MULTIVARIATE BEHAVIORAL RESEARCH, *41*(3), 261–293

Separating "Rotators" From "Nonrotators" in the Mental Rotations Test: A Multigroup Latent Class Analysis

Christian Geiser
Faculty of Psychology and Educational Sciences
University of Geneva

Wolfgang Lehmann
Department of Psychology
University of Magdeburg

Michael Eid
Faculty of Psychology and Educational Sciences
University of Geneva

Items of mental rotation tests can not only be solved by mental rotation but also by other solution strategies. A multigroup latent class analysis of 24 items of the Mental Rotations Test (MRT) was conducted in a sample of 1,695 German pupils and students to find out how many solution strategies can be identified for the items of this test. The results showed that five subgroups (latent classes) can be distinguished. Although three of the subgroups differ mainly in the number of items reached, one class shows are very low performance. In another class, a special solution strategy is used. This strategy seems to involve analytic rather than mental rotation processes and is efficient only for a special MRT item type, indicating that not all MRT items require a mental rotation approach. In addition, the multigroup analysis revealed significant sex differences with respect to the class assignment, confirming prior findings that on average male participants perform mental rotation tasks faster and better than female participants. Females were also overrepresented in the analytic strategy class. The results are discussed with respect to psychometric and substantive implications, and suggestions for the optimization of the MRT items are provided.

Correspondence concerning this article should be addressed to Christian Geiser, Faculty of Psychology and Educational Sciences, University of Geneva, 40, Bd. du Pont d'Arve, CH-1211 Geneva 4, Switzerland. E-mail: christian.geiser@pse.unige.ch

The Mental Rotations Test (MRT; Vandenberg & Kuse, 1978; redrawn version by Peters, Laeng, et al., 1995) is a frequently used paper-and-pencil test in the assessment of spatial abilities. The test uses the three-dimensional cube constructions originally introduced by Shepard and Metzler (1971) to measure mental rotation ability. A variety of studies have shown that the MRT is one of those spatial tests that produces the most robust and relatively large sex differences with males outperforming females (Linn & Petersen, 1985; Masters & Sanders, 1993; Moffat, Hampson, & Hatzipantelis, 1998; Peters, Chisholm, & Laeng, 1995; Quaiser-Pohl & Lehmann, 2000; Qubeck, 1997; Voyer, Voyer, & Bryden, 1995). Linn and Petersen (1985) as well as Voyer et al. (1995) reported mean effect sizes for mental rotation tasks of $d = 0.73$ and $d = 0.56$ in their meta-analysis, respectively, whereas in other spatial tasks (e.g., spatial perception; spatial visualization tasks) smaller or no gender differences at all are found.

However, the MRT is used not only in research on sex differences in spatial abilities but also in several other domains of psychological research—for example, to investigate the relationship between spatial abilities and math achievement (e.g., Casey, Pezaris, & Nuttall, 1992; Lehmann & Jüling, 2002; Pearson & Ferguson, 1989; Voyer, 1996). Moffat et al. (1998) showed that the MRT score was a good predictor for spatial route learning in humans. For applications in the field of neuroscience see, for example, Hausmann, Slabbekoorn, Van Goozen, and Cohen-Kettenis (2000) and Jordan, Heinze, Lutz, Kanowski, and Jäncke (2001).

It is interesting that, in spite of its widespread use, there is relatively little information available concerning the item properties of the MRT. Currently, we know of no detailed MRT item analysis. Therefore, it is not entirely clear whether all items of the test measure mental rotation appropriately. Research focusing on other spatial tests has shown that different test items might be prone to different solution strategies. For example, it has repeatedly been shown that in the Cube Comparison subtest of the German language Intelligence Structure Battery (Amthauer, 1953, 1970), different item types require the use of different cognitive strategies. Some of the test items can be better solved by a more elaborated spatial or "holistic" strategy—which involves mental rotation—whereas other problems do not require a spatial strategy, and can be solved using a nonspatial feature comparison approach (Glük, Machat, Jirasko, & Rollett, 2001; Hosenfeld, Strauss, & Köller, 1997; Köller, Rost, & Köller, 1994). The latter strategy is called "analytic" or "verbal" because it involves reasoning rather than mental manipulation of objects. The distinction between holistic and analytic strategies is very common in spatial ability research (see, e.g., Cooper, 1976; Just & Carpenter, 1985; Schultz, 1991). Given the fact that item analysis of other spatial tests have already shown that spatial problems can sometimes be solved by alternative solution strategies than those actually intended by the test authors (see, e.g., the analysis of cube comparison test items by Glück et al., 2001; Hosenfeld et al., 1997; Köller et al., 1994), it seems to be important to study the MRT items in detail as well. Prior studies in which the MRT was used support this view. For example, in the study of Peters, Laeng, et al.

(1995), participants reported that they had used different solution strategies in the MRT. Schultz (1991) used the Vandenberg and Kuse (1978) MRT version and other spatial tasks and was able to distinguish three different solution strategies for spatial problems. These differences were related to interindividual differences in the overall performance. Therefore, it seems possible that at least not all MRT items require a mental rotation approach (i.e., the items are probably not unidimensional). Taken together, it seems desirable to conduct a detailed item analysis to find out whether all individuals use mental rotations as solution strategy. More concrete, in the study presented here, we wanted to answer the following questions:

- Are there items in the MRT that can be solved by alternative strategies, and if yes, what are the characteristics of those items?
- Which solution strategies (other than mental rotation) can be successfully applied in the MRT?

Moreover, given the relatively large sex differences in MRT scores, we wanted to find out whether the use of different solution strategies is associated with gender differences. In past investigations, it was hypothesized that strategy differences may explain sex differences in spatial tests. The female disadvantage in spatial tasks was expected to be due to the predominant use of analytic strategies by females, whereas males were expected to prefer the more successful holistic approach. Research on this topic, however, yielded different findings. For the German cube comparison test the association between gender and strategy use was found to be insignificant (Hosenfeld et al., 1997). Burin, Delgado, and Prieto (2000) as well as Schultz (1991) also found no relationship between holistic versus analytic strategy choice and sex. In contrast, Peters, Laeng, et al. (1995) reported small but significant gender differences in strategy use in their redrawn MRT version. According to their results, males use a nonverbal strategy slightly more often than females. In addition, Qubeck (1997) showed that the male advantage was not present in all parts of the Vandenberg and Kuse (1978) version of the MRT. Given these equivocal findings, we wanted to clarify whether the superior performance of males can be explained by the use of more successful solution strategies. In the article presented here, we show how Multigroup Latent Class Analysis (MLCA) can be applied to answer the questions just discussed.

METHOD AND MATERIALS

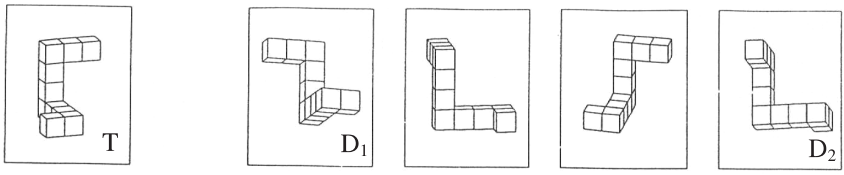
Participants

Altogether, 1,724 (865 female, 859 male) German volunteers participated in the study. There were 70.4% pupils (5th to 13th grade) of the German Bundesland

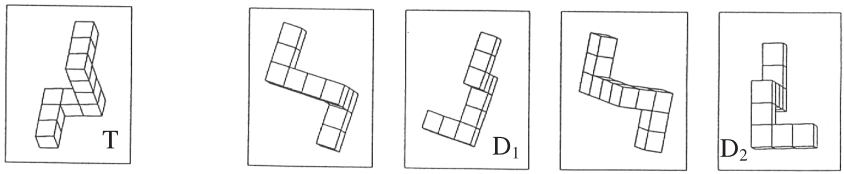
Sachsen-Anhalt, and 25.1% undergraduate students studying at the University of Magdeburg. For the item analysis there were 1,695 complete MRT data sets available. There were 1,693 participants (850 female, 843 male) who provided information on their age. Mean age was 16.9 (*SD* = 6.3) years for females and 16.7 (*SD* = 6.9) years for males. Participants received no reward for their participation except information about their personal test results.

MRT

We used a German language version of the redrawn Vandenberg and Kuse (1978) MRT by Peters, Laeng, et al. (1995; Form A). The MRT-A consists of 24 items that are administered in two sets (subscales) with 12 items, respectively. Each item consists of five three-dimensional block figures (two examples are depicted in Figure 1a and 1b). The block figure on the left (target, or T) has to be compared to the four similar constructions on the right-hand side. In each item, two of the four figures on the right are rotated versions of the target (correct alternatives), whereas the other two are distractor figures (D_1 and D_2). The two correct alternatives should be recognized and marked by the participants. An important aspect concerning the MRT item construction is that there are two different types of distractor figures. In most of the items the distractors are *mirror images* of the target, whereas in some items, the distractors are cube figures of *different shape*. The two item types are illustrated in Figure 1a and 1b. Figure 1a shows an MRT item in which the distractor



(a)



(b)

FIGURE 1 (a) Example of a MRT type I item. The distractor figures (D_1 and D_2) are mirror images of the target (T). (b) Example of a MRT type II item. The distractor figures (D_1 and D_2) are of different shape compared to the target (T).

constructions are mirror images of the target (*type I item*), whereas a *type II item* is depicted in Figure 1b. In the type II item, the distractors are different in shape from the target. Mixed forms (one distractor is a mirror image of the target and the other is of different shape) also exist in the test.

Procedure

First, participants read the instructions on the first page of the test. They were then given three "warming-up" items to become familiar with the test format. These sample items were administered without time limit, and participants were afterward informed about the correct solutions. Before they completed the actual test items, participants were told that they would receive credit for a problem only if *both* correct alternatives were marked. As we explain next, this scoring method is recommended because it provides a correction for guessing (Peters, Laeng, et al., 1995). For the actual test, we allowed 3 min per subscale, and the two subscales were separated by a pause of 2 min.

STATISTICAL ANALYSIS

MRT Scoring Method

We used the strict scoring method for the MRT that is recommended by the test authors. This procedure leads to binary item scores: An item score is 1 (i.e., "correct") if and only if *both* correct alternatives are marked by the partcipant, and the item is scored 0 (i.e., "incorrect") in all other cases; so items that were not reached are also scored as incorrect (0). According to this procedure, the maximum score is 24 (all items correct). This scoring method, which we refer to as *conventional score*, is generally applied because it has the advantage that the probability of receiving credit for an item by simply guessing is only $p = .17$.

In addition, we recorded how many items were reached by the participants to compute *ratio scores*. Ratio scores were computed by dividing the total number of items correct by the total number of items attempted. This scoring method is especially interesting because it provides a "correction" for the speededness of the test. Some studies have shown that the gender difference is reduced when women and men are compared according to ratio scores rather than conventional scoring methods, which do not consider how many items were reached (Goldstein, Haldane, & Mitchell, 1990; Stumpf, 1993; Voyer, 1997).

Latent Class Analysis

Following a similar approach that has been used in previous studies on solution strategies in spatial tests (Glück et al., 2001; Hosenfeld et al., 1997; Köller et al., 1994), we applied latent class analysis (LCA) to examine our research questions.

The main advantage of this approach is that the properties of each single item can be studied. In addition, whereas traditional methods of item analysis (e.g., factor analysis) assume that the population under study is homogeneous, LCA enables researchers to detect population heterogeneity with respect to solution strategies. Moreover, problems associated with the use of strategy self-report questionnaires are avoided. In the following section, we provide a short description of the LCA method (for a more detailed discussion, see, e.g., Clogg, 1995; Eid, Langeheine, & Diener, 2003; Langeheine & Rost, 1988).

LCA is a technique for the analysis of categorical outcomes (e.g., dichotomous test items as in the MRT), which assumes that the associations between items can be explained by the existence of several subgroups that cannot be observed directly (therefore called *latent* classes). Thus, LCA is a typological rather than a dimensional approach. Within one latent class, participants are assumed to have identical patterns of solution probabilities (i.e., the solution probability of a given item is the same for all individuals belonging to the same class). Between classes, however, there are differences with respect to the response probabilities. This means, for example, that Item A of Test X may be easy for all members of subgroup G1 but difficult for members of subgroup G2, whereas for Item B of the same test the reverse may be true (i.e., Item B has a high solution probability in subgroup G2 but a low probability in subgroup G1). This would be the case if, for example, Item A required solution strategy S1, which is used by the members of group G1 but not by the members of group G2, whereas Item B required use of strategy S2, which is only applied in G2. Hence, LCA is able to detect typological differences with respect to solution strategies. For assessment purposes, an individual can be assigned to the latent class for which her or his assignment probability is maximum. In the present case, participants can be classified according to the strategy that they used in the MRT. In addition, strategy choice as indicated by latent class membership can then be related to other variables (e.g., gender). In this way, sex differences with respect to solution strategies in the MRT can be studied.

In mathematical terms LCA models for dichotomous items have the general form

$$p(X_{vi} = 1) = \sum_{g=1}^{G} \pi_g \pi_{ig}, \tag{1}$$

where $p(X_{vi} = 1)$ is the probability that a randomly chosen individual v gives a correct response to item i, π_g refers to the (unconditional) probability that a randomly chosen individual belongs to latent class g and π_{ig} represents the conditional response probability for an item i in class g. Hence, the model parameter π_g refers to the size of class g and the parameter π_{ig} expresses that the solution probability for an item i is constant for all members of the same class g.

LCA, like exploratory factor analysis (EFA), can be seen as a data reduction method. Whereas in EFA one tries to explain observed interitem correlations by a reduced number of common factors (continuous latent variables), the aim of LCA is to explain interindividual differences in item response patterns by a reduced number of groups (latent classes, i.e., the values of a categorical latent variable). However, in most cases, the number of classes needed to appropriately account for the response patterns is not known beforehand. The appropriate number of classes can be determined by comparing the goodness of fit of several models with an increasing number of classes. There are a variety of indexes available for the evaluation of latent class models (see, e.g., Collins, Fidler, Wugalter, & Long, 1993; Eid et al., 2003). Information criteria (IC) like Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) are frequently used to compare the fit of competing LCA models. These indexes have the advantage that not only do they consider how well a model fits the data but they also reward more parsimonious models in contrast to more complex models in which many parameters are estimated. When comparing a series of models the model with the lowest IC value is selected. IC do not always agree with respect to the number of classes that should be chosen (Glück et al., 2001). AIC is known to prefer solutions with a relatively large number of classes (Lin & Dayton, 1997) and is especially recommended for small item numbers and large pattern frequencies, whereas BIC often points to more parsimonious models and should be preferred for sparse tables (large item numbers and small pattern frequencies; Rost, 1996). One disadvantage of the fit assessment based on IC is that these measures are rather descriptive and only meaningful in relation to other models. A possibility for testing model fit statistically is offered by the Pearson, the Cressie-Read and the likelihood ratio (LR) test statistics. These statistics are asymptotically chi-square distributed. Model fit is considered acceptable if the p value for the chi-square statistic does not fall below an a priori specified alpha level (e.g., $\alpha = .05$). Unfortunately, the assumptions under which these test statistics follow a chi-square distribution may be violated under the condition of sparse data. Data are considered sparse if there are many unobserved patterns, and this may be the case even with relatively few items in relatively large samples. Therefore, Langeheine, Pannekoek, and van de Pol (1996) as well as von Davier (1997, 2001b) recommend the use of a parametric bootstrap procedure if the asymptotic conditions for the test statistics are unlikely to hold. In this parametric bootstrap, the parameter values estimated for the target model are taken as true population values. Then, a large number of bootstrap samples (say, 500) are drawn from the hypothesized population model, and for each sample the model parameters are estimated and the values of the test statistics are computed. The so-obtained bootstrap distributions of the fit statistics are then used to evaluate the target model, that is, the test statistics computed for the target model are compared to the distributions of test statistics in the bootstrap samples. For example, if for 25 or more (i.e., 5% or more) out of 500 simulated bootstrap samples the fit statistics show larger values than those obtained for the real data, the model has not yet been rejected.

Furthermore, the Vuong-Lo-Mendell-Rubin likelihood ratio test (LMR test; Lo, Mendell, & Rubin, 2001) as well as an adjusted version of this test can be used to determine whether a model with a given number of latent classes (say, k classes) fits the data significantly better than a more parsimonious model with one class less (that is, $k - 1$ classes)[1]. Both tests are implemented in the computer program Mplus (Muthén & Muthén, 2004) for single group (but not for multigroup) LCA. A significant LMR test value (i.e., $p < .05$) for a model with k classes indicates that this model fits the data better than the $k - 1$ class model.

In our study, we used the computer program Mplus 3.12 (Muthén & Muthén, 2004) for the LCA. In addition, we fit the same models in PANMARK 3.09 (van de Pol, Langeheine, & de Jong, 1996). Both programs use the EM algorithm and maximum likelihood estimation to compute the LCA model parameters, standard errors, and fit statistics. The advantage of this "double-fitting" strategy was twofold. From a rather technical point of view this enabled us to better check whether the LCA solutions estimated by the two programs represented global likelihood maxima solutions and not local optima. Local solutions are quite common in mixture analysis, especially when models with a larger number of classes (say, $k > 5$) are fit to the data and/or the classes are not well separated. It is therefore recommended to try a large number of different starting values for each model to make sure that the estimation algorithm will find the global maximum of the likelihood rather than stop at a local optimum. If two different programs estimated different results for the same model this would indicate that (at least) one program did not find the global maximum.[2] On the other hand, using the two programs, we were able to investigate a larger variety of fit statistics for model selection (e.g., the LMR test is not implemented in PANMARK, whereas Mplus does not compute the Cressie-Read statistic and offers no parametric bootstrap facilities for LCA fit measures). Finally, we used the program WINMIRA 1.37 (von Davier, 2001b) to compute person fit statistics for the selected LCA models. These person fit statistics can help detect participants with aberrant response patterns who can have a large impact on model fit in LCA. Person fit values ≤ -2 indicate outliers who may show very unusual response patterns (details with respect to the computation of

[1]The conventional LR difference test procedure for nested models cannot be applied to test models with different numbers of classes against each other since the necessary regularity conditions are not met (for details, see McLachlan & Peel, 2000).

[2]To make sure that a true likelihood maximum would be reached for each model, we changed the default settings of Mplus for mixture analysis as follows. Instead of using only 10 random sets of starting values in the first stage of optimization, we always used 500 random sets of starting values. Second, we used the ending values of the 50 best solutions from Stage 1 (i.e., the solutions with the highest loglikelihood values) as starting values for the final stage optimization (the Mplus default is using only the single best solution from Stage 1). In addition, we increased the maximum number of initial stage iterations from the default of 10 to 50 iterations. In PANMARK, we also used up to 500 random sets of starting values.

person fit indexes in LCA can be found in an appendix to the WINMIRA manual; see von Davier, 2001a).

As this was an exploratory study, we fit models with different numbers of classes (i.e., the 1 to 8 class solution) to the MRT items. The two subscales (12 items per half) were analyzed separately, and we used dichotomously (binary) scored items for the LCA (according to the conventional score, see previous discussion). For model selection purposes, we compared the fit of the different LCA solutions. That is, the models were compared according to their AIC and BIC values, whereby we placed more emphasis on the BIC, as the data are relatively sparse in the investigation presented here. In addition, we examined the results of the LMR test and adjusted LMR (aLMR) test to choose on the number of classes (we only report the results of the aLMR test as the pattern of significance was exactly the same for the nonadjusted version of the test).

Furthermore, to investigate the absolute fit of a model, we ran parametric bootstrap analysis (based on 500 bootstrap samples for each model) for the Pearson and the Cressie-Read test statistics and computed bootstrap probability values for these statistics (as mentioned above, the bootstrap analysis were run in PANMARK as no such bootstrapping is available in Mplus version 3.12). We do not report the bootstrap results of the LR test because simulation studies of von Davier (1997) have shown that the bootstrap procedure does not work well with this test statistic.

One major goal of our study was to examine gender differences in strategy use in the MRT. Consequently, after selecting a latent class model for the entire sample, we tested for measurement equivalence across gender. Measurement equivalence requires the conditional (within class) response probabilities for the MRT items to be equal for both groups. Only if these probabilities are invariant over gender, the latent class model would be considered the same for both groups and meaningful comparisons of the class sizes could be made. To test for measurement equivalence, we performed unconstrained, semiconstrained, and fully constrained multigroup latent class analysis (MLCA) with gender as the grouping variable. In all MLCA, the same number of latent classes was chosen for both groups. However, in the unconstrained multigroup models, both the class sizes and the conditional response probabilities were allowed to differ across groups. In contrast, in the semiconstrained models, the class sizes were still allowed to vary but the conditional response probabilities in each class were constrained to be equal for females and males. Finally, in the most restrictive models (i.e., the fully constrained MLCA), both the class sizes *and* the conditional response probabilities were fixed to be the same in both groups. In comparing the fit of the semiconstrained model to the fit of the unconstrained model, one can investigate whether the class structures are the same for both groups. If the semiconstrained model does not fit worse than the unconstrained model, this would indicate that the assumption of measurement equivalence (equal conditional response probabilities in both groups) does not have to be rejected (i.e., the structure of the latent classes does *not* differ for fe-

males and males). Furthermore, in comparing the fully constrained model to the semiconstrained model, one can decide whether there are sex differences in strategy use, because in both models, the same class structures hold for both groups, and therefore, the same strategy classes exist. If the fully constrained model (which includes the restriction of equal class sizes across groups) fits worse that the semiconstrained model, one would conclude that there *are* sex differences in class sizes and therefore also in strategy use. In the study presented here, the different multigroup models previously discussed were compared according to their BIC values, and the model with the lowest BIC value was chosen.

RESULTS

Before discussing the LCA results we report overall scoring results and effect sizes for both scoring techniques previously described to verify the expected sex differences in the MRT. In addition, traditional reliability measures are given.

Overall Scoring and Sex Differences

To investigate the expected sex differences in the MRT performance we compared the overall scores of females and males. It turned out that males outperformed females no matter whether the conventional (strict) scoring method or the ratio score method was used. A detailed listing of the scores of males and females can be taken from Table 1.

As shown in Table 1, the effect size for the conventional MRT score (d = 0.72) was very similar to the results of the meta-analysis by Linn and Petersen (1985). The second finding is that the number of items attempted was larger for males than for females (d = 0.58). Third, the effect size for the gender difference was clearly reduced when ratio scores instead of the conventional scoring method were used (d = 0.52). This is in agreement with Goldstein et al. (1990) and Stumpf (1993) who also report a reduced gender effect when the MRT performance of males and females is compared using ratio scores. Note however, that the reliability of the ratio scores (.72) was somewhat lower than the reliability estimates for the conventional score.

LCA

Single group analysis. The AIC and BIC values for the 1 to 8 class solutions for the entire sample are printed in Table 2 for both subscales. The results showed that according to the BIC a model with five latent classes should be selected for both subscales (for this solution the BIC values are the lowest), whereas AIC pointed to a model with more classes for both subscales. As previously noted, it can be observed in many applications that the AIC reaches its lowest value for

TABLE 1
Sex Differences in the Overall Mental Rotations Test (MRT) Performance

					Reliability	
	N	*M*	*SD*	*Effect Size d*	*Split Half*	*Cronbach's α*
Conventional score						
Females	851	9.20	4.61			
Males	844	13.02	5.32	0.72	.80	.87
No. of items attempted						
Females	822	16.26	4.65			
Males	721	18.97	4.30	0.58	.79	—
Ratio scores						
Females	826	0.58	0.23			
Males	724	0.70	0.21	0.52	.72	—

Note. All mean differences are statistically significant ($p < .01$). Ratio scores were computed by dividing score 24 by the number of items attempted. The effect size *d* was computed by dividing the mean differences by the respective overall standard deviations. Split half reliability coefficients are Spearman-Brown corrected correlations between the two MRT subscale scores.

models with a relatively large number of classes, whereas BIC points to more parsimonious solutions. For the present case however, the decrease in AIC values was rather small for models with more than five classes. This is illustrated in Figure 2.

In Figure 2, the AIC and BIC values are shown for all LCA models. As can be seen in Figure 2, the differences in AIC values were only marginal for models with more than five classes. Therefore, we selected the five-class model, which appeared to be the most parsimonious description of the data, for both subscales. To further evaluate the five-class solutions, we investigated the results of the aLMR test. For Subscale 1, the aLMR test yielded a highly significant value for the comparison of four versus five classes (aLMR test value = 260.83, $p < .01$), indicating that the five-class model fit the data better than the 4-class model. However, the aLMR test was nonsignificant for five versus six classes (aLMR = 94.69, $p = .08$), showing that adding another class did not improve fit significantly. Therefore, the results of the aLMR test were in agreement with the BIC. The findings were somewhat different for Subscale 2. Again, five classes turned out to be better than only four (aLMR = 142.89, $p = .01$), but the aLMR test indicated that six classes were still better than five (aLMR = 61.86, $p < .01$). For further clarification, we used the Pearson chi-square and the Cressie-Read statistic as a third criterion for model selection. As the data are relatively sparse (large number of items relative to the sample size) these test statistics might not approximate the chi-square distribution. Therefore, we ran parametric bootstrap analysis for both five-class models in PANMARK. In each analysis, 500 bootstrap samples were simulated. Table 3 contains the values of the test statistics and their respective bootstrap probability values [*p*(B)].

TABLE 2
AIC and BIC Values for Different Latent Class Analysis Models

	Subscale 1		*Subscale 2*	
	AIC	*BIC*	*AIC*	*BIC*
1 class	22869.89	22935.12	25191.41	25256.64
2 classes	19690.54	19826.42	23199.81	23335.70
3 classes	19077.71	19284.25	22701.30	22907.84
4 classes	18630.68	18907.88	22474.45	22751.66
5 classes	18393.14	**18741.01**	22356.09	**22703.95**
6 classes	18323.48	18742.00	22319.59	22738.12
7 classes	18252.69	18741.88	22297.35	22786.54
8 classes	18234.86	18794.72	22284.10	22843.95

Note. Lowest BIC values are printed in boldface. N = 1,695. AIC = Akaike's information criterion; BIC = Bayesian information criterion.

It can be seen that both models do not show an acceptable fit according to the Cressie-Read statistic as the probability of finding a larger Cressie-Read value in a bootstrap sample than the value actually computed for the real data is smaller than .05. In contrast, the Pearson statistic indicates model misfit for Subscale 1 ($p < .01$) but provides an acceptable p value for the five-class solution for Subscale 2 ($p = .07$). To find out why a model fails the chi-square tests, it is often useful to investigate the impact of outliers on model fit. As already mentioned, outliers are participants who show unique response patterns. The program WINMIRA provides person fit statistics, which can be used to detect such outliers. In the present case, it turned out that several participants showed bad person fit values (≤ -2) associated with anomalous response patterns. Three typical examples of such aberrant patterns are given in Table 4 for the 12 items of Subscale 1. A 1 means that a participant has found the correct solution for an item, whereas 0 indicates a wrong response.

Table 4 shows that different kinds of deviant patterns occurred. The first row in Table 4 contains the pattern of a person who has not found the correct solutions for the first six items but for five of the last six items. There are two possible explanations. One possibility is that this participant has wrongly started with the second test page (there are six items on each page). Another explanation could be that this person has misunderstood the instructions of the test and wrongly marked the *false* alternatives (distractor figures) and not, as intended, the *correct* alternatives ("reverse score").

Another class of patterns seems to contain participants who guessed. An example is given in the second row of Table 4. As previously mentioned, the guessing probability for an MRT item is .17. Hence, on average, participants who simply guess can hit about two correct answers per subscale by chance. In the third exam-

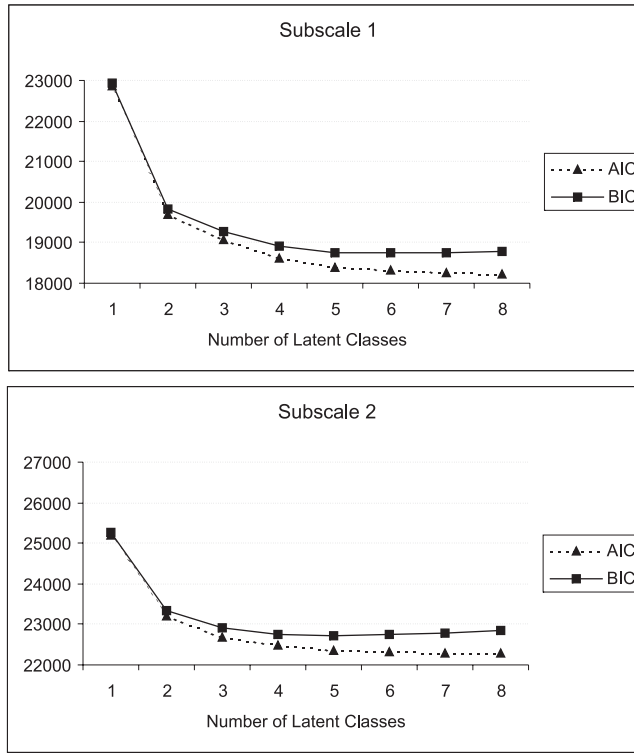


FIGURE 2 Plots of AIC and BIC values for the 1 to 8 class solutions.

ple (third row in Table 4), a participant has responded correctly to the first six and the last two items of Subscale 1 but not to Items 7 to 10. We suppose that participants switching to the last items when they realize that they run out of time may cause such unusual response patterns.

In a post hoc analysis we explored the impact of the aberrant response patterns on model fit. Therefore, we deleted the 41 worst-fitting cases (2.5 %) all of which had very unusual response patterns comparable to those illustrated in Table 4 and fitted the five-class solution to the reduced sample. The rank order of the BIC values remained the same for both subscales (i.e., the five-class solutions still showed the smallest BIC values for both subscales). The results of bootstrapping analysis indicated that in the reduced sample the five-class model fits the data well (N = 1,654; Pearson χ^2 p(B) value = .83; Cressie-Read p(B) value = .45). Similar results were obtained for Subscale 2. The initial misfit of the five-class model is therefore well explained by this special group of outliers who showed different kinds of idiosyncratic response patterns. We did not treat this relatively small group in a special

TABLE 3
Goodness of Fit Statistics for Both Five-Class Models

		Pearson χ^2		*Cressie-Read*	
	df	*Value*	*p(B)*	*Value*	*p(B)*
Five classes (Subscale 1)	4031	19761.78	< .01	4290.86	< .01
Five classes (Subscale 2)	4031	4329.89	.07	2948.28	< .01

Note. *p*(B) = probability value based on 500 parametric bootstrap samples, respectively.

TABLE 4
Examples of Aberrant Response Patterns (First Subscale)

Response Pattern (Subscale 1)		
Page 1	*Page 2*	*Hypothesized Cause of Strange Pattern*
0 0 0 0 0 0	1 1 0 1 1 1	Started with second test page/Instructions misunderstood
0 0 1 0 0 1	0 0 1 0 0 0	Guessed
1 1 1 1 1 1	0 0 0 0 1 1	Ran out of time and switched to the last items

Note. 0 indicates an incorrect item response. 1 indicates that the item was answered correctly.

manner and accepted the five-class solutions for the total sample (i.e., the outliers were not excluded). In the following section, we discuss the multigroup analysis that were performed in order to test for measurement invariance across gender.

Multigroup LCA. In the multigroup analysis, we investigated whether the five-class models chosen for the entire sample showed the same latent class structures and class sizes for both females and males. Therefore, for both subscales, we estimated and compared the three types of nested multigroup models previously discussed. The BIC values for the multigroup models are printed in Table 5. According to the BIC, for both subscales, the best model is the semiconstrained five-class model which assumes measurement equivalence (equal conditional response probabilities) across gender but permits unequal class sizes. This model is much more parsimonious than the unconstrained model in which 60 additional parameters (i.e., 60 additional conditional response probabilities) are estimated. Furthermore, the unconstrained solution showed very similar class structures for both females and males indicating that there were no strong structural differences. Therefore, we concluded that the assumption of measurement invariance across gender was justified. However, it seemed to be unreasonable to additionally assume equal class sizes for both groups as the benefit in terms of four additional *df* is associated with a significant increase in the BIC values for both subscales indicating nonnegligible sex differences in class proportions. Consequently, we will

TABLE 5
Bayesian Information Criterion (BIC) Values for Different Multigroup Models (Females vs. Males)

	Subscale 1			*Subscale 2*		
	Unconstrained	*Semiconstrained*	*Fully Constrained*	*Unconstrained*	*Semiconstrained*	*Fully Constrained*
Four classes	21348.08	21104.89	21265.06	25195.39	24960.38	25108.84
Five classes	21273.78	**20937.36**	21098.19	25239.08	**24918.35**	25061.13
Six classes	21348.74	20942.46	21099.18	25346.97	24949.64	25095.29

Note. Smallest BIC values are printed in boldface.

report the parameter estimates and standard errors for the semiconstrained five-class solution with equal conditional response probabilities but unequal class sizes for females and males.

Five-class model for Subscale 1. Table 6 shows class membership statistics for the five-class solution for the first subscale. In LCA, the probability of belonging to each of the different classes is calculated for all participants, and an individual can be assigned to the latent class for which her or his assignment probability is maximum. The mean assignment probabilities for all participants attached to the same class can be interpreted as reliability measures for the class assignment (see Table 6).

It can be seen that these mean probabilities are above .81 for all classes in both groups. Hence, the values for the present model indicate high classification reliabilities. Furthermore, it can be seen from the class sizes reported in Table 6 that there are sex differences with respect to the class assignment for all latent classes. This explains the higher BIC values observed for the comparison of the semiconstrained to the fully constrained multigroup models, because the latter assume equal class sizes for females and males (see Table 5). Before we discuss these sex differences in detail, we describe the structure of the five classes. The response profiles of the five classes in terms of conditional solution probabilities are shown in Figure 3 (for reasons of clarity, we do not present the standard errors of these estimates in Figure 3. However, the standard errors can be found in Appendix 2a for Subscale 1 and Appendix 2b for Subscale 2).

TABLE 6
Latent Class Membership Statistics for the Five-Class Solution (Subscale 1)

	Group Females			*Group Males*		
	Probability of Expected Class Membership			*Probability of Expected Class Membership*		
	M	*SD*	*Class Size (%)*	*M*	*SD*	*Class Size (%)*
Class 1 (26.8 %)	.90	.14	37.5	.86	.18	16.1
Class 2 (13.2 %)	.85	.15	14.9	.83	.17	11.5
Class 3 (28.6 %)	.87	.16	20.9	.86	.15	36.4
Class 4 (18.2 %)	.81	.17	21.7	.92	.08	14.7
Class 5 (13.1 %)	.88	.16	5.0	.93	.13	21.4

Note. Class proportions are based on the estimated model (not on estimated posterior probabilities).

The class profiles in Figure 3 show that there are three classes where members differ mainly in the number of items that they reach (Classes 3, 4 and 5). Members of Class 5 (13.1 %) have high solution probabilities for all items. Therefore, one could denote them as "high-performance class". In contrast, participants in the largest class (Class 3, 28.6 %) have high solution probabilities for the first 8 items but the majority does not reach the last 4 items. In Class 4 (18.2 %) there is an even earlier decrease in the solution probabilities that starts after the fourth item. For the last 6 items, the solution probabilities are close to zero in this class. One could denote these three classes as "speed classes" because they differ in the amount of items reached in the time limit. The other two groups (Class 1 and 2) show structurally different patterns. Class 1 (26.8 %) consists of participants who show weak performance, as indicated by the low solution probabilities for all but one item (Item 3: solution probability = .63). In Class 2 (13.2 %), a special response pattern is shown. Members of this group have high solution probabilities for Items 3, 4, 7 and 8 but relatively low probabilities (below .50) for the other items. It turned out that these 4 items, which are easy for the participants in Class 5, differ from the other 8 items with respect to the distractor figures that are used. Although the distractors consist of *mirror images* of the target figure in the majority of items (type I items; see Figure 1a), Items 3, 4, 7 and 8 have distractors that are *different in shape* from the target figure (type II items; see Figure 1b). We suspect that this makes it easy for participants to solve these items by a "non-mental-rotation" strategy. Instead of mentally rotating all four figures to find the correct solution, participants in Class 2 simply compare the shapes of the figures and exclude the distractor figures, because they have a different form than the target and thus do not come into question as correct solutions. Because this analytic feature comparison

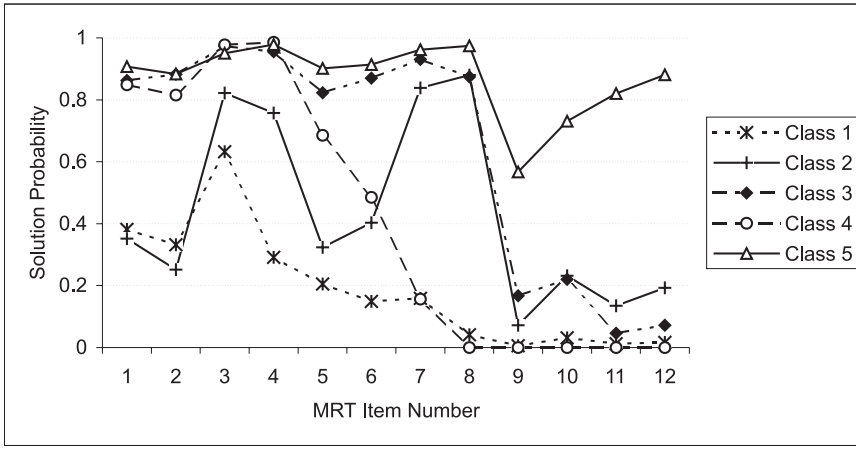


FIGURE 3 Latent class profiles for the five-class model (Subscale 1).

strategy does not work well for items that contain mirror image distractors (which cannot be excluded by simple shape comparisons), participants in Class 2, who seem to solely use the described nonrotation strategy, have high solution probabilities only for those four items that "allow" the successful use of this analytic strategy. In contrast, the other 8 items, which seem to require mental rotation, are difficult for this "nonrotation" class.

When taking a closer look at the sex differences with respect to the class assignment, it can be seen that differences were especially large for Class 1 (f = 37.5 %; m = 16.1 %), Class 3 (f = 20.9 %; m = 36.4 %), and Class 5 (f = 5.0 %; m = 21.4 %), indicating that females were more often assigned to the class with the worst performance (Class 1) and less often found in the high-performance Classes 3 and 5 than males. Furthermore, females were slightly overrepresented in the special strategy class 2 (f = 14.9 %; m = 11.5 %) and in Class 4 (f = 21.7 %; m = 14.7 %).

Five-class model for Subscale 2. The five-class model for Subscale 2 turned out to be quite similar to the five-class solution for the first test half. Table 7 contains the class membership statistics for this model. The mean class assignment probabilities were somewhat lower than for the first item set but with an average value of .82 (group female) and .81 (group male) still satisfactory. The size of the low-performance class number 1 (23.5 %) has slightly decreased. Furthermore, fewer participants were classified as Class 3 members (23.7 % vs. 28.6 %). In contrast, more participants are now found in Class 2 (17.3 % vs. 13.2 %), Class 4 (19.1 % vs. 18.2 %), and Class 5 (16.5 % vs. 13.1 %). The class profiles for the second test half are depicted in Figure 4.

TABLE 7
Latent Class Membership Statistics for the Five-Class Solution (Subscale 2)

	Group Females			*Group Males*		
	Probability of Expected Class Membership			*Probability of Expected Class Membership*		
	M	*SD*	*Class Size (%)*	*M*	*SD*	*Class Size (%)*
Class 1 (23.5 %)	.80	.16	34.3	.73	.16	12.5
Class 2 (17.3 %)	.84	.17	18.1	.79	.17	16.4
Class 3 (23.7 %)	.82	.17	18.5	.83	.16	29.0
Class 4 (19.1 %)	.75	.16	21.9	.76	.18	16.2
Class 5 (16.5 %)	.87	.13	7.1	.93	.13	25.9

Note. Class proportions are based on the estimated model (not on estimated posterior probabilities).

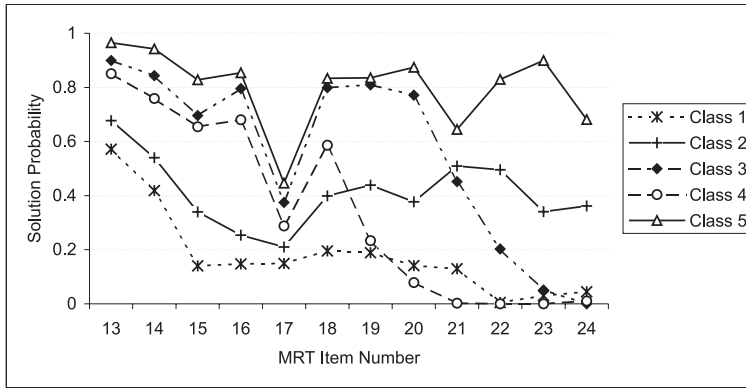


FIGURE 4 Latent class profiles for the five-class model (Subscale 2).

At first sight, the class profiles may seem different from those for the first subscale, especially because of the "break" at Item 17 that occurs in all classes. Item 17 therefore deserves closer consideration; we discuss this later. The second difference lies in the generally weaker performance of all classes. That means that the solution probabilities are lower in all classes compared to the five-class model for the first item set, probably due to fatigue effects.

A closer look at the class profiles reveals the similarities between the five-class models for both test halves. As for the first subscale, there are three speed classes (Classes 3, 4 and 5) that can be distinguished according to their speed-of-performance patterns. Again, participants in Class 5 show the highest mental rotation performance. Except for Item 17, the solution probabilities for all items are above .60. In Class 3, participants still show a relatively high performance that decreases after Item 20. Class 4 is again the slowest of the three speed classes, as the solution probabilities are above .50 only for the Items 13, 14, 15, 16, and 18.

A weak-performance class is also found in this model. In this group (Class 1), participants have very low solution probabilities that fluctuate around the guessing probability of .17, except for Items 13 and 14, where the probabilities are somewhat higher. Interestingly, we also find a non-rotation-strategy class in the second test half (Class 2). The easiest problems for this group are Items 13, 14, 21, and 22 that can be categorized as type II items. However, the solution probabilities in this group are not as clear cut as for the first item set. The reason could be that the structural differences of the distractors of Items 21 and 22 are not as obvious as for the type II items in the first item set. This makes it more difficult to use the described feature comparison strategy successfully for these items.

The pattern of sex differences with respect to class assignment was very similar to the results obtained for Subscale 1. Again, differences were especially large for Class 1 (f = 34.3 %; m = 12.5 %), Class 3 (f = 18.5 %; m = 29.0 %),

and Class 5 (f = 7.1 %; m = 25.9 %). However, females were also overrepresented in Class 2 and 4.

Latent class transition from Subscale 1 to Subscale 2. The structural similarities of the latent classes indicated that there is a high match between the five-class solutions for both MRT subscales. That is, mainly the same class structures are observed in both models. To further explore the concordance of both solutions we estimated latent transition probabilities (LTPs). LTPs can be used to find out how likely it is that individuals stay in the same latent class over time. On the other hand, using LTPs, it can also be assessed how likely it is that participants assigned to a given class at one occasion, say t_1, will be found in any of the other classes at another occasion (t_2). That is, it can be investigated whether there are participants who switch to a different class. In other words, LTPs can be interpreted as coefficients of class stability and class change. An LTP value of 1 (for the same class) would indicate perfect stability (all participants remain in the same class). In contrast, LTP values (for the same class) close to 0 would mean that it is very unlikely to stay in that class over time. Low LTPs for *different* classes show that it is unlikely that participants switch to another class from t_1 to t_2. In our case, assuming that the five classes are comparable across subscales, the question is, do participants stay in the same class for both MRT subscales? If LTPs for the same classes turned out to be smaller than 1, this would mean that class membership for Subscale 2 is not perfectly predictable from the class assignment for Subscale 1. In this case, it would be interesting to study interclass migrations. For example, it could be possible that some participants switch from one strategy (used to solve the items of Subscale 1) to another strategy for Subscale 2. To estimate LTPs for the five-class solution, we specified a longitudinal latent class model. In this model, both subscales (i.e., all 24 MRT items) were analyzed simultaneously. The model included two latent class variables (one for each subscale) with five latent classes each. In the model, the latent class variable for the items of Subscale 2 was regressed on the latent class variable for Subscale 1 (using a logistic regression). The model was specified as a multigroup model with equal conditional (within class) response probabilities across gender. The parameter estimates (conditional response probabilities) of the transition model were virtually identical to those obtained for the single latent class variable models for both subscales just discussed. Therefore, it seemed justified to interpret the latent transition probabilities for the five-class solution. These probabilities can be found in Table 8. The results show that all five classes are relatively stable across subscales (this can be seen from the LTP values in the diagonal of the transition matrix in Table 8, which are all above .50).

The most stable classes are the weak-performance Class 1 (LTP = .75) and the High-Performance Class 5 (LTP = .69). On the other hand, the least stable class is

TABLE 8
Estimated Latent Transition Probabilities Across Subscales

	Subscale 2				
	Class 1	*Class 2*	*Class 3*	*Class 4*	*Class 5*
Subscale 1					
Class 1	**.75**	.08	.03	.14	.01
Class 2	.29	**.63**	.03	.03	.02
Class 3	.00	.09	**.51**	.21	.19
Class 4	.07	.08	.25	**.58**	.02
Class 5	.00	.13	.17	.01	**.69**

Note. Probabilities are based on the estimated model (not on estimated posterior probabilities). Latent transition probabilities for the same class (class stability coefficients) are printed in boldface.

Class 3 (LTP = .51), where participants have relatively high transition probabilities for Class 4 (LTP = .21) and Class 5 (LTP = .19). This indicates that some participants in Class 3 show better performance for Subscale 2 (those switching to Class 5) whereas others perform less well (those switching to Class 4). When we take a look at the LTPs for the special solution strategy Class 2, it can be seen that this class is also quite stable across subscales as participants have a probability of .63 to stay in this class. It is also worth noting that of those participants who leave Class 2, most are found in the class with the worst performance (Class 1, LTP = .29). Only very few participants assigned to Class 2 for Subscale 1 switch to one of the higher performance classes (LTPs are only .02 through .03 for these classes).

Item 17. As already mentioned, Item 17 deserves a closer investigation because this item is especially difficult for all classes. To find out why this item is so difficult for most participants we had a look at the distractor figures of this problem. Item 17 is depicted in Figure 5. The distractor constructions are denoted D_1 and D_2, and the true alternatives T_1 and T_2.

The first distractor (D_1) obviously has a different shape than the target figure and therefore should be identified relatively easy as a false alternative. The second distractor (D_2) also has a different form than the target. However, this fact is much less obvious than for D_1. On the contrary, D_2 appears to be very similar to the target. In addition, the true alternative T_1 is rather hard to recognize. Indeed, although 90.4 % of all participants correctly reject D_1 as a false alternative, 57.6 % wrongly mark D_2 as a correct choice. This may explain the "gap" in the profiles of all classes at Item 17. Because of this specialty of Item 17, we excluded it from the general type I versus type II item classification (though, strictly speaking, it would have to be categorized as type II item).

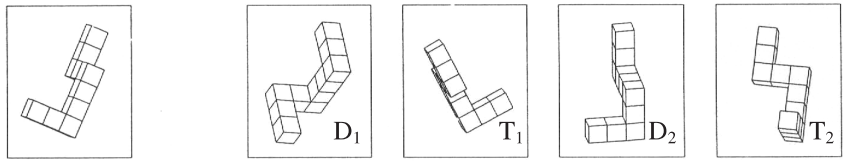


FIGURE 5 MRT Item 17.

DISCUSSION

In the study presented here, we investigated the psychometric properties of the MRT to find out whether all of the 24 MRT items require a mental rotation strategy (i.e., can only be solved by mental rotation). In addition, we were interested in the question of whether the relatively large sex differences observed in mental rotation tasks can be explained by the use of different solution strategies by females and males, respectively. The overall scoring results confirmed findings of prior studies which showed that males outperform females in the MRT. We found a relatively large effect size for the conventional MRT scoring procedure. This is in agreement with the meta analysis of Linn and Petersen (1985), as well as Masters and Sanders (1993), which revealed that the effect sizes in mental rotation have not substantially decreased in the past decades. However, in our data, the effect size was reduced when males and females were compared according to ratio scores instead of conventional scores, an effect that has already been shown by Goldstein et al. (1990), Stumpf (1993), and Voyer (1997).

The results of a multigroup LCA showed that a five-class solution with equal class structures for females and males is adequate for both subscales. Furthermore, the class structures were found to be very similar for both item sets. In both test halves, there are three subgroups that differ mainly in their speed of performance. Another class represents a low-performance group with low solution probabilities for most items. In a special class (Class 2), participants show high solution probabilities only for test items in which the distractor figures have a different shape than the target figure (type II items). Therefore, we assume that a special nonrotation strategy is used in this class. This analytic strategy appears to be only successful with type II items where the distractor figures can be excluded relatively easy by a feature comparison strategy. In contrast, using such a strategy, it is much harder to exclude the distractors of type I items, because here the distractors are mirror images of the target and thus much more similar to the target. For type I items a holistic mental rotation strategy seems to be more efficient. As participants in Class 2 have high solution probabilities only for type II items we suspect that they actively search for problems that do not require mental rotation. Although the solution probabilities for type I items are still above the chance level in Class 2, it is obvious

that members of this nonrotation strategy class have severe problems to find the correct solutions for items that require mental rotation.

It turned out that class membership was relatively stable across both subscales. However, latent transition probabilities for the same class were smaller than 1 for each class, indicating that there is no perfect stability of class membership across subscales. It was found that some participants switch to a higher performance class whereas others are found in a class with worse performance for the second subscale. This indicates that there are participants who benefit from the first test half as a kind of short-term training whereas others may show symptoms of fatigue or loss of motivation. Furthermore, there may be participants who switch from one strategy to another over time. Of interest, most participants who leave the special nonrotation strategy class are found in the class with the worst performance. On the other hand, participants in the nonrotation class have a very low chance of being found in one of the classes with higher performance. We therefore conclude that participants in Class 2 are relatively stable nonrotators, that is, these participants really seem to be unable (or unmotivated) to solve MRT problems using mental rotations.

Furthermore, we studied the relationship between class membership and gender and found that there are large differences in the frequency distribution of males and females concerning the five classes. Males are significantly more often found in Class 3 and Class 5. These groups show the highest MRT performance. In spite of the time limit of 3 min per item set, participants in Classes 3 and 5 attempt a high number of items and find the correct solution. On the other hand, females dominate in Classes 1 and 4. Class 1 is the group with the lowest performance, whereas participants in Class 4 show moderate performance but are quite slow. The fact that females are strongly overrepresented in this "low-speed" class is interesting, because it supports prior findings which showed that on average males rotate stimuli faster than females (Kail, Carter, & Pellegrino, 1979; Lohman, 1986). This also explains the reduced gender effect size for ratio scores, because this scoring method "corrects" for the speededness of a test. There are at least two possible explanations for the gender difference concerning class 4. On the one hand, it might be that many females are simply "slower rotators" than males. That is, participants in Class 4 might also use mental rotation as solution strategy but apply this strategy in a less efficient way. On the other hand, it seems also possible that participants in Class 4 are not slow rotators but use completely different solution strategies (e.g., analytic strategies), which are less efficient so that they need a lot more time. Finally, participants in Class 4 could use a mixture of both mental rotation and analytic strategies. All three interpretations are reasonable and on the basis of our findings, we cannot decide which is the true one. In either case, if they were given more time to complete the test, one would expect participants in Class 4 to show a higher MRT performance. Because there is an especially large number of females in this group, we would further expect an at least slightly reduced gender effect size for the conventional scores if the MRT was administered without time limit.

It is interesting that females were also slightly overrepresented in the Nonrotation Strategy Class 2, although this gender effect was not as large as for the other classes. This is in agreement with the results of other studies which also found that females might have a greater tendency than males to use analytic strategies that do not require mental rotation of stimuli.

The findings of this study have important implications for the application of the MRT in general. First—as already mentioned—the speed of performance is an important component that seems to explain some proportion of the robust sex differences in mental rotation tasks. The results of this study suggest that about one fifth of all females are disadvantaged by the speededness of the MRT. The response pattern of this group (Class 4) indicates that members would be able to solve more MRT problems if they were given more time to complete the test. This is an important aspect for selection processes. When speeded mental rotation tasks are for example used in personnel selection batteries, a large group of females may be disadvantaged.

Second, the discovery of the nonrotation strategy group (Class 2) confirms prior findings that participants use different strategies to solve spatial tasks. The solution strategy applied in Class 2 indicates that not all MRT items measure exclusively mental rotation. Participants in Class 2 are able to solve about one third of the items in each test half without the need to mentally rotate them. Therefore, we do not think that the MRT can be thought of as a "pure" measure of mental rotation ability. That is, MRT scores cannot be interpreted as "degree of mental rotation ability" for all participants. At least for one subgroup (Class 2 in our study) the MRT score has a different meaning, which could for example be denoted as "feature comparison ability." Although this group of participants appears to be relatively small (about 13–17 %), the fact that a large number of items does not require mental rotation is clearly problematic, especially because it can not be ruled out that participants assigned to one of the other classes use the same nonrotation strategy for type II items. This strategy appears to be a rather easy way to exclude distractor figures (much easier than mental rotation; see also discussion in Schultz, 1991). It is well known that high-performance participants are especially flexible in switching between different strategies. This flexibility is one of the strength of those participants. Therefore, it seems very likely that at least high-performance participants use the nonrotation strategy as well. In fact, type II items—at least at the beginning of each test half—are easier for most participants compared to type I items. The solution probabilities for the type II items 3, 4, 13 and 14 (range = .69 to .86 for all participants) are the highest in the whole test.

Unfortunately, we do not know why Vandenberg and Kuse (1978) used different distractor figures in the test construction. Peters, Laeng, et al. (1995) used the Vandenberg and Kuse MRT as a base for their updated MRT version but did not intentionally adopt the two classes of distractor constructions (Michael Peters, personal communication, 2001). We assume that the test authors did not in-

tend to use different items to measure different strategies; this may have happened rather accidentally. One suggestion to enhance the validity of MRT scores would therefore be to replace the distractors of all type II items and mixed item forms by mirror images of the target figure—that means convert all type II items and mixed forms into type I items. Although there might be a number of other solution strategies (other than mental rotation) that we could not discover in our investigation, such a reconstruction of the MRT seems useful to force participants to use a mental rotation strategy. In addition, we recommend replacing the distractors of the difficult Item 17.

From a methodological perspective, our findings show that LCA is a useful statistical tool to analyze speeded tests and cognitive tests in which the application of different solution strategies must be expected because of different item types. Using LCA, subgroups using different solution strategies can be identified. In addition, several LCA programs offer the possibility to specify latent class models as multigroup models and to incorporate covariates in the analysis. This makes it possible to study possible predictors of different strategies (such as gender) directly (without the need to first assign each individual to her or his most likely class). Furthermore, latent transition analysis can be performed to study strategy switching and change in performance over time.

For future research it seems interesting to study the influence of experience and training on the reported latent class structure. It is well known that training and pre-exposure have a large influence on spatial test performance. In our investigation we found that the class assignment was not perfectly stable over both test halves. The results indicate that at least some participants benefit from the first test half and switch to a higher performance class in the second test half. It would be interesting to find out whether the proposed five-class model changes when participants are tested the second time or after a mental rotation training. For example, does the nonrotation class vanish after an extensive mental rotation training?

REFERENCES

Amthauer, R. (1953). *Intelligenz-Struktur-Test (IST)* [*Intelligence Structure Test IST*]. Göttingen, Germany: Hogrefe.

Amthauer, R. (1970). *Intelligenz-Struktur-Test (IST-70)* [*Intelligence Structure Test IST-70*]. Göttingen, Germany: Hogrefe.

Burin, D. I., Delgado, A. R., & Prieto, G. (2000). Solution strategies and gender differences in spatial visualization tasks. *Psicológica, 21*, 275–286.

Casey, M. B., Pezaris, E., & Nuttall, R. L. (1992). Spatial ability as a predictor of math achievement. *Neuropsychologia, 30*, 35–45.

Clogg, C. C. (1995). Latent class models: Recent developments and prospects for the future. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling in the social sciences* (pp. 311–359). New York: Plenum.

Collins, L. M., Fidler, P. L., Wugalter, S. E., & Long, J. L. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research, 28*, 375–389.

Cooper, L. A. (1976). Individual differences in visual comparison processes. *Perception & Psychophysics, 19*, 433–444.

Eid, M., Langeheine, R., & Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis. *Journal of Cross-Cultural Psychology, 34*, 195–210.

Glück, J., Machat, R., Jirasko, M., & Rollett, B. (2001). Training-related changes in solution strategy in a spatial test: An application of item response models. *Learning and Individual Differences, 13*, 1–22.

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

Hausmann, M., Slabbekoorn, D., Van Goozen, S. H. M., Cohen-Kettenis, P. T., & Güntürkün, O. (2000). Sex hormones affect spatial abilities during the menstrual cycle. *Behavioral Neuroscience, 114*, 1245–1250.

Hosenfeld, I., Strauss, B., & Köller, O. (1997). Geschlechtsdifferenzen bei Raumvorstellungsaufgaben—eine Frage der Strategie? [Sex differences in spatial tasks—A question of strategy?] *Zeitschrift für Pädagogische Psychologie, 11*, 85–94.

Jordan, K., Heinze, H.-J., Lutz, K., Kanowski, M., & Jäncke, L. (2001). Cortical activations during the mental rotation of different visual objects. *Neuro Image, 13*, 143–152.

Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review, 92*, 137–172.

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

Köller, O., Rost, J., & Köller, M. (1994). Individuelle Unterschiede beim Lösen von Raumvorstellungsaufgaben aus dem IST- bzw. IST-70-Untertest "Würfelaufgaben". [Individual differences in solving spatial tasks from the IST- and IST-70 subtest "cube comparisons"] *Zeitschrift für Psychologie, 202*, 65–85.

Langeheine, R., Pannekoek, J., & van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research, 24*, 249–264.

Langeheine, R., & Rost, J. (1988). *Latent trait and latent class models*. New York: Plenum.

Lehmann, W., & Jüling, I. (2002). Raumvorstellungsfähigkeit und mathematische Fähigkeiten—unabhängige Konstrukte oder zwei Seiten einer Medaille? [Spatial ability and math ability—independent constructs or two sides of a medal?] *Psychologie in Erziehung und Unterricht, 49*, 31–43.

Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22*, 249–264.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56*, 1479–1498.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767–778.

Lohman, D. F. (1986). The effect of speed-accuracy tradeoff on sex differences in mental rotation. *Perception & Psychophysics, 39*(6), 427–436.

Masters, M. S., & Sanders, B. (1993). Is the gender difference in mental rotation disappearing? *Behavior Genetics, 23*, 337–341.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Moffat, S. D., Hampson, E., & Hatzipantelis, M. (1998). Navigation in a 'virtual' maze: Sex differences and correlation with psychometric measures of spatial ability in humans. *Evolution and Human Behavior, 19*, 73–87.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide. Third edition* [Computer software manual]. Los Angeles, CA: Muthén & Muthén.

Pearson, J. L., & Ferguson, L. R. (1989). Gender differences in patterns of spatial ability, environmental cognitions and math and English achievement in late adolescence. *Adolescence, 94*, 421–431.

Peters, M., Chisholm, P., & Laeng, B. (1995). Do engineering students show sex differences on a spatial ability test, and do such differences relate to performance in academic subjects? *Journal of Engineering Education, 84*, 69–73.

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

Quaiser-Pohl, C., & Lehmann, W. (2000). How can girls' spatial abilities be improved? — The role of experiences and attitudes in different academic subgroups. *International Journal of Psychology, 35*, 353.

Qubeck, W. J. (1997). Mean differences among subcomponents of Vandenberg's Mental Rotation Test. *Perceptual and Motor Skills, 85*, 323–332.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion* [Textbook test theory and test construction]. Bern, Switzerland: Huber.

Schultz, K. (1991). The contribution of solution strategy to spatial performance. *Canadian Journal of Psychology, 45*, 474–491.

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

Stumpf, H. (1993). Performance factors and gender-related differences in spatial ability: Another assessment. *Memory & Cognition, 21*, 828–836.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three dimensional spatial visualisation. *Perceptual and Motor Skills, 60*, 343–350.

van de Pol, F., Langeheine, R., & de Jong, W. (1996). *PANMARK 3. User's manual. Panel analysis using Markov chains – A latent class analysis program* [Computer software manual]. Voorburg, The Netherlands: Statistics Netherlands.

von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research-Online, 2*(2), 29–48. Retrieved February 17, 2005, from http://www.dgps.de/fachgruppen/methoden/mpr-online/issue3/art5/article.html

von Davier, M. (2001a). *Models available in WINMIRA — and some results on person fit. Appendix to the WINMIRA user manual.* Kiel, Germany: Institut für die Pädagogik der Naturwissenschaften.

von Davier, M. (2001b). *WINMIRA user manual* [Computer software manual]. Kiel, Germany: Institut für die Pädagogik der Naturwissenschaften.

Voyer, D. (1996). The relation between mathematical achievement and gender differences in spatial abilities: A suppression effect. *Journal of Educational Psychology, 88*, 563–571.

Voyer, D. (1997). Scoring procedure, performance factors, and magnitude of sex differences in spatial performance. *American Journal of Psychology, 110*, 259–276.

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.

APPENDIX A

Mplus Input for the 5-Class Multigroup Latent Class Model Depicted in Figure 3

```
TITLE:Semirestricted Multigroup LCA for MRT Subscale 1
       5-Class Model Depicted in Figure 3
! Comments can be inserted following exclamation marks

DATA:     FILE = mrt.dat;
! This command defines the data file

VARIABLE:   NAMES = SEX MRT1-MRT24;
              CATEGORICAL = MRT1-MRT12;
              USEVARIABLES = MRT1-MRT12;
              KNOWNCLASS = CSEX (SEX = 1 SEX = 2);
              CLASSES = CSEX (2) C (5);
! These statements define the names and the nature of the
! observed and latent variables used in the analysis
! SEX / CSEX is gender with categories 1 = female and 2 = male.
! MRT1-MRT12 are the binary items used as latent class indicators.
! As this is a multigroup analysis, gender is used as observed
!("KNOWNCLASS") variable CSEX with 2 "classes" (females vs. males).
! C is the (unobserved) latent class variable with 5 latent classes

ANALYSIS:   TYPE = MIXTURE;
              ESTIMATOR = ML;
              STARTS = 500 50;
              STITERATIONS = 50;
! LCA is requested using the TYPE = MIXTURE statement
! Maximum Likelihood (ML) estimation is chosen
! The remaining commands request 500 random sets of starting values
! in the first stage of optimization, 50 sets for the final stage
! and a maximum of 50 initial stage iterations (see also footnote 2)

MODEL:   %OVERALL%
           C#1 ON CSEX#1;
           C#2 ON CSEX#1;
           C#3 ON CSEX#1;
           C#4 ON CSEX#1;
! As this is a multigroup analysis, latent class membership
! has to be regressed on the KNOWNCLASS CSEX (gender)
```

```
! The following class specific statements are necessary
! in order to set the conditional response probabilities equal
! across gender. This is done by giving the same numbers for
! parameters to be held equal (numbers in parentheses).

%CSEX#1.C#1%
[MRT1$1] (1);
[MRT2$1] (2);
[MRT3$1] (3);
[MRT4$1] (4);
[MRT5$1] (5);
[MRT6$1] (6);
[MRT7$1] (7);
[MRT8$1] (8);
[MRT9$1] (9);
[MRT10$1] (10);
[MRT11$1] (11);
[MRT12$1] (12);

%CSEX#1.C#2%
[MRT1$1] (13);
[MRT2$1] (14);
[MRT3$1] (15);
[MRT4$1] (16);
[MRT5$1] (17);
[MRT6$1] (18);
[MRT7$1] (19);
[MRT8$1] (20);
[MRT9$1] (21);
[MRT10$1] (22);
[MRT11$1] (23);
[MRT12$1] (24);

%CSEX#1.C#3%
[MRT1$1] (25);
[MRT2$1] (26);
[MRT3$1] (27);
[MRT4$1] (28);
[MRT5$1] (29);
[MRT6$1] (30);
[MRT7$1] (31);
[MRT8$1] (32);
```

```
[MRT9$1] (33);
[MRT10$1] (34);
[MRT11$1] (35);
[MRT12$1] (36);

%CSEX#1.C#4%
[MRT1$1] (37);
[MRT2$1] (38);
[MRT3$1] (39);
[MRT4$1] (40);
[MRT5$1] (41);
[MRT6$1] (42);
[MRT7$1] (43);
[MRT8$1] (44);
[MRT9$1] (45);
[MRT10$1] (46);
[MRT11$1] (47);
[MRT12$1] (48);

%CSEX#1.C#5%
[MRT1$1] (49);
[MRT2$1] (50);
[MRT3$1] (51);
[MRT4$1] (52);
[MRT5$1] (53);
[MRT6$1] (54);
[MRT7$1] (55);
[MRT8$1] (56);
[MRT9$1] (57);
[MRT10$1] (58);
[MRT11$1] (59);
[MRT12$1] (60);

%CSEX#2.C#1%
[MRT1$1] (1);
[MRT2$1] (2);
[MRT3$1] (3);
[MRT4$1] (4);
[MRT5$1] (5);
[MRT6$1] (6);
[MRT7$1] (7);
[MRT8$1] (8);
```

```
[MRT9$1] (9);
[MRT10$1] (10);
[MRT11$1] (11);
[MRT12$1] (12);

%CSEX#2.C#2%
[MRT1$1] (13);
[MRT2$1] (14);
[MRT3$1] (15);
[MRT4$1] (16);
[MRT5$1] (17);
[MRT6$1] (18);
[MRT7$1] (19);
[MRT8$1] (20);
[MRT9$1] (21);
[MRT10$1] (22);
[MRT11$1] (23);
[MRT12$1] (24);

%CSEX#2.C#3%
[MRT1$1] (25);
[MRT2$1] (26);
[MRT3$1] (27);
[MRT4$1] (28);
[MRT5$1] (29);
[MRT6$1] (30);
[MRT7$1] (31);
[MRT8$1] (32);
[MRT9$1] (33);
[MRT10$1] (34);
[MRT11$1] (35);
[MRT12$1] (36);

%CSEX#2.C#4%
[MRT1$1] (37);
[MRT2$1] (38);
[MRT3$1] (39);
[MRT4$1] (40);
[MRT5$1] (41);
[MRT6$1] (42);
[MRT7$1] (43);
[MRT8$1] (44);
```

```
[MRT9$1] (45);
[MRT10$1] (46);
[MRT11$1] (47);
[MRT12$1] (48);

%CSEX#2.C#5%
[MRT1$1] (49);
[MRT2$1] (50);
[MRT3$1] (51);
[MRT4$1] (52);
[MRT5$1] (53);
[MRT6$1] (54);
[MRT7$1] (55);
[MRT8$1] (56);
[MRT9$1] (57);
[MRT10$1] (58);
[MRT11$1] (59);
[MRT12$1] (60);
```

APPENDIX B

TABLE B1
Conditional Response Probabilities and Standard Errors in the Five-Class Solution (Subscale 1)

	Class 1		*Class 2*		*Class 3*		*Class 4*		*Class 5*	
	Estimate	*SE*	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
Item 1	.38	.03	.35	.06	.86	.02	.85	.03	.91	.02
Item 2	.33	.03	.25	.06	.88	.02	.82	.03	.88	.03
Item 3	.63	.03	.82	.03	.97	.01	.98	.01	.95	.02
Item 4	.29	.03	.76	.04	.96	.01	.99	.02	.98	.01
Item 5	.21	.02	.32	.05	.82	.02	.69	.04	.90	.03
Item 6	.15	.02	.40	.05	.87	.02	.49	.05	.91	.02
Item 7	.16	.02	.84	.03	.93	.01	.16	.05	.96	.02
Item 8	.04	.01	.88	.05	.87	.03	.00	—	.97	.02
Item 9	.01	.00	.07	.02	.17	.02	.00	—	.57	.04
Item 10	.03	.01	.23	.04	.22	.03	.00	—	.73	.04
Item 11	.01	.01	.14	.03	.05	.02	.00	—	.82	.04
Item 12	.02	.01	.19	.04	.07	.02	.00	—	.88	.03

Note. Values are Maximum Likelihood estimates. SE = standard error. Dashes indicate that a standard error could not be computed because the corresponding parameter was estimated and fixed to 0 during estimation.

TABLE B2
Conditional Response Probabilities and Standard Errors in the Five-Class Solution (Subscale 2)

	Class 1		*Class 2*		*Class 3*		*Class 4*		*Class 5*	
	Estimate	*SE*	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
Item 13	.57	.03	.68	.05	.90	.02	.85	.03	.97	.01
Item 14	.42	.04	.54	.05	.84	.02	.76	.04	.94	.02
Item 15	.14	.03	.34	.05	.70	.03	.66	.05	.83	.03
Item 16	.15	.04	.25	.05	.80	.03	.68	.05	.85	.03
Item 17	.15	.03	.21	.03	.37	.03	.29	.04	.45	.03
Item 18	.20	.03	.40	.04	.80	.03	.59	.05	.83	.03
Item 19	.19	.03	.44	.05	.81	.03	.23	.05	.84	.03
Item 20	.14	.03	.38	.06	.77	.04	.08	.04	.87	.02
Item 21	.13	.03	.51	.05	.45	.04	.00	.02	.65	.03
Item 22	.01	.03	.50	.05	.20	.03	.00	—	.83	.03
Item 23	.03	.01	.34	.06	.05	.02	.00	—	.90	.03
Item 24	.04	.02	.36	.05	.00	—	.00	—	.68	.04

Note. Values are Maximum Likelihood estimates. Dashes indicate that a standard error could not be computed because the corresponding parameter was estimated and fixed to 0 during estimation.