*International Journal of Multidisciplinary Thought,*
CD-ROM. ISSN: 2156-6992 :: 1(06):265–274 (2011)


# A QUALITY STUDY OF THE SRINAKHARINWIROT UNIVERSITY ENTRANCE TEST (SWU-SAT) USING ITEM RESPONSE THEORY AND CLASSICAL TEST THEORY APPROACHES

**Suthiwan Pirasaksopon, Surachai Meechan, and Somkit Kitpoonwong***
*Srinakharinwirot University, Thailand*

The purpose of this study was to examine the quality of Srinakharinwirot University - Scholastic Aptitude Entrance Tests (SWU-SAT) using the classical test theory and item response theory method. In the first tryout, there were four subtests for each program on SWU-SAT, Science and Art program, analyzed by classical test theory. The result of the analysis found that the number of items achieving the p and r value of each test was 14, 26, 33 and 25 in Science Program and 21, 25, 28 and 29 in Art program. The result of the second tryout revealed that the number of items achieved the p and r criteria in CTT and the b and a in IRT on SWU-SAT51 in each test was not significantly different. The implementation of the tests for selecting students to study in Srinakharinwirot University academic 2009 year found that the item analysis of discriminant analyzed by CTT and IRT was positive significant relationship in Verbal, Spatial Test of Science program, except there were no relationship on Reasoning Test of Science program, Numerical Test and Reasoning Test of Art program. In contrast, there was negative significant relationship between that of difficulty as calculated by CTT and IRT. Furthermore, the differential item function on SWU-SAT51 found that the Reasoning Test of Science program had the most number of items in female group, and also in Verbal test of Art program.

**Keywords:** classical test theory, item response theory, differential item function

## INTRODUCTION

The Educational and Psychological Test Bureau was assigned by Srinakharinwirot University to construct and develop a standardized test called SWU-SAT (Srinakharinwirot University Scholastic Aptitude Test) used for selecting students to further study in the university. This project was aimed to continue from fiscal year 2008 to 2011 and it was run by a committee comprising experts from faculties or institutes of the university. In fiscal year 2008, the committee succeeded in creating SWU-SAT51 which contained verbal, numerical, reasoning, and spatial aptitude assessment. It was used in selecting high school students of Science and Arts Programs to pursue their Bachelor's Degree at Srinakharinwirot University through the direct admission system.

A significant stage of the test construction and development procedure was quality examination of empirical research tools at which test-takers' responses were analyzed in order to

find the difficulty and discrimination level of the tests. Nowadays the theory which is widely used in test analysis is Classical Test Theory. However, the theory's weak point is that the analysis results are most likely to vary upon samples. Unlike Classical Test Theory, Item Response Theory is able to reflects the relationship between test-takers' ability and probability in selecting correct answers. Moreover, tests' parameters which are difficulty and discrimination values will not vary upon samples. Therefore, Item Response Theory was also be used in this research. In addition to the tests' quality examination which relied on an analysis of the tests' parameters using the above two theories as well as a comparison of the number of appropriate items, the research also focused on whether the exam papers function differently when there was a difference in the test-takers' gender. The research results would be beneficial for the construction and development of SWU-SAT in the next fiscal years.

## METHOD

### The construction and development of SWU-SAT51

The tests were constructed and developed in accordance with determined factors and definitions by content experts and measurement specialists. The tests' questions and answers were then deliberately adjusted. There were four tests in a test: verbal, numerical, reasoning, and spatial. Forty items of each test were chosen.

### The First Tryout

SWU-SAT51 exams were taken in January 2008 by Mathayomsuksa Six students, academic year 2008, from all regions of Thailand. The samples obtained consisted of 1485 to 1723 students from both Science and Arts programs. The exam results were examined for the item difficulty and discrimination by Classical Test Theory approach. The number of SWU-SAT51 items for both Science and Arts program were found to meet the criterion with p-value ranging from .20 to .80 and r exceeding .20. Having considered each test of Science program tests, there were 14, 26, 33 and 25 items of the verbal, numerical, reasoning, and spatial tests found eligible. As for the Arts program items, there were 21, 25, 28 and 29 items of the verbal, numerical, reasoning and spatial tests that conformed to the criterion.

### The Second Tryout

The second tryout was conducted in June 2008 on Mathayomsuksa Six students and freshmen in the provinces randomly chosen from each region of the country. There were 815 to 1503 samples from both Science and Arts programs. The SWU-SAT51 test results were analyzed for descriptive statistics, reliability, and item difficulty and discrimination indices by using Classical Test Theory (CTT) and Item Response Theory (IRT) methods. The passed items of CTT method had the difficulty (p) between .20 and .80 with the discrimination (r) above .20. The passed items deriving from IRT approach had the difficulty (b) between -2 and +2 with the discrimination (a) between 0 and 2. The second tryout revealed that the spatial test of the Science program tests had the highest average value. The lowers were the reasoning, verbal, and numerical tests respectively. Also, the spatial test of SWU-SAT51 for Arts program had the highest average value. However, the verbal test was the next lower followed by the reasoning

and numerical tests respectively. It was noted that the average test scores of the spatial tests for both Science and Arts programs were higher than half of the total score, and only verbal test for Science program had a higher average test score than half of the total. This could be summarized that the SWU-SAT51 tests were likely to be difficult for the samples of the second tryout. The items with highest variation was the numerical ones. The next lower were the reasoning, spatial, and verbal items for both Science and Arts (see table 1). No difference was found in a comparison of the number of appropriate items of each skill tests between CTT and IRT methods (see table 2).

Before using IRT method, the tests' unidimensionality was assessed by means of principal component analysis and varimax rotation. The Eigen value of the first factor was higher than that of the second factor, and not different from the other factors. It could be concluded that each of the SWU-SAT51 tests for both Science and Arts can measure one dimension.

**The Implementation of SWU-SAT51 in Entrance Examination**

Thirty items were selected from each qualified aptitude tests derived from the second tryout. They were then used in Srinakharinwirot University's direct admission in academic year 2009. The verbal, reasoning, and spatial tests were taken by Science program students (Mathematics Learning Achievement Tests developed by certain faculties were used instead of the numerical aptitude tests). Students from Arts program must take all of the four aptitude tests. The examination results were inquired into correlation coefficients between the difficulty and discrimination values analyzed by CTT method and those analyzed by IRT method. The research results showed that the correlation coefficient between the discrimination values ($r_{pbi}$) obtained from CTT and IRT methods of each of the SWU-SAT51 tests was found to be positive at the .01 level of statistical significance, except that of the reasoning tests for both Science and Arts programs as well as the numerical tests for Arts program. In addition, there was no relationship between the discrimination values gained from the two methods. The correlation coefficient between the difficulty values resulted from CTT and IRT methods of all SWU-SAT51 tests was found to be negative at the .01 level of statistical significance (see table 3).

The results of Differential Item Functioning (DIF) analysis of the tests indicated that the verbal tests for Science program had only one DIF item. The reasoning tests had the highest number of DIF items once gender classification was considered. In the SWU-SAT51 tests for Arts program, DIF was least found in the numerical tests. When gender variables were involved, the verbal tests were found to have the greatest number of DIF items (see table 4).

## DISCUSSION

The purpose of this research was to examine the quality of SWU-SAT51 used to select students to further study at the Bachelor's Degree level in the academic year 2009 through the direct admission system of Srinakharinwirot University. At the beginning of the construction and development process of SWU-SAT51, a committee composed of experts and specialists conjointly determined the components and their definitions. There were 12 tests created but only seven of them were used in an entrance examination. Forty items of them were selected to test Mathayomsuksa Six students in January 2008 with an aim to find the item quality: difficulty and discrimination values by using Classical Test Theory. The test results were analyzed as to improve the ineligible items. Some items were adjusted if they were found to be vague and

unclear. After the revision, the tests were taken by another group of Mathayomsuksa Six students and some freshmen in June 2008.

**Table 1**. Mean, Standard Deviation, Coefficient Variation and Reliability.

| SWU-SAT51 | $\bar{X}$ | S | C.V. | $r_{tt}$ |
|---|---|---|---|---|
| **Science Program** | | | | |
| Verbal | 20.8849 | 4.5102 | 0.2160 | .6006 |
| Numerical | 15.1768 | 6.1341 | 0.4042 | .7870 |
| Reasoning | 18.3430 | 7.1884 | 0.3919 | .8378 |
| Spatial | 25.0344 | 7.7906 | 0.3112 | .8943 |
| **Arts Program** | | | | |
| Verbal | 16.4609 | 4.1461 | 0.2519 | .5451 |
| Numerical | 11.876 | 4.819 | 0.4058 | .6783 |

| | | | | |
|---|---|---|---|---|
| | 1 | 7 | | |
| Reasoning | 16.5165 | 5.8383 | 0.3535 | .7622 |
| Spatial | 24.2361 | 6.7696 | 0.2793 | .8587 |

**Table2 A comparison of the number of items that passed the criterion of quality analysis by means of CTT (p and r) and IRT (b and a) methods.**

| SWU-SAT51 | CTT | | Passed items | IRT | | Passed items | Z-test |
|---|---|---|---|---|---|---|---|
| | p | r | | b | a | | |
| **Science Program** | | | | | | | |
| Verbal | 36 | 20 | 22 | 28 | 40 | 28 | 1.150 |
| Numerical | 36 | 31 | 30 | 31 | 40 | 27 | 0.494 |
| Reasoning | 39 | 35 | 35 | 34 | 39 | 34 | 0.00 |
| Spatial | 31 | 40 | 31 | 40 | 39 | 39 | 2.366 |
| **Arts Program** | | | | | | | |
| Verbal | 31 | 20 | 18 | 22 | 40 | 22 | 0.671 |
| Numerical | 28 | 39 | 27 | 36 | 40 | 36 | 2.186 |
| Reasoning | 39 | 30 | 30 | 34 | 39 | 34 | 0.839 |
| Spatial | 28 | 39 | 27 | 40 | 35 | 35 | 1.874 |

**Table 3** Correlation coefficients between item parameters using CTT and IRT Methods.

| SWU-SAT51 | R between Rpbi- and a -value | R between p- and b-value |
|---|---|---|
| **Science Program** | | |
| Verbal | .927** | -.854** |
| Numerical | -.152 | -.936** |
| Reasoning | .893** | -.983** |
| Spatial | | |
| **Arts Program** | .850** | -.986** |
| Verbal | .301 | -.896** |
| Numerical | .253 | -.950** |
| Reasoning | .850** | -.986** |

**Table 4** A comparison of the results of DIF analysis of the SWU-SAT51 tests when the test-takers were classified by gender.

| SWU-SAT51 | Number of DIF items | Number of DIF items in percentage | The DIF items | DIF by gender |
|---|---|---|---|---|
| **Science Program** | | | | |
| Verbal | 1 | 3.33 | 8 | male |

| | | | | |
|---|---|---|---|---|
| Reasoning | 8 | 26.67 | 7, 8, 9, 10, 11, 13, 14, 15 | female |
| Spatial | 4 | 13.33 | 19, 22, 23, 27 | male |
| **Arts Program** | | | | |
| Verbal | 10 | 33.33 | 1, 4, 5, 8, 9, 11, 29<br>19, 20, 27 | female<br>male |
| Numerical | 3 | 10.00 | 7, 23<br>16 | female<br>male |
| Reasoning | 6 | 20.00 | 1, 2, 4, 12, 25<br>19 | female<br>male |
| Spatial | 5 | 16.67 | 14<br>16, 21, 22, 29 | female<br>male |

The second outcome showed that most of the tests were rather difficult. Average test scores were lower than half of the total score. There were only three tests that had average scores above half of the total: the spatial tests for both Science and Arts programs and the verbal tests for Science program. Moreover, the coefficients of variation of the numerical tests for both Science and Arts programs were found to be more variable than the other tests. As for reliability, the spatial tests for both Science and Arts programs had the reliability over .80. The lowest average scores belonged to the verbal tests for both Science and Arts programs. This might result from having insufficient knowledge and experience as the samples had been studying on Mathayomsuksa Six for only two months. Besides, the tested freshmen were not studying in a real scientific faculty even though they had studied in Science or Arts programs; thus, the average test scores were lower than half of the total score and there was much difference in the scores' variation. As mentioned earlier that the spatial tests possessed the highest average scores, this might be because the tests involved geometric pictures that tested the samples' abilities in viewing 3D and composite pictures, answering the spatial tests would not rely on any particular story. Without language ambiguity, the spatial tests were then not complicated for the test-takers. The fact that language is ambiguous became an obstacle in designing the tests. When the questions were unclear, the test-takers would not be able to understand the main point of the questions. Besides, Language use varied upon context and the age of the test-takers which might not correspond to the test-constructors'. Accordingly, the reliability of the verbal tests was quite low.

In comparing the number of items that passed the criterion, the passed items of CTT method must have the p-value between .20 and .80 with the r-value above .20. By IRT method the passed items must have the b-value between -2 and +2 with the a-value falling between 0 and 2. The comparison showed no difference between the two methods. This could be explained that all of the items were designed and verified by specialists. The questions and answers were improved and tried out. The tryout results were then analyzed and used in developing both questions and answers so they would be more obvious and effective. The analysis results were also consistent with Maneerat Buaban (2003) who studied the effect of the difference in question order towards the number of passed items gained from CTT and IRT methods. This research yielded the same result when the question arrangement was based on the content's difficulty values.

Regarding the use of the 30 appropriate items drawn from each aptitude test in Srinakharinwirot University's direct admission in academic year 2009 with an aim to investigate correlation coefficients of the difficulty and discrimination values derived from CTT and IRT methods, the discrimination values obtained from CTT and IRT methods were found to have positive relationship in the verbal and spatial tests of both Science and Arts programs at the .01 level of statistical significance. This meant that the discrimination values of the two methods acted in the same manner. In other words, an item that had a high $r_{pbi}$-value would have a high a-value as well. Relationship of the discrimination values was not found in the reasoning tests of both Science and Arts program as well as the numerical tests of Arts program. Because of the reasons that the texts used in the reasoning tests of Science program had some associative characteristics and the Science program students normally disfavored linguistic matters, the Science program students did not perform well in the reasoning tests. This might lead to the decrease of the correlation coefficients of the discrimination values gained from both methods and the absence of relationship.

The calculation of difficulty value based on the two methods resulted that the correlation coefficients were negative at the .01 level of statistical significance in every skill tests because the difficulty values of CTT and IRT were oppositely interpreted. By CTT method, the lower the p-value was, the more difficult the tests became. On the other hand, the more the b-value of IRT method approached -2 or +2, the easier or harder the tests would be. Accordingly, the negative correlation coefficients gained from CTT and IRT methods reflected negative relationship of the difficulty values of the two methods. What discovered from this research matched former research papers even though they involved other subject matters. Those were research papers of Watana Khudsee (1990), Tiemjit Puangsomjit (1995), and Sayhompoo Raksang (2004) who found that the correlation coefficients of the difficulty values derived from CTT and IRT methods had some relationship. This research was also consistent with Wiberg's (Wiberg. 2004) who theoretically investigated the Swedish citizens' driving tests and found that a- and $r_{pbi}$-value had a high relationship level at .753, b- and p-value had a highly negative relationship level at -.861. As for this research, the correlation coefficients of the tests, in which the relationship between a- and $r_{pbi}$-value were found, ranged from .850 to .927; and the correlation coefficients between b- and p-value ranged from -.854 to -.986. The level of relationship found in this research was higher than Wiberg's.

The investigation of differential item functioning of SWU-SAT51 revealed that the reasoning tests for Science program had the largest number of DIF items, especially in female students. The verbal tests for Arts program had the largest number of DIF items, especially found in female students. In taking the verbal and reasoning tests, the students had to rely on reading, understanding, and analyzing skills; the tests were, therefore, suitable for the female test-takers.

Some consistency was found among this research and the research of Kasorn Hwangchit (1996) and Ratchanok Yeesunesri (2001) who studied the differential functioning of Thai, English, and Mathematics tests, and found than the tests functioned differently according to the test-takers' gender.

## RECOMMENDATION

### Recommendation for implementation

The quality analysis results of SWU-SAT51 should be presented to experts and specialists who will be a committee in the following fiscal years as they can be used as tools for test construction and development. This will encourage an increase of a potentiality in achieving the tests, especially verbal and reasoning, of students with equal competency no matter which learning programs they belong to. Since the verbal and reasoning tests comprise textual questions, a student must be able to read and interpret in order to find a correct answer.

### Recommendations for future research

There should be a quality assessment of research instruments such as using factor analysis in detecting construct validity and the LISREL model in homogeneity testing.

There should be a study of the SWU-SAT51 test scores as to whether the scores had relationship with other variables such as graduate students' grade point average and O-net scores.

## REFERENCES

Allen, Mary J. & Yen, Wendy M. (1979). Introduction to Measurement Theory. Belmont California : Brook/Cole Publishing.

Hambleton, Ronald K. & Swaminathan, Hariharan. (1985). Item Response Theory : Principles and Applications). Boston : Kluwer - Nijhoff Publishing.

Kasorn Hwangchit. (1996). An analysis of Differential Item Functioning based on Mantel-Haenszel *procedure in Thai and English language graduate entrance examination tests*. M.A. Theses. Bangkok : Chulalongkorn University.

Maneerat Buaban. (2003). *The effects of item arrangements on qualities of Science Multiple-choice Tests when analyzed by using classical test theory and item response theory*. MA. Thesis. Mahasarakham : Mahasarakham University.

Rakchanok Yeesunesri (2001). *An analysis of differential functioning of items and tests based on dfit procedures in English and mathematics for university entrance examination*. M.A. Theses. Bangkok : Chulalongkorn University.

Sayhompoo Raksang. (2004). *A study of relationship of the item analysis using classical test theory (CTT) and item response theory (IRT)*. MA. Theses Bangkok: Srinakharinwirot University.

Sirichai Karnchawasee. (2002). *Modern Test Theories.* Faculty of Education: Chulalongkorn University.

Tiemjit Puangsomjit. (1995). *A study of the relationship between the test analysis by the classical model and rasch model in Thai Reading Comprehenstion multiple choice test*. MA. Thesis Bangkok: Srinakharinwirot University.

Watana Khudsee. ( 1990). *A study of the relationship between parameter gained from item analysis and test ability scores using classical test theory and item response theory of multiple-choice test.* MA. Theses Bangkok: Srinakharinwirot University.

Wiberg Marie. (2004). *Classical Test Theory vs. Item Response Theory : An evaluation of the theory test in the Swedish driving-license test*. UMEA UNIVERITET.

Weraphan Prombut. (1993). *A study of relationship between parameters which analyses by classical test theory and item response theory*. M.A. Theses. Phitsanulok : Naresuan University