AC 2011-2582: SCALING THE REVISED PSVT-R: CHARACTERISTICS OF THE FIRST YEAR ENGINEERING STUDENTS’ SPATIAL ABILITY

Yukiko Maeda, Purdue University

Yukiko Maeda is an assistant professor in the College of Education at Purdue University. She received her PhD in quantitative methods in education from University of Minnesota. Her research interests include survey and assessment design in educational research, and meta-analysis.

So Yoon Yoon, Purdue University, West Lafayette

So Yoon Yoon is a doctoral candidate in gifted education at Purdue University. She enjoys working with diverse students talented in STEM areas. Her current research interest is to scale an instrument to measure students’ spatial ability, specifically spatial visualization ability in 3-D mental rotation (the Revised PSVT:R) using Classical Test Theory (CTT) & Item Response Theory (IRT); to investigate associations between PreK-16 students’ spatial ability and their STEM performance; to conduct meta-analyses to examine gender differences in spatial ability and intervention effects in STEM education; and to identify PreK-16 students’ talents in STEM.
Measuring Spatial Ability of First-Year Engineering Students
With the Revised PSVT:R

Abstract

Past literature in engineering education has generally agreed that spatial ability plays a crucial role in determining students’ achievement in engineering courses, especially graphic and design courses. While various spatial tests are available, the Purdue Spatial Visualization Tests: Visualization of Rotation (PSVT:R) has been commonly used to predict students’ success in the engineering field for more than three decades. However, little attention had been given to the validity of the PSVT:R in measuring spatial ability. Therefore, the purpose of this study was twofold: (a) to characterize the item- and test-level functions of the Revised PSVT:R for incoming first-year engineering (FYE) students based on two fundamental measurement frameworks: Classical Test Theory (CTT) and Item Response Theory (IRT); and (b) to investigate its relationship with academic-related variables to provide validity evidence. Approximately 600 freshmen enrolled in the fall 2010 FYE Program in a large Midwestern public university completed the Revised PSVT:R. Students’ academic performance, such as SAT/ACT subject scores and high school core GPA, were retrieved from the university archives along with students’ demographic backgrounds. The results indicated that the revised PSVT:R measures a unidimensional subcomponent of spatial ability and the scores are reliable for measuring spatial visualization ability of FYE students. They also indicated that the test is relatively easy for this population.

1. Introduction

1.1. Spatial Ability

Spatial ability has received widespread attention since early 1900s because of its link to academic and vocational success. Several attempts have been made to define this construct. For example, Lohman (1996) defined it as “the ability to generate, retain, retrieve, and transform well-structured visual images” (p.112), while Carroll (1993) defined the same ability as an “ability in manipulating visual patterns, as indicated by the level of difficulty and complexity in visual stimulus material that can be handled successfully, without regard to the speed of task solution” (p.362). However, as further described by Lohman (1993), “It [spatial ability] is not a unitary construct. There are, in fact, several spatial abilities, each emphasizing different aspects of the process of image generation, storage, retrieval, and transformation” (p.4). For example, Michael, Guilford, Fruchter, and Zimmerman (1957) proposed the three-factor model of spatial ability: Spatial Visualization (SV), Spatial Relations and Orientation, and Kinesthetic Imagery, while McGee (1979) suggested a two-factor model with SV and Spatial Orientation (SO). Lohman (1988) agreed with basic components of McGee’s model of spatial ability and extended it further by distinguishing Spatial Relations (SR) from SO. Compared to other researchers, Carroll (1993) characterized the dimensions in spatial ability broadly by including other
peripheral components, such as Closure Speed, Flexibility of Closure, Perceptual Speed, and Visual Memory.

Although there is no consensus of dimensions in spatial ability, researchers seem to agree that spatial ability consists of at least two correlated, but theoretically separable, core dimensions: spatial visualization (SV) and spatial relation/orientation (SR). Moreover, Guay and McDaniel (1978) discussed that SV is also hierarchically structured: a low level SV only requires the visualization of two-dimensional (2-D) configurations while high level SV is characterized as requiring the visualization of three-dimensional (3-D) configurations, and the mental manipulation of these visual images.

1.2. Importance of Spatial Ability for Success in Engineering Programs

While the notion of spatial ability varies across studies because of its complexity and the ability is loosely defined in most studies, researchers generally agree that spatial ability plays a crucial role in determining students’ achievement in engineering courses, in particular graphic and design courses. For example, Baartmans (1990) found that a student’s score on a spatial ability test was the most powerful predictor for students’ success in an engineering design graphics course among 11 investigated variables (i.e., gender, teacher, experience with shop and draft training, solid geometry, construction toys, spatial ability, and ACT scores in English, mathematics, social science, and natural science). Shea, Lubinski, and Benbow (2001) found that spatial ability, like other academic abilities (such as verbal and mathematical skills), is a significant predictor of academic performance in the Science, Technology, Engineering and Mathematics (STEM) disciplines, including engineering. Field (2007) reported that the use of spatial ability scores as well as mathematics course grades improved the prediction of performance in undergraduate engineering design courses compared to the use of mathematics GPA as a single predictor.

Recently, the demand for spatial ability has further increased as computer technology rapidly advances; 3-D Computer-aid Design (CAD) applications have been introduced in graphic design in engineering, and a traditionally required skill for designing, drawing in 2-D space, is now being replaced by designing and manipulating in 3-D space using virtual models. Therefore, spatial ability has received more attention than ever for its role in predicting an individual’s academic success, in particular, success in first-year engineering (FYE) programs.

The FYE programs provide opportunities for students seeking degrees in engineering to learn and develop fundamental knowledge. During the first-year in the engineering program, freshmen take all crucial courses in STEM areas (such as calculus, physics, computer programming, and graphic courses) to develop the knowledge and skills necessary to proceed to advanced engineering courses. Budny, Bjedov, and LeBold (1997) found in their 17-year longitudinal study that 97% of freshmen who successfully completed the first-year requirements in the FYE program graduated from college, mostly with baccalaureate degrees in engineering (i.e., 89 %). This indicates that successful completion of the FYE program holds the promise for graduation with a college degree. The result also suggests the implication for providing effective instruction to FYE students. For example, providing appropriate remedial instruction for students with lower spatial ability may help those students to be better prepared to take fundamental STEM courses.
in the FYE program, which may result in the successful completion of their FYE program and subsequently increase the chance for graduation from college.

In fact, several studies by Sorby and her colleagues\textsuperscript{15, 16, 17, 18, 19} indicated the positive impact of interventions for improving spatial ability on performance in graphic engineering courses at Michigan Technological College. In their studies, several spatial ability tests were used as a placement test for the required graphic courses for FYE students. Students with low spatial ability were encouraged to take an introductory course in cultivating 3-D spatial visualization. After taking the course, students increased their spatial ability and later successfully completed graphic-related courses\textsuperscript{15, 19}. In the study by Sorby and Baartmans (2000)\textsuperscript{18}, they also indicated a possible link between positive outcomes through the intervention and an increase in students’ retention in the engineering program. Later studies\textsuperscript{20} further supported the positive relationship.\textbf{1.3. Spatial Ability Tests in Engineering Education Research}

Given evidence of the positive correlation between spatial ability and academic success, the spatial ability of FYE students is often measured (a) to predict their performance on engineering courses\textsuperscript{11, 21, 22} and (b) to identify students who may benefit from participating in a remedial intervention program\textsuperscript{16, 17, 18}. For these purposes, the valid selection of an instrument is a critical first step to providing useful information to support students in their academic success.

Several tests have been utilized in previous research with spatial ability partly due to the fact that there is no unitary definition of spatial ability; rather, it consists of several subcomponents of spatial ability\textsuperscript{5, 23}. Various spatial ability tests that are currently available include the Mental Cutting Test (MCT)\textsuperscript{24}, the Mental Rotations Test (MRT)\textsuperscript{25}, the Revised Minnesota Paper Form Board Test (RMPFBT)\textsuperscript{26}, and the Differential Aptitude Tests: Spatial Relations (DAT:SR)\textsuperscript{27} and the Purdue Spatial Visualization Tests: Visualization of Rotation (PSVT:R)\textsuperscript{28}. These have been used frequently in research on success in engineering programs because the particular ability measured by these tests seems to be tightly connected to the engineering profession\textsuperscript{29}.

The MCT\textsuperscript{24} aims to measure an individual’s spatial visualization ability. While taking the test, an individual solves 25 items of 3-D objects with a cutting plane indicating where to be cut through. The individual’s task is to imagine the trace of that cutting plane and select the correct image from the possible options. The MRT\textsuperscript{25} is a paper-and-pencil based test that consists of 20 items to measure spatial visualization by mentally rotating 3-D objects. The RMPFBT\textsuperscript{26} is another popular spatial visualization test originally developed in 1920s to “measure aspects of mechanical ability requiring the capacity to visualize and manipulate objects in space”\textsuperscript{30}. The test consists of 62 items with 2-D objects. However, solving a problem on this test does not involve the mental rotation of objects.

The DAT:SR\textsuperscript{27} is a subtest of a multiple aptitude battery (8 subtests) that requires examinees to indicate what an unfolded shape would look like when folded. Finally, the PSVT:R\textsuperscript{28} is a spatial visualization test involving the mental rotation of 3-D objects. The PSVT:R has been used primarily in research on educational settings in science, technology, engineering, and mathematics (STEM) disciplines for more than three decades. The test has been recognized as one of the most popular tests to measure students’ spatial visualization ability of mental rotation.
in engineering education\textsuperscript{10,11}. This is because, compared to other popular spatial tests for research in engineering education, the PSVT:R is unique in that the 3-D objects utilized in the test have inclined, oblique, and curved surfaces, which requires a higher level of spatial visualization ability compared to visualizing objects with simple cubic surfaces that are typically used in other popular tests\textsuperscript{31}.

1.4. Purdue Spatial Visualization Tests: Visualization of Rotation (PSVT:R)

Guay (1976)\textsuperscript{28} developed the Purdue Spatial Visualization Test (PSVT), which consists of three 12-item subtests entitled “Developments,” “Rotations,” and “Views,” respectively. The PSVT:R is an extended version of the subtest, “Rotations,” to measure 3-D mental rotation ability of individuals aged 13 or older in 20 minutes\textsuperscript{32}. The PSVT:R has 30 items consisting of 13 symmetrical and 17 nonsymmetrical figures of 3-D objects, which are drawn in a 2-D isometric format. A sample item of the PSVT:R is shown in Figure 1. In each item, the respondents’ task is to mentally rotate a figure in the same direction as indicated visually in the instructions, and then to select an answer among five possible options.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{sample_item.png}
\caption{A sample item from the Revised Purdue Spatial Visualization Tests: Visualization of Rotations (Revised PSVT:R)}
\end{figure}

While there have been abundant studies that utilize the PSVT:R, little attention has been provided about the accuracy of measuring spatial ability with the instrument. In addition, Yue (2006, 2007)\textsuperscript{33,34} recently identified 10 figural errors on 7 of 30 items of the PSVT:R. As a result, the instrument was revised by Yoon (2011)\textsuperscript{35} with the permission of the original author. The new instrument is distinguished from the original one in this study by referring to it as the revised version of the PSVT:R (Revised PSVT:R).

Thus, the purpose of the study was twofold: to investigate and characterize the functions of the Revised PSVT:R for incoming FYE students and to report its relationship with academic-related variables. The specific research questions used were:

1. How reliable are scores on the Revised PSVT:R for first-year engineering students?
2. To what extent do characteristics such as item difficulty and item discrimination vary across the items in the Revised PSVT:R?
3. To what extent is the Revised PSVT: R supported by the criterion-related validity evidence?
4. To what extent is the Revised PSVT:R supported by the construct-related validity evidence?
2. Methods

2.1 Data Source

The target population of this study includes all engineering freshmen enrolled in the FYE Program in a large Midwestern public university in the fall of 2010. The acceptance criteria to enter the FYE Program are the same as the general freshman admission criteria of the university. In other words, students are accepted if they provide evidence of high academic performance and demonstrate that their academic aspirations are a good fit with a designated program. Students who were accepted by the engineering program were invited to participate in an online survey by the School of Engineering Education during the summer of 2010, prior to their entrance to the FYE program. The Revised PSVT:R was administered as a part of the online survey. For the study, we retrieved the secondary data of the Revised PSVT:R scores gathered from 585 FYE students, as well as relevant academic and demographic information of those students. Of the 585 students, 480 (82.05%) were male and 105 (17.95%) were female. Their average age was 20.48 years old.

2.2 Data

The Revised PSVT:R raw scores. The Revised PSVT:R was administered online and respondents were given a maximum of 25 minutes to complete the test. In Yoon’s study (2011) with undergraduate students across all majors, more than 95% of individuals could complete the test within 25 minutes when no time limit was set. Thus, we considered the 25-minute time frame sufficient to control the speed effect of problem solving on their scores.

Academic related performance. The FYE students’ academic aptitude test scores, such as SAT and/or ACT composite scores, and high school GPA were retrieved from the archival record maintained by the university and used to evaluate the criterion-related validity of the Revised PSVT:R.

Demographic profiles. The online survey battery administered by the FYE program included several questions related to students’ demographic backgrounds, including gender. We retrieved gender information for this study.

2.3 Data Analyses

As preliminary analyses, a series of descriptive analyses and exploratory factor analysis (EFA) were first conducted on the total raw scores of the Revised PSVT:R. Means and standard deviations of the Revised PSVT:R total scores were computed with participants as a whole and by gender. In addition, frequency distributions of the total scores were produced by gender to examine any distributional differences by this factor. Then, an EFA was performed to investigate whether there exists a single underlying factor structure in the Revised PSVT:R, because the test intended to measure one sub-dimension of spatial ability (i.e., the spatial visualization ability of mental rotation). The Mplus 6.0 program was used for the analysis because the Revised PSVT:R produces a binary response for each item. The Mplus 6.0 employs the robust weighted
least squares estimator (WLSMV) method for estimation, which is a recommended estimation method to handle categorical data in factor analysis. This analysis also served in examining the unidimensionality of the data, which is the major assumption of IRT-based analysis. The estimated eigenvalues through EFA were summarized as the scree plot in Figure 2. As shown in the figure, the eigenvalue of the first factor was 9.32, and it shows a large difference from the eigenvalue for the second factor (2.67), represented by the steep slope of the line connecting the first and second factor. The eigenvalues for the rest of factors are similar and the line connecting these factors becomes horizontal. The results suggest that the Revised PSVT:R has a single factor structure, and the unidimensionality assumption for the IRT analysis was also satisfied.

![Scree plot of eigenvalues](image)

*Figure 2. Scree plot of eigenvalues*

Next, the item analyses applying both Classical Test Theory (CTT) and Item Response Theory (IRT) were conducted to address the first and second questions. The CTT and IRT are the two major frameworks frequently used in measurement research. The CTT defines the observed test scores with two components: a person’s true ability score and measurement error. This measurement framework has been used in a variety of testing situations because of the simplicity of its theoretical model, weak theoretical assumptions, and small sample size requirement for applying the framework in practice. In a CTT-based framework, the item difficulty (the $p$ value) is defined as the proportion of examinees who successfully answered a particular item. The item discrimination is typically defined as the point-biserial correlation between responses (correct or incorrect response coded as a dummy variable) on a particular item and the total raw scores. One of the largest drawbacks of applying the framework is that these item statistics depend on the sample characteristics utilized for the analysis and the examinees’ observed scores (e.g., the Revised PSVT:R total raw scores) will be also determined by the difficulty of the items used for testing.

On the other hand, the IRT is a measurement framework used to estimate the probability of obtaining a correct response on a particular item at examinee’s ability level independently of examinees’ group characteristics and the items used for testing. In the framework, a family of probabilistic models is mathematically defined with given item parameters, including item
difficulty, item discrimination and the guessing factor. Although IRT-based analysis shows some advantages over CTT-based analysis, the application of IRT is generally restrained by its strong theoretical assumptions and large sample size requirement.

Based on the CTT framework, Cronbach’s alpha coefficient of internal consistency was computed using IBM-SPSS 18. Item statistics for each item, such as item difficulty (i.e., percent-correct value) and item discrimination (i.e., item-total correlation) were also computed using BILOG-MG 3.0\textsuperscript{43}. Following the CTT-based item analyses, one parameter logistic (1-PL) IRT model was utilized to generate the item parameter estimates (i.e., an item difficulty parameter for each item and an item discrimination parameter applicable to all items) and the individual’s ability parameter estimates based on examinees' response patterns on the Revised PSVT:R. We chose the 1-PL model because it functions appropriately with a small number of examinees like the current sample size\textsuperscript{41}. The Chi-square item-fit index was used to assess the goodness-of-fit of the 1-PL IRT model to the sample data\textsuperscript{40}.

Further, Pearson product-moment correlation coefficients were computed between the Revised PSVT:R total scores and other academic variables (i.e., SAT composite or ACT composite score, and high school GPA) to address the third research question. Finally, a confirmatory factor analysis (CFA) was conducted to address the last research question. More specifically, the CFA was performed using the Mplus 6.0 program to test whether the theoretical factor structure of the construct measured by the Revised PSVT:R fit the sample data obtained from FYE students. The model fit was evaluated with a Chi-square test as well as multiple CFA fit indices, including, Root Mean Square Error of Approximation (RMSEA), Comparative fit index (CFI) and Tucker Lewis index (TLI).

3. Results

3.1 Descriptive Statistics of the Revised PSVT:R Total Scores

The frequency distribution of scores on the Revised PSVT:R was reported by gender in Figure 3. The Y-axis represents the percent of examinees within each gender group. The frequency distribution of male students ($N = 480$) is negatively skewed compare to the distribution of female students ($N = 105$). The score range of males is also wider than that of females. In addition, 4.4 % of males ($N = 23$), compared to 1.9% of females ($N = 2$) correctly solved all items on the Revised PSVT:R.
Means and standard deviations of the Revised PSVT:R total scores were reported in Table 1. On average, male students answered 77% of the items correctly ($M = 23.14$), while female students correctly answered 73% of the items on the Revised PSVT:R ($M = 21.92$). Considering the fact that the homogeneity of variance assumption for the independent sample $t$-test was satisfied, the independent sample $t$-test was performed to examine the gender differences in average total scores on the Revised PSVT:R. The result indicates that male students significantly outperformed female students ($t[583] = 2.214, p = .027$) with a mean difference of 1.12. However, Cohen’s $d$ effect size is small ($d = .239$).

Table 1. Descriptive statistics of raw score total

<table>
<thead>
<tr>
<th></th>
<th>$N$</th>
<th>$M$</th>
<th>$SD$</th>
<th>Min.</th>
<th>Max.</th>
<th>No. of Perfect Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>105</td>
<td>21.92</td>
<td>5.030</td>
<td>11</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>Male</td>
<td>480</td>
<td>23.14</td>
<td>5.111</td>
<td>4</td>
<td>30</td>
<td>21</td>
</tr>
<tr>
<td>Total</td>
<td>585</td>
<td>22.92</td>
<td>5.114</td>
<td>4</td>
<td>30</td>
<td>23</td>
</tr>
</tbody>
</table>

3.2 Item and Test characteristics of the Revised PSVT:R

Score reliability. Cronbach’s alpha coefficient of internal consistency for the Revised PSVT:R with the current sample was .84, which is considered to be reasonably high and indicates that at least 84% of the total score variance is due to true score variance. All 30 items of the test appeared to be worthy of inclusion because the removal of any items does not increase the score reliability.

Item characteristics. Item difficulty and item discrimination estimates were reported in Table 2. Looking at item difficulty estimates represented by percent-correct values, most of the items were answered correctly by more than half of the examinees, except for Item 22 (47.4 percent-
correct) and Item 30 (36.2 percent-correct). Under the CTT framework, these two items are considered to be difficult items for the given sample. On the other hand, Item 4 (93.5 percent-correct) is considered as the easiest among these 30 items. Item discrimination represented by item-total correlations range from .160 to .523, indicating that the discrimination power of the items varies, although all items appear to contribute to correctly discriminate examinees by the level of ability to some extent. For example, the correlation was low for the five easiest items (Item 1 through Item 5), meaning that these items are less useful in differentiating examinees by their levels of spatial visualization ability. This is because almost all examinees could identify the correct response on these items regardless of their level of spatial visualization ability. For the rest of the items, they discriminate moderately between examinees with low and high total scores.

Following the item analyses based on the CTT, the analyses with 1-PL IRT model were conducted. The estimate of item discrimination parameter ($a$) for all 30 items is 1.102. The estimate of item difficulty ($b$) for each item in the Revised PSVT:R is also reported in Table 2. Item difficulty parameters range from -2.872 (Item 4, the easiest item) to 6.42 (Item 30, the most difficult item) with the mean of -1.457, which means that the PSVT:R consists of relatively easy items. The result is consistent with the findings regarding item difficulty based on CTT. Chi-square item fit indices were also included in Table 2. Except for two items (i.e., Items 26 and 29), all the items fit well to the 1-PL IRT model.

<table>
<thead>
<tr>
<th>Item</th>
<th>Item Difficulty</th>
<th>Item Discrimination</th>
<th>IRT ($a = 1.102$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>92.1</td>
<td>.185</td>
<td>-2.667</td>
</tr>
<tr>
<td>2</td>
<td>92.3</td>
<td>.160</td>
<td>-2.691</td>
</tr>
<tr>
<td>3</td>
<td>92.6</td>
<td>.171</td>
<td>-2.740</td>
</tr>
<tr>
<td>4</td>
<td>93.5</td>
<td>.229</td>
<td>-2.872</td>
</tr>
<tr>
<td>5</td>
<td>87.5</td>
<td>.287</td>
<td>-2.144</td>
</tr>
<tr>
<td>6</td>
<td>85.0</td>
<td>.421</td>
<td>-1.918</td>
</tr>
<tr>
<td>7</td>
<td>89.4</td>
<td>.361</td>
<td>-2.334</td>
</tr>
<tr>
<td>8</td>
<td>82.2</td>
<td>.317</td>
<td>-1.706</td>
</tr>
<tr>
<td>9</td>
<td>92.1</td>
<td>.345</td>
<td>-2.667</td>
</tr>
<tr>
<td>10</td>
<td>84.3</td>
<td>.401</td>
<td>-1.862</td>
</tr>
<tr>
<td>11</td>
<td>76.2</td>
<td>.438</td>
<td>-1.311</td>
</tr>
<tr>
<td>12</td>
<td>72.5</td>
<td>.464</td>
<td>-1.093</td>
</tr>
<tr>
<td>13</td>
<td>66.3</td>
<td>.295</td>
<td>-0.769</td>
</tr>
<tr>
<td>14</td>
<td>81.5</td>
<td>.426</td>
<td>-1.657</td>
</tr>
<tr>
<td>15</td>
<td>75.7</td>
<td>.330</td>
<td>-1.280</td>
</tr>
<tr>
<td>16</td>
<td>80.5</td>
<td>.293</td>
<td>-1.585</td>
</tr>
<tr>
<td>17</td>
<td>67.7</td>
<td>.312</td>
<td>-0.838</td>
</tr>
<tr>
<td>18</td>
<td>84.6</td>
<td>.383</td>
<td>-1.890</td>
</tr>
<tr>
<td>19</td>
<td>77.3</td>
<td>.377</td>
<td>-1.373</td>
</tr>
<tr>
<td>20</td>
<td>78.5</td>
<td>.444</td>
<td>-1.449</td>
</tr>
<tr>
<td>21</td>
<td>75.0</td>
<td>.392</td>
<td>-1.239</td>
</tr>
</tbody>
</table>
The item characteristic curve (ICC) and its corresponding item information function (IIF) were reported for the easiest (Item 4) and the most difficult (Item 30) items in Figure 4. The ICC represents the probability of obtaining the correct response by examinees given the ability level. For example, ICC for Item 4 shows that the probability of obtaining a correct response is 0.5 when examinee’s ability level is -2.87, and the probability increases monotonically as the ability level increases. On the other hand, ICC for Item 30, which is the most difficult item in the test, indicates that the probability of obtaining a correct response is 0.5 at the ability level of 0.642. The ICC of Item 30 also indicates that the probability of identifying a correct response on Item 30 is close to 0 with the ability of -2.87, while the probability of correct response on Item 4 is 0.5 at the same ability level.
Figure 4. Item characteristic curve and item information function for the easiest item (item 4) and the hardest item (item 30)

Item information function (IIF) indicates how precisely a particular test item measures the target latent trait, which is spatial visualization ability in this study, at each level of the ability. For example, the IIF of Item 4 indicates that the information is maximized at -2.872, meaning that the item will produce the most precise estimate of the spatial ability at the relatively low ability level. However, the precision of the ability estimate will significantly decrease as one’s ability level increases. Similarly, the IIF of Item 30 indicates the maximum information at 0.642, meaning that the item functions best at the ability level in terms of the precision of ability estimation.

Figure 5 shows the test information function (TIF), which was produced by summing up IIFs of all items in the Revised PSVT:R. It shows that the test information is maximized at the ability level of -1.5, indicating that the Revised PSVT:R functions most precisely when it is used to measure the examinees with the relatively low ability level. This result is consistent with the fact that the 1-PL IRT item difficulty parameters are mostly located below zero (i.e., an average ability level) except two items (Items 22 and 30). Finally, the total raw score computed by counting the number of correct responses and their corresponding ability scores estimated with the 1-PL IRT model examinees’ ability score ($\theta$) was reported in Table 3.
Figure 5. Test information function (TIF) and standard error (SE) as a function of ability level ($\theta$).

Table 3. The compatible estimated ability score ($\theta$) for each raw total score

<table>
<thead>
<tr>
<th>Raw Total Score</th>
<th>$\theta$</th>
<th>Raw Total Score</th>
<th>$\theta$</th>
<th>Raw Total Score</th>
<th>$\theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>-2.88</td>
<td>15</td>
<td>-1.31</td>
<td>24</td>
<td>0.03</td>
</tr>
<tr>
<td>6</td>
<td>-2.45</td>
<td>16</td>
<td>-1.24</td>
<td>25</td>
<td>0.27</td>
</tr>
<tr>
<td>8</td>
<td>-2.23</td>
<td>17</td>
<td>-1.11</td>
<td>26</td>
<td>0.47</td>
</tr>
<tr>
<td>9</td>
<td>-2.15</td>
<td>18</td>
<td>-0.91</td>
<td>27</td>
<td>0.68</td>
</tr>
<tr>
<td>10</td>
<td>-2.02</td>
<td>19</td>
<td>-0.70</td>
<td>28</td>
<td>0.97</td>
</tr>
<tr>
<td>11</td>
<td>-1.83</td>
<td>20</td>
<td>-0.54</td>
<td>29</td>
<td>1.31</td>
</tr>
<tr>
<td>12</td>
<td>-1.62</td>
<td>21</td>
<td>-0.44</td>
<td>30</td>
<td>1.70</td>
</tr>
<tr>
<td>13</td>
<td>-1.46</td>
<td>22</td>
<td>-0.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>-1.37</td>
<td>23</td>
<td>-0.19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Criterion-related validity evidence. Pearson product-moment correlation coefficients among academic variables were reported in Table 4. The correlation coefficients between the Revised PSVT:R and aptitude test scores (SAT or ACT) were positive and the magnitude of the relationship is weak to moderate ($r = .292$ with ACT composite scores; $r = .318$ with SAT composite scores). On the contrary, the correlation between the Revised PSVT:R and high school GPA was negligible ($r = .068$).

Table 4. Pearson product-moment correlation coefficients among academic variables

<table>
<thead>
<tr>
<th></th>
<th>SAT Composite</th>
<th>High School GPA</th>
<th>Revised PSVT:R</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT Composite</td>
<td>.793**</td>
<td>.335**</td>
<td>.292**</td>
</tr>
<tr>
<td>SAT Composite</td>
<td></td>
<td>.422**</td>
<td>.318**</td>
</tr>
<tr>
<td>High School GPA</td>
<td></td>
<td></td>
<td>.068</td>
</tr>
</tbody>
</table>

Note. **$p < .05$.**
Construct-related validity evidence. A CFA was conducted to test the fit of the data obtained from the Revised PSVT:R to the hypothesized factor structure, a single factor model. Factor loadings are given in Table 5. Factor loadings were reasonably high, ranging from .313 to .709 with $M=.535$ and $SD=.106$. The Chi-square test is well known for its sensitivity to large sample size\(^{37, 40, 41}\). As expected, the Chi-square test ($\chi^2 = 670.01$, $df = 405$, $p < .001$) indicates a poor fit of the model. However, the other fit indices revealed evidence that the single factor model was a good fit to the sample data. For example, the RMSEA (Root Mean Square Error of Approximation) is 0.033, suggesting a good fit because it is less than 0.05\(^{44}\). The comparative fit index (CFI) and Tucker Lewis index (TLI) were 0.924 and 0.918, which indicate good fit as well because the values were close to 1.0\(^{37}\). Results from both the EFA and CFA support that the Revised PSVT:R measures a single factor.

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor Loading</th>
<th>Item</th>
<th>Factor Loading</th>
<th>Item</th>
<th>Factor Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.355</td>
<td>11</td>
<td>.629</td>
<td>21</td>
<td>.571</td>
</tr>
<tr>
<td>2</td>
<td>.313</td>
<td>12</td>
<td>.656</td>
<td>22</td>
<td>.416</td>
</tr>
<tr>
<td>3</td>
<td>.336</td>
<td>13</td>
<td>.427</td>
<td>23</td>
<td>.505</td>
</tr>
<tr>
<td>4</td>
<td>.464</td>
<td>14</td>
<td>.649</td>
<td>24</td>
<td>.480</td>
</tr>
<tr>
<td>5</td>
<td>.487</td>
<td>15</td>
<td>.477</td>
<td>25</td>
<td>.572</td>
</tr>
<tr>
<td>6</td>
<td>.642</td>
<td>16</td>
<td>.438</td>
<td>26</td>
<td>.709</td>
</tr>
<tr>
<td>7</td>
<td>.610</td>
<td>17</td>
<td>.440</td>
<td>27</td>
<td>.686</td>
</tr>
<tr>
<td>8</td>
<td>.508</td>
<td>18</td>
<td>.590</td>
<td>28</td>
<td>.536</td>
</tr>
<tr>
<td>9</td>
<td>.633</td>
<td>19</td>
<td>.542</td>
<td>29</td>
<td>.515</td>
</tr>
<tr>
<td>10</td>
<td>.618</td>
<td>20</td>
<td>.654</td>
<td>30</td>
<td>.596</td>
</tr>
</tbody>
</table>

4. Discussion

Engineers will often be required to visualize and represent their ideas involving abstract objects on paper or computer screens to communicate with others about the ideas graphically\(^{45, 46}\). Abundant research findings support a positive relationship between spatial ability, especially spatial visualization ability and/or mental rotation skills, and success in engineering courses\(^{7, 18, 21, 47}\). One of the frequently used spatial ability tests in the field of engineering is the Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT:R). However, our review indicated that no empirical study in which the item and test functions of the test were evaluated by utilizing psychometric theories is currently available. Thus, the study was conducted to investigate the psychometric properties of the Revised PSVT:R for use with FYE students after a revision of errors found by Yue (2006, 2007)\(^{33, 34}\). We first summarized our finding for each research question and then, if any, we indicated a potential future direction of research related to the question.

How reliable are scores on the Revised PSVT:R for first-year engineering students?

With the sample of 585 FYE students, we found a Cronbach’s internal consistency reliability of 0.849, which indicates high score reliability for the use of FYE students. While the magnitude of
reliability depends on the characteristics of the sample to whom the test was administered, the obtained reliability for the Revised PVST:R is compatible to those reported with other engineering student cohorts by Sorby and Baartmans ($r = .82$) (2000) and with general undergraduate cohorts (e.g., $r = .81$ reported by Alkateeb, 2004; $r = .86$, Guay & McDaniel, 1978).

**To what extent do characteristics such as item difficulty and item discrimination vary across the items in the Revised PSVT:R?**

Guay (1976) attempted to order items by difficulty in the test and the analyses of item difficulty conducted by applying both CTT and IRT frameworks that supported the item ordering by difficulty level in general. However, we found that the fourth item of the test is the easiest, while the last item is the most difficult as Guay intended. In addition, although the test functions well to discriminate the level of spatial visualization ability among FYE students, we found that, in general, items are relatively easy to solve for the population. Furthermore, the IRT-based analysis indicated that two items did not fit the 1-PL model. We acknowledge that item discriminations vary across items from the result of the CTT-based analysis and the fact that the examinee’s guessing may be unavoidable for selecting a response due to the multiple-choice item format. However, given the number of the sample size, we could not extend our IRT-based analysis by fitting the more complex IRT models to investigate differences in discrimination power and guessing probability for selecting a response among items. Thus, additional IRT-based analysis with complex models is worth conducting with larger sample size.

**To what extent is the Revised PSVT: R supported by the criterion-related validity evidence?**

Pearson’s correlation coefficients between the revised PSVT:R scores and ACT composite scores and SAT composite scores are .292 and .318, respectively. However, we speculate that the correlations between the PSVT:R and the subscores by subject area in these standardized tests may be different. In particular, it would be interesting to discover how ACT science subtest scores relate to the PSVT:R. Thus, additional analysis of the correlations between the Revised PSVT:R scores and sub-scores of these aptitude tests may be appropriate. Nevertheless, the results indicated that the scores on the Revised PSVT:R measure a different construct from the one measured by these aptitude tests.

One of the most important analyses related to the criterion-related validity, which we could not investigate in the current study, is to evaluate predictive validity of an inference made on the Revised PVST:R scores. Because one of the main purposes of using the test scores in a FYE program is to predict their future academic performance from the PSVT:R, it is worthwhile, for example, to investigate how the performance on the Revised PSVT:R is related to the first semester GPA in the FYE program. Furthermore, it may be interesting to establish evidence to support the use of the test scores to predict student retention in the engineering program. The evidence of predictive validity of the inferences will provide further implication for the use of the spatial ability scores in instruction and curriculum design to ascertain the role that spatial ability plays in students’ academic success.
To what extent is the Revised PSVT:R supported by the construct-related validity evidence?

While several studies reported converged-validity evidence to support construct-related validity \(8, 49, 50\), this study was significant because it tested the adequacy of the theoretical assumption of a single latent factor measured by the Revised PSVT:R. The results of EFA and CFA supported the unidimensionality of the construct measured by the Revised PSVT:R. We consider that all items in the Revised PSVT:R contribute to measuring one of the subcomponents in spatial ability, spatial visualization ability.

5. Conclusion

Cognitive assessments are used in a variety of instructional settings in the current educational system. On some occasions, we make serious educational decisions based on assessment results. For example, providing appropriate educational guidance for selecting courses based on assessment results is one of the main reasons assessments are used in educational settings.

The PSVT:R is one of the cognitive assessments used to measure examinees’ mental rotation ability and has been used to make inferences about examinees’ future academic success in STEM fields. The PSVT:R has been also used as a placement test for graphics courses in engineering education. However, without a full understanding of the nature of the test, fair use of the test in making educational decisions cannot be guaranteed. In particular, the validity of inferences drawn from test scores is incredibly important for interpreting test scores accurately and using them appropriately.

This study was conducted to investigate the psychometric properties of the Revised PSVT:R for use with FYE students. We focused on this population because the PVST:R has been used frequently with this population. Thus, we believe that identifying item functions specific to this population can help future users of the test enhance their understanding of test scores within this population. In summary, the study provided sound and detailed psychometric information regarding the Revised PSVT:R for current and future users of the test to support appropriate use and interpretation of the scores.

However, the study also raised an interesting issue. According to Hegarty and Waller (2004)\(^{52}\),

“Performance on tests of spatial abilities depends on execution of basic cognitive processes such as encoding a visual stimulus, constructing a visual image, retaining an image in working memory, transforming an image, and comparing a visual stimulus to an image in working memory.” (p. 136)

Mumaw, Pellegrino, Kail and Carter (1984)\(^{51}\) found that individual differences in spatial ability tend to occur at any step of the cognitive processing. They further indicated that the speed of cognitive processing when solving spatial tasks produces differences in individuals’ test scores. Although we could identify the level of item difficulty, further investigation is needed to understand what characteristics of the items contribute to the determination of item difficulty. Therefore, it may be interesting to scrutinize the features of 3-D objects in the Revised PSVT:R
and to investigate their relationship to the level of cognitive processing and processing speed. Finally, while this study may serve as a prototype for conducting measurement research on the Revised PSVT:R with other populations, additional studies are needed to address these unsolved questions.

Bibliography


