice sets. Thus, the fact that women improved their *d'* on occluded alternative items cannot be straightforwardly attributed to practice, because they did not practice with such items. If practice was at work, it must have been a general effect, rather than a specific one. Feng et al. (2006) suggest that improvement in women's MRT performance in their experiment was caused by improved spatial attention. Possibly, practice with MRT-like items improves spatial attention, as well as giving practice on some types of items. That no one improved *d'* on mirror image foil items was unexpected because such items were included in practice. An explanation for that phenomenon is not obvious at this time.

Because *d'* measures sensitivity in the absence of any control for speed, it changes a speed test score into a power test score. Thus caution is needed when interpreting the results.

Practice group women scored overall as well as practice group men. Spaced practice of six 15 min sessions appears to be enough to reduce the gender difference on MRT to non-significance on both items correct as well as sensitivity. It is not known, however, whether this gender equality can be sustained without continuing practice.

References

Feng, J., Spence, I. & Pratt, J. (2007). Playing an Action Video Game Reduces Gender Differences in Spatial Cognition. *Psychological Science, 18* (10), 850-855.

Heil, M., Rosler, F., Link, M., & Bajric, J. (1998). What is improved if a mental rotations task is repeated - the efficiency of memory access, or the speed of a transformation routine? *Psychological Research***,** *61,* 99-106.

Linn, M.C., & Peterson, A.C. (1985). Emergence and characterization of gender differences in spatial abilities: A meta-analysis. *Child Development, 56,* 1479-1498.

Murray, J.E., Jolicoeur, P., McMullen, P.A., & Ingleton, M. (1993). Orientation-invariant transfer of training in the identification of rotated natural objects. *Memory & Cognition, 21,* 604-610.

Peters, M. (1995). *Revised Vandenberg & Kuse Mental Rotations Tests: Forms MRT-A to MRT-D.* Guelph (ON), Canada: Technical Report, Department of Psychology,

Tarr, M.J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21,* 233-282.

Terlecki, M.S., & Newcombe, N.S. (2005). How important is the digital divide? The relation of computer and videogame usage to gender differences in mental rotation ability. *Sex Roles: A Journal of Research, 433,* 433-441.

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. (1995). Effect of practice on laterality in a mental rotation task. *Brain and Cognition, 29,* 326-335.

Voyer, D., & Hou, J. (2006). Type of items and the magnitude of gender differences on the mental rotations test. *Canadian Journal of Experimental Psychology, 60,* 91-100.

Voyer, D., Nolan, C., & Voyer, S. (2000). The relation between experience and spatial performance in men and women. *Sex Roles, 43,* 891-915.

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.

Wiedenbauer, G., Schmid, J., & Jansen-Osmann, P. (2007). Manual training of mental rotation. *European Journal of Cognitive Psychology, 19,* 17-36.

Wraga, M., Duncan, L., Jacobs, E.C., Helt, M., & Church, J. (2006). Stereotype susceptibility narrows the gender gap in imagined self-rotation performance. *Psychonomic Bulletin & Review, 13,* 813-819.