

## A Study on Developing a Spatial Ability Test for Myanmar Middle School Students

Nu Nu KHAING<sup>1)</sup>, Tsuyoshi YAMADA<sup>2)</sup>, Hidetoki ISHII

### 1. Introduction

Spatial ability plays an important role in our daily lives and we use it unconsciously. For example, by using spatial ability, we assemble furniture, we commute to school or work and so on. Regarding spatial ability, Gardner (1983) stated that spatial ability and spatial cognition were the basic building blocks that a child needed in order to develop higher level thinking skills.

Since the 2<sup>nd</sup> World War, the assessment of spatial ability has been used for personnel selection because it has been accepted that there is a strong relationship between spatial ability and an individual's achievement (Eliot & Smith, 1983). In recent years, many research studies have shown that spatial ability is an important component of success in a variety of scientific, technical, and mathematical related occupations (e.g., Hegarty & Waller, 2006; Humphreys, Lubinski, & Yao, 1993). Moreover, it has been given evidence that spatial ability links to creativity and achievements in science, math, medicine and engineering (e.g., Casey, Nuttall, Pezaris, & Benbow, 1995; Hegarty, Keehner, Khooshabeh & Montello, 2009; Humphreys et al., 1993). Besides, spatial ability becomes an important role in education these days. As a famous current spatial ability research, Shea, Lubinski & Benbow (2001) investigated the spatial ability of students from age 13 through to age 33 for their educational and vocational outcomes. According to their results, they concluded that spatial ability can be contributed to the prediction of educational tracks.

As mentioned above, the assessment of spatial ability becomes a primary interest for researchers, educators and teachers. However, Myanmar, one of the developing countries, has not yet become widely aware of the importance of spatial ability. Moreover, there are no typical spatial ability tests in Myanmar yet. So, it is necessary to develop a spatial ability test for Myanmar students to be able to classify children with reference to their ability, and to predict students for professional colleges and universities to some extent. This is the first reason for making this study.

Spatial ability tests are generally non-verbal tests. But for only this reason, it cannot be said that spatial ability tests are culture-fair tests. Childhood experiences and cultural factors play probably big part in explaining differences in spatial abilities. Barke & Engida (2001) have studied the students' spatial ability of German schools and Ethiopian schools. Their results tell us that cultural factors have influences on spatial ability, in agreement with the findings of earlier research work Berry (1971) that used other spatial ability tests.

Moreover, it has been heard about a problem concerned the culture influence on a standardized test, so-called 'Piano and xylophone problem', in Myanmar. It is saying about Myanmar culture and Wechsler Intelligence Scale for Children (WISC III). Completion test is one subtest of the WISC III. In which, children are shown artwork of common objects with a missing part, and asked to identify the missing part by pointing and/or naming. It has a question about picture of a piano. Interestingly, it was observed most of Myanmar children could not answer about piano correctly because they were not familiar with the piano. If, in spite of a piano, they were asked about a xylophone, a similar musical instrument, they would response correct answer well. (National Education Seminar in Myanmar, 2003, no press). Therefore, it is

---

1) Graduate Student, Graduate School of Education and Human Development, Nagoya University (Supervisor: Hidetoki Ishii)

2) Associate Professor, Graduate School of Education, Okayama University

clear that even pictorial items can cause such the above problem because of cultural differences. For the reasons mentioned above, it has become imperative to develop a well constructed spatial ability test for Myanmar students.

According to the previous literature, to construct a spatial ability test, there are two issues that need to be considered. The first issue is concerning with the comparability of various spatial ability tests and the second is concerning with the types of spatial ability tests which vary depending on the context and the purpose of each study.

The first issue is related to the definition of spatial ability. In fact, the definition of spatial ability is not unitary. Psychometric studies of spatial ability identified various definitions for spatial ability (e.g., Carroll, 1993; Eliot & Smith, 1983; Lohman, 1979, 1988; McGee, 1979). Depending on the various definitions of spatial ability, there are many instruments / tests which have been used in study. Eliot and Smith (1983) gave directions and example items for 392 spatial ability tests. Generally, the spatial ability tests require students to transform objects (such as by rotating or transposing them) mentally.

Despite a large number of paper and pencil measures of spatial ability which are known to exist, there is still confusion when we seek an appropriate spatial ability test for our purpose. This is because, according to previous literature, it was found that some researchers often applied different types of spatial ability tests for the same purpose. For example, to measure students' spatial abilities, Mohler (2008) used "Vandenberg Mental Rotations test" composed of only a single task – mental rotation tasks, while Kayhan (2005) used "Delialiglu spatial ability test", involving two tasks – a paper folding task and a surface development task.

Here, we should consider a psychometric factor that concerns the comparability of spatial test scores. Using different ability tests and comparing their results a certain problem arises with the test scores. Hambleton, Swaminathan, and Rogers (1991) have proposed that it is very difficult to compare the subjects who took different tests due to the test-dependent problem. It is also very difficult to compare items whose characteristics such as item difficulty and item discrimination obtained by using different groups of subjects because it may cause the sample-dependent problem.

The test-dependent problem and the sample-dependent problem can be found when test construction procedure is undertaken by utilizing only the item analysis of the Classical Test Theory (CTT) model. In the CTT model, an examinee's ability is defined only in terms of a particular test. Whether an item is hard or easy depends on the ability of the examinees being measured. Hence, it is very difficult to compare examinees who take different tests and to compare items whose characteristics (item difficulty and item discrimination, reliability, etc.) are obtained by using different groups of examinees (Hambleton et al., 1991). The Item Response Theory (IRT) model overcomes the above limitations of CTT analysis by providing information on how examinees at different ability levels on a trait have performed on an item.

Therefore, it is clearly meant that if we use only the CTT model for item analysis, item selection and test construction, it may cause the sample-dependent and test-dependent problems. Therefore, for this study, it was decided to apply not only the CTT model as classical item analysis but also the IRT model, mainly to be able to develop a spatial ability test systematically.

The second issue is concerning the types of spatial ability tests. According to previous literature, spatial ability tests can be found in two types; single-task tests and multiple-task tests. Single-task tests consist of only one spatial task such as Vandenberg Mental Rotations Tests (Mohler, 2008; Vandenberg & Kuse, 1978), while multiple-task tests are composed of two or more different spatial task items. One example of a multiple-task test is the Spatial Aptitude test, developed by Psychometric Success, which requires a set of different tasks: shape matching, group rotation, combining shapes, cue views in 3-dimensions, maps and plans, and other solids in two and three dimensions.

In the above situations, we notice that if we use only a single-task test to measure the spatial ability, the test will measure only one aspect. In other words, by using a single-task test we cannot measure a broad range of spatial ability. Therefore, development of more multiple-task spatial ability tests should be considered.

Moreover, according to Eliot and Smith (1983), using the single task test is likely to be boring or tedious because it consists entirely of items of a single type. Therefore, they often suggested that a composite test (a multiple task test) consisting of several different types of

items is more likely to be interesting, as well as producing a better measure of the major spatial group factor. These reasons clearly lead to the need to develop a new multiple-task spatial ability test.

Based on the above literature review, the purposes of the study are to:

- Develop a new spatial ability test to properly measure the spatial abilities of the Myanmar middle school students,
- Construct the test with items of multiple spatial tasks using the two-parameter logistic IRT model.

Concerning the spatial ability definition, this study defines spatial ability as “the ability to generate, retain, retrieve, and transform well-structured visual images”, which is proposed by Lohman.

In addition, middle school students are selected as subjects for this study because Barke (1993) found that around the age of 14 years, middle school ages, spatial ability is developed to a point that students interpret the two-dimensional drawings of cubes, tetrahedrons or octahedrons in a spatial way. Shea et al. (2001) have recommended that if the students know their level of spatial ability from their middle school ages, it will help them to develop their spatial skills by practising and selecting major subjects at professional colleges and universities.

## 2. Procedure and Methods

### 2.1. Sample of the Study

A total of 798 middle school students (388 boys and

410 girls) voluntarily participated from nine schools in the Yangon City Development Area. Their ages ranged from eleven to fifteen years. There were 430 students aged 13, 350 aged 14 and 18 others.

### 2.2. Planning the test

The first step is to prepare a table of test content specification with the reference volume of the International Directory of Spatial Tests, which has been sorted by Eliot & Smith (1983). In the Directory, spatial ability tests are sorted and grouped into ten task categories according to the perceived similarity of their test stimuli and requirements. For this research, it was planned to develop the test items of ten spatial tasks: 1) copying, 2) maze tasks, 3) embedded figure tasks, 4) visual memory tasks, 5) form completion tasks, 6) form rotation tasks, 7) blocks tasks, 8) block rotation tasks, 9) paper folding tasks and 10) surface development tasks.

At first, some spatial ability tasks were carefully selected from the Directory. Tasks from each category which were suitable for the age range of 11-15 year-old students were selected. The characteristics of the tests, i.e., test instruction, time allowed, item format, were carefully studied. After that, the table of task content specifications with the 10 spatial tasks was constructed as shown in Table-1.

After drawing the table of task content specifications, pictorial items of the test are developed originally by authors. Table 2 describes sample test items of ten spatial

**Table 1 Task Content Specifications of the Test**

No.	Name of Tasks	Item Numbers	Item Response types	Amount of Items	Duration (min)
1	Maze	Sp-1~Sp-8	Free Response	8	2
2	Coping	Sp-9~Sp-20	Free Response	12	6
3	Visual Memory	Sp-21~Sp-35	Matching	15	2
4	Embedded Figure	Sp-36~Sp-47	Matching	12	4
5	Block Counting	Sp-48~Sp-63	Free Response	16	2
6	Paper Formboard	Sp-64~Sp-72	Multiple Choice	9	2
7	Figure Rotation	Sp-73~Sp-80	Multiple Choice	8	2
8	Paper Folding	Sp-81~Sp-86	Multiple Choice	6	2
9	Block Rotation	Sp-87~Sp-91	Multiple Choice	5	2
10	Surface Development	Sp-92~Sp-95	Multiple Choice	4	4
	Total	Sp-1~Sp-95		95	28

Note: time duration and amount of items of each task are different depending on the referenced tests.

tasks.

1. **Maze.** Students need to find a path through a maze quickly, and to draw a pencil line through each maze without crossing any printed lines.
2. **Copying.** Students need to copy a figure superimposed upon a framework of crosses onto a similar framework of dots. Students must keep the pattern in mind so that they can quickly find it in a square of dots.
3. **Visual Memory.** Students have to memorize the shapes from the first table (table A) and then choose the same shapes from the second table (table B). Students should fill in the correct number on the blank part of the answer sheet.
4. **Embedded Figure.** Students must identify or draw a simple given figure which is embedded in a more complex figure. For example, there are three simple figures, and students have to decide which figure is embedded in two more complex figures. This item type is matching.
5. **Block Counting.** Students need to count the total number of blocks in each pile. Two different types of blocks are used but only one type of block is used in any one pile. This item type is free response.
6. **Paper Formboard.** Each problem has a numbered figure to the left and four lettered figures to the right. Students must find the lettered figure made of exactly the same pieces that are in the numbered figure. This item type is multiple-choice with four options as follows.
7. **Figure Rotation.** Students have to indicate which of four figures, when mentally turned or rotated, are different from a given figure. This item type is multiple-choice with four options.
8. **Paper Folding.** Students must imagine the folding and unfolding of pieces of paper. In each problem, the figures to the left represent a piece of paper being folded. One of the four figures to the right of the vertical line shows where the holes that are in the paper will be when it is completely unfolded. Students have to decide which one of these figures is correct. This item type is also multiple-choice with four options.
9. **Block Rotation.** Each problem consists of five drawings, and four of them are the same. Students must indicate which block, when mentally turned

or rotated, is different from a given block or object.

This item type is multiple-choice with four options.

10. **Surface Development.** Each problem consists of a pattern which can be cut out and then folded on the dotted lines to form a closed 3-dimensional figure. Students must decide which of the lettered edges will fall along the edge marked with the arrow after the pattern is folded. This item type is multiple-choice with four options.

Test instructions were also prepared. These were (1) to answer all items (i.e., test items), (2) to use only pencils and erasers but not rulers, (3) to answer only the tasks that the teachers allowed, (4) not to read the next task unless the teacher permitted it, and (5) to listen to and follow the instructions carefully.

Before administering the test, the content was checked to see whether item pictures and test instructions were clear, with the authors and ten 3<sup>rd</sup> year graduate students, from the Department of Educational Psychology of Okayama University, who had learned about test constructions. After discussion, some items were modified.

With respect to ambiguous meaning and conformity with junior high school levels, the test taker review was done by three Myanmar middle school students. They were asked whether they would have enough time to complete each task, and whether they could understand the instructions well. Based on this review, time duration was adjusted and some items were rewritten. Finally, 95 items were ready to be used in the study.

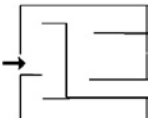

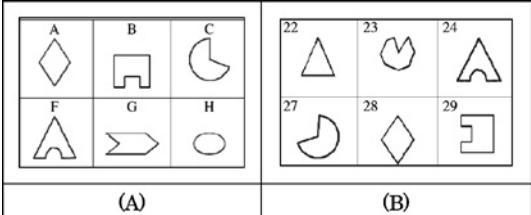
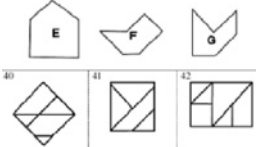
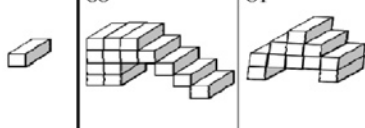
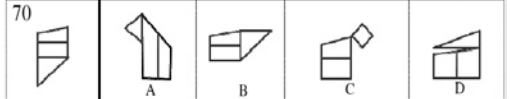
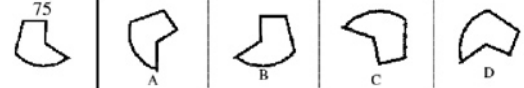
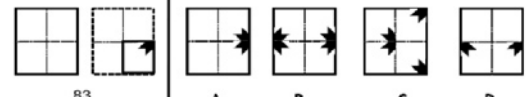
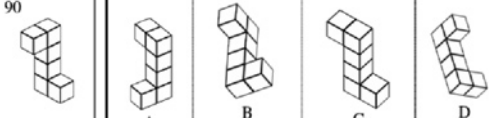
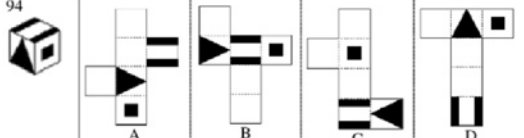
### 2.3. Data Collection

The whole test of 95 items was administered to 798 middle school students in February 2010 in Myanmar. The total time duration of the test was 45 minutes: 28 minutes for test-taking and the remaining time for reading test instructions. The total time of 45 minutes is the length of one lecture period for middle school students in Myanmar. Responses were scored 1 if answered correctly and 0 if answered incorrectly.

## 3. Data Analysis and Results

The procedure of the data analysis was performed with the reference to Hambleton, Swaminathan & Rogers (1991) and Myint (1997) as follows.

Table 2 Sample Items of Ten Tasks

1. Maze.	<p>2</p> 
2. Copying.	<p>13 15 17</p> 
3. Visual Memory.	 <p>(A) (B)</p>
4. Embedded Figure.	
5. Block Counting.	<p>60 61</p> 
6. Paper Formboard.	<p>70</p> 
7. Figure Rotation.	<p>75</p> 
8. Paper Folding.	<p>83</p> 
9. Block Rotation.	<p>90</p> 
10. Surface Development.	<p>94</p> 

### 3.1. Classical Item Analysis

As a first item analysis, classical item analysis was conducted in which p-value (difficulty index in CTT) and point-biserial correlation (the discrimination index) of the 95 spatial items were calculated.

The p-value is the proportion of examinees who answer the item correctly. It was found that 43 items had very high p-values (more than 0.90). These items were very easy and the answer was probably too obvious for the examinee group. Therefore, they were discarded from the test because they could induce a ceiling effect. In addition, one item (Sp-17) from the Coping task had a low p-value (less than 0.20). It meant that it possessed a high item difficulty, and it could cause a floor effect, thus it was removed.

The point-biserial correlation ( $\rho_{pbis}$ ) reflects item-total test correlation. Among the remaining 51 of the 95 items, there were no negative items with  $\rho_{pbis}$ . Sp-7 possessed 0.27 of  $\rho_{pbis}$  but it had 0.87 of p-value. Another 13 items were lower  $\rho_{pbis}$  (less than 0.3). A low point-biserial correlation implies that the examinees who get the item correct tend to do poorly on the overall test, and that the examinees who get the items wrong tend to do well on the test. Therefore, these 13 items were removed, and as a result, there were 38 items left in the test.

### 3.2. Investigate whether it was a non-speeded test

Even though its intended purpose is to measure the level of ability, when a test has restrictive time limits, the items at the end of the test may measure the construct of test-taking speed more than the items at the beginning. Therefore, it is essentially to investigate whether it was a non-speeded test.

Non-speeded test administration was investigated by calculating the ratio of the variance of the number of omitted items to the variance of the number of items answered incorrectly (Hambleton, et. al., 1991). As a result, it was observed that the ratios of variances of items in the Maze task and the Block Counting task were far greater than zero. It signaled that the items in the Maze task and the Block Counting task were very easy and they were done quickly to complete all the items by the examinee group.

In fact, seven items of the Maze task and all the items in the Block Counting task have been eliminated since section 3.1 because they have very low p-values and very low item-total test correlation. In this step, the Sp-7 item that was left in the Maze task was also removed because of its condition in speeded test administration.

Up to this step, 37 items were remaining in the test. These items possessed both fair p-values (0.20 to 0.90) and fair point-biserial correlation coefficients (0.3 to 1) simultaneously.

Table 3 shows the obtained items (37 items) of six spatial tasks; they are 10 items from Copying, 6 items from Embedded Figure, 7 items from Paper Formboard, 6 items from Figure Rotation, 3 items from Paper Folding, and 5 items from Block Rotation.

### 3.3. Check the assumption of Unidimensionality

The assumption of unidimensionality is a common one for test constructors since they usually desire to construct unidimensional tests to enhance the interpretability of a set of test scores (Myint, 1997). To assess the unidimensionality of the data, a scree plot of the eigenvalues of the tetrachoric correlation matrix was graphed

**Table 3** Item Results of the Test Development Process

No.	Name of Tasks	Item Numbers	Item Response types	Constructed Items	Discarded Items	Remaining Items
1	Coping	Sp-9~Sp-20	Free Response	12	2	10
2	Embedded Figure	Sp-36~Sp-47	Matching	12	6	6
3	Paper Formboard	Sp-64~Sp-72	Multiple Choice	9	2	7
4	Figure Rotation	Sp-73~Sp-80	Multiple Choice	8	2	6
5	Paper Folding	Sp-81~Sp-86	Multiple Choice	6	3	3
6	Block Rotation	Sp-87~Sp-91	Multiple Choice	5	0	5
Total						37

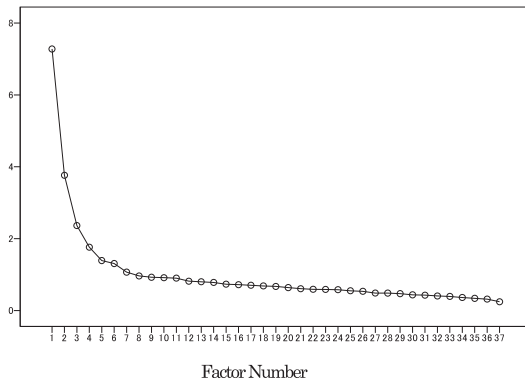


Figure 1 Scree plot of the eigenvalues

(Figure 1).

Figure 1 illustrates the largest eigenvalue of the factors is over two times larger than the second largest eigenvalue. Theoretically speaking, it may be said that the result can not satisfy enough with unidimensionality assumption. However, Hambleton et al., (1991, p.9) insisted that the unidimensionality assumption could not be strictly met because several cognitive, personality, and test-taking factors might affect test performance, at least to some extent. According to their recommendation, what is required for the unidimensionality assumption to be met adequately by a set of test data is the presence of a dominant component or factor that influences test performance. Therefore, it was assumed that it had a reasonable unidimensionality for this study.

### 3.4. Check preliminary selection of promising IRT Models

In order to apply an IRT model for this study, a preliminarily selected two-parameter model was confirmed with two facts. These were (1) the variation in p-values and point-biserial correlations were large and thus, one-parameter model was not suitable to apply, and (2) the sample size of the study was less than 1000, it should not be applied a three-parameter model (these rules followed the recommendation of Lord, 1968).

### 3.5. Estimate ability and item parameters

As a next step, the items were calibrated with a two-parameter logistic (2PL) IRT model. In this model the probability of correct response is modeled by:

$$P_i(\theta) = \frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \quad i = 1, 2, \dots, n$$

where  $\theta$  is the examinee's ability,  $\alpha$  is the item discrimination (also called the item slope),  $b$  is the item difficulty.

The item parameters of the 37-items were estimated using marginal maximum likelihood (MML) in BILOG-MG 3 (Zimowski, Muraki, Mislevy & Bock, 2003). According to MML procedure, at first, the ability parameters were integrated out, and the item parameters were estimated. Therefore, the average ability of examinees was set to 0, with a standard deviation of 1 by assuming that the distribution of  $\theta$  is standard normal distribution. With the item parameter estimates determined, the ability parameters were estimated. Table 4 provides the estimates of the item difficulty ( $b$ ) and item discrimination ( $\alpha$ ) of each item.

For item discrimination ( $\alpha$ ), a higher value indicates that the item discriminates between high and low proficiency examinees better. For this test, the variability of  $\alpha$  values ranges from 0.49 to 1.64 and the mean is 0.81. Therefore, it was concluded by a consideration of their discrimination powers, the items were fairly good items to provide appropriate discrimination for the whole test.

The item difficulty ( $b$ ) column indicates that easier items have lower (negative) difficulty indices and harder items have higher (positive) indices. On this test, the variability of  $b$  values ranges from -1.77 to 1.31 and the mean is -0.48. It was observed that 72% (26 items) of the items had negative  $b$  values. Therefore, it could be said that the test with these items was relatively easy.

### 3.6. Investigate the invariance of Parameter Estimates

After estimated parameters, the model data fitness was investigated. Although there were many approaches for assessing the goodness of fit (Hambleton et al, 1991), investigation method of the invariance of parameter estimates were conducted in this research.

#### 3.6.1. Invariance of item parameter estimates

To investigate the invariance of the item parameter estimates of the 37 items, the sample 798 students were grouped into two random equivalent groups. By using the BILOG-MG software again, two IRT analyses were conducted separately for the two groups to obtain item

**Table 4** Item Parameters for 37 Items

Item	<i>a</i>	<i>b</i>	Item	<i>a</i>	<i>b</i>
Sp-9	0.68	0.83	Sp-68	0.62	-0.32
Sp-11	0.68	0.57	Sp-71	0.78	-1.77
Sp-12	0.51	0.38	Sp-72	0.57	-0.37
Sp-13	0.74	0.92	Sp-73	0.82	-0.65
Sp-14	0.81	0.72	Sp-74	0.71	-0.36
Sp-15	0.49	-0.30	Sp-75	0.58	-0.56
Sp-16	0.60	0.80	Sp-76	0.86	-0.47
Sp-18	0.65	0.48	Sp-77	0.53	-0.05
Sp-19	0.55	0.64	Sp-78	0.74	-0.42
Sp-20	0.69	1.31	Sp-79	0.66	0.12
Sp-40	1.42	-1.17	Sp-80	0.64	-0.77
Sp-41	1.39	-1.23	Sp-83	0.55	-0.87
Sp-42	1.56	-1.19	Sp-84	0.65	-1.07
Sp-44	1.44	-1.44	Sp-86	0.72	-0.93
Sp-46	1.60	-1.33	Sp-87	0.77	-1.36
Sp-47	1.64	-1.38	Sp-88	0.73	-1.34
Sp-64	0.76	-1.19	Sp-89	0.65	-1.19
Sp-66	0.89	-1.67	Sp-90	0.59	-0.02
Sp-67	0.59	-1.23			

parameter estimates. Then, a plot of difficulty estimates (*b*-values) estimates of two random equivalent groups for 37 items was graphed (Figure 2).

As shown in the Figure, it was observed that the  $R^2$  is 0.97 and the correlation ( $r$ ) of two difficulty estimates was above 0.98. The difficulty estimates lay along a straight line with a few scattered. Moreover, it was found that the results of discrimination estimates also caused a straight line graph like those of the difficulty estimates. Therefore, it was concluded that the item parameters held the invariance property.

### 3.6.2. Invariance of ability parameter estimates

To investigate the invariance of ability parameters estimates across different samples of items, the ability parameters of the 798 examinees were compared for two randomly equivalent samples of test items based on examinee performance on the odd-numbered items and on the even-numbered items. Figure 3 demonstrates the ability estimates based on equivalent test halves (Odd vs. Even Items).

As shown in Figure 3 it was observed that the  $R^2$  is 0.88,

the correlation ( $r$ ) of two difficulty estimates is 0.93, so it was concluded that the invariance of ability parameters over the test was present.

### 3.7. Test Characteristic Function and Test Information Function

The test characteristic curve (TCC) for the 37-item test was graphed (Figure 4) to learn the peculiarities of the test as a measuring instrument. The TCC shows how test scores on each test are related to the ability  $\theta$  of the examinee (Hambleton et al., 1991). The TCC is also the true score ( $\tau$ ) of an examinee with an ability  $\theta$  in IRT.

By looking at Figure 4, it is visually clear that the test is discriminating well among examinees with the range of ability level  $-2.0$  to  $+1.0$  but is discriminating poorly among examinees with extremely low or high  $\theta$ . Since the ability distribution of the examinees was assumed as a standard normal distribution, the test was desired to provide maximum discrimination or information in the range of  $-2$  to  $+2$ . However, it was observed that this test could provide maximum discrimination only in the ability range of  $-2.0$  to  $+1.0$ .



原

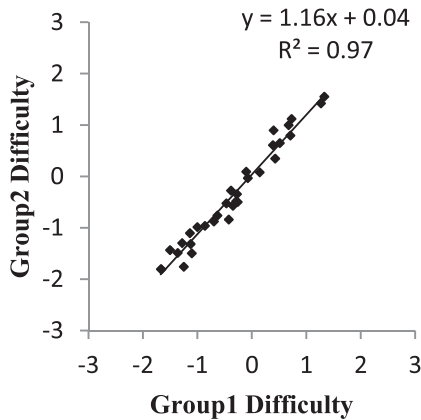


Figure 2 Plot of item difficulty estimate values

著

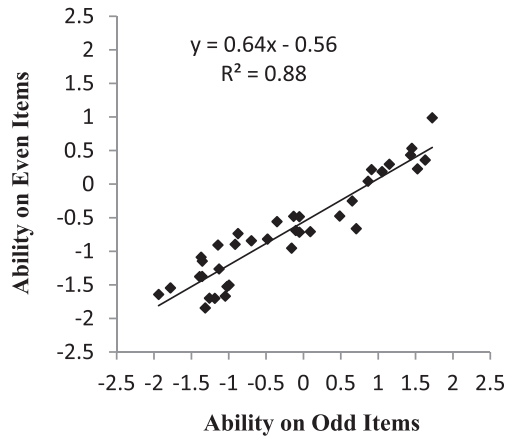


Figure 3 Plot of the Ability Estimates Based on Odd vs. Even Items

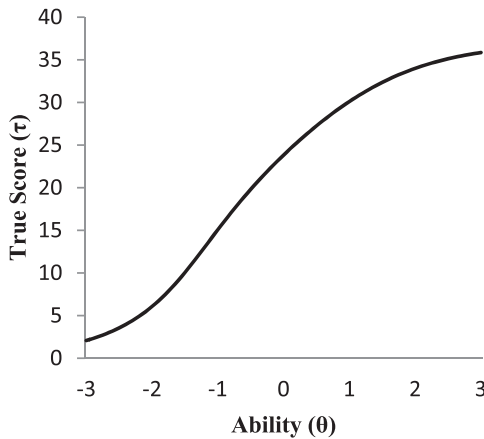


Figure 4 Test characteristic curve for the test with 37 items

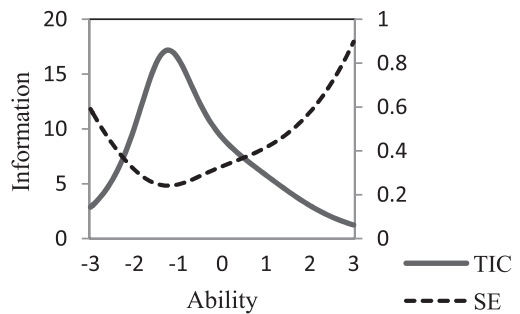


Figure 5 Test information curve for the test composed of 37 items

In order to know precisely the maximum amount of information obtained from the test, the test information function was calculated. In IRT, the information function is used to know the standard error of the test and its reliability. The standard error of the test is the inverse of the square root of information, so that the greater the information, the smaller the standard error and the greater the reliability (DeMars, 2010). Figure 5 illustrates the test information curve (TIC) of the 37-items test. SE is the standard error of estimation.

The TIC shows that the test has smaller standard errors across the ability scale from  $-2.3$  to  $+0.5$ , and larger standard errors at the low and high ends of the scale. The

maximum amount of information  $I(\theta) = 16.96$  is at  $\theta = -1.35$ . Smaller standard errors are associated with highly discriminating items for which the correct answers cannot be obtained by guessing (Hambleton et al., 1991, p.95). Therefore, this test will be most suitable for examinees whose spatial ability  $\theta$  range is  $-2.3$  to  $+0.5$ , but it cannot discriminate well the students who have higher ability levels (above  $\theta = +0.5$ ) and lower ability extreme levels (less than  $\theta = -2.3$ ).

#### 4. Discussion and Further Research

In this research, a spatial ability test was developed as a multiple-task spatial ability test to measure more

aspects of spatial ability than single-task tests. The test items were analyzed systematically by applying the CTT and the IRT analysis in order to solve sample-dependent and test-dependent problems and in order to express at the item level rather than at the test level. Moreover, in this research, the treatment of reliability and error of measurement through test information function (TIC) was presented.

Some limitations were found in this research. First, as explained in section 3.5, it was found that the obtained test information curve functioned only from the range of  $-2.3$  to  $+0.5$ . Therefore, it can be said that the test is an easier test for the students. With these items, it may not provide enough information to the participants of Myanmar middle school students yet. In order to accurately measure the spatial ability of Myanmar students, it is still necessary to fill more difficult items and to arrange them from the easy items to difficult items across the ability scale, until the test information function (TIC) ranged  $-2$  to  $+2$  is achieved.

Second, Hambleton et al. (1991, p.5) argued that an IRT model does not require strictly parallel tests for assessing reliability. The reason is that when a given IRT model fits the test data, the test results possess invariance properties. But it provides only internal consistency reliability or homogeneity, directly related to the 'unidimensionality' of the test. In addition to this, the consistency of test scores is also of considerable importance in evaluating a test as a measurement instrument. Therefore, it is still required to investigate test-retest reliability, also known as 'stability of the test', in this study. It remains for further study.

Third, the sample selection of the students for the research was conducted at Yangon City Development Area in Myanmar. Moreover, this research was performed using only the middle school students at the above schools. Therefore, it cannot be said that the sample is fully representative of the population of Myanmar middle school students. In addition, the sample sizes were just around 800 students. It is not known if the results will hold up for larger sample sizes yet. This issue is still to be investigated as a further study.

Fourth, it is essential to validate with the other spatial ability tests (for example, mental cutting test (MCT), mental rotation test (MRT), etc.). Moreover, it is still necessary to undertake an international comparison of

the test for future validation of the test itself. Moreover, a new test can develop by using some of the spatial tasks of this test and new additional items until the desired test information function of  $-2$ ~ $+2$  is reached.

The final limitation of this research is the assumption of unidimensionality. This research has applied CTT item analysis and IRT item analysis in the test development. To apply IRT item analysis, assumption of unidimensionality should be held (DeMars, 2010; Hambleton et al, 1991). However, in this study it cannot be said that the test data of the study satisfied enough the assumption of unidimensionality (refer to Figure 1). This fact might be that the test items had different item response types of spatial tasks, depending on the reference test directory. It is questionable that if the spatial ability test was developed with the same item response types in all tasks, would it have been more satisfactory with the unidimensionality assumption. Therefore, as a further study, it will be necessary to investigate whether or not the unidimensionality assumption will be satisfied more than this study if the same item response types in all tasks are used in the test.

## 5. Conclusion

To sum up, in this paper, a spatial ability test for Myanmar middle students was developed. The test has some limitations and it is still under development. Other further studies are necessary to investigate. Further work can focus on longitudinal research with an item pool. It is hoped that this contribution will aid spatial ability research to some extent.

## Acknowledgements

We would like to give special thanks to Dr. Aye Aye Myint (Yangon Institute of Education, Myanmar) for her cooperation in test administration. Moreover, we wish to express our deep gratitude to all principals and participants of this study.

## References

- Barke, H.-D. (1993). Chemical education and spatial ability. *Journal of Chemical Education*, 70, 968.
- Barke, H.-D., Engida, T. (2001). Structural chemistry and spatial ability in different cultures. *Chemistry Education: Research and Practice in Europe*, 2(3), 227-239.

- Berry, J. W. (1971). Ecological and cultural factors in spatial perceptual development. *Canadian Journal of Behavioural Science*, 3, 324-329.
- Carroll, J. B. (1993). *Human Cognitive Abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Casey, M. B., Nuttall, R., Pezaris, E., & Benbow, C. P. (1995). The influence of spatial ability on gender differences in mathematics college entrance test scores across diverse samples. *Developmental Psychology*, 50, 179-184.
- DeMars, C. (2010). *Item Response Theory*. Oxford University Press.
- Eliot, J. C., & Smith, I. M. (1983). *An International Directory of Spatial Tests*. Windsor, England: NFER-Nelson.
- Gardner, H. (1983). *Frames of Mind*. New York: Basic Book Inc.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications, Inc.
- Hegarty, M., Keehner, M., Khooshabeh, P., & Montello, D. R. (2009). How spatial ability enhances, and is enhanced by, dental education. *Learning and Individual Differences*, 19, 61-70.
- Hegarty, M., & Waller, D. (2006). Individual differences in spatial abilities. In P. Shah & A. Miyake (Eds.). *Handbook of Visuospatial Thinking*. Cambridge University Press, 121-169.
- Humphreys, L. G., Lubinski, D., & Yao, G. (1993). Utility of predicting group membership and the role of spatial visualization in becoming an engineer, physical scientist, or artist. *Journal of Applied Psychology*, 78, 250-261.
- Kayhan, E. B. (2005). Investigation of High School Students' Spatial Ability. *Master Thesis*, Middle East Technical University, Turkey.
- Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 28, 989-1020.
- Lohman, D. F. (1979). *Spatial ability: A review and re-analysis of the correlational literature*. Technical Report, Stanford, CA: Aptitudes Research Project, School of Education, Stanford University.
- 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). Examining the spatial ability phenomenon from the student's perspective. *Engineering Design Graphics Journal*, 72, 1-15.
- Myint, A. A. (1997). Investigation of the numerical reasoning ability of Myanmar high school students. 東京大学大学院教育学研究科紀要, 37, 165-176.
- Shea, D. L., Lubinski, D., & Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *Journal of Educational Psychology*, 4(3), 207-230.
- Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations: A group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, 47, 599-604.
- Zimowski, M., Muraki, E., Mislavy, R. J., & Bock, R. D. (2003). BILOG-MG 3: Item analysis and test scoring with binary logistic models. Chicago, IL: Scientific Software. [Computer software.]

(Accepted: 2011.9.30)

ABSTRACT

A Study on Developing a Spatial Ability Test for Myanmar Middle School Students

Nu Nu KHAING, Tsuyoshi YAMADA, Hidetoki ISHII

Recently, many researchers have investigated that spatial ability affects a lot of professional fields and it can predict success in many life areas. Then, it becomes more important to measure spatial ability on career formation. However, in Myanmar, it was not widely aware of the importance of spatial ability, and there was no typical spatial ability test yet. Therefore, in this paper, a spatial ability test for Myanmar middle students was developed. To develop the test, classical item analysis was conducted and item response theory (IRT) was applied. As the IRT model, two parameters logistic model (2PL) was utilized. Consequently, a test composed of 37 items was developed. The test information function showed that the accuracy of ability estimates was sufficient on the range of lower ability level ( $-2.3$  to  $+0.5$ ). Therefore, it was concluded that the constructed test was an easier test for the students.

Key words: spatial ability, IRT, two parameter logistic model, test information function