Research and Evaluation in Education, 2(2), 2016, 165-180
Available online at: http://journal.uny.ac.id/index.php/reid
Research Article

DETERMINING STANDARD OF ACADEMIC POTENTIAL BASED ON THE INDONESIAN SCHOLASTIC APTITUDE TEST (TBS) BENCHMARK

***[1]Idwin Irma Krisna; [2]Djemari Mardapi; [3]Saifuddin Azwar**
[1]Center of Educational Assessment, Jl. Gunung Sahari Raya Block B No.4, Gn. Sahari Selatan,
Kemayoran, Jakarta Pusat Municipality, 10610, DKI Jakarta, Indonesia
[2]Graduate School of Universitas Negeri Yogyakarta, Jl. Colombo No. 1, Karangmalang,
Caturtunggal, Depok, Sleman, 55281, Yogyakarta, Indonesia
[3]Faculty of Psychology of Universitas Gadjah Mada, Jl. Sosio Humaniora, Bulaksumur,
Caturtunggal, Depok, Sleman, 55281, Yogyakarta, Indonesia

Abstract

The aim of this article was to classify The Indonesian Scholastic Aptitude Test or *Tes Bakat Skolastik (TBS)* results for each subtest and describe scholastic aptitudes in each subtest. The subject of this study was 36,125 prospective students who took the selection test in some universities. Data analysis began by estimating testees' ability using the Item Response Theory, and benchmarking process using the scale anchoring method applying ASP.net web server technology. The results of this research are four benchmarks (based on cutoff scores) on each subtest, characters which differentiate potential for each benchmark, and measurement error on each benchmark. The items netted give a description of the scholastic aptitude potential clearly and indicate uniqueness so that it could distinguish difference in potential between a lower bench and a higher bench. At a higher bench, a higher level of reasoning power is required in analyzing and processing needed information so that the individual concerned could do the problem solving with the right solution. The items netted at a lower bench in the three subtests tend to be few so that the error of measurement at such a bench still tends to be higher compared to that at a higher bench.

Keywords: *Indonesian Scholastic Aptitude Test (TBS), benchmark, scholastic aptitude*

How to cite item:

Krisna, I., Mardapi, D., & Azwar, S. (2016). Determining standard of academic potential based on the Indonesian Scholastic Aptitude Test (TBS) benchmark. *Research and Evaluation in Education, 2*(2), 165-180. doi:http://dx.doi.org/10.21831/reid.v2i2.8465

***Corresponding Author.**
e-mail: idwinirma@gmail.com

Introduction

Selection in relation to the entry of new students into a university has always become an important issue in several countries. It is related to the criteria used in the acceptance of new students who would study at the university. The ratio of the number of student candidates to the small student capacity causes the universities to be compulsorily selective in choosing the candidates that would be their new students. Besides, effectiveness in the selection of new students is also an important matter in higher educational system because the quality of student candidates has an effect on the internal efficiency and quality of the educational program offered (Harman, 1994, p. 313). Effectiveness would be attained when the selection system has an accuracy in prediction so that it would have an effect on efficiency in the economic aspect.

The selection activity for the entry of new students into the university in Indonesia generally uses an achievement test as a reference for decision-making. An achievement test is designed to measure the result of a learning or training program conducted in a controlled condition (Anastasi, 1988, p. 411). Only in 2009 the test of academic potential started to be used as complement of the achievement test. Though 2009 was the year when the test of potential was nationally started to be in use, several universities had started using it earlier.

The test of potential as part of the entry test at the university is also used by developed countries, such as the United States of America, which uses a test of potential called the Scholastic Aptitude Test (SAT). Sweden has developed the Swedish Scholastic Aptitude Test since 1977 (Wedman, 1994, p. 5). Also known as SweSAT, it is designed as a selection test that is fair and in line with the future success of student candidates if they are accepted as new students at the university. One of the institutions in Indonesia developing the test of potential is Centre of Educational Assessment, Office of Research and Development, Ministry of Education and Culture Republic of Indonesia. The development has been conducted since 1990 and since 2000 the test of academic potential has been named *Tes Bakat Skolastik* ('Scholastic Aptitude Test').

The construction of the test of potential or *Tes Bakat Skolastik* (TBS) is based on understanding of intelligence. Some research indicates that the test of potential has a relation with the intelligence test. The results of research by Frey and Detterman (2003) show that the correlation between SAT scores with those of several IQ tests ranges from 0.53 to 0.83, with this giving strong evidence that SAT could also serve as intelligence test. Intelligence and aptitude are cognitive abilities possessed by every individual (Cohen & Swerdlik, 2002, pp. 257, 301). Intelligence refers to the intellectual ability which generally functions in various fields of achievement, while aptitude is a more specific ability used in certain fields of achievement only (Berk, 2000, pp. 316-319). Aptitude serves to predict one's future success which requires special ability.

The test of potential measures learners' reasoning ability more than their memory. The reasoning process is a more specific part in the thinking process, with one, in reasoning, more frequently using the principles of logic (Galotti, 2004, pp.391-392). Reason is used to make a conclusion based on information obtained. In reasoning, each individual has his or her own ways. Psychologists continuously make explorations on general principles related to human experience not restricted to only one type of reasoning.

Conclusion-drawing models which are related to one's logic and thinking process were developed by Johnson-Laird (Solso, 2001, pp.428-429). Some findings related to one's way in reasoning indicate the use of premises in the form of phrases or in the form of illustrations. The reasoning abilities which are measured in TBS consist of verbal reasoning and mathematics applying reasoning concepts. In line with the research by Olatoye and Aderogba (2011), numerical and verbal reasoning could together explain the variance amounting to 38.8% in an aptitude test and the coefficient of correlation between verbal and numerical reasoning of up to 0.713. Numerical ability is the same in domain as verbal ability and general aptitude.

Since 2001, TBS has been used by several state universities as a part of selection. TBS is also used in the selection of new employees at private agencies and several ministries. TBS consists of three subtests, namely, verbal, quantitative, and reasoning subtests. The three subtests measure the same ability, namely, reasoning, presented in the form of verbal logic, mathematical logic, and reasoning ability in evaluating the correctness of a conclusion. The three subtests indicate a sufficiently significant correlation and the highest correlation is between the quantitative and reasoning subtests (Azwar, 2008, p. 12). The development of the TBS items is done in several stages, starting with the stage of writing the item grid through to the stage of item storage in the bank of items.

The process of data analysis determines the test items fit to enter the test-item bank. The item analysis uses the Item Response Theory (IRT) model. The estimation of the testees' ability (called latent trait) in IRT is based on their response to test items. The IRT model specifically describes the relation between ability and item characteristic on one side and the testee's response to the test items on the other. IRT has models which are not limited to types, depending on the number of parameters used to describe the test items. Measurement in psychology and education is usually of the same dimension with different test items and also with different groups.

TBS uses different test packages to measure different groups of people but the dimension measured is the same. Hopefully, the scores obtained from the test could be used to compare one group with another. For that purpose, the processing of test results uses the model called IRT 1 PL or the Rasch model. The Rasch model is called the one-parameter logistic model because it contains only one parameter related to the test item, namely, level of difficulty. Therefore, this model is known as a simple model in IRT (Embretson & Reise, 2000, p.67). Even if there is another factor having an effect on the results, when we measure something that is certain such as the right solution to a test item, only one of the attributes of the two factors is needed.

With empirical data as the basis, it would be difficult to separate the concepts of level of difficulty and level of human ability. Rasch gives a contribution in relation to this matter by providing the Rasch model formula using the concepts of statistical mathematics. The dependent variable is the probability of a person to successfully answer the test item i, with the probability shown as P(Xis =1). The logistic function for the Rasch model is as follows:

$$P(X_{is} = 1 \mid \theta_s, \beta_i) = \frac{\exp(\theta s - \beta i)}{1+\exp(\theta s - \beta i)} \qquad (1)$$

In which θs is a person's ability, βi is the level of item difficulty, and exp (θs – βi) is the natural antilog of the difference in score between the person's ability and level of item difficulty. The level of item difficulty would be obtained at the time the person's ability to answer correctly has the probability of 0.5. The higher the level of item difficulty, the higher the level of the person's ability to answer correctly with a probability of 0.5. Consistent movement to the right along the ICC (Item Characteristic Curve) indicates increasingly more difficult items and increasingly higher levels of ability.

Up to now, improvements have been done continually and used as part of decision making. All this time the test results announced have been in the form of only scores without the accompaniment of interpretations of the scores. The formulation of test result interpretation depends on definitions of the content, level, and cutoff score, which specifically describe the ability descriptor that would be used as reference for policy makers (Ferrara, Svetina, Skucha, & Davidson, 2011, p.5). Deciding the cutoff score is part of determining benchmarks in the test results. Setting a bench could be done by deciding a score to be used as reference. In setting a cutoff score, consistency is required among educational policy makers and psychometry experts (Bejar, 2008, p.4). At the level of higher education, benchmarks could be used as tools in preparing students for the next teaching and learning process and their chances in pursuing a career. The benchmark of SAT based on combined scores is 1550 (Wyatt, Kobrin, Wiley, Camara, & Proestler, 2011, p.13). Participants attaining the benchnark (of 1550) have the advantages of, among others, a greater possibility of enrolling at a higher educational institution with a length of study time of 4 years (rather than 2 years), more possibility of survival up to the second and third academic years, and having a higher FYGPA (First-Year Grade Point Average).

One of the important components in benchmarking is a set of performance standards based on testees' response to questions (Resnick, Nolan & Resnick, 1995, p.454). A description of someone's cognitive process could be obtained by using as a basis for his or her response to a test item given. The process of developing the test and the procedure of establishing performance standards are, in nature, prospective (based on the descriptor level to orient the test development), progressive (based on the content and performance standard articulated at each level), and predictive (using the descriptor level of performance and standards based on theory and empirical evidence). With the setting of the two systems, decision makers could give accurate decisions based on existing information by using the right measurement and evaluation methods.

The method which was employed to define the level of someone's ability and the cutoff score related to the level is called standard setting. Standard setting is a part which is integrated into the development of a test instrument (Cizek & Bunch, 2007, p.247). Standard setting is a method which is related to the coherence between the educational policy and the evaluation system in a country. The activity of determining the standard or cutoff score could be done by using some methods. Most of the existing standard setting methods could be categorized as continuum models.

Continum models are divided in type into models focusing on the test, or test-centered models, and models focusing on testees or examinee-centered models (Jaeger, 1989, p.492). The determination of a minimum completeness criterion is determined not only through government policies, but also by the participants based on tests and based on measuring instruments (Mardapi, Hadi, & Retnawati, 2015, p.39). The test-centered model is based on experts' judgment concerning the test used. The experts judge the ability needed at each test item to estimate someone's ability according to the standards that have been set. The examinee-centered model is based on experts' judgment in grouping people according to level of ability by using several external criteria outside the test scores. The standard setting method is developed to overcome problems in setting performance standards. One of the methods that could be used in setting the standards is the scale anchoring method.

The scale anchoring method is one of the methods of interpreting measurement results (in the form of scores) which describes the ability and competence possessed by the learners at several different values in a scale. The making of the description is related to behavior scale or item mapping. Item mapping starts with the concept of content referencing which is introduced by Bock, et al. (Kelly, 2002, p.377). Content referencing describes IRT and recommends the procedure in which items are placed on a scale of response probability to describe students' ability and comprehension. Item mapping is widely used in interpreting assessments in a large scale like the Young Adult literacy Survey, National Adult Literacy Survey (NALS), National Assessment of Educational Progress (NAEP) and TIMSS (Trends in International Mathematics and Science Study). The continuation of the content referencing and item mapping is called the scale anchoring method (Beaton & Allen, 1992, pp.195-198). In this method, several points in a scale are chosen and then the test items fitting those points are identified.

A test item would be declared fitting a point (or anchor point) when most of the learners related to the anchor point could do the item concerned but those related to the point under it could not do it. The distribution of the test item group that could be answered by most learners at each different point is then studied to obtain a description of ability at each point. Additional critera would be used when there are only a few suitable test items so that the description of someone's ability would be more enriched.

Scale anchoring provides normative information of the knowledge that someone masters based on a construct being measured (Beaton & Allen, 1992, p.192). The basic idea of scale anchoring is to know the ability that someone possesses at a certain point in a scale based on response to an item given and to pay attention to other responses at an adjacent point. The description of someone's ability at determined points might well be unreliable so that one should be careful in determining the points. The points chosen are called anchor points or anchor levels. According to Forsyth (1991), percentiles could be used to determine the points. The procedure could be widely applied on various scales with the purpose of grouping or making a certain characteristic in someone's ability even with a test which is noncognitive in nature or at least the scale is an ordinal one (Beaton & Allen, 1992, p.191). Thus, the method could also be used with the test of potential.

Four numbers established as international benchmarks used as anchor points are the 25th, 50th, 75th, and 90th percentiles (Martin, Mullis, Beaton, Gonzales, Smith & Kelly, 1997; Mullis, Martin, Beaton, Gonzales, Smith & Kelly, 1998). With the percentile value as the basis, corresponding learners' scores are determined. There is a possibility for several test scores of learners not to be exactly the same as these scores so that ranges of scores plus and minus five are given. This range contains students' scores that are homogenous and sufficiently concentrated at each anchor point where adjacent levels are sufficiently far from each other so that it would hopefully enable recognition of inter-level distinction.

The percentage of correct answers to each item is calculated and the criteria as a reference for the inclusion of the items in the benchmark category are determined. A response probability of 50% would result in an item at an anchor point with the students answering correctly and those, otherwise, equaling each other. A response probability of 80% would result in an item that could be answered correctly by 80% of the students but it would possibly become considered an easy item. In order to overcome it, it is determined that the item is interpreted as being mastered when the response probability is 65%.

In scale anchoring, the anchor item at each level hopefully could distinguish adjacent anchor points. To determine that, criteria are needed to identify the item that should be chosen to consider performance at more than one anchor point. Additional criteria for percentage of students attached to a certain anchor point and that of those attached to the anchor point right under it need to be determined. A criterion determined is that the response probability is less 50%, meaning that the students answering the item concerned incorrectly would be more than those answering it correctly. Anchor items are items that reflect learners' conceptual knowledge and comprehension at different scale points expressed with a high value of probability.

The description at each benchmark level had better imply that the students who reach that point have a great possibility of understanding and doing the item concerned. The definition at each level should be carefully considered so that it could distinguish the levelof someone. Ideally, when the evaluation program has a clear definition of a level and intends to use the level, its establishment is done early in the process of test development (Perie, 2008, p.16). It would help the test planner and the test user and also the policy maker in reporting the level of a test that could distinguish the ability and knowledge that someone has.

Measurement results would be more meaningful when the form of the report is easily understood by various circles, able to give policy makers accurate information, and able to minimize the occurrence of errors in interpretation. Interpretation of the potential at each level is greatly needed by *Puspendik* (Centre of Educational Assessment) in the course of reporting test results to stakeholders. A report of test results which includes a description of someone's potential would be more meaningful when it also gives information about errors in measurement. Accurate calculations of measurement errors could be done by using the IRT approach (Geisinger & McCormick, 2010, p.40).

IRT not only gives estimations of test item and testee parameters but also considers what the precision of each parameter estimated is like. The use of information as a term in this context was first raised by Fisher in 1922 to indicate precision of estimation (Keeves & Alagumalai, 1999, p.35). The function of what is termed information in this case is to describe to what extent the model which has been chosen (1PL, 2PL, or 3PL) is able to give information concerning traits-level estimation along a latent-traits scale**.** Thus, the effectiveness of test or test item measurement at each ability level would be able to be measured.

Mathematically, the information function of an item (IF) fulfills the following equation.

$$I_i\,(\theta) = \frac{[P_i{}'\,(\theta)]^2}{P_i(\theta)Q_i(\theta)} \qquad (2)$$

I_i (θ) is the information of item i on θ, $P_i{}'(\theta)$ is a derivation of $P_i(\theta)$ on θ, $P_i(\theta)$ is the response function of the item, and $Q_i(\theta) = 1 - P_i(\theta)$. (Hambleton, Swaminathan, & Rogers, 1991). Equation (2) would be more simple when calculated by using Equation (3).

$$I_i\,(\theta) = \frac{2.89\,a_i^2(1-c_i)}{c_i+e^{1.7a_i(\theta-b_i)}[c_i+e^{1.7a_i(\theta-b_i)}]^2} \qquad (3)$$

When using IRT 1 PL, a =1 and c = 0.

Based on formula (3), information would be higher in level when the value of b is close to θ. The information function of the test is the accumulation of the information function of the items and mathematically fulfills Equation (4).

$$I(\theta) = \Sigma\, I_i\,(\theta) \qquad (4)$$

An amount of information given by the test on θ would be inversely proportional to a precision of ability estimation called the standard error of measurement (SEM). The relation between SEM and test information is expressed as in Equation (5).

$$SE\,(\theta) = \frac{1}{\sqrt{I(\theta)}} \qquad (5)$$

Based on Equation (5), the standard error (of measurement) and the test information are inversely related, with the greater the information, the smaller SE would be. The magnitude of measurement error depends on the number of test items in the test and the quality of the test.

This writing would discuss the classification of each subtest of TBS by using the scale anchoring method, describe the scholastic aptitude of each subtest of TBS according to test item grouping, and estimate the measurement error at each bench. The description of the scholastic aptitude potential and standard error at each benchmark level could be adopted as references for the test developers so that the development of TBS items becomes more effective and efficient.

Method

In its design, the research which was concerned here was descriptive in nature so that it could describe and interpret an object in accordance with the reality in existence. By the means of descriptive research, a description of someone’s potential in line with the benchmark level determined could be obtained. The data used in the research were obtained from the Center of Educational Assessment, Institute of Research and Development, and Ministry of Education and Culture.

The data originated in the results of the selection test for new students’ entry into state universities in Indonesia. The subjects put under analysis were 36,125 in number. The data were dichotomous in form, with any right answer given the score of 1 and any wrong answer given the score of 0. In addition to those raw data, the data that were qualitative in nature were also compiled in the form of an interpretation of analysis results by means of holding FGD (focus group discussion).

The test instrument in the form of the scholastic aptitude test used consisted of 12 different test packages. Each package consisted of 30 verbal subtest items, 20 quantitative subtest items, and 31 reasoning subtest items. The verbal subtest measured the verbal logic ability, namely, the ability in solving problems verbal in nature and containing language elements. There were 4 verbal abilities measured, namely, synonymy, antonymy, analogy, and reading comprehension. The quantitative subtest measured reasoning or numerical logic ability, namely, the ability to solve problems related to numbers by using basic mathematical concepts. The quantitative subtest consisted of sub-subtests of number sequences, arithmetic and algebra, and geometry. The reasoning subtest measured individual logic abilities, including the ability to evaluate the truth of a conclusion and the ability to use logic to construct a conclusion. The reasoning subtest was divided into three sub-subtests, namely, those of logical, diagrammatic, and analytical reasoning.

The benchmarking process was initiated with an estimation of human ability by means of the IRT approach using the Winsteps program. Participants' ability was expressed in the form of a logit scale ranging from -4 to 4. The values would further be converted by using the mean of 300 and the standard deviation of 50. The benchmark setting was conducted by using the scale anchoring method with the technology of the web server ASP.net.

The benchmark setting was executed through four stages. The first stage was of the setting of the cutoff score at each bench. The cutoff scores used in the research concerned here were percentile 25 (bench 1), percentile 50 (bench 2), percentile 75 (bench 3), and percentile 90 (bench 4). The second stage was of the grouping of test items according to cutoff scores. After the cutoff scores were set, the data of test participants' responses to test items were obtained. With those data as a basis, the proportion of correct answers to each test item was determined. The third stage was of deciding the test items entering the benches according to the criteria that had been set. A test item would belong to a certain bench when most responses answer the item correctly at the bench and answer it wrongly at the bench under it. The fourth stage was of deciding the descriptor of potential at each bench. The descriptor setting was done by holding FGD attended by resource persons competent at TBS development.

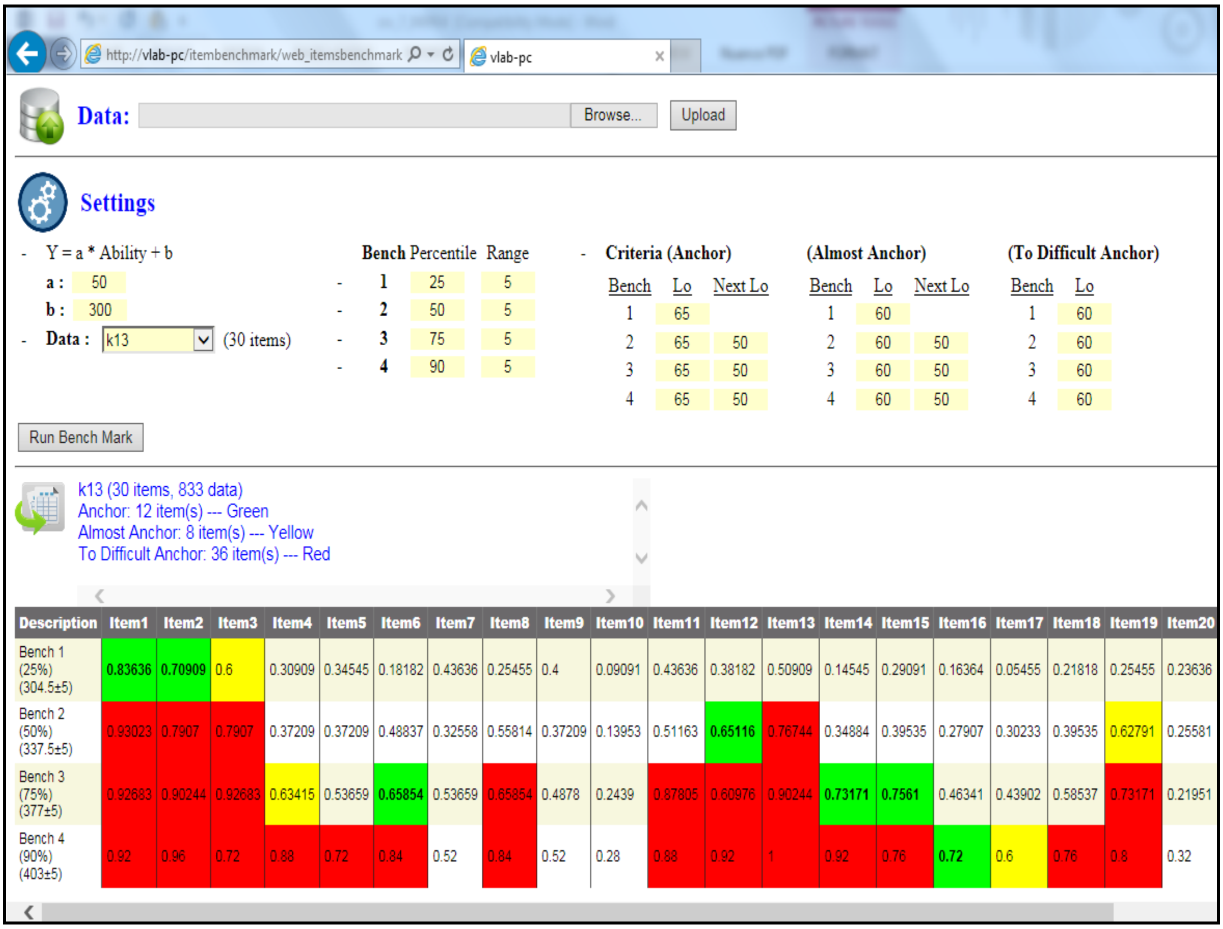
The analysis stage after the benchmark process was that of calculating the standard error of measurement. It was calculated based on the test information formula. The test information was obtained based on the group of test items at each level obtained from the benchmarking process above. The test information was determined at the value of θ. The θ value was obtained based on the score obtained at the bench point (or cutoff score).

Findings and Discussion

The IRT analysis with the Winsteps program had the purpose of making a conversion table that would be used as a basis in determining testees' ability. The IRT analysis was also done to discard test items that did not fit according to the 1 PL model. The analysis was done on several packages and each package had a conversion table differing from that of any of the other packages. After testees' ability was determined, the next thing was making a file of the person's response and ability. The file would be used in determining the classification of TBS items by using the c# program. The classification of TBS items had the purpose of mapping the test items according to the levels discussed in previous sections. The results of the analysis on the program classifying the TBS items can be seen in Figure 1.

In Figure 1, the benches are in the column of description. At each bench, the value of ability is presented according to the percentile. The grouping of test items at each bench is presented in three colors, green indicating the test items meeting the bench criteria, yellow indicating those meeting the almost-anchor criteria, and red indicating those meeting the criteria of being too difficult to anchor.

Figure 1 presents the results of the analysis on the quantitative subtest of package 13. Of 30 test items, 12 meet the bench criteria, eight meet the almost-anchor criteria, and five meet the criteria of being too difficult to anchor. The classification of the results of each subtest of TBS is presented in Table 1.



- Y = a * Ability + b

a : 50

b : 300

- Data : k13 (30 items)

	Bench	Percentile	Range
-	1	25	5
-	2	50	5
-	3	75	5
-	4	90	5

- Criteria (Anchor) | (Almost Anchor) | (To Difficult Anchor)

Bench	Lo	Next Lo	Bench	Lo	Next Lo	Bench	Lo
1	65		1	60		1	60
2	65	50	2	60	50	2	60
3	65	50	3	60	50	3	60
4	65	50	4	60	50	4	60

k13 (30 items, 833 data)
Anchor: 12 item(s) --- Green
Almost Anchor: 8 item(s) --- Yellow
To Difficult Anchor: 36 item(s) --- Red

Description	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10	Item11	Item12	Item13	Item14	Item15	Item16	Item17	Item18	Item19	Item20
Bench 1 (25%) (304.5±5)	0.83636	0.70909	0.6	0.30909	0.34545	0.18182	0.43636	0.25455	0.4	0.09091	0.43636	0.38182	0.50909	0.14545	0.29091	0.16364	0.05455	0.21818	0.25455	0.23636
Bench 2 (50%) (337.5±5)	0.93023	0.7907	0.7907	0.37209	0.37209	0.48837	0.32558	0.55814	0.37209	0.13953	0.51163	0.65116	0.76744	0.34884	0.39535	0.27907	0.30233	0.39535	0.62791	0.25581
Bench 3 (75%) (377±5)	0.92683	0.90244	0.92683	0.63415	0.53659	0.65854	0.53659	0.65854	0.4878	0.2439	0.87805	0.60976	0.90244	0.73171	0.7561	0.46341	0.43902	0.58537	0.73171	0.21951
Bench 4 (90%) (403±5)	0.92	0.96	0.72	0.88	0.72	0.84	0.52	0.84	0.52	0.28	0.88	0.92	1	0.92	0.76	0.72	0.6	0.76	0.8	0.32

Figure 1. Results of analysis on the program classifying TBS items

Table 1. Classification of TBS items at each bench

Subtes	Bench	Cut off score	Anchor	Almost anchor	Too difficult to anchor
Verbal	1	320	60	15	0
	2	340	6	5	31
	3	360	14	14	45
	4	380	6	14	40
Quantitave	1	280	5	3	0
	2	310	6	7	3
	3	340	11	12	23
	4	370	12	17	25
Reasoning	1	340	52	11	0
	2	370	19	9	24
	3	400	18	16	18
	4	430	10	8	25

With the results of analysis as the basis, the test items meeting the anchor criteria at each bench turned out to be only a few in number. The verbal test items that could be retained at benches 1, 2, 3, and 4 are 13.6%, 1.4%, 3.2%, and 1.4% of the 440 test items analyzed. However, at each bench, the test items used in making the description of potential could be netted. The quantitative subtest items that the anchor criteria of benches 1, 2, 3, and 4 could net are 1.8%, 2.2%, 3.9%, and 4.3% of the 280 items analyzed. The reasoning subtest items that the anchor criteria of benches 1, 2, 3, and 4 could

Table 2. Description of potential in the verbal subtest

Bench	Verbal Subtest			
	Synonymy	**Antonymy**	**Analogy**	**Reading**
1 **270**	The individual is able to determine a word in daily general use that is the equivalent of a word in a set of word choices that are also in daily general use and tend to have no similarity in meaning with each other.	The individual is able to identify a word in daily general use that is the antonym of a word in a set of word choices which include one of the equivalents of the identified word above.	The individual is able to identify analogies related respectively to subject/concept-place, object-product, and concept-example.	The individual is able to know, mention, or reexplain something in a discourse presented.
2 **290**	The individual is able to determine a word not in sufficient daily general use which is the equivalent of a word in a set of word choices that are in daily general use and a part of them have a similarity in meaning.	The individual is able to identify a word in daily general use which is the antonym of a word in a set of word choices that have significantly different meanings, are in general use, and tend to include no equivalent of the identified word above.	The individual is able to identify analogies related respectively to concept-function and otherwise, and concept-ownership and otherwise.	The individual is able to comprehend, interpret, and express with different words/sentences something in a discourse presented.
3 **320**	The individual is able to determine a word that is rarely used in daily life which is the equivalent of a word in a set of word choices that are in daily general use with most of them having a similarity in meaning.	The individual is able to identify a word rarely used in daily life which is the antonym of a word in a set of word choices that have significantly different meanings, are in general use, and tend to include no equivalent of the identified word above.	The individual is able to identify analogies that are respectively categorical in nature and related to two concepts that are mutually complementary.	The individual is able to apply and analyze something in a discourse presented.
4 **340**	The individual is able to determine a word that is rarely used in daily life which is the equivalent of a word in a set of word choices that are rarely used in daily life with most or all of them having a similarity in meaning.	The individual is able to identify a word that is rarely used in daily life which is the antonym of a word in a set of word choices that have a similarity in meaning, tend to be rarely used, and include no equivalent of the identified word above.	The individual is able to identify analogies that are respectively synonymy and antonymy in nature.	The individual is able to make a synthesis and an evaluation of something in a discourse presented.

net are 17.6%, 6.4%, 6.1%, and 3.4% of the 296 items analyzed. The description of the testees' potential at each sub-subtest would be made more in-depth by using other criteria. However, the items meeting the anchor criteria remain being the main references. The setting of benches used the scores around the mean of the results of analysis on all the packages above.

The classification of TBS items per subtest based on four international benchmark percentile values (Kelly, 2002, p.378) resulted in four benches and a group of test items at each bench. Based on the analysis, the items netted at benches 3 and 4 turned out to be greater in number compared to those retained at benches 1 and 2. It could be explained because the test packages analyzed were those used for selection needs. In constructing test items into a test instrument, one had better pay attention to the purpose of giving the test and be able to anticipate the distribution of the testees' ability. A test whose purpose is to be a selection instrument should be able to net individuals with high levels of ability.

Table 3. Description of potential in the quantitative subtest

Bench	Quantitative Subtest		
	Number Sequence	Arithmetic & Algebra	Geometry
1 **240**	The individual is able to determine the pattern of a number sequence which is a combination of numerical operations.	a. The individual is able to calculate a numerical operation (addition/ subtraction) on whole numbers and exponent numbers. b. The individual is able to solve a linear equation of two variables which is presented in a story form and uses whole numbers. c. The individual is able to solve a statistics problem in a diagrammatic form.	The individual is able to solve a problem in both picture and story form involving simple two- or three-dimensional geometrical shapes.
2 **280**	The individual is able to solve a number sequence with 1 or 2 jumps each time and by using combined numerical operations, such as, the addition of an exponented number each time.	a. The individual is able to solve a numerical computation using a combination of numerical operations (addition, subtraction, multiplication, or division) on whole numbers. b. The individual is able to solve a computation related to a story involving a number set. c. The individual is able to solve an algebraic computation with one variable. d. The individual is able to solve a problem with statistics (the mean score) presented in graph or diagram form.	a. The individual is able to solve a problem in plane geometry involving a combination of two two-dimensional shapes. b. The individual is able to determine the volume of a solid shape (cubic or rectangular) having different units.
3 **320**	The individual is able to solve a number sequence with 1 or 2 jumps each time by using a combination of numerical operations so that a pattern of another number sequence is formed in the number sequence.	a. The individual is able to calculate a computation in number or story form with a numerical operation (addition/subtraction/division/multiplication) on whole or rational numbers. b. The individual is able to solve a problem in algebra consisting of three unknown variables. c. The individual is able to solve a problem with statistics (the mean score) presented in story form. d. The individual is able to solve a problem in a story involving arithmetic by making an equation consisting of 2 variables and using rational numbers.	a. The individual is able to calculate the magnitude of an angle in a geometrical shape. b. The individual is able to analyze information in a picture and use the information in problem solving. c. The individual is able to calculate the area/volume of an object whose case fits daily life.
4 **360**	The individual is able to solve a number sequence with 1 or 2 jumps each time which is formed from a multi-level pattern.	a. The individual is able to calculate a numerical computation by means of a combination of numerical operations and a combination of number types (whole, exponented, or rational). b. The individual is able to determine the relation involving two or three variables and using rational numbers. c. The individual is able to analyze and process information of a problem in statistics presented in story or graph form. d. The individual is able to use logical and mathematical reasoning to solve a problem in arithmetic.	a. The individual is able to use logical and mathematical reasoning to solve a problem in two-dimensional geometry in combined-shapes or story form. b. The individual is able to use logical and mathematical reasoning to solve a problem in three-dimensional geometry in combined-shapes or story form.

Test items chosen for selection needs are to be able to estimate testees having the ability fitting cutoff scores with the probability of 0.5 in answering items correctly (Hambleton & Swaminatan, 1985, p.229). Thus, the proportion of items with moderate and high levels of difficulty is greater when compared to that of easy items when constructing TBS items. One of the considerations is that the test takers are prospective students at several state universities in Indonesia. An effect resulting from this condition is that the construction of the description of potential at bench 1 becomes less perfect.

Based on the classification of test results in a previous section, the next step was constructing the description of potential at each bench. At this stage, the researcher was helped by 10 people who were competent at their field. Before the description making, a preceding FGD was held between the researchers. The description of the potential in the verbal, quantitative, and and several resource persons. reasoning subtests can be seen in Table 2, Table 3, and Table 4.

With the description of potential at each bench of the verbal, quantitative, and reasoning subtests as the basis, it was found that each bench had unique features. The unique features of the verbal subtest were, among others, (1) differences in degree of generality in the use of words in daily life and combination in category of answer choices, (2) a pattern of relation occurring in items on analogy, and (3) cognitive activity ranging from memorization through to evaluation in items on reading comprehension. Hayes (1989) breaks down cognitive activity into several stages: identifying the problem, representing the problem, planning the solution,

Table 4. Description of potential in reasoning subtest

Bench	Reasoning Subtest		
	Logical	Diagrammatic	Analytical
1 **280**	The individual is able to determine a conclusion based on two premises containing an argument that is, in nature, general/universal/a common postulate.	The individual is able to determine the function of part relation between objects differing in type/form/ function/characteristic.	The individual is able to use the information needed for problem solving.
2 **310**	The individual is able to determine a conclusion based on two premises containing an argument which is assumptive (supposition/assumption) in nature and does not apply universally/generally/commonly.	The individual is able to determine the function of part relation between objects that are the same in classification but different in function /form/characteristic.	The individual is able to analyze and determine the information needed for problem solving.
3 **340**	The individual is able to determine a conclusion based on two premises containing an argument which is assumptive in nature and does not apply generally on all answer alternatives using the two premises.	The individual is able to determine the function of part relation between living creatures or between abstract concepts (like constructs of profession/status/ condition/characteristic).	The individual is able to analyze and process the information needed for problem solving.
4 **370**	The individual is able to determine a conclusion based on two premises containing an argument which is hypothetical in nature.	The individual is able to determine the function of part relation between objects-living creatures- concepts simultaneously.	The individual is able to analyze and process the information needed for problem solving with various possible solutions.

executing the plan, evaluating the plan, and also evaluating the solution. According to the TOEFL descriptor IBt for the reading test, between low and high levels, there is a difference in understanding the sentence which is expressed explicitly or implicitly, factually or abstractly, and in the complexity of a concept (Gomez, Noah, Schedl, Wright & Yolkut, 2007, pp.424-437). The higher the bench of the TBS item on discourse, the more the need for the evaluation stage of cognitive activity for solution. Examples of items per bench of the verbal subtest are presented as follows.

Example 1. Item of Bench 1 in the Verbal Subtest on reading.

Oceans have enchanted the human race for thousands of years – perhaps since people stood on the shore thinking of where waves came from and what was there beyond the far horizon. But at those times the sea was also something feared. There reigned storm gods, horrible creatures, and catastrophes. Only after centuries do human beings dare to traverse it far to the middle until the land is out of sight.

The sea still enchants us though many of its secrets have been revealed. We fly across it without hesitation. Various cargo ships traverse it, transporting food, fuels, raw materials, and factory products. Modern fishing ships hunt fishes and process them on board. But in various places there are still many traditional fishermen using nets from sailing ships or sailboats. For scientists studying the sea, the last 30 years have yielded interesting and abundant new information. As if in a detective story, gradually clues are collected – from rocks at the sea bottom and fossils on land, from modern volcanoes and traces of magnetism in ancient rocks. And from all those emerges a picture of a past gigantic geological force – still changing the sea bottom even now. Imagine a landscape with mountains greater than the Himalayas, plains defeating Africa and Asia in vastness, and trenches that could swallow mountains. That landscape exists – at the ocean bottom – made by an awesome force that has been tearing the Earth's rocky crust, and then shaking it and turning it inside out repeatedly for millions of years.

The idea that continents shift is nothing new. It was first expressed 130 years ago. But at that time it was considered outrageous and ridiculous and the idea was ignored. With the passing of time, there was increasingly more proof until the invention of the echo sounder and the equipment to grip and open the curtain in the 1960s.

Martin Bramwell "***Ocean***"

What made the condition at the bottom of the ocean?
a. The movement of the Earth
b. A gigantic geological force*
c. The power of a sea-bottom creature
d. Huge animals of the sea bottom
e. Waves left by large boats

Example 2. Item of Bench 2 on Antonymy in the Verbal Subtest

REDUCTION
a. profit
b. dividend
c. demand
d. addition*
e. advantage

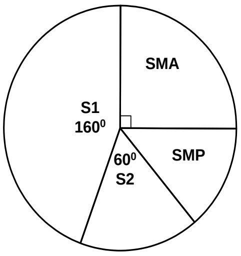
Example 1 is a test item on discourse of bench 1. According to that example, it is hoped that someone could mention a fact, definition, or concept found in the discourse without having to do any analyzing activity. The fact to be mentioned according to the discourse is the process forming the condition at the ocean bottom. Example 2 is a test item on antonymy of bench 2. The word *reduction* is a word in common use in education. It has the sense of descent or decrease. The answer choices given are also common in nature and tend to have different meanings.

The description of the quantitative subtest is differentiated into those of number sequence, arithmetic and algebra, and geometry. Each bench also has its own specific characteristics. The characteristics are, among others: (1) a complexity in the pattern forming a number sequence, with the higher the bench, the more complex the pattern occurring in the series; (2) the mathematical operation, number type, number of variables in equations and item material in the sub-subtest of arithmetic and algebra; and (3) the geometrical shapes in pictorial and narrative items.

However, the cognitive process at bench 3 and that at bench 4 are almost the same in complexity so that no consistent increase occurs. The results of research by Ferrara, et al. (2011) also indicate that the description of the cognitive and language process in items on mathematics does not consistently rise at levels 3, 4, and 5. A descriptor that could not describe the ability that should be mastered at each level causes lack of clarity of what ability someone should possess at each level. Several examples of the quantitative subtest items at each bench are as follows.

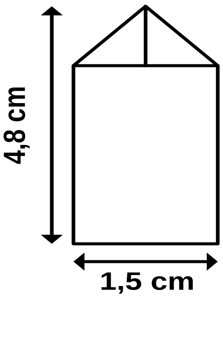
Example 3. Item of Bench 1 on Arithmetic and Algebra in Quantitative Subtest

The diagram describes the level of final education of every head of the family in an RT (*Rukun Tetangga* or 'neighborhood community') named RT. 03. If the number of the families in RT. 03 is 72, how many of them are those whose final education was SMP (*Sekolah Menengah Pertama* or 'junior high school')?



a. 10
b. 12*
c. 15
d. 18
e. 32

Example 4. Item of Bench 2 on Geometry in the Quantitative Subtest



The actual height of a house is 7 m. If, in a drawing of the house, its height is 4.8 cm and its width is 1.5 cm, its actual width is ….

a. 2.188 m*
b. 2.240 m
c. 21.88 m
d. 22.40 m
e. 224 m

In Example 3, an item of arithmetic and algebra at bench 1, the potential measured is in solving a statistics item in the form of a circular diagram. The item becomes an easy one for an individual because the individual directly uses the information obtained from the diagram without having to make a mathematical equation. The item on geometry in Example 4 is also a geometry item of bench 2. It is hoped that, in dealing with the item, an individual could solve a problem in geometry concerning an already modified two-dimensional drawing of an object.

Based on the description of potential in reasoning previously discussed, the specific characteristics distinguishing the benches from each other are: (1) the nature of the premise in the sub-subtest on logic and the drawing of a conclusion based on two premises given; (2) the characteristic of the subject-concept relation in the diagram item; and (3) the cognitive process in the sub-subtest on analyticality, starting from the using until processing information in order to obtain a solution. The following are examples of items of certain benches in the reasoning subtest.

Example 5. Item of Bench 1 on Logic

All motorcycle riders must wear a yellow helmet.

All female motorcycle riders wear gloves.

a. A number of motorcycle riders do not wear a yellow helmet though they wear gloves.
b. All motorcycle riders do not wear gloves.
c. A number of motorcycle riders wear neither a helmet nor gloves.
d. There are motorcycle riders who wear a yellow helmet but they do not wear gloves.*
e. A number of motorcycle riders do not wear a helmet and do not wear gloves.

Example 6. Item of Bench 2 on Analyticality

The results of a geological survey in several regions in Africa indicate that there are several volcanoes that are still active with a división as follows. A volcano which is highly active has an activeness scale above 7, one which is moderately active has an activeness scale ranging from 4 to 7, and one which is of a low level in being active has an activeness scale of less than 4. It is found that Mount H has an activeness scale of 5, Mount K has an activeness scale 4 points higher than that of Mount H, Mount A is below Mount K with a difference of 3 points below its activeness scale while Mount W has an activeness scale 5 points above that of Mount S, which has an activeness scale 3 points below that of Mount A.

So the right statement is … .

a. Mounts H and W are volcanoes with moderate activeness.
b. Mounts K and A are volcanoes which are highly active.
c. Mount H is higher in level of activeness than Mount W.
d. Mounts A and H are volcanoes with moderate activeness.*
e. Mount K is lower in level of activeness than Mount W.

Examples 5 and 6 are respectively items on logicality and analyticality. The premise given in Example 5 is general and factual in nature and describes the obligation of motorcycle riders. With that condition, it could be easier for an individual to draw a conclusion. The item on analyticality in Example 6 measures an individual in analyzing and determining the information being used. The individual could determine the activeness of a volcano based on the criteria available and make an analysis to decide which volcano fits the criteria the most. The error of measurement at the university level was also determined based on the value of the information function of the test. The results of the analysis on the error of measurement at the benches can be seen in Table 5.

The mean of the error of measurement at each bench in the verbal subtest is 0.22, that in the quantitative subtest is 0.31, and that in the reasoning subtest is 0.21. The error of measurement of the quatitative subtest at bench 1 is still sufficiently high; it is also caused by the smallness of the number of items retained at the bench. The case is different from the reasoning subtest; there the error of measurement at bench 1 is the smallest compared to that at any of the other benches. According to the classification results, the items retained at that bench are sufficiently great in number.

Table 5. Error of Measurement

Sub / Bench	Verbal	Quantitative	Reasoning
Bench 1	0.25	0.47	0.19
Bench 2	0.22	0.33	0.22
Bench 3	0.20	0.24	0.20
Bench 4	0.22	0.18	0.25

With the analysis on the error of measurement at each bench as the basis, it is found that the error of measurement at a the lower bench tends to be higher except that at bench 1 of the subtest on reasoning. A factor causing it is that at a lower bench the items retained are only a few in number so that the test information given is also little in amount. The small value of the test information causes the error of measurement to become greater. The proportion of items with low difficulty levels is smaller compared to that of items with high difficulty levels because the test package is used as a selection instrument. Another factor is the existence of difference between human ability and the difficulty level of the items retained so that the resulting test information becomes little in amount.

Conclusion and Suggestion

Conclusion

With the results of the analysis and discussion as the basis, the following conclusion could be drawn. (1) The classification of TBS with the method of scale anchoring is able to group the TBS items into four benches. The items netted give a description of the scholastic aptitude potential clearly and are able to distinguish difference in cognitive process between a lower bench and a higher bench. (2) The description of the potential at each bench indicates uniqueness so that it could distinguish difference in potential between a lower bench and a higher bench. At a higher bench, a higher level of reasoning power is required in analyzing and processing needed information so that the individual concerned could do the problem solving with the right solution; (3) The items netted at a lower bench in the three subtests tend to be few so that the error of measurement at such a bench still tends to be higher compared to that at a higher bench.

Suggestion

Based on the objective, significance, and conclusion of the research, it is suggested that other researchers who are interested in the benchmarking process conduct related research on another class subject. This suggestion is offered with the consideration that each class subject possesses its own uniqueness.

References

Anastasi, A. (1988). *Psychological testing (6th ed.)*. New York, NY: Macmillan.

Azwar, S. (2008). Kualitas tes potensi akademik versi 07A [The quality of the academic potential test version 07A]. *Jurnal Penelitian dan Evaluasi Pendidikan, 12(2),* 231-250. Retrieved from http://journal.uny.ac.id/index.php/jpep/article/view/1429/1217

Beaton, A.E. & Allen, N.L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, summer, 17(2), 191-204.

Bejar, I. I. (2008). Standard setting: What is it? Why is it important?. *R&D Connections*, 7.

Berk, L. (2000). *Child development (5th ed.)*. Massachusetts, MA: Allyn and Bacon.

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

Cohen, R.J. & Swerdlik, M.E. (2002). *Psychological testing and assessment: An introduction to test and measurement (5th ed.)*. Boston, MA: McGraw-Hill.

Embretson, S. & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Ferrara, S., Svetina, D., Skucha, S. & Davidson, A.H. (2011). Test development with performance standards and achievement growth in mind. *Educational Measurement: Issues and practice*, 30(4), 3-15.

Forsyth, R.A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice*, 10(3), 3-9, 16.

Frey, M.C. & Detterman, D.K. (2003). *Scholastic assessment or g? The relationship between the scholastic assessment test and general cognitive ability*. Case Western Reserve, OH: Department of Psychology.

Galotti, K.M. (2004). *Cognitive psychology in and out of the laboratory (3rd ed.) pp.391-392).* Belmont, CA: Wadsworth.

Geisinger, K.F. & McCormick, C.M. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice*, spring, 1, 38-44.

Gomez, P.G., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing, 24, 417-444.*

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hambleton, R.K. & Swaminathan, (1985). *Item resnse theory: Principles and applications.* Boston, MA: Kluwer-Nijhoff.

Harman, G. (1994). Student selection and admission to higher education: Policies and practices in the Asian region. *Higher Education*, 27(3), 313-339.

Hayes, J. R. (1989). *The complete problem solver (2nd ed)*. Hillsdale, NJ: Erlbaum.

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York, NY: American Council on Education/Macmillan.

Keeves, J.P. & Alagumalai, S. (1999). New approaches to measurement. In G.F. Masters & J.P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp.23-42). New York, NY: Pergamon.

Kelly, D.L. (2002). Appplication of the scale anchoring method to interpret the TIMSS achievement scales. In D.F. Robitaille & A.E. Beaton (Eds), *Secondary analysis of the TIMSS data*. New York, NY: Kluwer Academic Publishers.

Mardapi, D., Hadi, S., & Retnawati, H. (2015). Menentukan kriteria ketuntasan minimal berbasis peserta didik. *Jurnal Penelitian dan Evaluasi Pendidikan*, 19(1), 38-45. doi:http://dx.doi.org/10.21831/pep.v19i1.4553

Martin, M.O., Mullis, I.V.S., Beaton, A.E., Gonzalez, E.J., Smith, T.A., & Kelly, D.L. (1997). Science achievement in the primary school years: IEA's Third International Mathematics and Science Study (TIMSS). *Chestnut Hill, MA: Boston College.*

Mullis, I. V. S., Martin, M. O., Beaton, A.E., Gonzalez, E. J., Kelly, D. L., & Smith, T.A. (1998). *Mathematics and science achievement in the final year of secondary school: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: Boston College.

Olatoye, R.A. & Aderogba, A.A. (2011). Performance of senior secondary school science students in aptitude test: The role of student verbal and numerical abilities. *Journal of Emerging Trends in Educational Research and Policy Studies (JETERAPS), 2(6),431-435.*

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice*, Winter, 27(4),15-29.

Resnick, L.B, Nolan, K.J., & Resnick, D.P. (1995). Benchmarking education standards. *Educational Evaluation and Policy Analysis*, 17(4), 438-461.

Solso, R. (2001). *Cognitive psychology (6th ed, pp.428-429).* Boston, MA: Allyn and Bacon.

Wedman, I. (1994). The swedish scholastic aptitude test: Development, use, and research. *Educational Measurement: Issues and Practice*, Winter, 13, 5-11.

Wyatt, J., Kobrin, J., Wiley, A., Camara, W.J., & Proestler, N. (2011). SAT benchmarks: Development of the college readiness and its relationship to secondary and postsecondary school performance. *College Board: Research Report*, 5, 5-30.