Comparison Of Item Difficulty Index National Exams Package (Analysis Using Item Response Theory)

Yuliatri Sastra Wijaya*

Universitas Negeri Jakarta
*Corresponding author : yuliatri_sl@yahoo.com

Abstract

The Assessment of learning outcomes conducted by the government's National Exam (UN) is used to view of students' abilities in a pure and able to function as an indicator of the successful educational process. Given the importance of the UN, the matter is arranged should be able to measure what should be measured, providing reliable measurement results, and reflect the students' abilities. The devices that used in UN made a few packets. View of the problem that given adequate with equal abilities among students and then analyze of the package using Item Response Theory. Item Response Theory frees responders and items of interdependence, so the item is no longer difficult Index dependent (invariant) the ability of the respondent, the respondent is no longer dependent ability (invariant) to the level of hard items. Scale is used of the scale WIT on Rasch models. The result obtained from this study is there no difference in difficulty index three package parameters as object of research. Conclusion,that the package used in the UN is equal.

Keyword : Evaluation, a national exam, difficulty, Item Response Theory, scale WIT

INTRODUCTION

Philosophically, education assessment is an attempt to evaluate the performance and outcomes of learning activities. Assessment of the process is performed to assess the quality and improvement of teaching and learning activities, while assessment intended to assess the achievement of results competencies of learners. Assessment should be able to improve the learners' learning activities or assessment as learning (AaL), teaching teachers or assessment for learning (AfL) and evaluate the achievement of learners at the conclusion of certain levels or assessment of learning (AoL).

Implementation of educational assessment in an effort to prepare the country shall be founded on the values of the nation's most fundamental, as reflected in the Pancasila. Values that include religiosity, humanity, unity, democracy, and justice. Based on the values contained in Pancasila, the appraisal needs to uphold justice, equality and objectivity. That is, in the assessment of individual learners are treated equally, not to favor or disfavor any person or group of learners are assessed. In addition, assessment of learners should not discriminate based on religious background, socioeconomic status, culture, language, and gender. Implementation of the assessment, both process and outcome, it must be carried out with full honesty without engineering.

National exam (UN) is one of the external assessment used by the government to collect data on student learning achievement, the extent of achievement of learners achieving Competency Standards (SKL). UN function is as: (1) quality control tool of education nationwide, (2) driving an increase in the quality of education, (3) a material consideration in determining graduation and achievement predicate learners, and (4) acceptance of a material consideration in the selection of new students on higher education

Examine the function (1) it is expected that with the holding of the UN national education quality can be controlled. This means that schools with low test results need to be treated for the provision of assistance to improve the quality of processes and issues related to the subject. So the UN is not just a mapping function on the grouping of school quality and school quality are not alone but how to provide assistance to schools with a low value of the UN.

Function to (2) of the UN is a driving force to improve the quality of education. Expected to follow the UN gradually educational unit can be consciously and continuously improve the quality of its graduates. Conscious and continuous improvement becomes very important because if the school felt pressured to improve its quality (competing under pressure) then the opposite happened, the study of students is decreasing. This is in line with research Ladd and Fiske (2003) in New Zealand that uses 181 principals and 361 teachers as respondents found that 49.5% of respondents considered the model with the pressure of competition has a positive effect on the quality of student learning, 40% think of no effect, and 10.5% considered negative effect on the quality of student learning.

Function to (3) of the UN is a material consideration in determining graduation and student achievement predicates. The UN must ensure that learners' achievements predicate can be determined accurately. This is important in order to avoid mismatch values obtained from the UN with the ability to actual learners.Function to (4) of the UN is a material consideration in the selection of new students receiving training at a higher level.

The above description shows that the UN is still needed because one of its functions to improve the quality of education. If until now the quality of education in Indonesia has not been evenly distributed, meaning the implementation of the UN must be addressed, both in quality and quality of questions. UN should always strive to be better from year to year. Meanwhile, the execution of quality improvement in 2013, is to multiply the variation of the test package, from 5 to 20 packets. This policy only to the level of SMA / MA / SMK for participants to work on the problems of different tests in one exam room to foster self-reliance and self-confidence of pupils. National exam is one form of summative assessment or assessment of learning (AOL), which according to Andrade and Crizek[1] (2010) must meet two criteria, namely: (1) at the end of levels, and (2) the goal is to group students based on mastery of competencies learned.

In the testing program, especially on a large scale, the preparation of some of the tests are equivalent is one of the important activities as one of its tasks is to maintain the security of the test device. At some stage of equality some test devices can be created when developing the test itself, but usually varies between a test device with the other tests, especially in terms of level of difficulty. This can be overcome by conducting equivalence between the test device with the proper and correct manner. Often found in schools, participants of different trials should be measured by different tests even though the tests are not necessarily equivalent and is expected to measure the nature and demands of achieving comparable results[2] (Tumilisar, 2006: 3). Although to some extent a test of equality can be pursued at the time of making the tests themselves, but in general the degree of variation between tests difficult to keep going[3](Swediati, 1997: 1). In addition, the equivalence tests need to consider that devising tests truly parallel is not easy. So make two tests empirically the same, never be perfectly parallel, reliable or unidimension, so that the resulting Silverback Silverback-can not be compared [4](Grounlund, 1985: 169). If the test results are used to determine the increase in class or program majors, of course it is not fair because it does not do score equivalence for the different tests. Therefore, it is important to do-Silverback Silverback adjustments so that participants of different trials, applying different tests can be compared.

The problem can be overcome by performing equalization score obtained from participants who took the tests. Statistical process known as an equalization method (equating), has been developed to address this problem. In other words, equalization is a process to determine the relationship between the scale score of two or more tests in order to test-Silverback Silverback is treated fairly. The activity equivalent test can be done by developing a system conversion system unit tests to test another unit so that once converted score derived from two sets of tests to be equivalent and interchangeable. This activity can be done by using classical test theory and the theory of the response items. In this article the discussion is focused on the application of the IRT.Thus, the problem in this paper is: whether the UN package that is used between a package with another package equivalent?.

LITERATURE

Item Response Theory (IRT)

Item response theory is also named as a latent trait or item characteristic curve theory. For ease of understanding, the term is used here only an item response theory [5](Dragon, 1992: 160).) Which is an assessment of the test and Silverback items are based on assumptions relating to the parameters of the test items and abilities. According to Hambleton, Swaminathan, Rogers [6](1991: 7) Item response theory that refers to two basic postulates, namely, (a) response to a test taker test items can be predicted or explained by a group of factors called latent traits or characteristics or abilities , and (b) the relationship between the response of the test participants and collection of the properties underlying the item responses can be described by a monotonically increasing function called Item Characteristic function or Item Characteristic Curves=ICC.

On IRT, correct answers are given opportunities of students, characteristics or parameters of item, and the characteristics or parameters of the test participants connected through a model formula to be followed either by a group or groups of test items test takers. That is, the same item for a different test participants should be subject to the rules of the convention, or the person taking the exam is the same for different test items also must adhere to the rule. In such a process there was what is called invariance between test items and test takers. In the modern measurement, item difficulty level is not directly tied to the ability of the respondent.

According to Lord [7](1990: 121) that the invariance test item parameters through a group of test participants is the most important characteristic of the response theory item. It is generally stated that the test item difficulty index as the ratio of correct answers that it is difficult to conceive of how the index test difficulty can be invariant to the group of test takers of different levels of ability.In the modern measurement, item difficulty level is associated directly with item characteristics. Tara difficult point in modern measurement lies in: P (q) = Pmin + 0.5 (Pmaks - Pmin) = Pmin + 0.5 (1 - Pmin). High and low ability have the same level of hard items. Respondents' ability and level of difficult item to be independent. Modern measurements can be used to match the ability of respondents to the level of hard item Theory of item necessary to ascertain the response characteristics of the models used items. Models can take the shape of the item characteristic parameter (1P), two-parameter (2P), three-parameter (3P), or other models. On this occasion the model discussed is a parameter (parameter level of difficulty = b) the Rasch

models. IRT and the item of the respondents liberating interdependence, so extent difficult item is no longer dependent on the ability of respondents. The power of the respondent no longer depends on the level of item difficulty. Independence through difficult points in the item and the ability of respondents to select items that matched respondents 8(Xing and Hambleton, 2004: 7). In this example of a match between the level of item difficulty and ability of respondents, then: if the difficulty level of item is known, the ability of respondents can be found. If the ability of respondents notes, item difficulty level can be set.

In IRT, item difficulty level and the different test items remain the same, even though the test items completed by different groups of test takers. To that end, the response theory item parameters to produce a model that links point to the power of the trial participants. According to Hambleton [9] (1991: 9) assuming a theoretical model of the response to item depth is used, so that only one power as measured by the test items. It was named after unidimension. A concept that connects unidimension is what is called local independence (local independence).

Another assumption in the theoretical model of the response characteristics of the items is a function that specifically describes the relationship between unobservable ability variable with variable abilities observed. Such assumptions also involve the relevant characteristics of the test items on the public presentation of the test participants on a test item . The big difference between the models in IRT sharing is in the number and the types and characteristics of participants are assumed to perform tests. So in essence the response items with those assumptions, then each case must be represented by an Item Characteristic Curve (ICC). ICC is a statement related to the probability of success of candidates according to their ability

Matching Models

Matching the model is to compare the characteristics of the model chose in this case Rasch models with the data from the field . In this examination of the possibility of there items that matched the model and there are items that do not fit the model . The items involved in the calculation are the items that matched the models.

Test equating

Equalization by Peterson , Kolen and Hover [10](1989 : 221) is defined as the process used to ensure Silverback Silverback - resulting from the administration of the tests can be used interchangeably and Crocker and Algina [11](1986 : 457) , equality is defined as a process to establish Silverback Silverback - equivalent in the two instruments . Silverback equalization is an empirical procedure is necessary to transform a set of tests to Silverback Silverback other test devices . Because it is an empirical procedure based on data equalization Silverback Silverback test. Makes about equal in two or more packages , certainly not easy or even impossible , because surely there's a difference . This is because it is almost impossible tests that compose the multi package actually parallel[12] (Petersen , Kolen , & Hoover , 1989) . Although the compiler tests using the same test specifications in writing and items will only change the numbers , there is no guarantee that the level of difficulty of the items will be the same . Moreover, if different is the keyword and the substance of the answer choices . According to Angoff [13] (1971) e equalization method is divided into two categories , namely : 1) equalization eki percentile , and 2) linear equalization (linear equating). The first category is revised by Silverback Silverback test comparison between X and Y to be equivalent if the rank order of percent of each group is the same . Furthermore, to equalize the score in 2 different tests , then both should be given to the proficiency level tests examine the same group . Later in the second class , it is assumed that x on test X and y in Y test has direct relationships / alignment (linearly related) . According Tumilisar [14], the equalization method is the search for ways of equalizing the two Silverback relationship tests of two different research instruments using certain statistics , and data collection has done with specific data collection design .

Process of equalization of multiple device test (equating) can be done in two ways , namely horizontally equivalency and vertically equivalency . Equalization process is obtained from two different test devices but measure the same thing is called horizontal equivalency . The process of equalization of the two groups of participants of different tests in level / level of education , but given the same problem called vertical equalization [15](Crocker & Algina , 1986) . Basically equating aims to equalize Silverback Silverback by comparing obtained from working on a test device with Silverback obtained from other test devices that work through the process on both devices score equivalency tests (Hambleton & Swaminthan , 1991)[16] . According to Zhu [17] , on test A and test B can be compared if they meet four conditions , namely : 1) measure the same abilities or characteristics . 2) after equalization , the frequency distribution Silverback on A test should be the same as the frequency distribution Silverback on test B , so sekor on test A and test B can be interchangeable after equalization ; 3) equivalency test must be free of data or job candidates in the equalization process , and conversion from equalization should apply to all similar situations , and 4) transformation should be the same regardless of which test is used as a base or reference conversion , meaning that the same interpretation should be good score equivalency test from test a to B or from B test to test A.

In the equivalent group design used two groups of participants equivalent (K1 and K2) and two sets of tests (X and Y) . K1 work group participants and group X test the participants work on devices K2 test Y.

According to the Theory of Responsibility equalization method ItemEqualization method according to the theory of the response function to determine item conversion constants . This is considering that the equivalence

between two devices or more tests can be performed if the conversion constants are recognized . Conversion value is then replaced in the equation at the scale employed in the purpose of equalization . There are several methods that can be used equivalence tests and the factors that affect the accuracy of the method equivalency test . In principle , there are four item responsiveness equivalent test methods , namely : regression , the mean and standard deviation , mean and deviation standard strong and characteristic curve (Angoff, 1982)[18]. In addition to equalization method , known also 4 kinds of design . (Dragon : 2013 ; 353)[19] . The design will be used is the design of A , which is like the picture below In this paper, which will be compared is the parameter b (item difficult) by applying the mean and standard deviation . Under the plans this means there are two tests that will compare the difficulty often aloof sphere . It called for the same metric . Equivalency degree was a difficult item to a predetermined metric .
Equalization can be compiled into a table as a table of concordance between the metrics that will distarkan with other metrics . Equalization involves at least two score which will Sekor X to Y is a benchmark score . The model has only one parameter Rasch item difficulty level of the parameter b , so that only the required translation and rotation is not necessary so . Equalization coefficients

$b * Y = bX + d$ with d coefficients .

Of equalization for research models
$b * Y = bX + d$ so that
$\mu_{bY} = \mu_{bX} + d$, $\sigma_{bY} = \sigma_{bX}$
Of this equation is obtained

$d = \mu_{bY} - \mu_{bX}$
$b^*_Y = b_X + (\mu_{bY} - \mu_{bX})$

In addition, the difficulty level of the item has a value of -4 (item very easy) to +4 (very difficult item) . In order item does not have a negative value then the value of item in the transfer to the WIT scale made by Benjamin Wright with formula
$Wb = 9.1\ b + 100$

Preparation of the National Examination Test preparation (UN) needs to be well planned and coordinated . Performed with a large -scale UN held every year . According Kumaidi [20] to develop and test a number of items that are " defesible " the procedures necessary be regulated development , in the sense of the test development process (and writing items) beginning with the development or design of the test grating , which is preceded by surgery curriculum that includes all the information about the test. The initial step in developing a test is set specification tests , which contain descriptions that depict the overall characteristics that must be possessed of an examination . Clear specification will ease in writing about , and anyone who writes about will result in the same relative degree of difficulty . Preparation of test specifications includes the following activities : 1) define test objectives , 2) develop a test grating ; 3) selecting a test , and 4) determine the length of the test (Setiadi, 2009: 167)[21] .Furthermore , according to study states that each year the problems that used to be made by a special committee formed for the national exam , so that each year must be spent substantial funds for revision purposes such questions . For security purposes are also needed some alternative package tests (parallel form) , in which the questions in a package with other packages being an equal level of difficulty because simply because they are based on the same lattice without empirical data based on trial results matter in the field.This test involves the design development specialists (including teachers) field of study , so that when the test design has been completed then the design of the tests should be validated , through the review of experts and peers, so it really fit with the message of the curriculum . To overcome the excessive variations items , with the understanding indicator items , it is better developed what is called by Nitko [22] as specification items (item specification) . This specification concerns the description of the boundaries and guidelines that must be adhered to by the author items.

METHODOLOGY

This study aims to: see equality tests used in the National Exam. The study population was a high school student responses in Jakarta on the 2012 National Examination for eye Mathematics test. Stratified sampling using proportional random sampling. Sample size of 5000 respondents. The instrument used was 3 devices IPS high school math skills test consisting of 40 items with a form of objective test 5 answer choices. The data is processed by a score of respondents IRT concepts using software Bilog, item parameters obtained further difficulty level three packages equated using IRT approach. Requirements for using the IRT approach, is the test of model fit.Three further test whether the test device stara using one-way ANOVA, with trials preceded normality and homogeneity.

RESULTS

Description of the data

Results of parameter estimation using Rasch IRT models as follows:Descriptively using the mean WIT scale package

$\mu 12 = 81.765$, $\mu 25 = 81.777$, $\mu 39 = 81.767$.

The test model is a requirement that must be passed if it will be using the IRT approach. The results of the model test concluded all items (40 points) fits the model stated, means to 40 points can be used for processing using IRT. The next step is to find the equalization formula using Rasch IRT models, models of the mean and standard deviation equalization, with the kind of design Equalization is done by taking 12 packages as its foundation, meaning that packets 25 and 39 at the 12 with a package to balance the formula:

b * Y = bX + d with d coefficients.

Of equalization for Rasch models

b * Y = bX + d so that

$\mu_{bY} = \mu_{bX} + d$, $\sigma_{bY} = \sigma_{bX}$

$d = \mu_{bY} - \mu_{bX}$

$b^*_Y = b_X + (\mu_{bY} - \mu_{bX})$

b * Y = outcome equivalency of 25 packs or 39 packs to 12

for 25 to 12 packages per item formula

b * Y = b25+ (81.777 - 81.765)

b * Y = b25 +0.01274

for 39 to 12 packages per item formula

b * Y = b39 +(81.767 -81.765)

b * Y = b39+0.00296

result of equalization in can mean the following

$\mu 12 = 81.765$, $\mu 25 = 81.790$, $\mu 39 = 81.780$

Test requirements for ANOVA analysis of the test for normality, and homogeneity. Normality test results obtained from the three bundles are normally distributed population, are the third homogeneity test package is derived from a homogeneous population. While the ANOVA analysis found that there was no difference between the item difficulty level 12 packets, with 25 packets, and the packet 39. This means that 3 pack equivalent national exams used.

CONCLUSION

IRT is an alternative option that aims to break away from dependence on a given test with sample test participants. In this case, although the problems are worked out by a clever student or students who are less intelligent, an indication of the level of difficulty of a problem remains unchanged. Besides the application of IRT in equivalency test requires the fulfillment of assumptions unidimension and local independence through the model test. Steps to perform equivalency tests based on IRT activities, namely: 1) estimate the parameters, 2) estimates the response theory item scale with a linear transformation to CDT scale, and 3) to equalize the score level hard items. The method used is difficult item level equalization method mean and standard deviation. Final analysis results obtained three packages used, namely packet 12, 25, and 39 equivalent.

REFERENCES

Chong Ho Yu dan Sharon E. Osborn. 2005. Test Equating by Common Items and Common Subject: Concepts and Applications. *Practical Assessment, Research & Evaluation.* X (4): 187-198.

Crocker, Linda, & Algina, James. 1986. *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston, Inc.

Gronlund, Norman. E. 1985. *Measurement and Evaluation in Teaching*. New York: Macmillan Publishing Company.

Hambleton, Ronald K, Swaminathan, H., dan Jane Rogers, H. 1991. *Fundamentals of Item Response Theory*. London: SagePublications.

Hambleton, Ronald K., dan Swaminathan, H. 1985. *Item Response Theory: Principle and Applications.* Boston: Kluwer Nijhoff Publishing.

Kolen, Michael J., dan Robert L. Brennan. 2004. *Test Equating, Scaling, and Linking: Methods and Practices*. New York: Springer.

Kumaidi. 2000. Standardisasi Butir Soal. *Jurnal Pendidikan dan Kebudayaan.* V(5): 132-143.

Livingstone, S. A., Doran, N. J. dan Wright, N. K. 1990. What Combination of Sampling and Equating Methods Work Best?. *Applied Measurement in Education.* III (2): 73-95.

Lord, Frederick, M.1990. *Aplications of Item Response Theory to Practical Testing Problems.* New Jersey: LawrenceErlbaum Associates, Publishers.

Mary J.Allen and Wendy M Yen, 1989, *Introduction to Measurement Theory,* California: Broke.

McDonald, Roderick P. 1991. *Test Theory: A Unified Treatment.* New Jersey: Lawrence Erlbaum Associatiates Publisher.

Naga, Dali, S. 1992. *Pengantar Teori Sekor Pada Pengukuran Pendidikan*. Jakarta: Besbats.

Naga, Dali, S. 2012 *Teori Sekor Pada Pengukuran Mentaln*. Jakarta: PT. Nagarani CitraYasa.

Peraturan Pemerintah No. 19 Th 2005 Tentang: Standar Nasional Pendidikan (SNP). Bandung: Citra Umbara.

Peterson, N.S., Kolen, M.J., dan Hoover, H.D. 1989. Scaling, Norming, and Equating. In R.L. Linn (Ed), *Educational Measurement*. New York: Macmillan.

Sukirno, D. S. 2007. Penyetaraan Tes UAN: Mengapa dan Bagaimana. *Jurnal Cakrawala Pendidikan.* XXVI (3): 305-321.

Swediati, Nonny. 1997. *Metode untuk Penyetaraan* (*Equating*) *Sekor Tes Secara Klasik.* Pusat Pengujian Balitbang Dikbud: Jakarta.

Tumilisar, A.V.J. 2006. Akurasi Relatif Penyetaraan Sekor Tes untuk Sampel Berukuran 300 Ditinjau dari Metode Penyetaraan dan Teknik Penghalusan. *Jurnal Pendidikan Penabur*. V (6): 1-19.

Zhu, W. 1998. Test Equating: What, Why, How?. *Research Quarterly for Exercises and Sport*. Wayne State University.