Journal of Research Methodology, Volume 20, Number 1 (January-April 2007)

วารสารวิธีวิทยาการวิจัย ปีที่ 20 ฉบับที่ 1 (มกราคม-เมษายน 2550)

Using Rasch Model and Bookmark Standard Setting Procedure to Establish a Cutscore on STOU-TBS Test

Sungworn Ngudgratoke
Ratchaneekool Pinyopanuwat
Nalinee Na Nakorn

ABSTRACT

The Sukhothai Thammathirat Open University (STOU)'s test of basic skill (STOU-TBS) measures three domains of basic skill: mathematics, verbal, and logical reasoning. The primary objective of this study was to establish a cutscore on the STOU-TBS test that decides whether an examinee passes the STOU-TBS test by demonstrating the desired level of proficiency. The cutscore on the STOU-TBS was achieved by using the Rasch measurement model in conjunction with the bookmark standard setting procedure. The data used for the bookmark standard setting procedure was a STOU-TBS test data from the 2004 administration. The accuracy of the derived cutscore was evaluated using the Posterior Probability of Passing (PPoP) curve. The result was that the final cutscore was 1.118 which was an appropriate cutscore because the classification error rates associated the final cutscore were acceptably low. More specifically, the false passing and false failing rates were 0.0026 and 0.0972 respectively, and the total classification error rate was 0.0998.

Introduction

Like Scholastic Achievement Test (SAT) commonly used for college admission purpose by almost universities in the United States, the Sukhothai Thammathirat Open University's test of basic skill (STOU-TBS) measures three components which are mathematics, verbal, and logical reasoning, These components are considered the desirable basic skills of STOU undergraduates. The latest version of the STOU-TBS test was developed in 2004. For this version, each component consists of 20 dichotomously scored items. The STOU-TBS test was mainly used to diagnose if new prospective freshmen possess desired level of basic skill a significant factor facilitating them to study successfully in the distance higher education context. New STOU students who take the STOU-TBS test and receive low score on the test are encouraged to take one or more remedial courses, depending on skill for which they are incompetent, in order to enhance their knowledge and skills before taking regular courses in their program.

The processes of development and refinement of the STOU-TBS have been carefully undertaken so as to meet the Standard for test development (APA, AERA, & NCME, 1999). Following guidelines of test development as indicated in the Standard enables STOU-TBS to be appropriate and meaningful quantitative snapshots of student's skills. To ascertain that the scores on the STOU-TBS test function well in diagnosing students' skill, it is necessary to evaluate the psychometric properties of the STOU-TBS test. Typically, reliability and validity evidence need to be documented to be used to evaluate psychometric property of the test. According to the Standard, validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests (AERA, APA, & NCME, 1999) and reliability refers to the consistency of measurements when the testing procedure is repeated on a population of individuals or groups. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. High reliability coefficient indicates that test scores are free from measurement error (Brennan, 2001). STOU test developers have documented such evidences and showed that STOU-TBS was a valid and reliable measure of basic skill.

Considered as an early stage of development of the STOU-TBS, STOU used classical test theory to analyze student's responses to STOU-TBS test in order to develop scale scores and student performance on the STOU-TBS test was determined by simply a total score which is the total count of correct responses. Wright and Mok (2000) highlight the ineffective interpretation of count scores because they are always illusion. They suggested that, to develop a meaningful metric, "the counts must be incorporated into a stochastic process which constructs inferential stability" and the Rasch measurement model can provide a logit metric suited to this purpose.

To classify students into two performance standard categories namely pass and fail, it is necessary to obtain an appropriate cutscore that could be used as a cut point for making a classification decision about performance of examinees who took the STOU-TBS test. Standard setting methodologies have been developed and used to establish one or more cutscores on test that examinees must meet or exceed to demonstrate that they have met a performance standard (Cizek, & Bunch, 2007; Reckase, 2006), where the term performance standard is sometimes used interchangeably with terms such as cutscore, standard, and passing score (Cizek, & Bunch, 2007).

Even though several standard setting methodologies have been used, the uses of standard setting methods that use item response theory (IRT) as an integral ingredient have increased exponentially in recent years. There are two major standard setting methods that combine the advantages of IRT with experts' judgments. One is the item-mapping method (Wang, 2003) and another is the bookmark method (Mitzel et al., 2001). Developed independently, these two methods share three significant characteristics. The first common characteristic is that both methods attempt to provide less cognitively complex approaches for judges to determine a cutscore, as compared to the Angoff method, by integrating IRT methodologies into standard setting process (Wang, 2003). The second similar characteristic is that judges are not required to provide percentage estimates of item mastery for the minimally competent examinees (Wang, 2003; Cizek, & Bunch, 2007). The last similarity which is the most unique is that the uses of both methods require items to be reordered according to their item difficulties estimated through the applications of an IRT model and then judges are instructed to select the item from a group of items arranged in

difficulty order such that a group of minimally qualified examinees has a chance of answering less than a selected response probability (RP). An item's difficulty that is determined consensually from a group of judges is a cutscore. Note that the way to present items to a group of judges and how many items to be picked up by judges are somewhat different between the item-mapping the bookmark methods.

This present study used the bookmark standard setting together with the Rasch model to establish a cutscore on the STOU-TBS. The bookmark standard setting was used because it has been specifically developed for educational assessment setting (Wang, 2003) and it is well known and widely used in several large-scale assessment program (Huynh, 2006).

The bookmark procedure is named because participants or judges express their judgments by entering markers in a special designed booklet consisting of a set of items placed in difficulty order, with items ordered from easiest to hardest. This booklet is called an ordered item booklet (OIB). The bookmark standard setting has two key activities. The first one is that it is necessary to create an OIB in which items are usually reordered from easiest items to hardest items. The second major activity is that judges or standard setting participants are asked to review items in the OIB and then they will be instructed to put a bookmark at the first item in the OIB at which the chances of a minimally qualified examinee answering correctly drop below 0.67. This implies that participants indicate that the items before the bookmark item (marker) represent content that a minimally qualified examinee would be expected to master at the RP specified (Cizek, & Bunch, 2007). The RP criterion of 0.67 has been widely used in the field of educational testing because it has been proved that it would maximize information carried in the correct response (Huynh, 2006). Moreover, for each individual judge, the item before the bookmark item is used to establish a cutscore. That is, the median of cutscores provided by all judges is used as a cutscore.

The Purpose of This Study

The purpose of this study was to establish the cutscore on the STOU-TBS test by using Rasch measurement model and the bookmark standard setting procedure. Before

doing so, we assessed the fit of data to Rasch measurement model and the items that fit Rasch measurement model were then used for the bookmark standard setting procedure. Then accuracy of the established cutscore was evaluated using the Posterior Probability of Passing (PPoP) curve (Wainer, Wang, Skorupski, and Bradlow, 2005). According to PPoP curve, the false passing and false failing rates as well as their sum were calculated using PPoP curve to be used as indices for assessing the accuracy of the cutscore.

Method

Sample

The STOU-TBS test consisting 60 dichotomously scored items was administered to 2,318 freshmen in 2004 at the 36 regional test centers. The sample consisted of 1,330 female (57.4%) and 998 male (42.6%). Table 1 shows that examinees had diverse backgrounds in terms of gender and field of study.

Table 1 Characteristics of STOU students taking STOU-TBS test in 2004

School	Gender		Total
	Male	Female	
Liberal Arts	48	85	133
Communication Arts	34	74	108
Educational Studies	36	106	142
Business Administration	226	574	800
Law	232	128	360
Health Sciences	16	19	35
Nursing	5	22	27
Economics	13	30	43
Human Economy	6	102	108
Political Science	221	124	345
Agricultural Extension and Cooperation	129	49	178
Science and Technology	22	17	39
Total	**988**	**1,330**	**2,318**

Standard Setting Procedures

The followings described how a cutscore was derived from a panel of judges through Rasch model and the bookmark standard setting procedure.

1. Item and Person calibration

This step involves analyzing the dichotomously scored item response data to create the scale using WINSETPS (Linacre, & Wright, 1999). Any individuals who receive all items right or all items incorrect will be eliminated from the analysis because their responses provide no information about the difficulty of the items. In the present study, the dichotomous Rasch measurement model is used and this model is expressed by:

$$\ln\left(\frac{P_{ni}}{1-P_{ni}}\right) = B_n - D_i \qquad (1)$$

or

$$P_{ni} = \frac{\exp(B_n - D_i)}{[1+\exp(B_n - D_i]} \qquad (2)$$

where P_{ni} is the probability of person n with ability B_n succeeding on item i which has difficulty level D_i (Wright, & Mok, 2000).

To evaluate if data fit to the Rasch measurement model, Smith (2000) recommends that the standardized fit index (Z) computed by a cube-root transformation of the mean square value be used to interpret item fit, rather than the mean square values. The corresponding Z values represent standardization of the mean squares models to approximate the unit normal distribution. A rule of thumb for assessing item fit has been discarding any item with Z value greater than +/-2.0 (Schumacker, 2004). A Z value greater than 2.0 would indicate an unexpected or irregular response pattern across items, i.e., noise or lack of unidimensionality. A Z value less than 2.0 would indicate possible redundancy in item response, i.e., a lack of expected stochastic fit or violation of local item independence. The Z standardized fit index is used to identify both misfitting item and misfitting person response patterns.

2. Construct validity and reliability analysis

In this study, construct validity evidence for items that had acceptable fit indices was obtained by performing principal component analysis (PCA) using WINSTEPS. Typically the objective of this analysis was to assess if unidimensionality was present in the selected set of items. Construct validity is supported if the unidimensionality assumption holds. In addition to PCA, item and person reliability coefficients were obtained directly from the WINSTEPS analysis.

3. Bookmark standard setting Procedure

The bookmark standard setting procedure was used to derive a cutscore on the STOU-TBS from a group of experts. Specifically, a three-round standard setting procedure was used. However, before starting the standard setting procedure, judges were asked to discuss the characteristics of the minimally competent examinees, after an introduction to the bookmark standard setting procedure and STOU-TBS. Note that the standard setting was to establish a single cutscore for classifying examinees into two categories: pass and fail. Intuitively, a minimally competent examinee is an examinee who obtains low score on STOU-TBS test but high enough to pass the exam. In other words, it is believed that a minimally competent examinee has a score on STOU-TBS above or equal to the intended cutscore (ICS). Later, they discussed the item contents among themselves.

Before implementing the bookmark standard setting procedure, it was necessary to assemble OIB, The standard setting facilitators calculated the location of items on the ability sale using the following equation (Cizek & Bunch, 2007):

$$B_n = D_i + .708. \qquad (3)$$

Note that the equation (3) was derived from (2) by replacing P_{ni} with .67 and then solved for B_n. RP criterion of 0.67 was used in this study because item information is maximized at RP criterion of 0.67 (Huynh, 2006).

To set the passing score, seven judges consisting of educational measurement experts, high school mathematics teachers, and STOU faculty were used to set a cutscore on the STOU-TBS test. The panel of judges was broken into two groups of three and four

judges. Again, three rounds of bookmark standard setting were organized. For the first round, each judge read individually through a booklet of items, and then was instructed to place his/her bookmark on the first item that has a probability of answering correctly less than 0.67.

For the second round, judges explained why they placed bookmarks where they did. Other members of the group were invited to disagree, if they felt the placement was inappropriate. Following the discussion, judges placed their bookmarks a second time and then a cutscore for each judge was obtained. For the last round, all two groups convened as one group. Each judge shared his or her bookmark placement from the second round with the large group, followed by further discussion of rationale for particular bookmark placements. The judges then placed their bookmarks third and a final. A final cutscore was the median of judges' cutscores from the third round.

4. Evaluation of the cutscore

This study used the Posterior Probability of Passing (PPoP) that was proposed by Wainer, Wang, Skorupski, & Bradlow (2005) to assess quality of the final cutscore that was chosen through expert judgments. The PPoP is a Bayesian method for evaluating passing score. This method usually first begins with fitting data to an item response theory (IRT) model such as Rasch model using Markov chain Monte Carlo (MCMC) procedure. MCMC with Gibbs sampler provides samples from posterior distribution of examinee proficiency (θ). To calculate the probability of an examinee scoring above the passing score, a sample of proficiency (θ) from posterior for each examinee will be drawn and then we can count how many times the sampled proficiency is above the passing score (Wainer, Wang, Skorupski, & Bradlow, 2005). Then the probability of passing ($P(\theta)$) for that examinee is obtained by dividing that count by the number of draws. For example, if 1,000 draws were drawn independently of one another and 600 are above the passing score, $P(\theta)$ is 0.6.

In this study, we fitted data to the Rasch model using WinBUGS 1.4 (Spiegelhalter, Thomas, Best, & Lunn, 2003). Every 5th draw of the remaining 1,000 proficiency parameters for each examinee was drawn after the first 1,000 were discarded. To plot the PPoP curve, calculated $P(\theta)$ and then plotted $P(\theta)$ against examinee proficiency (θ).

The PPoP curve was also used to calculate the false passing and false failing rates. The sum of false passing and false failing rates is the total error rate. In any infinite sample, the probability of the false pass is $\int_{-\infty}^{c} P(\theta)dF(\theta)$, where c is the cutscore and $F(\theta)$ is the ability distribution, and the probability of false fail is $\int_{c}^{\infty} [1-P(\theta)]dF(\theta)$ (Wainer, Wang, Skorupski, & Bradlow, 2005). However, in finite sample, the integrals need only to be replaced by summations over the population of examinees. In this study we used SAS program to compute the false passing and false failing rates.

Results

1. Item and Person Calibration

Item and person calibration was performed using WINSTEPS and the result showed that 19 out of 60 items had reasonably good fit statistics. 41 items did not fit the Rasch measurement model and were not used in the bookmark standard setting procedure. Table 2 shows estimates of difficulty of 19 items estimated by WINSTEPS. The transformed values (θ_i) in the third column of Table 1 were the location on the ability scale at which the probability of a correct response was .67. These numbers were used to assemble the OIB. Note that these numbers were transformed using the equation (3).

Table 2 Item Difficulty and location of 19 items

Item Number	Difficulty(D_i)	θ_i
7	1.19	1.898
56	1.11	1.818
39	1.03	1.738
25	0.89	1.598
45	0.71	1.418
29	0.69	1.398
37	0.41	1.118
23	0.34	1.048
38	0.33	1.038
50	0.21	0.918
58	0.12	0.828
27	0.03	0.738
55	-0.01	0.698
54	-0.06	0.648
35	-0.33	0.378
8	-1.11	-0.402
32	-1.38	-0.672
20	-1.64	-0.932
3	-2.52	-1.812

2. Construct validity and reliability analysis

Construct validity of the 19 items that were chosen based on their adequate fit indices was evaluated by examining construct validity or unidimensionality assumption. Construct validity was assessed by performing principal component analysis (PCA) using WINSTEPS. As seen in Figure 1, the latent trait PCA analysis revealed that four dimensions accounted for variation in the data. However, the first factor which was the Rasch factor dominantly accounted for such variation (eigen value=13.4), while the second, third and fourth factors had relatively low eigen values (1.4, 1.2, and 1.2, respectively). This result implied that for the data consisting of selected 19 items measuring basic skill was unidimensional. In other words, there was an evidence to conclude that the selected 19 items measure a single dimension.

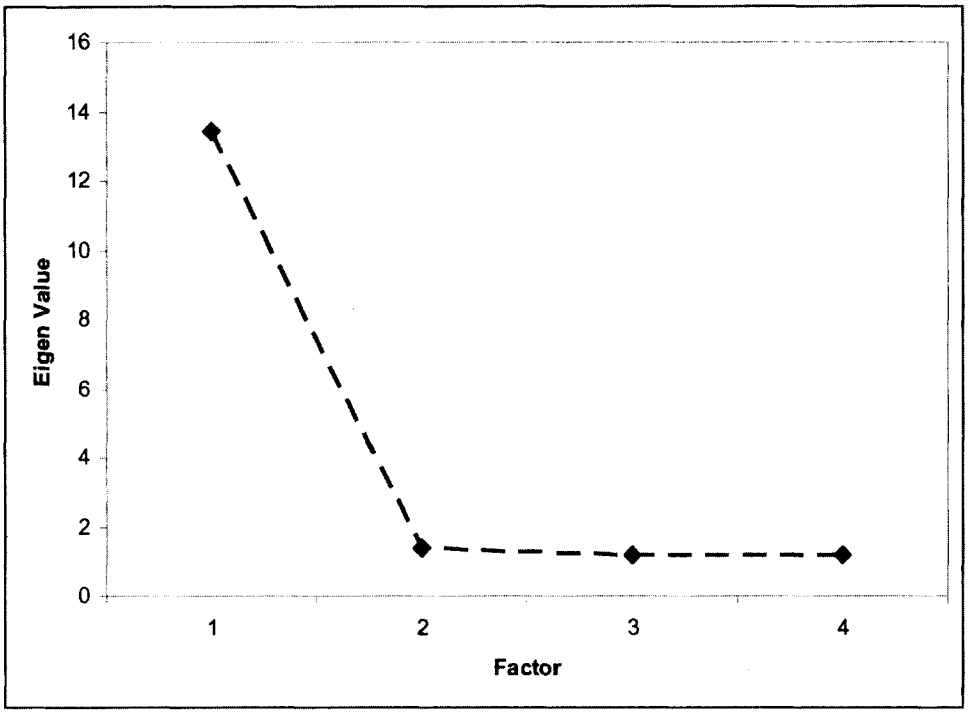

Figure 1 Latent Trait PCA Scree Plot

The WINSTEPS program provides reliability information shown in Figure 2. The person reliability index equals 0.57 which was interpreted similar to a Cronbach Alpha reliability coefficient, so this means the examinees responded in a fairly consistent fashion across the set of 19 items. The item reliability index equals 1.00 (with item separation of 19.10) which is very good and indicates adequately dispersed items on the scale of basic skill.

SUMMARY OF 2313 MEASURED (NON-EXTREME) ABILITIES

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	6.9	19.0	-.69	.57	1.00	.0	1.00	.0
S.D.	2.9	.0	.91	.09	.23	.9	.39	.9
MAX.	18.0	19.0	3.21	1.08	1.82	3.3	5.23	3.6
MIN.	1.0	19.0	-3.39	.50	.49	-2.2	.17	-1.8

REAL RMSE .60 ADJ.SD .69 SEPARATION 1.14 ABILIT RELIABILITY .57
MODEL RMSE .58 ADJ.SD .71 SEPARATION 1.24 ABILIT RELIABILITY .60
S.E. OF ABILITY MEAN = .02

SUMMARY OF 19 MEASURED (NON-EXTREME) BASICS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	845.4	2313.0	.00	.05	1.00	-.1	1.00	-.1
S.D.	428.4	.0	.98	.00	.02	1.1	.04	1.0
MAX.	1921.0	2313.0	1.19	.06	1.04	1.9	1.08	1.7
MIN.	375.0	2313.0	-2.52	.05	.96	-1.7	.94	-1.9

REAL RMSE .05 ADJ.SD .98 SEPARATION 19.10 BASIC RELIABILITY 1.00
MODEL RMSE .05 ADJ.SD .98 SEPARATION 19.21 BASIC RELIABILITY 1.00
S.E. OF BASIC MEAN = .23

Figure 2 Item and Person Reliability Information

3. Bookmark standard setting

The bookmark standard setting took place in January 2007. The meeting of a group of experts consisting of 2 high school mathematics teachers, 3 STOU faculty who have taught statistics and educational measurement courses, and 2 experts in verbal and logical assessment was organized. The first part of this meeting centered on developing a performance level label (PLL) for a passing score. Judges mentioned that the minimally competent examinees should be able to (1) do mathematics computations, (2) interpret and understand information received through a various formats such as texts, numbers, pictures, and graphs, (3) explore simple relationships among objects presented in the item questions.

Table 3 shows the medians of cutscores from Round 1, 2 and 3. Even though the Round 2 and 3 cutscores were exactly same, it was evident that in round 2 there was a less consensus among participants about the appropriate cutscore. When Round 2 was completed,

a facilitator provided participants the proportion of examinees who passed the exam by using the cutscore obtained from round 2 as a cut point, participants discussed about appropriateness of the cutscore, and they viewed that the cutscore was very rigorous. The result of Round 3 shows that nearly all participants lowered their bookmarks, especially judge 1 and 2 changed their cutscores from 1.398 to 1.118. The result of Round 3 also showed a greater consensus among participants and for this Round the final cutscore was 1.118. However, even though participants lowered their bookmarks, the final cutscore was still rigorous; about 2.33% of examinees passed the STOU-TBS exam when the final cut was used as a cut point.

Table 3 Outputs from Round 1, 2 and 3 of Bookmark Standard Setting Procedure

Participant ID No.	Round 1		Round 2		Round 3	
	Item at Cut	Cutscore	Item at Cut	Cutscore	Item at Cut	Cutscore
1	45	1.418	29	1.398	37	1.118
2	29	1.398	29	1.398	37	1.118
3	38	1.038	23	1.048	38	1.038
4	38	1.038	37	1.118	37	1.118
5	29	1.398	23	1.048	37	1.118
6	37	1.118	38	1.038	37	1.118
7	29	1.398	37	1.118	37	1.118
		Median = 1.398		Median = 1.118		Median = 1.118

To show graphical representation of the consensus among judges, figure 3 showed that the consensus was a downward trend. That is, the median of judges' cutscores dropped from 1.398 to 1.118 from Round 1 to Round 3. The medians of judges' cutscores for Round 2 and 3 were unchanged but greater consensus was achieved in Round 3. Figure 3 also showed that the judges who set very high very high standard in round 1 tended to lower their bookmarks in round 2, while the judges who set relatively low standard did not change their bookmarks dramatically.

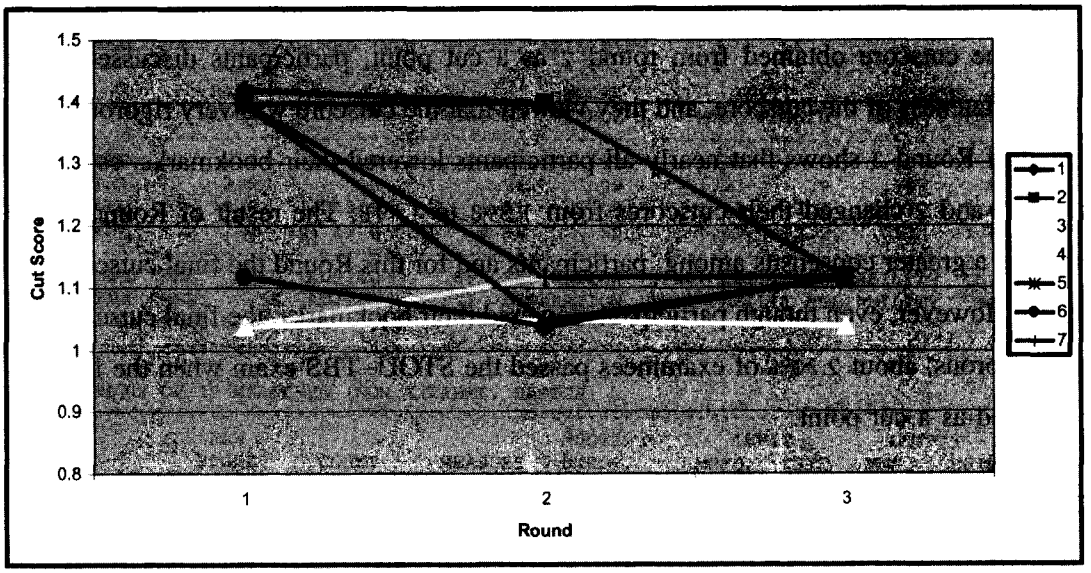

Figure 3 Judges's bookmarks from Round 1 to 3

4. Evaluation of the cutscore

The PPoP curve in Figure 4 had an S-shape. This curve shows that examinees with higher proficiency had more chance of passing the exam than examinees with lower proficiency. The point on the probability of passing ($P(\theta)$) where the PPoP curve crossed with the cutscore of 1.118 was approximately 0.5, meaning that examinees with proficiency equals to the cutscore has approximately 50 % chance of passing the exam. The slope of the PPoP curve measures "the reliability of the decisions made on the basis of the test and that particular passing score" (Wainer, Wang, Skorupski, & Bradlow, 2005). As seen in Figure 4, the slope of the PPoP curve at the cutscore was steep, meaning that the derived cutscore could be used to make a good classification decision on classifying examinees into pass and fail categories.

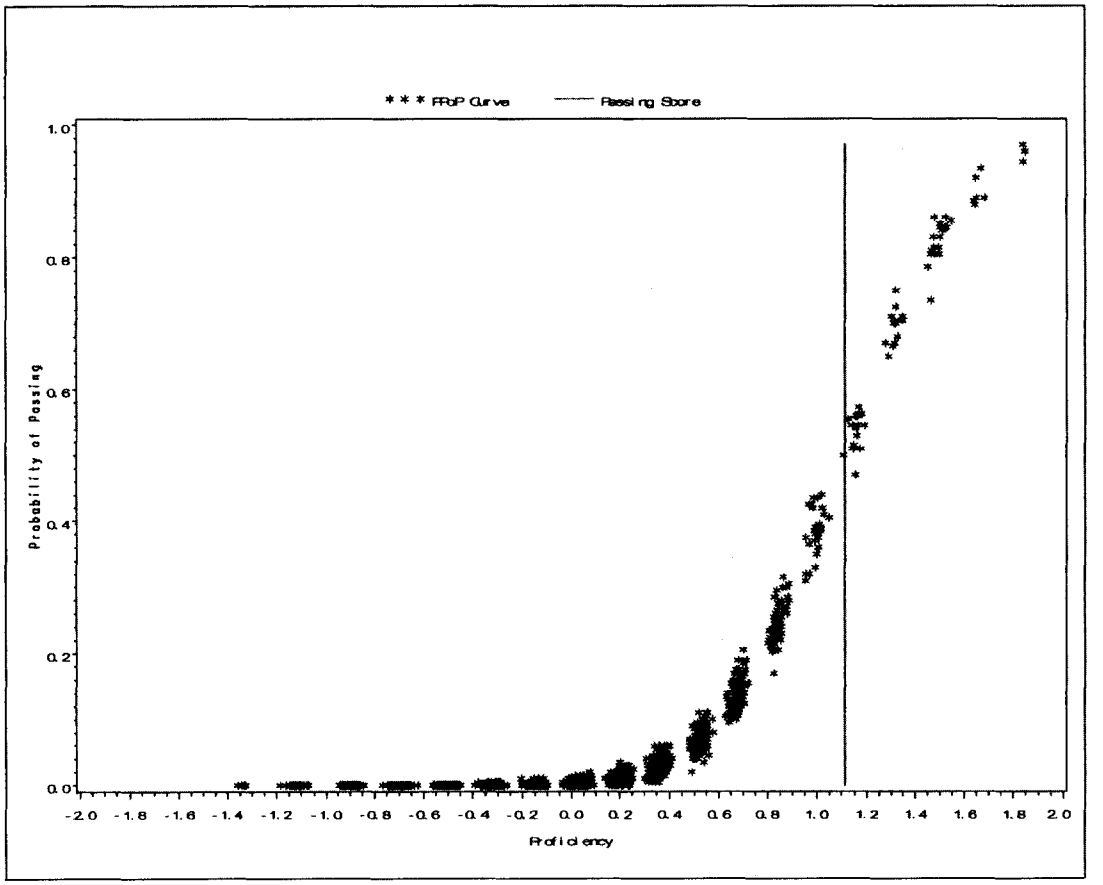

Figure 4 The probability of passing curve for the basic skill assessment of the STOU-TBS

We used the PPoP curves to evaluate the accuracy of the cut score. The result indicated that the probability of a false passing was 0.0026 which was very small, meaning that a chance of passing for examinees scoring below the cutscore is about zero. The probability of a false fail was 0.0972, meaning that approximately 9.72% of examinees with ability above or equal to the cutscore would be falsely classified as fail. The total error rate was 0.0998 (0.0026+0.072).

Because the final cutscore was rigorous, we have tried to lower the performance standard by choosing a cutscore of 0.918 which was slightly lower than 1.118, and then the accuracy of the 0.918 was evaluated using the PPoP curve. Note that when a cutscore of 0.918 was used, about 4.18 percent of examinees passed the exam. The result showed

that when the cutscore of 0.918 was used, the probability of a false passing was 0.0131 and the probability of a false fail was 0.1148. These resulted in the total error rate of 0.1279. This result indicated that the cutscore of 0.918 would produce higher classification errors than the cutscore of 1.118. This indicated that the final cutscore from the bookmark standard setting was relatively more appropriate.

Conclusions and Discussions

This present study used the Rasch model and the bookmark standard setting procedure to establish a cutscore on the STOU-TBS test that decides whether an examinee passes the STOU-TBS test by demonstrating the desired level of proficiency. The cutscore was aimed to be used for classifying STOU students into two performance levels (pass and fail). Three rounds of bookmark standard setting were organized to yield a greater consensus among seven bookmark standard setting participants. After the final cutscore was achieved, the Posterior Probability of Passing (PPoP) of curve was used to evaluate the accuracy of the final cutscore. The major result was that the final cutscore was 1.118 which was a rigorous cutscore because it was evident that approximately 2% of student taking STOU-TBS passed the exam when the final cutscore of 1.118 was used as a cut point.

For the present study, the Rasch model and the bookmark standard setting procedure provided an acceptable cutscore as evident by low classification errors that were obtained by using the PPoP curve. The final cutscore had total classification error of 0.0998. This implies that when the cutscore of 1.118 was used to classify examinees into pass and fail categories, about 10% of students would be misclassified. The false passing rate was close to zero, meaning that examinees with ability less than the cutscore would have no chance of passing the exam. However, the false failing rate was 0.0972, implying that about 10% of students having ability more than or equal to the cutscore could be falsely classified as fail. It was also evident that the total classification error was contributed by the false failing rate by very large degree; however, false passing rate has an ignorable effect contributed to the total classification error. This was not surprised because the cutscore was very rigorous.

Note that this result was obtained when a response probability (RP) of 0.67 was used to establish the cutscore. In practice, there is no agreement among users of the item-mapping and the Bookmark methods about which appropriate RP criterion we can use to obtain the best cutscore. Wang (2003), in Rasch model context, preferred RP criterion of 0.5 because "the item information will be maximized when the probability of success is 0.5." However, Huynh (2006) noted that the item information is maximized at RP criterion of 0.67. The results from this study found that the total classification error when RP criterion of 0.67 was used was lower than that when RP criterion of 0.5 was used. Future bookmark standard setting research should be conducted by using both RP criterion of 0.5 and 0.67 so that the classification errors resulting from the two different RP criteria can be compared meaningfully.

Even though the evaluation of the final cutscore showed that the bookmark standard setting procedure and the Rasch model yielded an acceptable final cutscore on the STOU-TBS. This study has limitations. First, only 19 items were used because other 41 items did not fit the Rasch model. The selection 19 items from the STOU-TBS to be used for setting a performance standard may not completely represent the hypothetical construct the STOU-TBS is supposed to be measuring. The STOU-TBS has been designed to measure the three major skills: mathematics, verbal, and logical reasoning. This test specification implies that when STOU-TBS test data were calibrated using the Rasch model, the unidimensional assumption of the Rasch model would not be satisfied. Thus, it was obvious that some STOU-TBS items did not fit the Rasch model because many items of STOU-TBS seem to measure more than one construct. The further study should be conducted by fitting the Rasch bifactor model to data rather than the unidimensional Rasch model in order to relax the stringent assumption of unidimensionality. Alternatively, use the Rasch model to estimate difficulty and ability parameters for each component of the STOU-TBS and then bookmark standard setting procedure can be implemented to set a cutscore for the individual components by the STOU-TBS.

Another limitation of the standard setting employed in this study was that a small group of bookmark standard setting participants was used. As mentioned earlier, the final

cutscore was psychometrically acceptable; however it was rigorous. That is, a small proportion of students could pass the exam and thus students and the university administrators may not be pleased with the test results. Therefore, we anticipate that more standard setting participants might bring more experiences and discussions among participants and then a more appropriate cutscore could be achieved.

References

American Psychological Association, American Educational Research Association, and National Council on Measurement (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Brennan, R. L. (2001). An Essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*(4), 295-317.

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standard on tests.* Thousand Oaks, CA: Sage.

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice, 25*(2), 19-20.

Linacre, J. M., & Wright, B. D. (1999). *A user's guide to WINSTEPS.* Chicago: MESA Press.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In C. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum.

Reckase, M. D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice, 25*(2), 4-18.

Shumacker, R. E. (2004). Rasch measurement using dichotomous scoring. *Journal of Applied Measurement, 5*(3), 328-349.

Smitt, E.V. (2000). Metric development and score reporting in Rasch measurement. *Journal of Applied Measurement, 1*(3), 303-326.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2003). WinBUGS *1.4 user manual* [Computer software manual]. Cambridge, UK: MRC Biostatistics Units.

Wainer, H., Wang, X. A., Skorupski, W. P., & Bradlow, E. T. (2005). A Bayesian Method for evaluating passing scores: The PPoP Curve. *Journal of educational measurement, 42*(3), 271-281.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of educational measurement, 40*(3), 231-253.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.

Wright, B. D., Mok, M. (2000). Rasch models overview. *Journal of Applied Measurement, 1*(1), 83-106.