TESTING OF SPATIAL ABILITY: CONSTRUCTION AND EVALUATION OF A NEW INSTRUMENT

Petr KVĚTON, Martin JELÍNEK, Dalibor VOBOŘIL

Institute of Psychology, Academy of Sciences of the Czech Republic, p.r.i.
Veveří 97, 602 00 Brno, Czech Republic
E-mail: kveton@psu.cas.cz

Abstract: The aim of the paper is to describe the process of development and evaluation of a newly designed spatial ability test. It consists of two consecutive studies. In Study I (N = 267) we proposed 35 items equally divided into five subsets. The items were designed with respect to theoretically described spatial ability dimensions (spatial perception, orientation, visualization, relations, and mental rotation). Even though a five factor structural model fitted the data reasonably well, on the principle of parsimony we agreed on a unidimensional model. Items with the best parameters (n = 25) were considered as the final version of the test. In Study II (N = 124) we verified that there is no significant impact of the administration media (paper/pencil vs. computer-based). The test-retest stability with a six-week interval was acceptable (r = 0.796), and so was the internal consistency (Cronbach's α = 0.752). We have found a modest correlation (r = 0.470) with the Spatial Reasoning subtest of the Intelligence Structure Test.

Key words: spatial ability, testing, psychometrics

INTRODUCTION

The concept of spatial ability is hard to define. In general, spatial ability enables the individual to deal with problems and tasks, which require estimation, prediction, or assessment of spatial relations between individual objects or figures (Eliot, Smith, 1983). At the turn of the 20th century, researchers started to acknowledge spatial ability as an independent factor separate from general intelligence (Mohler, 2008). The efforts to identify specific factors within the construct of spatial ability can be traced back to the middle of the 20th century, the era of outstanding theoreticians of intelligence such as Thurstone, Guilford, or Zimmerman (Hegarty, Waller, 2005). Since those days, researchers have developed many definitions of spatial ability together with various measures attempting to capture this phenomenon. These authors reached relatively diverse conclusions about the number and the nature of spatial ability components.

McGee (1979) significantly contributed to the clarification of the topic by providing a comprehensive review. Based on the results of available factor analytic studies, he concluded that all of the different factors found by various authors reflected two fundamental dimensions, which he called spatial visualization and spatial orientation. Visualization incorporates the ability to mentally manipulate (rotate, twist and invert) visual stimuli.

Acknowledgements: This study was supported by project nr. P407-11-2397, Czech Science Foundation, and by RVO: 68081740.

DOI: 10.21909/sp.2014.03.663

Orientation involves comprehension of the way elements are arranged within a visual stimulus pattern and dealing with changes in orientation of variably depicted spatial configurations. Despite McGee's thorough analysis, however, the controversies about the structure of spatial ability still remain. Some researchers identify three spatial factors – for example, Lohman (1988) names spatial visualization, spatial relations, and spatial orientation, whereas Linn and Petersen (1985) refer to spatial perception, spatial visualization, and mental rotation. Maier (1994) combines these factors and proposes a five-factor model of spatial ability, comprising Spatial Perception (the ability to correctly determine horizontal or vertical position of an object despite confusing visual information); Spatial Visualization (the ability to create mental images of the inner configurations of spatial objects, or modifications of the configurations); Mental Rotation (the ability to rotate visual images of planar or 3-D objects); Spatial Relations (the ability to comprehend spatial configurations of objects or their parts, and their mutual relations); and Spatial Orientation (the ability to orient oneself in any spatial situation). However, the author himself notes that there are strong interrelations between these factors, and they often cannot be strictly differentiated.

It is obvious that there is no unequivocal consensus about the nature and structure of the spatial ability construct. Generalizability of results from various studies is further complicated by the fact that theoretical considerations mostly depend on empirical studies involving different spatial ability tests constructed by various item principles, on which factor analysis was applied (Hegarty, Waller, 2005). Moreover, there is an ongoing debate concerning analogue versus analytic character of item solving strategies. Spatial ability is considered to be an analogue process in principle, in which mental imagery reflects an actual physical manipulation of objects (Embretson, 2007). Yet, some tasks commonly used in spatial ability testing can be solved not only by employing the target visual-analogue processes, but also by creating propositions (Paivio, 2009). Solving strategies based on propositions are referred to as verbal-analytic strategies (Embretson, 2007). In some cases, limited competence in solving spatial ability tasks by visual-analogue strategies can be compensated by employing verbal-analytic hints. This issue was thoroughly examined in our previous study using data from a spatial ability subtest of a university admission test (Jelínek, Květon, Vobořil, 2013).

In the current study, we proposed item principles, which cover spatial ability elements suggested by Maier (1994), as his comprehensive theoretical approach incorporates most of previously considered spatial ability dimensions. The goal was to design and validate a complex test of spatial ability. The paper consists of two parts. Study I addresses the issue of construct dimensionality and provides evaluation of psychometric properties of the proposed test items. Based on the results, we chose the appropriate items for the final version of the test. Study II was performed to gain information about psychometric properties of the final version of the test. Since there is a demand to administer psychodiagnostic methods using paper/pencil and computer, we created both versions of the test. The relevant literature mentions that in case of performance testing with graphical stimuli the results of the methods can

get influenced by the administration media (Aspillaga, 1996; Květon, Klimusová, 2002). Therefore, we used a complete equivalence design to examine the impact of administration media and to evaluate psychometric properties of the final version of the test.

STUDY I

METHOD

Instrument Development

Based on a review of existing tests of spatial ability, e.g. directories and compendia (Eliot, Smith, 1983; Muchinsky, 2004; Svoboda, 2010; Ekstrom, French, Harman, 1976), as well as our previous experience with designing items for spatial subtest of the admission test at Masaryk University, Brno (Květon et al., 2012; Jelínek, Květon, Vobořil, 2013), we proposed five item prototypes (see Figure 1), one for each of Maier's (1994) dimensions. Prototype 1 was designed to capture the Spatial Relations dimension (identifying relations between patterns on the sides of a cube depicted as a net); Prototype 2 corresponds to the Spatial Visualization dimension (visualization of a cross section of a solid 3-D object, produced by a plane); Prototype 3 is expected to capture the Spatial Perception dimension (estimation of the correct orientation of an object according to gravity – this is supposed to indicate the

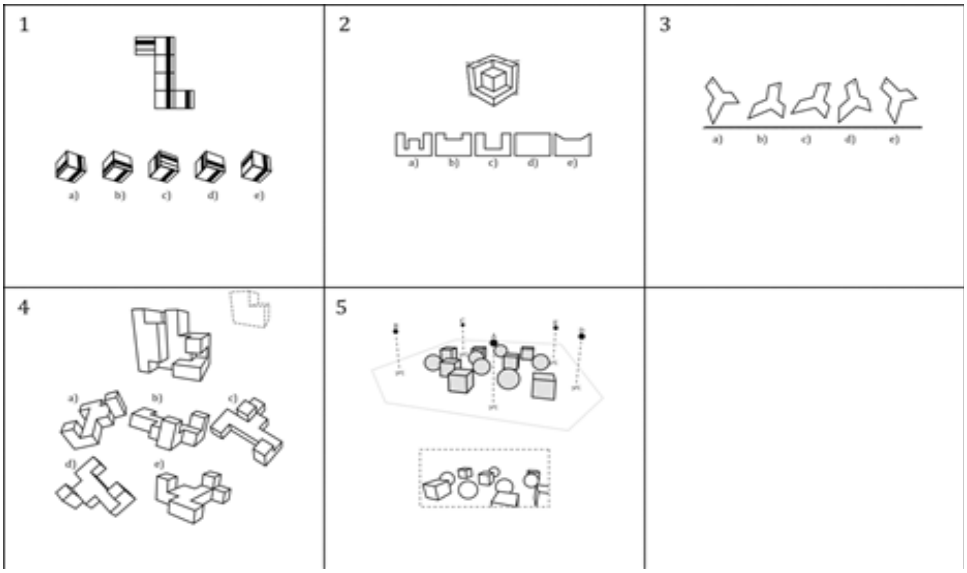

Figure 1. Prototypes of spatial ability items. Item #1 instruction: Decide which option matches the unfolded cube; Item #2: Decide which option corresponds to the cut through the object as indicated in the depiction; Item #3: Decide which freely hanging flat object is in the correct position (with respect to the horizontal ground); Item #4: Decide which part will complete the object (to the gray template shape); Item #5: Decide from which point the depicted configuration can be observed.

respondent's sense of verticality); Prototype 4 was designed to test Mental Rotation abilities (mental rotation of objects to determine which of them completes a depicted figure to obtain the indicated target structure); and, finally, Prototype 5 represents the Spatial Orientation dimension (determining the correct position from which a target scene can be observed – an indicator of the respondent's ability to orient himself/herself by a pattern of spatial objects).

When designing the prototypes, we were inspired by various tests: Surface Development Test (Ekstrom, French, Harman, 1976), Cube Comparison Test (Ekstrom, French, Harman, 1976) for Spatial Relations tasks; Water Level Task (Linn, Petersen, 1985), Rod and Frame Test (Witkin et al., 1977) for Spatial Perception tasks; Mental Cutting Test (CEEB, 1939), Object Aperture Test (Eliot, Smith, 1983), Mental Cutting Test "Schnitte" (Quaiser-Pohl, 2003) for Spatial Visualization tasks; Vandenberg's Test of Three-Dimensional Spatial Visualization, Guay's Visualization of Rotations (Eliot, Smith, 1983) for Mental Rotation tasks; Barratt-Fruchter's Chair Window Test (Eliot, Smith, 1983) for Spatial Orientation tasks.

In order to assess the content validity of these item prototypes, the items, together with verbal descriptions of the dimensions, were presented to a group of ten experts in the respective field (two from the Institute of Psychology, Academy of Sciences of the Czech Republic, five from Masaryk University, and three from Charles University in Prague). To avoid a situation in which dimensions would be assigned to prototypes on the basis of progressive elimination, and bearing in mind the possibility of within-item multidimensionality, we decided to instruct the experts to assign a maximum of two dimensions per item prototype. The summary of the expert assessment is shown in Table 1.

We set the threshold at which an item was considered significantly loaded by a dimension to agreement of at least four experts. The mode of choices for each prototype corresponded to our initial expectations. However, two of the prototypes were also quite frequently assigned to a different dimension (Prototype 1 and Prototype 4, both to Spatial Visualization). Based on this evidence, we modified our assumption and suggested that Prototype 1 covers the ability of Spatial

Table 1. Summary of expert evaluation of item prototypes

	Prototype 1	Prototype 2	Prototype 3	Prototype 4	Prototype 5
Spatial Perception	0	0	10	0	0
Spatial Visualization	6	9	0	6	2
Mental Rotation	1	0	2	10	2
Spatial Relations	8	2	0	1	2
Spatial Orientation	0	0	1	0	10

Visualization and grasping Spatial Relations; Prototype 2 corresponds to Spatial Visualization; Prototype 3 to Spatial Perception; Prototype 4 to both Mental Rotation and Spatial Visualization; and Prototype 5 to Spatial Orientation. From each prototype item we derived six more items, which were varied in terms of difficulty. This way we obtained an item pool comprising 35 items, with seven items in each subset.

Instrument Description

The web-based test consisted of 35 items arranged into five subsets. The format of all items was multiple choice (five options) with one single correct answer. Each item was presented on a separate page, which also displayed a test progress indicator (item number/35). The test allowed the respondents to move freely back and forward and change their selected answers if they wanted to. Before moving to the ability test itself, the respondents were asked to provide the necessary personal data (sex, age, education). The test ended with a voluntary comment box. The entire web application was available in two language versions – Czech and English.

Sample and Procedure

Data[1] for Study I were collected in March 2013 using a web-based form of administration. The invitation to participate in the study was advertised on the official webpage of the authors' home institution and also their personal Facebook pages. The call for participation was supplemented with a request for further spreading of the call. This way we collected a database of 683 unique records. To reduce biases potentially arising from online data collection, we performed several steps before the analysis. In the first step, we excluded all respondents who stopped working before reaching the end of the test and also those who did not provide information about their age and/or gender. This left us with a sample of 294 subjects. In the second step, we excluded additional 17 respondents who omitted more than 50% of items (to eliminate those who might have only taken the test out of curiosity and were not sufficiently motivated to make serious attempts at solving the tasks). After that, we excluded one respondent who admitted to random responding (in the open feedback field at the end of the test) and also respondents with unrealistic test-taking times (one respondent who took less than 2 minutes and three respondents who took more than 3 hours). Finally, we excluded respondents identified as extreme cases based on the test-taking time (more than 3 times the interquartile range from upper or lower quartile). The eventual sample size was thus reduced to 267 subjects. The median of test-taking time was 30 min and 55 s (mean = 2096.5 s, SD = 1035.0 s, min = 450 s, max = 5806 s). A detailed description of the sample is provided in Table 2.

[1] All participants were informed about the nature of the research and knowledgeably and voluntarily decided to participate in our study. The participants were instructed that completion of the research questionnaire expresses their willingness to participate in the study. All data were analyzed and presented anonymously. The research project and data collection procedure was approved by the Institutional Board of the Institute of Psychology, Academy of Sciences of the Czech Republic.

Table 2. Sample characteristics

Sex	Males	Females			
	46.8%	53.2%			
Preferred language†	Czech	English			
	263	4			
Education*	Basic	High School	University		
Total (male/female)	9.0% (7.6%/10.2%)	45.7% (49.6%/42.3%)	45.3% (42.9%/47.4%)		
Age (y.)	Minimum	Maximum	Mean	Median	SD
	13	75	32.07	29	11.00

Note: * Expressed in valid percent, 11 respondents did not state their education
† Based on respondent's choice of the test's language version

RESULTS

Dimensionality of the Construct

In the first stage of data analysis, we focused on the structure of the spatial ability construct. The raw data collected through the web-based application were converted into dichotomous variables (right/wrong answer). Hence, the confirmatory factor analysis was based on tetrachoric correlation matrix, computed on dichotomously coded vectors in the R software (R Core Team, 2012) using the PSYCH package (Revelle, 2013). Based on the theoretical background and expert assessment we proposed an initial model with five factors, as depicted in Figure 2.

The model as a whole was found to be marginally acceptable (based on fit indices showed in Table 3). However, the inspection of regression coefficients revealed that, contrary to our expectations, the relations of the Visualization factor to items i23 to i28 were considerably weak (in fact, these regression weights were the lowest and close to zero). Therefore, we proposed a modified five fac-

Table 3. Fit indices for three proposed models of the spatial ability construct

	Root mean square of residual (RMR)	Normed fit index (NFI)	Parsimonious normed fit index (PNFI)	Goodness of fit index (GFI)	Parsimonious goodness of fit index (PGFI)
Initial 5-factor model	0.018	0.917	0.829	0.938	0.801
Modified 5-factor model	0.018	0.916	0.838	0.937	0.809
Hierarchical model	0.019	0.914	0.852	0.935	0.824
Unidimensional model	0.020	0.898	0.845	0.924	0.821

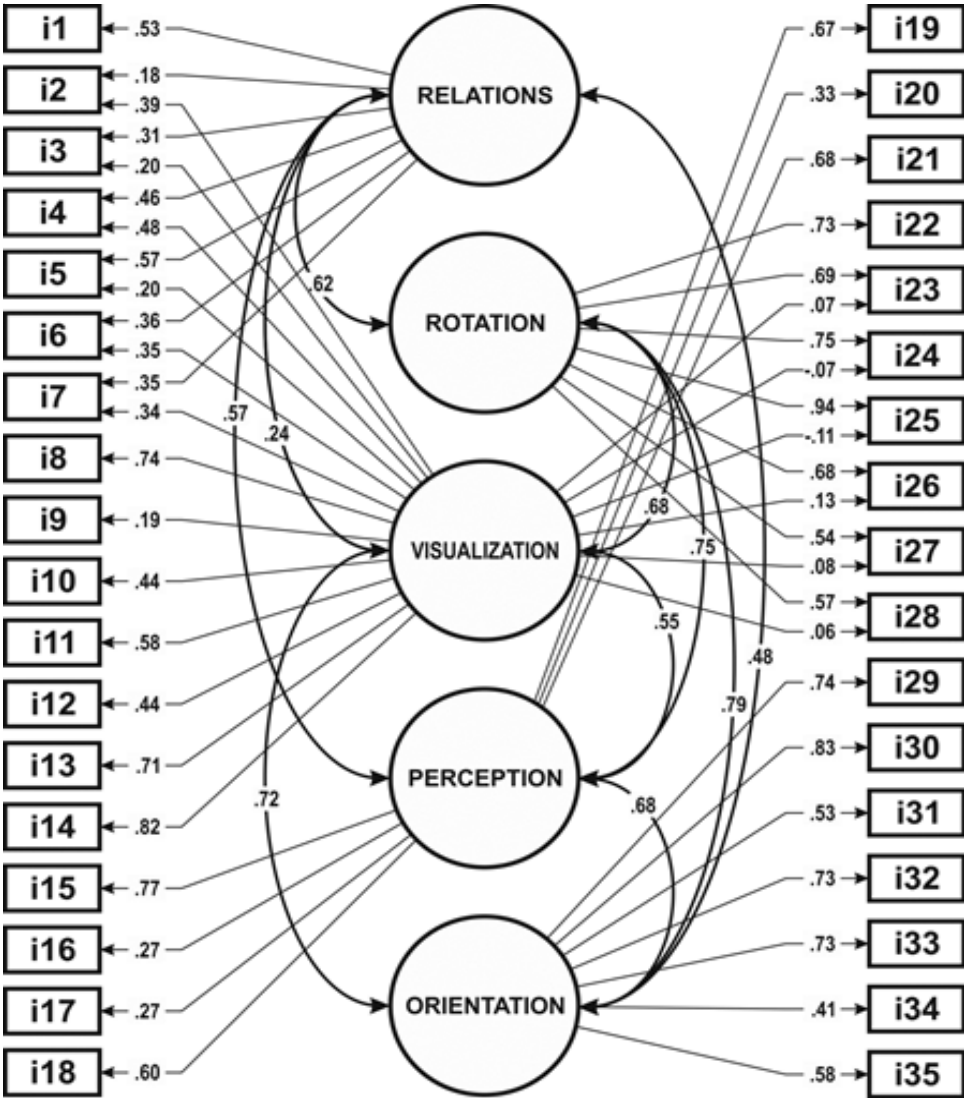

Figure 2. Five factor model of the spatial ability construct. CFA was performed in AMOS 20.0.0 with unweighted least squares estimation. Numerical values represent standardized regression coefficients and correlations. Regression weights between Visualization and i1 and Visualization and i22 were removed to make the model identifiable (Hessen, Dolan, Wicherts, 2006). Error components were omitted from the diagram for the sake of clarity.

tor model, where the relations between Visualization and items i23 to i28 were omitted (see Figure 3). As indicated by the RMR, NFI, and GFI indices, the modified model is comparable to the initial model. The variants of the NFI and GFI indices, which take into account the model's parsimony slightly favors the modified model.

With regard to existing interrelationships between factors and theoretical assumptions about the mutual intertwining of the spatial ability components, we proposed hierarchical model with five specific factors and one general factor (see Figure 4). This model showed comparable fit as previous two models based on GFI and NFI indices and proved better acceptance based on parsimonious indices PNFI and PGFI. In the final step we proposed also unidimensional model (see Figure 5), which showed slightly weaker model fit acceptance based on the values of NFI and GFI, but their parsimonious variants were comparable to the hierarchical model.

With respect to the results of the model building process in the CFA framework we decided to consider both five-dimensional and unidimensional scoring of the test in the following psychometric analysis.

Item Analysis

The scores were calculated as a simple sum of correct answers. Descriptive statistics are summarized in Table 4. The distribution of raw scores did not show notable departure from normal distribution both in case of individual subsets and overall score. Reliability of the total score is sufficiently high, but individual subsets do not reach acceptable level of internal consistency.

The structural analyses did not indicate (by means of regression coefficients in all models) wrongly functioning items to be present in the test. But when looking at the item characteristics summarized in Table 5, it is evident that several subsets include items with lower values of item-total or item-subset total correlation. Taking the value of 0.3 as a cut-off, the several items are below this limit. When considering item-subset total correlation, items 1, 2, 3 from subtest 1, items 9, 12 from subtest 2, items 16, 17 from subtest 3, and item 34 from subtest 5 are below the 0.3 limit. When considering the item-total correlation, items 1, 2, 3 from subtest 1, items 9, 10, 12 from subtest 2, items 16, 17, 20 from subtest 3, and item 34 from subtest 5 are below the limit. To assure content balance, we decided to keep the same number of items in each block and therefore chose to remove two items from each block. This decision was further justified by the fact that the expected reliability of a shortened, 25-item test, calculated using the Spearman-Brown prophecy formula, was 0.822, which is still considered acceptable (the reliability of a 20-item long test falls below the 0.8 level, to 0.787).

From the first block, we excluded items 1 and 3. From the second block we omitted items 9 and 12, and items 16 and 17 from the third block. In block 4 we identified two pairs of items of similar difficulty – the first pair consisted of items 22 ($p = 0.78$) and 23 ($p = 0.79$), the second pair consisted of items 24 ($p = 0.67$) and 26 ($p = 0.66$). To maintain the spread of difficulties across the subset, we decided to keep one item from each pair (namely items 22 and 26). For the same reason, we removed item 30 from block 5, which was paired by difficulty ($p = 0.86$) with item 29 ($p = 0.86$). The second item removed from the last block was item 34, selected due to its weaker characteristics in comparison to the remaining 5 items in the block.

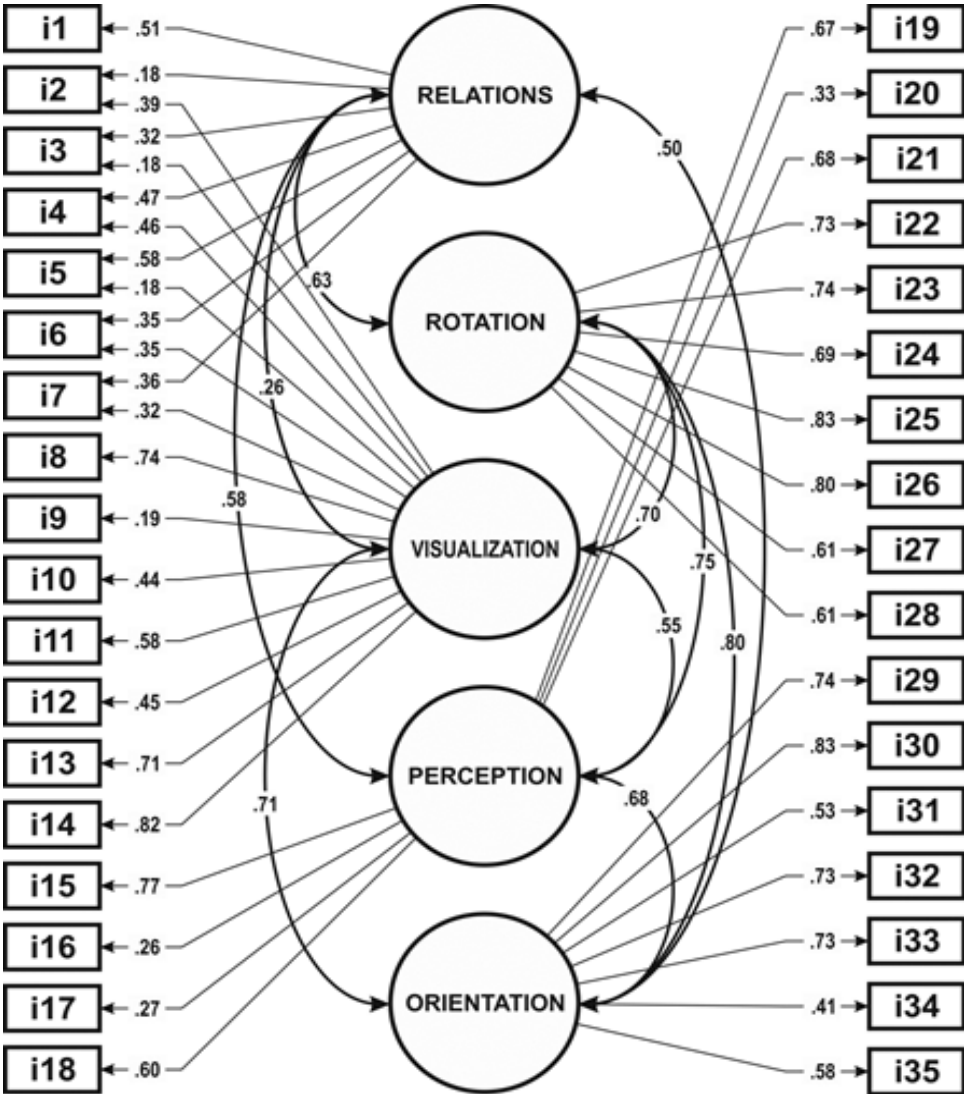

Figure 3. Modified five factor model of the spatial ability construct. CFA was performed in AMOS 20.0.0 with unweighted least squares estimation. Numerical values represent standardized regression coefficients and correlations. Regression weight between Visualization and i1 was removed to make the model identifiable (Hessen, Dolan, Wicherts, 2006). Error components were omitted from the diagram for the sake of clarity.

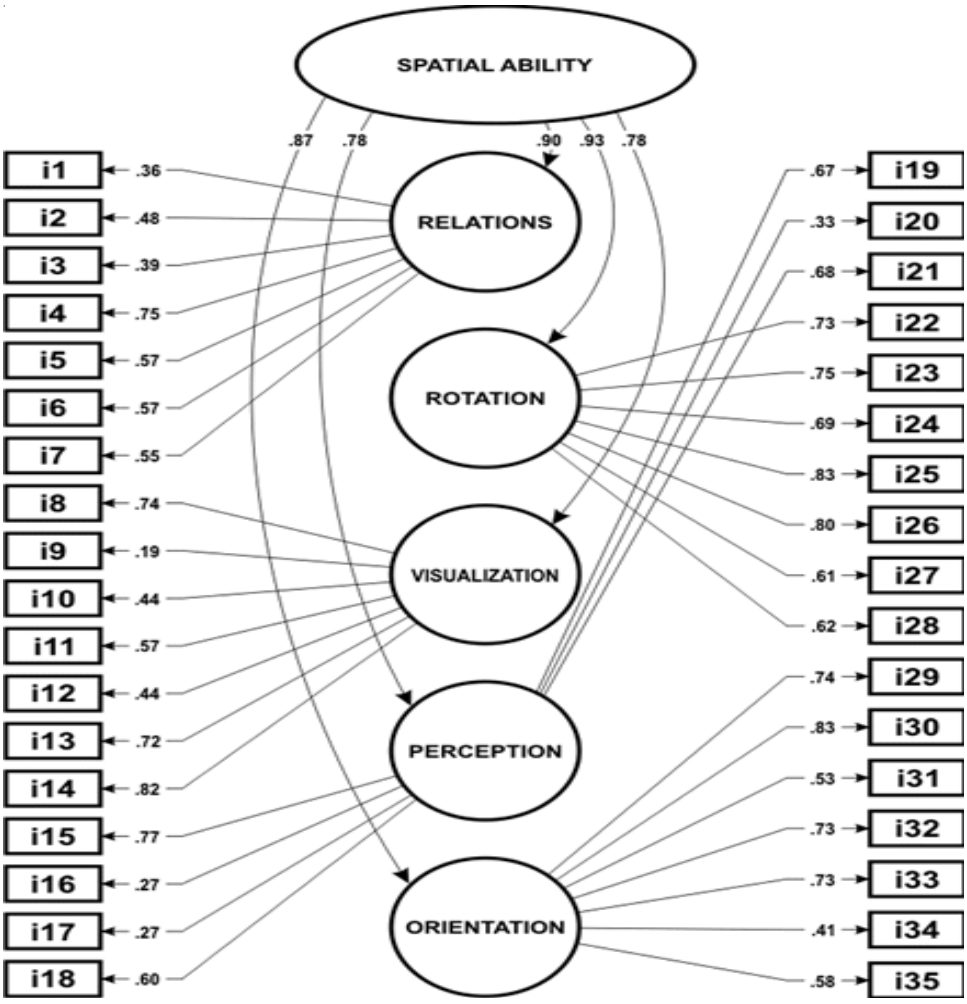

Figure 4. Hierarchical model of the spatial ability construct. CFA was performed in AMOS 20.0.0 with unweighted least squares estimation. Numerical values represent standardized regression coefficients and correlations. Regression weight between Visualization and i1 was removed to make the model identifiable (Hessen, Dolan, Wicherts, 2006). Error components were omitted from the diagram for the sake of clarity.

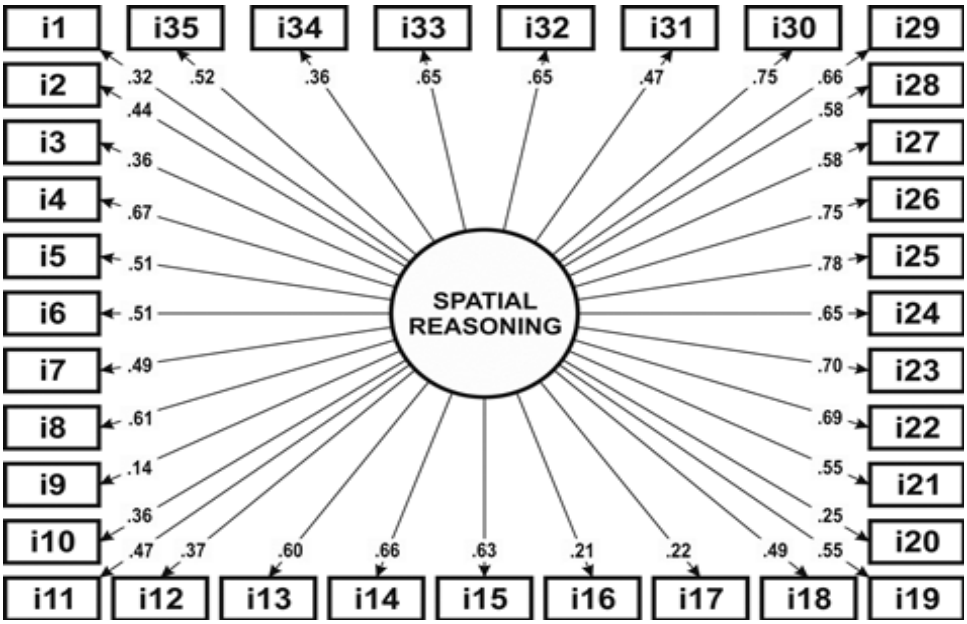

Figure 5. Unidimensional model of the spatial ability construct. CFA was performed in AMOS 20.0.0 with unweighted least squares estimation. Numerical values represent standardized regression coefficients. Error components were omitted from the diagram for the sake of clarity.

Table 4. Descriptive characteristics of the test and its subsets

	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Total
Min	0	0	0	0	0	4
Max	7	7	7	7	7	34
Mean (SD)	4.75 (1.66)	3.37 (1.86)	4.02 (1.76)	4.37 (2.10)	4.18 (1.81)	20.68 (6.87)
Skewness	-0.54	0.24	-0.37	-0.47	-0.26	-0.23
Kurtosis	-0.47	-0.80	-0.48	-0.86	-0.55	-0.78
Cronbach's α	0.571	0.622	0.577	0.763	0.671	0.866

Table 5. Parameters of the proposed items

Item	p	Item-subset total correlation	Item total correlation
1	0.92	0.24	0.17
2	0.70	0.26	0.30
3	0.66	0.15	0.27
4	0.72	0.44	0.49
5	0.58	0.37	0.37
6	0.70	0.32	0.38
7	0.46	0.31	0.36
8	0.71	0.37	0.45
9	0.58	0.20	0.10
10	0.57	0.33	0.28
11	0.45	0.34	0.34
12	0.49	0.27	0.27
13	0.25	0.33	0.39
14	0.32	0.54	0.47
15	0.79	0.33	0.44
16	0.43	0.21	0.15
17	0.44	0.17	0.17
18	0.58	0.34	0.37
19	0.79	0.35	0.38
20	0.47	0.35	0.19
21	0.52	0.36	0.43
22	0.78	0.44	0.47
23	0.79	0.47	0.47
24	0.67	0.51	0.47
25	0.59	0.61	0.58
26	0.66	0.57	0.55
27	0.48	0.38	0.43
28	0.40	0.42	0.41
29	0.86	0.38	0.41
30	0.86	0.42	0.47
31	0.45	0.38	0.35
32	0.33	0.44	0.46
33	0.61	0.43	0.48
34	0.71	0.29	0.26
35	0.36	0.36	0.38

Note: p – CTT difficulty estimate

Table 6. Description of the four groups in complete equivalence design

Group	Measurement 1	Measurement 2	N (females)
A	paper/pencil	paper/pencil	32 (24)
B	paper/pencil	PC	32 (23)
C	PC	paper/pencil	30 (22)
D	PC	PC	30 (22)

The above procedure yielded a 25-item long test of spatial ability, further referred to as the Spatial Reasoning Test (SRT). This test can be considered as a measure of different aspects of spatial ability or as a measure of general factor of spatial ability. With respect to the results of psychometric analysis, we recommend to build test interpretations on the basis of total score. The following Study II employs a complete equivalence design (Bartram, 1994) in order to verify psychometric characteristics of the test in a controlled setting and to examine the effect of administration media.

STUDY II

METHOD

Sample and Procedure

The research sample[2] in Study II consisted of 124 students (91 females; mean age = 22.04 y., SD = 3.66 y.) enrolled on Psychology at Masaryk University in Brno. Data was collected in retest design with a 6-week delay. The first data collection took place in April 2013. We created four groups with comparable gender distributions, who completed the SRT in different administration forms (paper/pencil vs. computer-based) in first and second wave. The administration was performed in supervised group sessions. All computer stations were equipped with standard 20" LCD monitors set to a 1920x1080 native screen resolution. Description of the individual groups is provided in Table 6.

As a part of the first measurement, all of the respondents also completed the spatial reasoning subtest SP from Amthauer's Intelligence Structure Test IST-70 in a paper/pencil form. The SP subtest was administered immediately after the SRT test.

Instruments

SRT consists of 25 items described in Study I of this paper. The computer version was of the same design as the one used for the web-based data collection in Study I. The paper/pencil version was printed with each item on a separate page. Answers were marked on a separate answer sheet. The total administration time was 30 minutes. Because our intention was to position the test closer to the power tests on the power-speed continuum, we derived the administration

[2] All participants were informed about the nature of the research and knowledgeably and voluntarily decided to participate in our study. The participants were instructed that completion of the research questionnaire expresses their willingness to participate in the study. All data were analyzed and presented anonymously. The research project and data collection procedure was approved by the Institutional Board of the Institute of Psychology, Academy of Sciences of the Czech Republic.

time from Study I data, where there was no time limit for test completion. An average item solving time was calculated from the upper quartile of the actual test completion times (2611/35 = 74.6 s). This figure was then multiplied by the number of items in the final version (74.6*25 = 1865 s), and the result was rounded to 30 minutes.

The SP subtest from IST-70 (Amthauer, 1973) originally consists of 20 items. For the purpose of this study, we used a shortened version of the subtest (odd items only). The time limit was reduced with respect to the number of items (6 minutes).

RESULTS

The Effect of the Order and Form of Administration

Repeated measures GLM was used to determine the overall effect of the order and form of administration on the test scores, supplemented by test of between-subject effects from multivariate GLM to identify the effect of the form of administration on the test scores for the first and the second measurement separately. Descriptive statistics for the individual groups and measurement sessions are summarized in Table 7.

The test of between-subject effects revealed no significant differences either in the first ($F(3,120) = 0.478$) or the second ($F(3,120) = 0.607$) measurement. There was an overall significant difference, however, between the two subsequent measurements (Wilks' $\lambda = 0.772$; $F(1,120) = 35.467$; $p < 0.01$), with the total SRT score being higher in the second measurement (partial $\eta^2 = 0.228$). No significant effect of interaction between time and group membership was found (Wilks' $\lambda = 0.990$; $F(1,120) = 0.395$). As there were no significant differences between the individual groups, the form of administration was considered irrelevant in all subsequent analyses.

Psychometric Characteristics of the Final Version of the Test

The stability of the SRT test was sufficient ($r = 0.796$). The internal consistency was acceptable (Cronbach's $\alpha = 0.752$). The criterion validity of the test was assessed by examining its relationship with the SP subtest score from Amthauer's Intelligence Structure

Table 7. Descriptive statistics of SRT scores for the individual groups and measurement sessions

	Group	Mean (SD)
Measurement 1	paper/pencil	12.85 (4.66)
	paper/pencil	14.06 (4.62)
	PC	13.77 (4.23)
	PC	13.90 (4.29)
	total	13.64 (4.43)
Measurement 2	paper/pencil	14.41 (5.36)
	PC	15.91 (4.88)
	paper/pencil	15.63 (4.28)
	PC	15.03 (4.66)
	total	15.24 (4.80)

Test. Correlation between SRT and SP was 0.470 (adjusted correlation coefficient for reliability was 0.674 when internal consistencies of both instruments were taken into account – Cronbach's $\alpha_{srt} = 0.752$; Cronbach's $\alpha_{sp} = 0.647$). Table 8 provides item statistics for the SRT test.

Table 9 shows orientation guidelines for a norm-referenced interpretation for the population of university students. Median for the total score was 13; the lower and upper quartiles were 11 and 17, respectively. We have found differences in the overall ($t = -4.540$, $df = 122$, $p < 0.01$, Cohen's $d = 0.93$) as well as in individual subsets ($t_{subset\ 1} = -1.212$, $p = 0.228$; $t_{subset\ 2} = -3.329$, $p < 0.01$; $t_{subset\ 3} = -1.915$, $p = 0.058$; $t_{subset\ 4} = -4.720$, $p < 0.01$; $t_{subset\ 5} = -3.085$, $p < 0.01$) performance of males and females. Because of this finding we provide separate results for males and females.

Table 8. Item parameters of the SRT test

Item	p	Item-total corr*
2	0.50	0.40
4	0.75	0.29
5	0.55	0.24
6	0.62	0.17
7	0.52	0.40
8	0.68	0.45
10	0.49	0.25
11	0.47	0.33
13	0.30	0.36
14	0.38	0.50
15	0.76	0.37
18	0.53	0.33
19	0.85	0.26
20	0.32	0.07
21	0.60	0.16
22	0.79	0.27
25	0.55	0.24
26	0.70	0.39
27	0.52	0.40
28	0.41	0.15
29	0.91	0.13
31	0.35	0.24
32	0.26	0.25
33	0.55	0.37
35	0.29	0.40

Note: p – difficulty estimate; * corrected

Table 9. Raw scores of SRT with cumulative percentages

Raw score	Cumulative %	Cumulative % (f)	Cumulative % (m)
5	1.6	2.2	
6	4.8	6.6	
7	7.3	9.9	
8	11.3	14.3	3.0
9	17.7	22.0	6.1
10	22.6	26.4	12.1
11	37.1	44.0	18.2
12	47.6	54.9	27.3
13	54.8	63.7	30.3
14	59.7	69.2	33.3
15	67.7	79.1	36.4
16	72.6	82.4	45.5
17	77.4	86.8	51.5
18	83.1	90.1	63.6
19	86.3	93.4	66.7
20	88.7	94.5	72.7
21	96.0	97.8	90.9
22	100.0	100.0	100.0
Mean (SD)	13.64 (4.43)	12.63 (4.03)	16.42 (4.36)

Note: f – females; m – males

DISCUSSION

The present study describes the development and psychometric evaluation of a new spatial reasoning test SRT. Theoretical background of the study was provided by Maier's theory of spatial ability. The theory, which distinguishes five components of spatial ability, is based on the major theories of intelligence, as well as on empirical evidence from numerous studies and meta-analyses (Maier, 1994). The item pool of the test was designed to cover all of the components – Spatial Orientation, Spatial Visualization, Spatial Perception, Mental Rotation, and Spatial Relations.

Study I was focused on the structure of spatial ability, as there is considerable debate about the nature and number of dimensions within this construct (Hegarty, Waller, 2005; McGee, 1979; Mohler, 2008). When designing the items according to Maier's theory, individual subsets of items were created to uniquely represent each dimension, as we expected only between-item multidimensionality. Five item prototypes were evaluated by a panel of ten experts. Based on their evaluation, we took into account potential within-item multidimensionality and proposed the first and the subsequently modified model comprising five latent dimensions. In the modified model, items from the first subset were

loaded not only by the Spatial Relations factor, but also by the Spatial Visualization factor. We can hypothesize that the item principle can be solved either by identifying relations between patterns on the sides of the cube, or by visualization of the object as a whole. These two strategies can possibly also work complementary to each other. With respect to mutual relations between dimensions, we also tested a hierarchical model with one general factor and five components and a unidimensional model. These two models seemed to be most promising. We might speculate about the existence of one general factor of spatial ability with several specialized components. Follow-up psychometric analysis was performed both for multidimensional and unidimensional scoring. With respect to lower levels of internal consistency in case of the five dimensions, we suggest to use the test as unidimensional measure of spatial ability.

When considering if all suggested items should be included in the final test, we primarily took into account item characteristics. This way we identified poorly functioning items in subset 1, 2, and 3. These items showed lower levels of item discrimination in comparison with other items in the respective subset. To maintain content balancing, we decided to retain an equal number of items in each subset to maintain content balance. Content balancing is a common requirement (Leung, Chang, Hau, 2003), even for tests, which are considered unidimensional and provide a single score. The topic is often discussed in the context of computerized adaptive testing, where dimensionality is a key issue (Luecht, 1996; Flaugher, 2000; Jelínek, Květon, Vobořil, 2011). Study I was based on data obtained through web-based application. The online data collection is attended by loss of control over the testing situation. Even though we employed several procedures (i.e. , amount of omitted items; control of time donation per item) to identify and exclude potentially biased data, still it partly limits the interpretations of results.

In Study II, we administered the 25-item test in two sessions with alternating conditions of administration to achieve complete equivalence design. Although some researchers suggest that there might be some effect of media in case of performance tests with graphical stimuli (Federico, 1991; French, Beaumont, 1990; Květon et al., 2007), we found no effect of administration media (paper/pencil vs. computer-based) in the present study. In our opinion, it might be the case that modern computer display technology provides higher image quality, which is no longer limiting for perception in comparison with paper print. Also, negative influence of other factors such as computer anxiety or lack of computer experience on performance in computerized tests was well documented in earlier studies (Mahar, Henderson, Deane, 1997; Heinssen, Glass, Knight, 1997). However, with respect to the composition of our sample, we did not expect these factors to interfere with the computerized test results. We might assume that these issues are diminishing with the progressing penetration of computers into everyday life. Neither was there any interaction effect between administration media and the order of administration: In all groups, we found similar improvement in performance, which can be explained by the learning effect.

Criterion validity of SRT test was evaluated by examining the relationship with the Spatial Reasoning subtest from the standardized Intelligence Structure Test IST-70. We

found a moderate correlation between the two tests (r = 0.470), which partly supports the criterion validity of our test, especially when considering the low level of reliability of the shortened version of the SP (Cronbach's α = 0.647) and the fact that the SP consists of a single task principle, which does not fully match any of our five item prototypes. We found gender difference in performance in the test. The difference varied when looking at individual subsets. This finding is in accordance with relevant literature, which reports different effect sizes for different task principles (Voyer, Voyer, Bryden, 1995). Several studies consistently report gender differences in mental rotation ability (Masters, Sanders, 1993; Debelak, Gittler, Arendasy, 2014), which in our case was the biggest of all subsets, as expected (Linn, Petersen, 1985). Due to the gender differences found, we provided orientation guidelines for interpretation separately for both genders. However, these guidelines should be treated carefully because the research sample consisted of students from only one field of study with lower share of men.

CONCLUSIONS

Spatial Reasoning Test SRT is a measurement tool assessing the ability to comprehend spatial relations and mentally manipulate spatial objects. It can be equivalently used in either a paper/pencil or a computerized form of administration. Although the underlying structure of the test can be considered unidimensional, the test is nevertheless content-balanced to cover the basic components of spatial ability.

Received March 7, 2014

REFERENCES

AMTHAUER, R., 1973, *Test štruktúry inteligencie T-S-I* [Intelligence structure test I-S-T]. Bratislava: Psychodiagnostické a didaktické testy.

ASPILLAGA, M., 1996, Perceptual foundations in the design of visual displays. *Computers in Human Behavior*, 12, 587-600.

BARTRAM, D., 1994, Computer-based assessment. In: C.L. Cooper (Ed.), *International review of industrial and organizational psychology* (pp. 31-69). London: Wiley.

CEEB, 1939, *Special aptitude test in spatial relations*. USA: College Entrance Examination Board.

DEBELAK, R., GITTLER, G., ARENDASY, M., 2014, On gender differences in mental rotation processing speed. *Learning and Individual Differences*, 29, 8-17.

EKSTROM, R., FRENCH, J., HARMAN, H., 1976, *Manual for kit of factor referenced cognitive tests*. Princeton, NJ: Educational Testing Services.

ELIOT, J., SMITH, I.M., 1983, *An international directory of spatial tests*. Highlands, NJ: NFER-Nelson.

EMBRETSON, S.E., 2007, Mixed Rasch models for measurement in cognitive psychology. In: M. von Davier, C.H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York: Springer.

FEDERICO, P.A., 1991, Measuring recognition performance using computer-based and paper-based methods. *Behavioral Research Methods, Instruments, Computers*, 23, 341-347.

FLAUGHER, R., 2000, Item pools. In: H. Wainer, N.J. Dorans, D. Eignor, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg, D. Thissen (Eds.), *Computerized adaptive testing: A Primer (2nd edition)*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

FRENCH, C., BEAUMONT, J.G., 1990, A clinical study of the automated assessment of intelligence by the Mill Hill Vocabulary test and the Standard Progressive Matrices test. *Journal of Clinical Psychology*, 46, 129-140.

HEGARTY, M., WALLER, D.A., 2005, Individual differences in spatial abilities. In: P. Shah, A. Miyake (Eds.), *The Cambridge handbook of visual spatial thinking*. Cambridge: Cambridge University Press.

HEINSSEN, R.K. Jr., GLASS, C.R., KNIGHT, L.A., 1997, Assessing computer anxiety: Develop-

ment and validation of the Computer Anxiety Rating Scale. *Computers in Human Behavior*, 3, 49-59.

HESSEN, D.J., DOLAN, C.V., WICHERTS, J.M., 2006, The multigroup common factor model with minimal uniqueness constraints and the power to detect uniform bias. *Applied Psychological Measurement*, 30, 233-246.

JELÍNEK, M., KVĚTON, P., VOBOŘIL, D., 2011, *Testování v psychologii: Teorie odpovědi na položku a počítačové adaptivní testování* [Testing in psychology: Item response theory and computerized adaptive testing]. Praha: Grada publishing, a.s.

JELÍNEK, M., KVĚTON, P., VOBOŘIL, D., 2013, Skryté aspekty v testování prostorové představivosti: Identifikace uplatňovaných stylů řešení položek [Hidden aspects in spatial ability testing: Identification of respondents' item solving strategies]. *Československá Psychologie*, 57, 297-306.

KVĚTON, P., JELÍNEK, M., VOBOŘIL, D., KLIMUSOVÁ, H., 2012, Rozbor volby odpověďových kategorií v testu prostorové představivosti s využitím teorie odpovědi na položku [Analysis of response categories preference in spatial reasoning test using IRT Nominal Categories Model]. *Československá Psychologie*, 56, 31-40.

KVĚTON, P., JELÍNEK, M., VOBOŘIL, D., KLIMUSOVÁ, H., 2007, Computer-based tests: The impact of test design and problem of equivalency. *Computers in Human Behavior*, 23, 32-51.

KVĚTON, P., KLIMUSOVÁ, H., 2002, Metodologické aspekty počítačové administrace psychodiagnostických metod [Methodological issues of computerized administration of psychodiagnostic methods]. *Československá Psychologie*, 46, 251-264.

LEUNG, C.K., CHANG, H.H., HAU, K.T., 2003, Computerized adaptive testing: A comparison of three content balancing methods. *Journal of Technology, Learning, and Assessment*, 5, 2-15.

LINN, M.C., PETERSEN, A.C., 1985, Emergence and characterization of gender differences in spatial abilities: A meta-analysis. *Child Development*, 56, 1479-1498.

LOHMAN, D.F., 1988, Spatial abilities as traits processes, and knowledge. In: R.J. Sternberg (Ed.), *Advances in the psychology of human intelligence*. Hillsdale, NJ: Erlbaum.

LUECHT, R.M., 1996, Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, 20, 389-404.

MAHAR, D., HENDERSON, R., DEANE, F., 1997, The effects of computer anxiety, state anxiety, and computer experience on users' performance of computer based tasks. *Personal and Individual Differences*, 22, 683-692.

MAIER, P.H., 1994, *Räumliches Vortellungsvermögen* [Spatial imagination]. Frankfurt am Main: Peter Lang GmbH.

MASTERS, M.S., SANDERS, B., 1993, Is the gender differences in mental rotation disappearing? *Behavior Genetics*, 23, 337-341.

MCGEE, M.G., 1979, Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin*, 86, 889-918.

MOHLER, J.L., 2008, A review of spatial ability research. *Engineering Design Graphics Journal*, 72, 19-30.

MUCHINSKY, P.M., 2004, Mechanical aptitude and spatial ability testing. In: M. Hersen (Ed.), *Comprehensive handbook of psychological assessment, 4 Volume Set*. Hoboken, NJ: John Wiley, Sons, Inc.

PAIVIO, A., 2009, *Imagery and verbal processes*. New York, Hove: Psychology Press.

R CORE TEAM, 2012, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Retrieved from: http://www.R-project.org/

REVELLE, W., 2013, *Psych: Procedures for personality and psychological research, Version = 1.3.10*. Northwestern University, Evanston, Illinois, USA. Retrieved from: http://CRAN.R-project.org/package=psych

QUAISER-POHL, C., 2003, The Mental Cutting Test "Schnitte" and the Picture Rotation Test – two new measures to assess spatial ability. *International Journal of Testing*, 3, 219-231.

SVOBODA, M., 2010, *Psychologická diagnostika dospělých* [Psychological assessment in adulthood]. Praha: Portál.

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.

WITKIN, A., MOORE, C.A., GOODENOUGH, D.R., COX, P.W., 1977, Field-dependent and field-independent cognitive styles and their educational implications. *Review of Educational Research*, 47, 1-64.

TESTOVÁNÍ PROSTOROVÝCH SCHOPNOSTÍ: STAVBA A HODNOCENÍ NOVÉ METODIKY

P. K v ě t o n, M. J e l í n e k, D. V o b o ř i l

Souhrn: Předložená studie popisuje vývoj a psychometrické zhodnocení nově navrženého testu prostorové představivosti. Studie je rozdělena do dvou navazujících částí. V části I (N = 267) je představeno 35 položek rovnoměrně rozdělených do pěti subsetů. Položky byly navrženy s ohledem na pětifaktorovou teorii prostorové představivosti, která zahrnuje dimenze prostorové percepce, prostorové orientace, prostorové vizualizace, mentální rotace a chápání prostorových vztahů. Ačkoli pětifaktorový model prokazoval přijatelnou shodu s daty, na principu parsimonie jsme upřednostnili model obecného faktoru. Položky s nejlepšími charakteristikami (n = 25) byly vybrány do finální verze testu, jejíž psychometrické charakteristiky byly ověřeny v rámci části II (N=124). Výsledky v tomto testu vykazují uspokojivou míru test-retest stability (r = 0,796) při šestitýdenním intervalu mezi měřeními a vnitřní konzistence (Cronbachova α = 0.752). Nebyl prokázán vliv media administrace (počítač vs. papír/tužka) a byl nalezen středně těsný vztah (r = 0,470) se subtestem Inteligenčního strukturního testu zaměřeného na prostorovou představivost.