Performance factors and gender-related differences in spatial ability: Another assessment

HEINRICH STUMPF

The Johns Hopkins University, Baltimore, Maryland

To explore whether test-taking styles (*performance factors*) can contribute to explaining gender-related differences on tests of spatial ability, 15 spatial tests were administered to three samples of subjects. On each test, number-correct scores and ratio scores (number of items solved divided by the number of items attempted) were computed. In accordance with previous research findings, the use of ratio scores significantly reduced the magnitude of the gender-related differences on the Mental Rotations Test. For most of the remaining tests, however, the reduction of the gender-related score difference was small. It was concluded that the difference reduction for the Mental Rotations Test was specific to the format of this test. In common spatial tests, performance factors may account for a small portion of gender-related variance, but the bulk of this variance must be attributed to other factors.

Gender-related differences in spatial test scores have been frequently reported and extensively discussed (e.g., Benbow, Stanley, Kirk, & Zonderman, 1983; Halpern, 1992; Harris, 1981; Hyde, 1981; Linn & Petersen, 1985; Maccoby & Jacklin, 1974). Various theories have been put forth to explain these differences. The theories include biological approaches stressing the importance of genetic, maturational, or hormonal factors or the role of differential cortical lateralization, as well as environmental approaches stressing the role of differential socialization experiences (for an overview of a variety of these models, see, e.g., Halpern, 1992). Two important observations in the ongoing discussion are that (1) although the differences occur with great regularity, they are often relatively small (Hyde, 1981, 1993), and (2) different types of spatial ability items elicit different magnitudes of those differences (Linn & Petersen, 1985). In particular, tasks requiring mental rotation appear to elicit the largest gender-related differences.

More recent analyses (Emanuelsson & Svensson, 1986; Feingold, 1988; Hilton, 1985; Rosenthal & Rubin, 1982; Stumpf & Klieme, 1989) have revealed indications of a long-term trend in which the gap between the average performances of males and females on spatial tasks is narrowing, but some of these studies have met with criticism (Halpern, 1989). Evidence that such differences can be found on at least some spatial tasks is still being reported (Blum & Hensgen, 1991; Oosthuizen, 1991).

In an important contribution to the discussion, Goldstein, Haldane, and Mitchell (1990) have recently presented findings that could have considerable consequences

both for our understanding of gender-related differences in the cognitive domain and for the development of educational measures to deal with them. The authors argued that, in the discussion of explanations for the male advantage in many tests of spatial ability, two sets of variables have received too little attention so far.

Virtually ignored in this controversy . . . are the effects of experiential or task factors (or, more generally, performance factors) on the spatial task behavior of the two sexes . . . [T]ask variables—such as time limits or task difficulty—and experiential factors—such as previous task exposure or expectations of task success—undoubtedly influence performance and may contribute to the sex differences reported in the literature. Of particular interest is the observation of Maccoby and Jacklin (1974) that males tend to perform tasks quickly, while females work more slowly and carefully. . . .

In addition to a slower speed of responding, females may have lower levels of confidence in their ability to perform various tasks. . . . Consequently, girls might be reluctant to guess if uncertain of the correct answer, leading females to commit more errors of omission than males. Females would then be penalized by scoring systems that emphasized the number of items correct rather than the ratio of correct items to items attempted. (Goldstein et al., 1990, p. 547)

Thus, performance factors reflect the experiences that a subject has had in solving tasks. It is believed that these experiences mold a person's attitude toward the problem of solving such tasks, thus affecting the person's perception of a task's difficulty, the likelihood that the person will be successful in tackling the task, his/her preferences for such a task over others, and his/her perceived competence in solving the task or the degree of confidence with which the task is approached.

To examine the role of such performance factors, Goldstein et al. (1990) administered the Mental Rotations Test (MRT; Vandenberg & Kuse, 1978) to two samples of

The author is indebted to John Eliot, Carol Mills, Julian Stanley, Sherri Tissot, and an anonymous reviewer for helpful comments on earlier drafts of this paper. Requests for reprints should be sent to H. Stumpf, Center for Talented Youth, The Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218.

mathematically talented college students. They found the typical large difference between the performances of males and females when the test was administered and scored in the conventional way. When the test was administered without a time limit, however, or when the ratio of items solved to items attempted was used as the score, the score differences between the sexes were no longer statistically significant, although males still earned higher scores.

The results of Goldstein et al. (1990) are consistent with psychosocial theories of cognitive gender differences. On the basis of the finding that achievement-related self-esteem is more important to the male's sex-role stereotype than to the female's stereotype (e.g., Intons-Peterson, 1988), some of these theories (see Halpern, 1992, chap. 6, for an overview) predict that males approach spatial tasks with more confidence and have a lower threshold for giving an answer despite uncertainty about its correctness.

Parallels to both findings of Goldstein et al. (1990) have been reported in the literature. Gallagher (1989), as well as Gallagher and Johnson (1992), for example, found that score differences between the sexes on a spatial task were reduced when the test was administered without a time limit. Kail, Stevenson, and Black (1984) observed that males worked faster on spatial items, but that females gave correct answers on the problems about as often as males.

The first finding, indeed, gives rise to some skepticism with respect to the practical implications of the traditional male advantage on spatial tests. If, for instance, "women in mathematical or scientific careers take a few minutes longer than men on an occasional visual-spatial task, they may suffer little or no substantive disadvantage" (Goldstein et al., 1990, p. 550).

The observation that females have less confidence in their ability and might be more reluctant to guess if they are uncertain of the correct answer to a problem would also have important implications. It might mean that females are put at a disadvantage by the conventional way of scoring tests without differing from males on the latent trait that is supposed to be measured by these tests. The question whether females really have less confidence in their spatial ability, however, cannot be answered with certainty at present. Lunneborg and Lunneborg (1984) reported such a finding, but Olson, Eliot, and Hardy (1988) were unable to replicate it. Moreover, Olson et al. identified a few space-related activities that women perform with more confidence than men. It should also be noted that not even the results of Goldstein et al. (1990) demonstrate that females are less inclined to guess when they are uncertain of the correct answer. If this were the case, females should have higher ratio scores than males, but Goldstein et al. found only that the male advantage was reduced when ratio scores were used.

Goldstein et al.'s (1990) results, however, do show that the females in their samples had marked fewer answers on the MRT than had the males, and that the gap between males and females was considerably smaller when the ratio score instead of the number-correct score was used. Thus, the more general notions that females have different styles

of responding to the items than males do and that these styles operate to the disadvantage of females when the spatial tests are scored in the conventional way appear to be supported by their data. Put differently, there is evidence that the scores of males and females on at least one type of spatial item are differentially affected by what Goldstein et al. call *performance factors*, although the nature of these factors is largely unknown.

Although it might be tempting to apply the findings of Goldstein et al. (1990) more generally to other areas in which cognitive gender-related differences have been observed or even to the overall motivation of males and females in approaching cognitive tasks, such an attempt is explicitly avoided here. This caution is based in part on empirical findings. Using a mathematical reasoning test, Mills, Ablard, and Stumpf (1993) were able to replicate the observation of gender-related differences in this area, but were unable to replicate the reduction of effect size observed by Goldstein et al. with the MRT.

It is obvious that the findings of Goldstein et al. (1990) would be of considerable consequence for our understanding of gender-related differences in spatial ability if they were replicable and generalizable across other types of spatial tasks and if the findings were not artifacts due to the method used by the authors. The generalizability and validity of their findings, however, may be questioned for several reasons.

The test Goldstein et al. (1990) used—the MRT—differs in format from the usual types of spatial tasks. In every item of the MRT, the subject has to select two figures (out of four) that are congruent with a standard stimulus (but rotated in space); in most ordinary spatial tests, the subject has to select one option out of five. Test-taking behavior on the MRT might be affected by this and thus be untypical of the way subjects approach more conventional types of tasks.

Other potentially serious problems of the study by Goldstein et al. (1990) have to do with some properties of the type of scores used by the authors—ratio scores. Such scores have some serious flaws, two of which will be mentioned here. First, they do not tell us enough about the subjects' standing on the trait to be measured and are no real alternative to conventional scores. A psychometrician might be interested in the fact that a subject has the ratio score 1.00 on a test, but he/she will certainly also be interested in whether the subject attempted 1 item and got it right or attempted 20 items and got all of them correct. If ratio scores were used routinely, it would take little wisdom on the part of the subjects to mark an answer on one item appearing to be easy, get it right, and get the maximum possible score.

This problem did not affect the results of Goldstein et al. (1990), because the subjects were unaware of the way in which the test was going to be scored. A second problem, however, might have affected their results directly. This possible flaw is that ratio scores might be less reliable than conventional scores (e.g., Carroll, 1993, p. 445). If this were the case in the data of Goldstein et al., the lack of a substantial difference between the sexes on the

ratio scores could be explained by the fact that the scores simply lack the systematic variance that would show gender-related differences.

The present research is an attempt to replicate the findings of Goldstein et al. (1990) and to examine the generalizability of their results across more conventional spatial tasks. In order to replicate the findings, the MRT was given to a group of subjects similar to the sample in Goldstein et al. In order to examine the generalizability of the findings, an attempt was made to assess the construct of spatial ability and its various subskills as comprehensively as possible and to compare the conventional and ratio scores on the tests used. In both cases, the reliability of the two types of scores was assessed.

A large number of potential measures of spatial ability might be used in a comprehensive assessment of the construct. Eliot and Smith (1983) tried to survey and collect all existing tests of spatial ability and found several hundred of them. Thus, it might appear hopeless to try such a comprehensive assessment, but Eliot (1980, 1983; see also Eliot, Stumpf, & Tissot, 1992) found that the existing tests with very few exceptions can be classified into 16 categories. These categories, listed in Appendix A, can be grouped in turn into two divisions, a recognition and a manipulation division. Recognition tasks require the perception, retention, or transformation of visual forms within a two-dimensional plane, whereas manipulation tasks require the mental manipulation of visual shapes across a two-dimensional plane (Eliot, 1983, p. 11).

Eliot's classification is based on perceived similarities among the existing spatial tasks. The category of figure-ground tasks, for instance, is based on the observation that measures like the Hidden Figures Test CF-1 from the Kit of Reference Tests for Cognitive Factors (French, Ekstrom, & Price, 1963), the Flexibility of Closure Test from the Comprehensive Ability Battery (Hakstian & Cattell, 1975), the Spatial Test 1 from the NFER collection of tests (Smith, 1950), and the Hidden Figures Test NFT-04B of the Structure of Intellect Tests (Guilford, 1968) are all variations of the Gottschaldt (1926) figures. Eliot (1983) provides a more detailed account on the derivation of the categories. The content of tasks typical for 14 of the 16 categories is described in Appendix D.

Relying on Eliot's classification, one should be able to represent most of the major ways in which spatial ability has been operationally defined so far in one test battery, if a very large one.

METHOD

The CTY Spatial Test Battery

In 1991 and 1992, an experimental version of such a battery, the CTY Spatial Test Battery (STB; Eliot & Stumpf, 1992), was developed at the Center for Talented Youth at The Johns Hopkins University in Baltimore. This battery has 14 subtests and three Forms, AA, BB, and CC. This paper will deal with Form AA and one subtest of Form BB only. The number of subtests was limited to 14 for each form in order to make the administration of the battery feasible. When the battery was compiled, the category of perceptual speed tasks was deleted, and the copying and maze categories were combined into one. All subtests considered in this paper had been published and used before in different contexts.

Appendixes B and C give overviews of the tests included in Form AA of the STB for the recognition and manipulation divisions. In the tables, the subtests appear in the order of the categories they are designed to cover. The numbers of the subtests in the test booklets are given in the first column. The Battery has two parts, the first of which contains Subtests 1-6. Part 2 includes Subtests 7-14. The administration of one part takes about one hour and a half. The content of the items of the various tests is briefly described in Appendix D. Sample items for most of the subtests are given in the directory of spatial tests by Eliot and Smith (1983).

Procedure

Forms AA, BB, and CC of the CTY Spatial Test Battery were administered to a sample of students in summer courses offered by the Center for Talented Youth (CTY). These courses provide fast-paced instruction in mathematics, natural sciences, and the humanities. The courses take place at colleges, where the students live for about 3 weeks. The courses are open to high school students who have reached at least the seventh grade and attained scores on the Scholastic Aptitude Test that meet certain criteria. These requirements depend on the age of the students. To qualify for courses in mathematics and science, for instance, the youngest applicants (up to 13 years and 6 months) need a score of at least 500 on the mathematical part of the SAT and a combined score of at least 930 on SAT-M and SAT-V. The requirements for the oldest students (16 years and 6 months to 17 years) are a score of 670 on the mathematical part of the SAT and a score of 1270 on the overall SAT.

The STB was administered at one site where the CTY courses were offered to about 950 students. An announcement was made in all classes, informing students that they were requested to report to assigned rooms at a particular day and time for the purpose of participation in a research study. The announcement for the study implied that the students would take a test, but the content of the test was not specified in advance. Participation was voluntary. Students who did not choose to participate were assigned to a study hall. As stated above, only the data on Form AA of the STB and the second subtest of Form BB, the MRT, will be considered here.

The STB was administered in large classrooms by CTY teachers who had been given test administrators' handbooks (Eliot et al., 1992) describing the administration procedure. This handbook included the instructions to be read to the students. The teachers were given a training and orientation session prior to testing.

Subjects

A total of 233 subjects (146 males and 87 females) took the MRT as part of Form BB of the STB; Part I of Form AA was taken by 355 students (201 males and 154 females). The number of students completing Part II of Form AA was 218 (115 males and 103 females). The age of the students ranged from 12 to 17 years. The fact that there are more males than females in these samples reflects the proportion of males enrolled in the CTY courses in the same year (55%; the percentage of males enrolled in CTY summer classes each year typically ranges from 53% to 57%).

RESULTS

In a first step of data analysis, an index for the overall difficulty of the 15 tests used here was obtained. We computed the average p value for the items in each test.

The average p values for the items of the MRT and the subtests of the STB, Form AA, are summarized in Table 1. This table also displays the reliability coefficients for the conventional and ratio scores on the 15 tests.

Table 1
Reliability Estimates and Average *p* Values for the Spatial Tests Used

No.	Name	Reliabilities		
		Conventional	Ratio	<i>p</i>
	Mental Rotations Test	.88	.85	.54
10	Path Finding Test	.91	.81	.78
3	Hidden Figures Test	.80	.82	.37
6	Learning Figures Test	.66	.65	.41
11	Gestalt Completion Test	.68	.71	.64
8	Paper Formboard Test	.94	.91	.74
12	Figural Rotation Test	.91	.84	.83
7	Block Counting Test	.97	.94	.51
4	Intersections Test	.83	.77	.49
2	Eliot-Price Mental Rotations Test	.87	.86	.71
14	Identical Blocks Test	.91	.88	.68
5	Paper Folding Test	.82	.81	.81
9	Pattern Assembly Test	.97	.95	.65
13	Surface Development Test	.94	.90	.49
1	Guay's Visualization of Views Test	.82	.82	.55

As is evidenced by the table, most of the tests had an adequate difficulty level in our sample, although some of them (Path Finding, Figural Rotations, Paper Formboard, Eliot-Price Mental Rotations, Paper Folding) turned out to be rather easy.

In the second step of data analysis, reliability estimates for both the conventional and the ratio scores on the STB subtests were computed. These estimates are odd-even split-half reliability coefficients corrected by the Spearman-Brown formula.

The reliability estimates of the MRT are .88 for the conventional score and .85 for the ratio score. According to these estimates, the conventional score on the MRT is highly internally consistent in the present sample. The split-half reliability for the ratio score is only slightly lower. Given the similarity of the two estimates, any reduction in gender-related score difference from the conventional to the ratio score is unlikely to be an artifact of lower reliability of the latter score. The same applies to most of the subtests of the STB, Form AA. In most cases, the consistency coefficients for the conventional

scores are somewhat higher than those of the ratio scores, but the differences are so small that they are of little practical consequence when the two types of scores are compared across the tests. The largest difference was obtained for the Path Finding Test. In this case, the fact that the gender-related difference on the ratio score was smaller than the difference for the conventional score (see Table 2) might in part be due to the somewhat lower reliability of the ratio score. It is more likely, however, that this shrinkage in difference is due to a ceiling effect on the ratio score on this test. Both males and females got correct more than 95% of the items they attempted.

In a third step of data analysis, each of the STB subtests was scored separately for males and females. Both the number-correct score and the ratio of items solved to items attempted were computed. An item was considered to be attempted if the subject had marked an answer on the answer sheet or in the test booklet for that item. To compare the scores of males and females, *t* tests were performed and effect size indices (*d* coefficients; Cohen, 1977) were computed. In order to evaluate whether the gender-related differences on the conventional and ratio scores differed significantly, both scores for each test were separately standardized to a mean of 100 and a standard deviation of 10. These standard scores were then subjected to an analysis of variance. A mixed design was chosen, with gender as a between-subjects variable and type of score (conventional or ratio) as a within-subjects variable. This analysis yielded an estimate of the significance of the change in the gender-related difference from the conventional score to the ratio score (interaction effect of the factors gender and type of score).

Table 2 displays the mean scores of the subsamples of males and females on both scores, as well as the associated *t* values (two-sided *t* tests) and effect size estimates (*d* coefficients) across the 15 tests used here. Table 3 summarizes the results of the analyses of variance for the 15 tests. As far as the MRT is concerned, there is a highly significant difference in favor of males on the conven-

Table 2
Conventional and Ratio Scores of Males and Females on the Spatial Tests Used

No.	Name	Males				Females							
		Conventional		Ratio		Conventional		Ratio		Conventional		Ratio	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>d</i>	<i>t</i>	<i>d</i>
	Mental Rotations Test	12.02	4.28	71.75	20.39	8.77	4.28	60.99	24.51	5.64†	.74	3.59†	.48
10	Path Finding Test	20.39	3.90	96.87	5.85	18.78	4.30	96.63	6.66	2.91†	.39	.28	.04
3	Hidden Figures Test	7.81	3.79	50.39	25.57	6.95	3.87	47.32	26.72	2.10*	.22	1.10	.12
6	Learning Figures Test	8.37	3.70	43.84	19.36	8.16	3.92	42.87	20.39	.54	.06	.45	.05
11	Gestalt Completion Test	12.90	3.01	78.28	16.35	12.72	2.80	77.18	15.72	.47	.06	.50	.07
8	Paper Formboard Test	19.34	5.68	83.02	20.50	17.79	6.19	80.36	21.75	1.93	.26	.93	.13
12	Figural Rotation Test	26.19	3.83	91.17	10.33	23.62	5.69	86.98	14.96	3.95†	.52	2.43*	.33
7	Block Counting Test	44.17	15.53	78.40	18.99	37.68	12.67	71.92	17.59	3.36†	.45	2.60*	.35
4	Intersections Test	13.07	5.24	58.98	22.06	11.01	5.13	51.08	22.14	3.69†	.39	3.33†	.35
2	Eliot-Price Mental Rotations Test	15.56	3.70	79.39	18.36	12.39	4.59	64.88	22.02	7.19†	.72	6.74†	.68
14	Identical Blocks Test	14.78	5.22	78.37	23.62	12.06	4.71	67.62	22.31	4.02†	.51	3.42†	.46
5	Paper Folding Test	16.56	2.84	83.88	13.92	15.60	2.94	79.68	14.36	3.10†	.33	2.78†	.29
9	Pattern Assembly Test	34.98	12.43	73.31	22.08	30.45	12.22	66.99	22.31	2.70†	.38	2.10*	.28
13	Surface Development Test	16.22	7.99	65.67	25.85	12.88	7.55	57.46	27.08	3.15†	.42	2.27*	.31
1	Guay's Visualization of Views Test	12.82	4.88	62.27	23.19	9.96	4.03	48.71	18.88	5.90†	.60	5.90†	.60

Note—Total degrees of freedom: Sample size -2. * $.01 < p < .05$. † $p < .01$.

Table 3
Results of the Analyses of Variance

Spatial Tests	Gender $\times$ Score <i>F</i> Ratio
Mental Rotations Test	9.54†
Path Finding Test	9.59†
Hidden Figures Test	1.05
Learning Figures Test	.02
Gestalt Completion Test	.00
Paper Formboard Test	3.16
Figural Rotation Test	4.29*
Block Counting Test	.81
Intersections Test	.32
EP Mental Rotations Test	.29
Identical Blocks Test	1.44
Paper Folding Test	.90
Pattern Assembly Test	3.35
Surface Development Test	1.27
Guay's Visualization of Views Test	.15

Note—*df* = 1,231 for the Mental Rotations Test, 1,351 for the tests of Part I of the STB, and 1,216 for the tests of Part II. * .01 < *p* < .05. † *p* < .01.

tional score. The magnitude of this difference is about three quarters of a standard deviation. The ratio score (reported as a percentage in Table 2 and throughout) also shows a significant gender-related difference in favor of males, but this difference is smaller than the one mentioned before, about half a standard deviation. The results of the analysis of variance reported in Table 3 indicate that the reduction in difference from the conventional to the ratio score is highly significant.

For most of the tests in Form AA of the STB, too, the mean conventional scores of the males exceed the respective scores of the females. This effect is significant for 11 of the 14 tests. The largest effect was obtained for the Eliot-Price Mental Rotations Test, which is similar to the MRT. The Learning Figures, Gestalt Completion, and Paper Formboard Tests did not show significant gender-related score differences. The effect size estimates (*d* coefficients) for the ratio scores in general are somewhat lower than those for the respective conventional scores, but still the gender-related differences are significant for nine of the tests.

Although most of the tests in Form AA of the STB showed significant gender-related differences in favor of males, it should be noted that these differences are smaller in the recognition division than in the manipulation division of the battery. The average *d* for the recognition division is .25 with respect to the conventional and .12 with respect to the ratio scores. As far as the manipulation division is concerned, the average *d* is .48 for the conventional scores and .42 for the ratio scores. The two divisions thus differ considerably with respect to their gender sensitivity, and the recognition division shows a somewhat larger reduction of gender sensitivity than the manipulation division does, as one moves from conventional to ratio scores.

According to the results of the analyses of variance reported in Table 3, within Form AA of the STB, the reduction in gender-related score difference from the conventional to the ratio scores (as expressed by the inter-

action of the factors of gender and type of score) is statistically significant only in two cases, the Path Finding and Figural Rotation Tests, although the interaction effect approaches significance in a few other cases.

Although most of the difference reductions are small, the number of tests in the STB, Form AA, showing such a reduction in the amount of *d* (12) by far exceeds the number of tests showing an increase in *d* from the conventional to the ratio score (1). This difference is highly significant [sign test; $\chi^2(1) = 9.31$, *p* < .005]. This indicates that performance factors do play a role in the STB, although their effect is very small.

DISCUSSION

The present study replicated the frequent finding of an advantage for males on a variety of spatial tests. The various tests used here, however, differed in their degree of gender sensitivity. In general, the male advantage was relatively small on the tests of the recognition division, but larger on the measures of mental manipulation. This finding indicates that the gender-related differences often found in the domain of spatial ability are not primarily a matter of differences in the basic processes of perceiving and retaining visual forms. Rather, they appear to occur primarily in the complex, multistep processes involved in spatial visualization in general and in mental rotation in particular (Halpern, 1992; Linn & Petersen, 1985).

Given this pattern of results, one might argue that the amount of gender difference found on a spatial test is a function of the difficulty of that test. This hypothesis would be consistent with the fact that the MRT was relatively difficult in the present sample and produced the largest gender difference. The hypothesis would also explain the large gender-related differences on Guay's Visualization of Views Test, but it would be inconsistent with the observation that two of the most difficult tests in the battery, the Hidden Figures and Learning Figures Tests, produced only small gender-related differences.

A more viable argument is that gender-related differences are linked to the content of the tasks. This is in accord with previous findings (Linn & Petersen, 1985). In particular, mental rotation as a content appears to be responsible for gender-related differences. Not only does this explain the large differences found on the two mental rotations tests in this study, it may also account for the differences on tasks like the Identical Blocks Tests, which also involve mental rotation. It might even account for the finding that the Figural Rotation Test produced the largest gender-related difference in the recognition division of the STB. At present, it is not clear what other elements of task content contribute to gender-related score differences on the STB, but this will be addressed in future structural analyses of the battery.

As far as performance factors and the MRT are concerned, the present study replicated two basic findings of the research by Goldstein et al. (1990): First, there was a large advantage for males on this test when conventional scores were used. Second, the gender-related difference

on the MRT was significantly smaller when ratio scores were employed.

Goldstein et al. (1990) had conducted two experiments in which they used two versions of the conventional score on the MRT. In their first experiment, they counted the number of shapes the subjects had correctly identified as congruent with the standard stimuli. In their second experiment, they counted the number of tasks in which the subjects had correctly identified both shapes that were congruent with the standard shape. The latter procedure is commonly used in scoring the MRT (Vandenberg & Kuse, 1978). The effect size estimate (d) was .85 for the conventional score and .29 for the ratio score in the first experiment. When the same scoring procedure was applied to the present MRT data, d coefficients of .71 and .36 were obtained for the conventional and the ratio scores, respectively. Thus, in this respect, there is a rather close correspondence between the present results and the ones of Goldstein et al.

In their second experiment, Goldstein et al. (1990) obtained a d of .83 for the conventional score and a d of .53 for the ratio score. Again, the present result of $d = .74$ for the conventional and $d = .48$ for the ratio score agrees well with the findings of Goldstein et al.

The present results on the Eliot-Price Mental Rotations Test, which is highly similar to the MRT, however, are at odds with some of these findings. It is true that this test showed the largest gender-related difference within Form AA of the STB, but it exhibited only a very small reduction of gender sensitivity in the ratio score. The most likely explanation for this is not the content of the two tests, but their different formats. There appears to be a weak tendency of males to be more inclined to give answers despite being uncertain of their correctness. In the MRT, the effect of this tendency on the conventional score is much stronger than in other spatial tests, especially if the MRT is scored by counting the number of shapes identified correctly. In this case, the probability of getting an answer correct by chance is .5, which makes the test vulnerable to guessing. This probability is much larger than the respective probability in conventional spatial tests and in the Eliot-Price Test. In the latter test, the subject has to select the solution out of four options.

While the attempt to replicate the findings of Goldstein et al. (1990) on the MRT proved to be successful, the results obtained with Form AA of the STB indicate that the effect observed by Goldstein et al. (1990) is smaller in ordinary spatial tests than in the MRT and fails to reach significance for most of the tests even in the relatively large samples considered here. Overall, in many cases in the present data the gender-related difference in the ratio score was smaller than the one in the conventional score. The reduction in difference, however, was insignificant for most of the STB tests. The largest reduction in Form AA of the STB was found for the Path Finding Test, where the d dropped from .39 to .04, but this reduction can be explained in part by a ceiling effect and the somewhat lower reliability of the ratio score in this case. For the remaining subtests, the reduction was much smaller,

although reaching statistical significance in one case, the Figural Rotation Test. Thus, even if one accepts ratio scores as valid indicators (despite the flaws mentioned above), they do not help much to explain gender-related differences on spatial tests.

Of course, differences in basic characteristics of the samples tested should always be considered as alternative explanations if the findings of two studies disagree. Goldstein et al. (1990) tested academically talented college students; the subjects of the present study were highly talented high school students. One basis for comparing the samples is their scores on the MRT. An average p value of .45 for the MRT items in the second experiment of Goldstein et al. and an average p value of .54 for the present study suggest that our subjects had a slight advantage on the skill of mental rotation. This difference, however, is certainly not large enough to indicate fundamental qualitative differences in spatial ability between the samples. Moreover, any qualitative gap in ability between the samples would be at odds with the close correspondence found in respect to the differences between the sexes on the various scores on the MRT.

It is true that the percentage of males in our sample completing Form BB was somewhat larger than in the general CTY student population. This might have been due to differential self-selection by males and females. For example, more males might have come to the testing because males are more confident with respect to their performance on tests. But if self-confidence had really caused differential self-selection in this sample, males and females would have been more *similar* in self-confidence than in the other two samples. This possibility, however, is contradicted by the fact that when the conventional and ratio scores were compared, the largest reduction of gender-related difference was observed in this same sample.

The results of Goldstein et al. (1990) are consistent with psychosocial theories of gender differences, in particular with the view that sex role stereotypes enable males to feel more confident than females about their performance on spatial tasks. This confidence might lead males to a riskier way of responding, which, in turn, would boost their conventional scores on such tasks. Although there is little doubt that such stereotypes exist (e.g., Halpern, 1992; Intons-Peterson, 1988), our results indicate that their effect on test-taking behavior is not large.

Thus, when we argue that performance factors cannot explain the bulk of gender-related variance on spatial tasks, we do not rule out other environmental or psychosocial theories as explanations of gender-related score differences on spatial tasks. It would be inappropriate to base any speculations with respect to the nature-nurture controversy on our results.

Although it is clear that performance factors cannot explain much of the gender-related variance, this does not mean that they should be ignored in further studies. On the contrary, a more precise definition and a more detailed knowledge of performance factors and alternative ways of measuring them are worthy of continued research. Future studies on performance factors should focus more

closely on the process of solving spatial tasks. Presenting items individually and assessing the latencies of subjects' responses, as well as using ratings to assess subjects' levels of confidence when they take spatial tests are promising ways of operationally defining performance factors more precisely.

In the end, performance factors could be one piece in the mosaic that might explain gender-related differences in spatial ability. It is likely, however, that other factors, such as differential training on spatial tasks, or even differential cortical lateralization (see Halpern, 1992, for an overview) will explain equal, if not larger amounts of, gender-related differences in spatial ability.

REFERENCES

- BENBOW, C. P., STANLEY, J. C., KIRK, M. K., & ZONDERMAN, A. B. (1983). Structure of intelligence in intellectually precocious children and their parents. *Intelligence*, 7, 129-152.
- BLUM, F., & HENSGEN, A. (1991). Vergleiche einzelner Teilnehmergruppen beim fünften Termin des besonderen Auswahlverfahrens [Comparisons of various groups of participants at the fifth regular date of testing]. In G. Trost (Ed.), *Test für Medizinische Studiengänge (TMS): 15. Arbeitsbericht* [Test for medical studies: Fifteenth research report] (pp. 22-72). Bonn: Institut für Test- und Begabungsforschung.
- CARROLL, J. B. (1993). *Human cognitive abilities*. New York: Cambridge University Press.
- COHEN, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- COLLEGE ENTRANCE EXAMINATION BOARD (1951). *Special aptitude-spatial relations tests*. New York: Author.
- DEVILLIERS, L., OOSTHUIZEN, S., & KRUGER, J. P. (1977). *Senior aptitude tests for Indian South Africans*. Pretoria, South Africa: Human Sciences Research Council.
- EDUCATIONAL TESTING SERVICE (1959). *Comparative guidance and placement program*. Princeton, NJ: Author.
- EKSTROM, R. B., FRENCH, J. W., HARMAN, H. H., & DERMEN, D. (1976). *Kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.
- ELIOT, J. (1980). Classification of figural spatial tests. *Perceptual & Motor Skills*, 51, 847-851.
- ELIOT, J. (1983). The classification of spatial tests. In J. Eliot & I. M. Smith (Eds.), *An international directory of spatial tests* (pp. 11-15). Windsor, U.K.: NFER-Nelson.
- ELIOT, J., & PRICE, L. (1976). *A test of three-dimensional spatial visualization*. College Park, MD: University of Maryland, Institute for Child Study.
- ELIOT, J., & SMITH, I. M. (1983). *An international directory of spatial tests*. Windsor, U.K.: NFER-Nelson.
- ELIOT, J., & STUMPF, H. (1992). *CTY spatial test battery*. Baltimore, MD: Center for Talented Youth.
- ELIOT, J., STUMPF, H., & TISSOT, S. (1992). *CTY spatial test battery: Administrator's handbook*. Baltimore, MD: Center for Talented Youth.
- EMANUELSSON, I., & SVENSSON, A. (1986). Does level of intelligence decrease? A comparison between thirteen-year-olds tested in 1960, 1966, and 1980. *Scandinavian Journal of Educational Research*, 30, 25-37.
- FEINGOLD, A. (1988). Cognitive gender differences are disappearing. *American Psychologist*, 43, 95-103.
- FOUCHE, F. A., & ALBERTS, N. F. (1978). *Senior aptitude tests*. Pretoria, South Africa: Human Sciences Research Council.
- FRENCH, J. W., EKSTROM, R. B., & PRICE, L. A. (1963). *Kit of reference tests for cognitive factors*. Princeton, NJ: Educational Testing Service.
- GALLAGHER, S. A. (1989). Predictors of SAT mathematics scores of gifted male and female adolescents. *Psychology of Women Quarterly*, 13, 191-203.
- GALLAGHER, S. A., & JOHNSON, E. S. (1992). The effect of time limits on performance of mental rotations by gifted adolescents. *Gifted Child Quarterly*, 36, 19-22.
- GOLDSTEIN, D., HALDANE, D., & MITCHELL, C. (1990). Sex differences in visual-spatial ability: The role of performance factors. *Memory & Cognition*, 18, 546-550.
- GOTTSCHALDT, K. (1926). Über den Einfluss der Erfahrung auf die Wahrnehmung von Figuren [On the influence of experience on the perception of figures]. *Psychologische Forschung*, 8, 261-275.
- GUAY, R. B. (1976). *Visualization of views test*. West Lafayette, IN: Purdue Research Foundation.
- GUILFORD, J. P. (1968). *Structure of intellect tests*. Orange, CA: Sheridan Psychological Services.
- HAKSTIAN, A. R., & CATTELL, R. B. (1975). *Comprehensive ability battery*. Champaign, IL: Institute for Personality & Ability Testing.
- HALPERN, D. F. (1989). The disappearance of cognitive gender differences: What you see depends on where you look. *American Psychologist*, 44, 1156-1157.
- HALPERN, D. F. (1992). *Sex differences in cognitive abilities* (2nd ed.). Hillsdale, NJ: Erlbaum.
- HARRIS, L. J. (1981). Sex variations in spatial skill. In L. S. Liben, A. H. Patterson, & N. Newcombe (Eds.), *Spatial representation and behavior across the life span* (pp. 31-54). New York: Academic Press.
- HILTON, T. L. (1985). *National changes in spatial-visual ability from 1960 to 1980* (Report No. RR-85-27). Princeton, NJ: Educational Testing Service.
- HYDE, J. S. (1981). How large are cognitive gender differences? A meta analysis using omega and d. *American Psychologist*, 36, 92-101.
- HYDE, J. S. (1993). What is the big difference? Gender and abilities [Review of *Sex differences in cognitive abilities*]. *Contemporary Psychology*, 38, 183-184.
- INSTITUT FÜR TEST- UND BEGABUNGSFORSCHUNG (1986). *Der neue TMS* [The new test for medical studies]. Göttingen: Hogrefe.
- INTONS-PETERSON, M. J. (1988). *Gender concepts of Swedish and American youth*. Hillsdale, NJ: Erlbaum.
- KAIL, R., STEVENSON, M., & BLACK, K. (1984). Absence of a sex difference in algorithms for spatial problem solving. *Intelligence*, 8, 37-46.
- LINN, M. C., & PETERSEN, A. C. (1985). Emergence and characterization of sex differences in spatial ability. *Child Development*, 56, 1479-1498.
- LUNNEBORG, C. E., & LUNNEBORG, P. W. (1984). Contribution of sex-differentiated experiences to spatial and mechanical reasoning abilities. *Perceptual & Motor Skills*, 59, 107-113.
- MACCOBY, E. E., & JACKLIN, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.
- MILLS, C. J., ABLARD, K. E., & STUMPF, H. (1993). Gender differences in academically talented young students' mathematical reasoning: Patterns across age and subskills. *Journal of Educational Psychology*, 85, 340-346.
- OLSON, D. M., ELIOT, J., & HARDY, R. C. (1988). Relationships between activities and sex-related differences in performance on spatial tests. *Perceptual & Motor Skills*, 67, 223-232.
- OOSTHUIZEN, S. (1991). Sex-related differences in spatial ability in a group of South African students. *Perceptual & Motor Skills*, 73, 51-54.
- RICHARDSON, BELLows, & HENRY STAFF (1950). *Three dimensional test*. Washington, DC: Author.
- ROSENTHAL, R., & RUBIN, D. B. (1982). Further meta-analytic procedures for assessing cognitive gender differences. *Journal of Educational Psychology*, 74, 708-712.
- SMITH, I. M. (1950). *Spatial test 1*. Windsor, U.K.: NFER-Nelson.
- STUMPF, H., & KLIEME, E. (1989). Sex-related differences in spatial ability: More evidence for convergence. *Perceptual & Motor Skills*, 69, 915-921.
- THURSTONE, L. L. (1937). *Psychological tests for the study of mental abilities*. Chicago, IL: University of Chicago Press.
- VANDENBERG, S. G., & KUSE, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual & Motor Skills*, 47, 599-604.

APPENDIX A
The Categories of Eliot's Classification
of Figural Spatial Tasks

Recognition Division	
Category 1	perceptual speed tasks
Category 2	copying tasks
Category 3	maze tasks
Category 4	figure-ground tasks
Category 5	visual memory tasks
Category 6	Gestalt resolution tasks
Category 7	paper formboard tasks
Category 8	figural rotation tasks
Manipulation Division	
Category 9	block tasks
Category 10	intersection tasks
Category 11	block rotation: internal
Category 12	block rotation: external
Category 13	paper folding tasks
Category 14	pattern assembly tasks
Category 15	surface development tasks
Category 16	perspectives tasks

Note—Adapted from Eliot et al. (1992, p. 6).

APPENDIX B
Subtests of the CTY Spatial Test Battery, Form AA, Recognition Division

No.	Name	Category	Author(s)
10	Comparative Guidance and Placement Program: Form QPG, Book 3: Path Finding Test	copying and maze tasks	Educational Testing Service (1959)
3	Kit of Reference Tests for Cognitive Factors: Hidden Figures Test CF-1	figure-ground tasks	French, Ekstrom, & Price (1963)
6	Learning Figures Test	visual memory	Institut für Test- und Begabungsforschung (1986)
11	Kit of Factor-Referenced Tests, Gestalt Completion Test: Cs-1	Gestalt resolution tasks	Ekstrom, French, Harman & Dermen (1976)
8	Senior Aptitude Test for Indian South Americans: Form B (Paper Formboard Test)	paper formboard tasks	DeVilliers, Oosthuizen, & Kruger (1977)
12	Senior Aptitude Tests: Subtest 7: Spatial 2D (Figural Rotation Test)	figural rotation tasks	Fouche & Alberts (1978)

Note—The number in the left column indicates the position of the subtest in the battery.

APPENDIX C
Subtests of the CTY Spatial Test Battery, Form AA, Manipulation Division

No.	Name	Category	Author(s)
7	Thurstone's Primary Mental Abilities Tests: Subtest 17 (Block Counting Test)	block tasks	Thurstone (1937)
4	Special Aptitude-Spatial Relations: Subtest 1 (Intersections Test)	intersection tasks	College Entrance Examination Board (1951)
2	Eliot-Price Mental Rotations Test	block rotation (external)	Eliot & Price (1976)
14	Special Aptitude-Spatial Rotations: Subtest 4, Identical Blocks Test	block rotation (internal)	College Entrance Examination Board (1951)
5	Kit of Reference Tests for Cognitive Factors: Paper Folding Test, Vz-2	paper folding tasks	French, Ekstrom, & Price (1963)
9	Three Dimensional Test: RBH-0125 Subtest (Pattern Assembly Test)	pattern assembly tasks	Richardson, Bellows, & Henry Staff (1950)
13	Special Aptitude-Spatial Relations: Subtest 3 (Surface Development Test)	surface development tasks	College Entrance Examination Board (1951)
1	Guay's Visualization of Views Test	perspective tasks	Guay (1976)

Note—The number in the left column indicates the position of the subtest in the battery.

APPENDIX D
Description of the Content of the Spatial Tests Used in the Study

Test	Description
Mental Rotations	The subject has to identify two out of four block shapes that are congruent with a standard shape but rotated in space.
Path Finding	The subject has to trace several curved lines through a pattern of distracting lines.
Hidden Figures	Simple geometrical figures are embedded in more complex designs and have to be identified.
Learning Figures	Irregular shapes, parts of which are blackened, are presented in the memorization phase. In the reproduction phase, the subject has to remember the blackened parts and specify their positions.
Gestalt Completion	Incomplete drawings of familiar objects are presented. The subject has to identify the objects.
Paper Formboard	The subject has to determine which of several simple geometric figures fit together to form a more complex design.
Block Counting	Piles of blocks are presented. The subject has to find out for some of the blocks how many other blocks they touch in the pile.
Intersections	Simple three-dimensional blocks are presented. The subject has to find out what the blocks would look like when cut by a plane.
Figural Rotation	The subject has to identify one simple figure out of five that cannot be rotated to be congruent with the remaining four.
Eliot-Price Mental Rotations	The subject has to identify one out of four block shapes that is congruent with a standard shape but rotated in space.
Identical Blocks	The subject has to identify one irregularly shaped block out of five other blocks that is congruent with a model block but rotated in space.
Paper Folding	Each item shows a piece of paper being folded and punched with a hole. The subject has to determine what the paper would look like when unfolded.
Pattern Assembly	The subject has to find out how a pattern can be folded to form a three-dimensional object (box, pyramid, etc.).
Surface Development	The subject has to find out how a hollow three-dimensional object can be unfolded to yield a two-dimensional pattern.
Guay's Visualization of Views	The subject has to determine what an irregular three-dimensional object would look like from a different perspective.

(Manuscript received October 23, 1991;
revision accepted for publication April 14, 1993.)