Learning and Individual Differences 21 (2011) 129–132

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif

Sex differences on the mental rotation test: An analysis of item types

Douglas A. Bors [a,*], François Vigneau [b]

[a] Department of Psychology, University of Toronto at Scarborough, 1265 Military Trail, Toronto, Ontario, Canada M1C 1A4
[b] École de psychologie, Université de Moncton, Moncton, Nouveau-Brunswick, Canada E1A 3E9

ARTICLE INFO

Article history:
Received 14 September 2009
Received in revised form 20 September 2010
Accepted 26 September 2010

Keywords:
Gender effect
Mental Rotation Test
Rasch analysis

ABSTRACT

Replicating a finding now common in the literature, the present study revealed a significant difference between the performance of men ($M = 19.66$; $SD = 5.34$; $SK = 0.52$) and the performance of women ($M = 14.85$; $SD = 6.06$; $SK = -0.38$, Cohen's $d = 0.90$) on the Mental Rotation Test (Vandenberg & Kuse, 1978). In an attempt to identify determinants of the observed sex differences, hypotheses related to MRT item types (structural, mirror, occluded, and non-occluded) first suggested by Voyer (Voyer & Hou, 2006) were tested. Although the results in part supported the notion that different types of items may contribute to the production of sex differences, the bulk of the sex differences remained unexplained. In terms of factor structure, it is argued that the MRT is best conceptualized as a consistent set of non-independent items all contributing to the sex differences. Furthermore, neither the MRT as a whole nor any of the proposed subsets of items were found to be sex biased. In fact, the performance of men was significantly superior to that of women on all of the test's items. Finally, the sex differences were not found to be related to the degree of angular disparity in correct response choices. This last finding suggests that the sex differences observed on the MRT may not be primarily related to the ability to rotate objects mentally.

Various researchers have reported differences in performance between men and women on the Mental Rotation Test (MRT; Vandenberg & Kuse, 1978), with men, on average, outperforming women. Expressed in terms of Cohen's (1988) d, the size of the sex differences on the MRT ranges from about $d = 0.60$ to around $d = 1.00$. For example, in two studies reported by Voyer and Saunders (2004), the size of the sex effect was $d = 0.95$ and $d = 1.03$. In a separate study, Voyer and Hou (2006) reported data that translate into an effect size of $d = 0.73$.

In the present paper, we examine sex differences hypotheses based on item types. These hypotheses focus on subsets of items that are said to be more difficult for women than for men, and hence mainly responsible for the sex differences on the MRT. Voyer and Hou (2006) described four subsets of items associated with this approach. The first two subsets, called structural and mirror items, are based on differences in the nature of the items' distractors. Structural items are items whose distractor configurations are structurally different from the configuration of the target figure. Mirror items are items whose distractors are simply a mirror image of the target figure. It has been hypothesized (Voyer, Rodgers, & McCormick, 2004) that structural items would be more easily solved than mirror items.

The other two subsets of MRT items described by Voyer and Hou (2006) involve the distinction between occluded and non-occluded items. Occluded items contain figures (correct responses or distractors) in which parts of the configuration are occluded, either fully or partially, as a result of the two-dimensional representation of an object rotated in the three-dimensional space. Non-occluded items are deemed to involve occlusion only to a limited extent. Occlusion is assumed to make the task more difficult because it could more easily result in misperceptions of the figures.

In an attempt to validate the item type hypotheses, Voyer and Hou (2006) did not find support for the structural/mirror item distinction with respect to sex differences (no main effect of distractor-type, no sex by distractor-type interaction). However, their results supported the occluded/non-occluded distinction, with a main effect of occlusion and a sex by occlusion interaction both reaching significance. The magnitude of the sex differences, expressed as Cohen's ds, were 0.86 for occluded items and 0.49 for non-occluded items.

Positing individual differences with respect to item types on the MRT assumes that the test is multidimensional. The dimensionality of a test can be assessed in various ways, the most common of which is factor analysis. However, factor analysis, because of its reliance on distributional normality, is not suitable to assess the dimensionality of tests comprised of dichotomously scored items, particularly when the items have a large range in difficulty. Item response modeling, and Rasch modeling in particular, represent approaches adapted to the assessment of the dimensionality of dichotomously scored instruments. Specifically, these techniques have the advantage of being insensitive to the shape of item distributions. Rasch analysis was not developed for the explicit purpose of investigating the dimensionality

* Corresponding author. Fax: +1 416 287 7642.
E-mail addresses: bors@utsc.utoronto.ca (D.A. Bors), vigneaf@umoncton.ca (F. Vigneau).

1041-6080/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.lindif.2010.09.014

of psychometric instruments. However, given that unidimensionality is an essential assumption of Rasch analysis, proponents of the approach have developed means of testing this postulate. In the present article, we use Rasch analyses to investigate the dimensionality of the MRT as a whole and of the various item clusters that have been defined by Voyer and Hou (2006).

In addition to item types, the present study will also examine the possible effects of the degree of angular disparity on sex differences. Angular disparity is here defined as the difference, relative to the main vertical axis, between the orientation of the target figure and the orientation of a correct response choice. Given that men have been found to be faster than women on tasks of mental rotation (Tapley & Bryden, 1977), it is surprising that, to date, there does not appear to have been any studies examining the relationship between sex differences on the MRT and the extent of rotation required to identify the correct responses. If speed of rotation reflects a subject's ability to rotate, sex differences should be more pronounced on those items with a greater degree of angular disparity than on those items with smaller degrees of disparity. Such a finding would not only locate an important source of the sex difference on the MRT but also support the notion that the MRT is in fact a test of mental rotation ability.

1. Method

1.1. Subjects

The data reported and analyzed were collected from administration of the MRT to 624 undergraduate students (407 women, 217 men) from the Université de Moncton and the University of Toronto. Subjects ranged in age from 17 to 58 years ($M = 21.0$, $SD = 4.3$); 5.4% of them were older than 25. Mean age was similar across sexes (men: $M = 21.1$, $SD = 4.1$; women: $M = 20.9$, $SD = 4.4$, $t(622) = 0.50$, $p = 0.62$).

1.2. Procedure

Subjects were administered with the instructional items, the practice items, and the 24-item MRT (in the version redrawn by Peters et al., 1995) using the standard Vandenberg and Kuse's (1978) instructions and timing. These instructions invite the subjects to refrain from guessing unless they had a good idea that their choice is correct. Ten minutes was allotted for the 24 test items, and the subjects were informed when there were 5 min and 2 min remaining.

The MRT was administered in large groups of volunteers ranging approximately from 30 to 70 subjects. Some of the subjects were given credits in an introductory psychology course for their participation in the study. Unless otherwise stated, scoring in the present study was according to Vandenberg and Kuse (1978), that is, a point was given for each item on which both correct alternatives were identified, and no points for anything else. Scores from this scheme thus range between 0 and 24.

2. Results

As expected, there was a significant difference between the performance of men ($M = 14.85$; $SD = 6.06$; $SK = -0.\ 38$) and the performance of women ($M = 9.66$; $SD = 5.34$; $SK = 0.52$) on the MRT, $t(395.5) = 10.6$, $p < 0.05$ (t corrected for heterogeneity of variance). The magnitude of the sex differences expressed as a Cohen's d was 0.93. This is equivalent to a correlation between sex and score of $r = 0.41$. As expected, controlling for responsiveness (by scoring correct items as a proportion of items attempted) produced some attenuation in the sex effect. The correlation between sex and score, however, was only partially reduced (from 0.41 to 0.30), leaving the bulk of the sex effect unaccounted for.

As can be seen from Table 1, not only was there a sex effect for the MRT total score, but there also was a significant sex effect for each and every item on the test (r ranged from 0.10 to 0.33). Although the sex effect was significant for all items, the size of the sex effect increased with increases in item difficulty, $r = 0.64$.

Table 1
Voyer and Hou's (2006) MRT item classification, proportion correct for women (407) and men ($N = 217$), and item correlations between performance and sex.

Item	Type	Women	Men	*r*	Item	Type	Women	Men	*r*
1	M; NO	0.60	0.79	0.20	13	S; NO	0.55	0.74	0.18
2	M; NO	0.60	0.79	0.19	14	S; O	0.40	0.66	0.24
3	S; NO	0.85	0.92	0.10	15	M; O	0.22	0.52	0.30
4	S; NO	0.71	0.85	0.16	16	M; NO	0.25	0.59	0.33
5	M; NO	0.55	0.81	0.26	17	S; O	0.09	0.29	0.26
6	M; NO	0.67	0.82	0.16	18	Mixed; O	0.30	0.54	0.23
7	S; NO	0.67	0.87	0.21	19	M; NO	0.24	0.48	0.25
8	S; NO	0.71	0.87	0.18	20	M; NO	0.19	0.44	0.26
9	M; O	0.22	0.50	0.28	21	S; NO	0.15	0.36	0.23
10	M; O	0.43	0.62	0.18	22	S; NO	0.16	0.37	0.23
11	M; O	0.24	0.54	0.30	23	Mixed; NO	0.14	0.36	0.26
12	M; NO	0.53	0.80	0.27	24	S; NO	0.15	0.31	0.19

Note. S: structural; M: mirror; O: occluded; NO: non-occluded. All correlations are significant at $p \leq 0.01$.

The Cronbach's alpha for the full 24-item MRT was 0.91. The alphas for the structural (10 items) and mirror (12 items) subsets were 0.77 and 0.85, respectively. Alphas for the occluded (7 items) and non-occluded (17 items) subsets were 0.77 and 0.87, respectively.

2.1. Rasch analyses

Rasch analyses are designed to be free of the problems associated with dichotomously scored variables and such analyses make no assumption concerning the shapes of the distributions of the test items. It is thus possible to test for a single predominant factor by examining those assumptions necessary for Rasch analysis.

All Rasch analyses reported in this article were performed using the conditional maximum likelihood (CML) estimation procedure as implemented in RSP (Rasch Scaling Program; Glas & Ellis, 1994). Logits were used as initial estimates.

As suggested by Van den Wollenberg, 1982 (see also Glas, 1988, and Glas & Ellis, 1993), a distinction should be made between two assumptions central to the Rasch model. The distinction is between unidimensionality in terms of difficulty and unidimensionality in terms of ability. We tested difficulty unidimensionality with the test statistic $Q1$, which is sensitive to the degree to which the item difficulty functions are parallel (unidimensional difficulty). A significant $Q1$ indicates that the functions deviate significantly from the assumed parallel structure. Ability unidimensionality was tested using the $Q2$ statistic. As a measure of dimensionality, $Q2$ is based on the assumption that if there is only one factor upon which the items are scaled, then the subjects' residual scores on all items should be uncorrelated. A significant $Q2$ indicates that there remain significant correlations among at least some of the items, and that unidimensionality cannot be safely assumed.

As pointed by Glas and Ellis (1993), $Q1$ and $Q2$ are highly sensitive to the subject sample size. For this reason, they recommend testing the significance of the $Q1$ and $Q2$ test statistics with very small p levels, such as 0.001. This prescribed level was used in all interpretations of $Q1$ and $Q2$ reported in the present article.

When all 24 items from the MRT were included in the Rasch analysis, the scale as a whole was not found to be homogeneous. That is, it appeared that at least some of the items differed on a dimension other than difficulty. This was illustrated by the fact that the item difficulty functions varied across performance-level groups (or intersecting item response functions: $Q1 = 393$, $df = 92$, $p < 0.001$). Multidimensionality was further suggested by substantial residual inter-item correlations, reflected in a high value of $Q2$ ($Q2 = 3322$, $df = 1260$, $p < 0.001$).

Table 2
Q1 and Q2 statistics for various structural, mirror, occluded, and non-occluded item subsets.

	Statistic	*df*	*p* value
Structural ($k=10$; $n=587$)			
Q1	177	27	0.0000
Q2	600	140	0.0000
Mirror ($k=12$; $n=547$)			
Q1	126	44	0.0000
Q2	705	270	0.0000
Occluded ($k=7$; $n=454$)			
Q1	54	18	0.0000
Q2	88	56	0.0042
Non-occluded ($k=17$; $n=586$)			
Q1	344	64	0.0000
Q2	1989	595	0.0000
Occluded mirror ($k=4$; $n=349$)			
Q1	16	3	0.0010
Q2	1	4	0.8743
Non-occluded mirror ($k=8$; $n=494$)			
Q1	117	21	0.0000
Q2	285	80	0.0000
Non-occluded structural ($k=8$; $n=562$)			
Q1	192	28	0.0000
Q2	673	100	0.0000

Among the seven subsets of items examined (structural, mirror, occluded, non-occluded, occluded mirror, non-occluded mirror, and non-occluded structural), only the small subset of 4 occluded mirror items was found to be homogeneous in terms of both the Q1 and Q2 criteria (see Table 2).

It should be noted that several subsets comprised of a mixture of items from the various item-type categories, based on a principle of proximity, were also found to be Rasch-homogeneous. In particular, the analysis of subsets based on item position revealed that the first eight items in the test (four structural items, four mirror items) achieved unidimensionality (see Table 3). Furthermore, a subset comprised of three occluded and three non-occluded items (items 5, 6, 7, 9, 10, and 11) was also found to be homogeneous ($Q1=32$, $df=15$, $p=0.007$; $Q2=41$, $df=36$, $p=0.266$).

Maybe more importantly, Rasch analyses failed to reveal any sex bias on the MRT as a whole or in any of the seven main subsets examined. In other words, the hypothesis that the item difficulty functions of groups defined by sex were parallel could not be rejected for the test as a whole (Q1 using sex as the splitting variable $=29$, $df=23$, $p=0.1723$) or in any of the item subsets reported in Table 2 (all $ps>0.05$). This means that the sex differences observed on the MRT cannot be attributed to just a few biased items in the test.

2.2. Sex differences on item subsets

2.2.1. Structural/mirror

An analysis of variance of MRT performance was conducted with the structural/mirror item categorization as a within-subject factor and sex as between-subject factor. Following Voyer and Hou (2006), the structural and mirror raw scores were standardized on the total number of items of each type. The descriptive statistics for the four conditions are shown in Table 4. As expected, there was a main effect of sex, $F(1,622)=119.06$, $MSe=0.11$, $p<0.01$, $d=0.92$). Although small, there also was a main effect of item type, mirror items being significantly more difficult than structural items, $F(1,622)=4.08$, $MSe=0.02$, $p=0.04$, $d=0.17$). The item type by sex interaction was also significant, $F(1,622)=20.91$, $MSe=0.02$, $p<0.01$), indicating that the men outperformed the women more on the mirror than on the structural items.

Table 3
Q1 and Q2 statistics for three subsets of adjacent items.

	Statistic	*df*	*p* value
Items 1 to 8 ($k=8$; $n=400$)			
Q1	29	14	0.0108
Q2	68	60	0.2113
Items 9 to 16 ($k=8$; $n=400$)			
Q1	71	14	0.0000
Q2	142	60	0.0000
Items 9 to 16 ($k=8$; $n=400$)			
Q1	57	14	0.0000
Q2	196	60	0.0000

2.2.2. Occlusion

A second analysis of variance was conducted with the occluded/non-occluded (again standardized on the total number of items of each type) item categorization as a within-subject factor and sex as the between-subject factor. The descriptive statistics for the four conditions are shown in Table 4. As expected, there was once again a main effect of sex, $F(1,622)=126.29$, $MSe=0.12$, $p<0.01$, $d=0.94$). There also was a main effect of item type, occluded items being significantly more difficult than non-occluded items, $F(1,622)=377.75$, $MSe=0.02$, $p<0.01$, $d=1.63$). The item type by sex interaction was also significant, $F(1,622)=9.76$, $MSe=0.02$, $p<0.01$), indicating that the men outperformed the women more on the occluded than on the non-occluded items.

2.2.3. Degrees of angular disparity

The effect of angular disparity (the smallest degree of angular disparity about the vertical axis; see Caissie, Vigneau, & Bors, 2009) on sex differences was examined by dividing the 48 correct response choices into two groups (high versus low angular disparity) using a median split (120°). The hypothesis being examined was that if men were faster and better at mental rotation than women, then the sex differences would be greater on items with a high degree of angular disparity. The mean score for the low angular disparity items was 13.98 ($SD=5.72$) (men: $M=16.87$, $SD=5.50$; women: $M=12.45$, $SD=5.22$) and the mean for the high angular disparity items was 14.10 ($SD=5.39$) (men: $\mathrm{M}=16.77$, $SD=5.31$; women: $M=12.67$, $SD=4.88$). The Cronbach's alpha for the group of low disparity responses was 0.89 and was 0.88 for the high-disparity responses. An analysis of variance was conducted with the degree of angular disparity (low versus high) as a within-subject factor and sex as the between-subject factor. Again, a main effect of sex was found, $F(1,622)=100.88$, $MSe=50.92$, $p<0.01$, $d=0.84$. However, there was no effect of angular disparity, $F(1,622)=0.47$, $MSe=2.71$, $p=0.49$, $d=0.06$, and no interaction, $F(1,622)=2.65$, $MSe=2.71$, $p=0.10$. Sex differences were practically the same (about $d=0.84$) across the two angular disparity conditions.

As a further test of the absence of an effect of angular disparity on performance, correlations between correct response choices difficulty (percent correct) and angular disparity were calculated across the 48 correct response choices of the MRT. The overall correlation between difficulty and angular disparity was $r=0.02$. This correlation was unaffected by statistical control of item position in the test (partial $r=-0.05$, $df=45$). These general results also applied to both sex groups when analyzed separately (men: $r=0.01$, partial $r=-0.06$; women: $r=0.02$, partial $r=-0.04$). Finally, the sex difference in correct response choice difficulty (p for women subtracted from p for men) also showed no association with correct response choice angular disparity ($r=-0.05$).

Table 4
Means and standard deviations for men and women in four item type conditions.

	Structural	Mirror	Occluded	Non-occluded
Men	0.62 (0.25)	0.64 (0.28)	0.53 (0.32)	0.66 (0.25)
Women	0.45 (0.22)	0.40 (0.26)	0.28 (0.26)	0.46 (0.23)

3. Discussion

The size of the sex differences ($d = 0.93$) found in the present study was within the range of that typically reported in either timed or untimed administrations of the MRT. This means that about 16% of the variance in MRT scores can be attributed to or associated with sex differences, a small effect. The scores of the lowest scoring men are as low as the scores of the lowest scoring women, and the scores of the highest scoring women are as high as those of the highest scoring men. However small, the observed sex difference certainly is reliable and, in the present study, manifested on each and every item. Also, Rasch analyses indicated that the sex differences do not seem to arise from an inherent sex bias of some specific MRT items.

The Cronbach's alphas revealed that the MRT as a whole and the various item clusters proposed by Voyer and Hou (2006) were internally reliable. However, despite the internal reliability of the test as a whole and of the examined item subsets, only the subset defined as occluded mirror items proved to pass the test of homogeneity necessary for conceiving of the items as being a reflection of a single underlying factor. In terms of factor structure, the MRT thus seems to be best conceptualized as a set of non-independent items whose inter-correlations do not allow for a significant single factor, or a clear set of factors, to emerge.

Unlike Voyer and Hou (2006), and although the effect was small, we found men to significantly outperform women more on the mirror items than on the structural items. Furthermore, as did Voyer and Hou (2006), we found the men to outperform the women more on the occluded items than on the non-occluded items. Two comments must be made about these results. First, although the Rasch analyses provided evidence undermining the items subsets identified by Voyer and Hou (2006), the analyses of sex differences provided supporting evidence for the differential predictive validity of those item categories. Second, although the sex differences were mediated by item type, effect sizes related to these differences are not large enough to account for the overall sex effect. That is, although a portion of the typical sex effect may be the product of the presence on the MRT of certain types of items (namely, mirror and occluded), these types of items cannot be viewed as primary sources of the sex differences. The fact remains that men significantly outperformed women on all clusters of items and, for that matter, on all items.

Although much of the results were unexpected, the most surprising finding pertained to the degree of angular disparity. Although men outperformed women in both high and low disparity conditions, the degree of angular disparity was unrelated to sex differences. This finding not only raises questions about the nature of sex differences on the MRT, it also raises questions about the nature of the MRT itself. It must be kept in mind that the consistent sex effect is being used as a tool to uncover the types of items that may be responsible for the sex differences. If one set of items is particularly more difficult for women than it is for men, that set can be examined for some quality that may characterize it. In this case, by examining degree of angular disparity, we are hoping to measure a subject's ability to mentally rotate a three-dimensional figure. If men tend to have more of this ability, their advantage should manifest itself more in those cases where the degree of angular disparity is greatest. This was not found to be the case. Thus, if the assumption that relates speed of rotation, angular disparity, and mental rotation ability is not erroneous, then we must conclude that either the items are not primarily a test of mental rotation, or that there are no differences between men and women with respect to this ability. We are still left with the problem of identifying the source of the robust sex differences on the MRT and we are yet to confirm what it is actually measuring.

Acknowledgment

This research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Caissie, A. F., Vigneau, F., & Bors, D. A. (2009). What does the Mental Rotation Test measure? An analysis of item difficulty and item characteristics. *The Open Psychology Journal*, *2*, 94−102.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.) Hillsdale, N.J.: Lawrence Earlbaum Associates.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, *53*, 525−546.

Glas, C. A. W., & Ellis, J. L. (1993). RSP user's manual: Rasch scaling program. Groningen, The Netherlands: iec ProGamma.

Glas, C. A. W., & Ellis, J. L. (1994). Computer software. Groningen, The Netherlands: iec ProGamma.

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.

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.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika*, *47*, 123−139.

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

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., Rodgers, M. A., & McCormick, P. A. (2004). Timing conditions and the magnitude of sex differences on the Mental Rotation Test. *Memory & Cognition*, *32*, 72−82.

Voyer, D., & Saunders, K. A. (2004). Sex differences on the mental rotation test: A factor analysis. *Acta Psychologica*, *117*, 79−94.