IJER Indonesian Journal of Educational Review
p-ISSN 2338-2017 | e-ISSN 2335-8407
Vol.4, No.1 , July 2017, p 84-92

Available online at
http://journal.unj.ac.id/unj/index.php/ijer

CLASSICAL TEST THEORY OF INNAPROPRIATE INDEX SCORE'S ACCURACY COMPARISON USING CONFUSION MATRIX ACCURACY PROPORTION IN EDUCATIONAL MEASUREMENT

Wardani Rahayu[1], Eko Wahyudi[2].
State University of Jakarta
wardani.rahayu@unj.ac.id
ekowahyudi@gmail.com

Abstract

The objective of this study is to determine the Donlon Fischer aacuracy method, Jacob and SHL inappropriate score using confusion matrix accuracy proportion. This research was conducted by analyzing secondary data in response to the physics national examination results throughout the West Kalimantan Province in 2012 using 1545 research samples. The sampling of the research is done by using the random sampling technique of 1545 population. The research method is experimental method which is done by comparing the proportion of accuracy obtained from confusion matrix. The hypothesis was tested by using a differences in proportions Z test. Test results showed that the SHL inappropriate index score is more accurate than Donlon Fischer and Jacob innapropriate index score. This research is expected to find an accurate method in detecting the inappropriate score used as an evaluation parameter for educational measurement.

***Keywords**: Confusion Matrix, Inappropriateness Index, Classical Test Theory*

The development in the field of educational measurement has now became one of serious concern. There are various ways to conduct an educational measurement in order to overcome problems that often arise in relation with exam's phenomenon at school. Good test measurement reflects a good education. Measuring is an important activity as the first step to detect a person's ability. Person's ability interpretation process measurement in term of his/her attributes or behavior in descriptive manner into numerical quantitative data formation for example a number formation according to certain rules (Mehrens and Lehmann, 1972; Nunnaly, 1970; Nitko, 2001; Ebel and Frisbie, 1991; Miller, 2009 ; Barrow, 1979). Thus, if the measurement process is disturbed, the numbers that describe a person's attribute does not reflect the real state.

Good educational measurement requires a good test instrument. The test is a set of items or instruments that have been standardized to provide numbers on the behavior of individuals or certain attributes according to systematic and objective procedures to demonstrate learning achievement that can be used in decisions making on the learning process carried out by educators. (Tyler, 1971; Brown,

1976; Anastasi, 1998; Ebel, 1991). If the measurement bias is caused by index performance that is different when being applied in two different samples (eg: boys and girls), but the different individual performance when the index is different called individual bias. (Sumintono and Widhiarso, 2014). So eventhough the test tools that being used is already good, inconcordance measurement results with actual students ability still can happened.

Learning evaluation in schools can be either done by formative or summative tests that can indicate progress and mastery of knowledge by the learners. The test results serves as a report of learning outcomes and graduation that says the success in one's learning process at the end of their education. Most people considering learning achievement only from high score, rather than on the process. It results student's stressor to always obtaining high scores. But these scores are not necessarily obtained with a good process. In addition, students consider tests as a failure, not as a measurement to show the evaluation as a results of the study.

A person whose psychological state does not allow him/her to take the test properly, will obtain scores with peculiarities in the pattern of test answers. Such psychological conditions may be the result of high anxiety such as fear of not passing the test. This resulted in an inconsistency of the answer of the test takers based on his/her ability compared to an idealized model. Such response pattern can be said as an inappropriate response or show thinking consistency or it may result from cheating (Sumintono and Widhiarso, 2014). The state of inappropriate response patterns caused the measurement results to fail to show the true capabilities that being measured which may interfere with measurement process in education.

The inappropriate response patterns can also be interpreted with innapropriate score. Innapropriate score may occur among test takers with high ability but fail to answer a simple item, as well as on test participants with low ability but able to correctly answer the problems deemed as difficult (Naga, 1992). It may also called as *appropriateness* which refers to a method for detecting test takers whose tests score fails to measure latent characteristics to be measured so that the pattern of responses produced by test participants is inappropriate (Hulin, 1983). Inappropriate index is also a matching simple calculation models as the most common psychometric test item by item response pattern (Levine and Rubin, 1976). Inappropriate score does not depend on the grain. Even good test problems may cause inappropriate score as a result of unwell condition of test takers.

Such deviating pattern may occur in test participants affected by several factors such as anxiety, carelessness, unfit health condition, or unfamiliarity with the new system used by the respondent when undergoing the measurement. Although it is possible for them to guess, it will be uncertain due to anxiety. Guessing may occur in test takers not knowing the material content and multiple

choice questions may also allow guessing. Therefore, a special study on the deviation patterns caused by carelessness and anxiety should also be considered.

There are students who get a spuriously low score due to answering at the wrong number of problems. For example answer number 7 is placed at number 6 or vice versa. Because the time is running, test participants fail to check the answers and ultimately resulted in a low score. There are test participants who received a spuriously high score than their actual ability (Hulin, Dasgrow, and Pearson, 1983), Rudner (1983) uses the term accuracy assessment.

Statistically, inappropriate score can be detected by analyzing the pattern of responses. Each participant who answered the item in turn has a distribution from a low to a high ability capability derived from the calculation of raw scores into scores which the participants test illustrates as the ability of the test taker. Raw scores were ranked from the largest to the smallest that show the ability of the test taker from highest to lowest. Along with this rank, it should be the correct response pattern according to the order of test items as well.

Each item that has been answered by the participants of a particular test will produce scores. Score that has been answered correctly by the participants is compared to the number of test takers produce the level of difficulty proportions. This proportion has a value from 0 to 1. The proportion of zero means no one correctly answered the item while the proportion of 1 means the item correctly answered by all participants of the test. About the difficulty level can be arranged from the easiest to the most difficult. Items that have a high proportion indicate that it was easy, while the items which have a lower proportion indicate that it is difficult.

Inapropriate score can be detected by building an item's score interaction in the form of the proportion of with participants' score test. If the participant's test matrix is not in accordance with his ability, it will be indicated that the score is inappropriate. Test participants who had high scores should be able to answer the problems with difficulty level ranging from a low to a certain degree of difficulty correctly. It should be applied to the participants test with low scores who only be able to answer easy questions to a certain level of ability.

In classical test theory, there are several methods used to detect inappropriate score, among others, the methods of Sato modified by Harnisch and Linn, methods indexes U of Van der Flier, personal bi-serial of Donlon and Fischer, norm conformity Index of Tatsuoka and Tatsuoka, and index agreement and disagreement by Kane and Brennan (Harnisch and Linn). Jacob uses the term Jacob's weighted average in detecting inappropriate index (Hulin, 1983).

Some of the classic inappropriate score have different characteristics, but all calculations to analyze the deviation pattern of response are based on level of difficulty and the ability of the test taker grains. Previous research comparing the inappropriate score index is SHL-Fisher Donlon on math test score achievement of VII grade junior high school students shows that there is no difference between

inappropriate index tested using SHL and Donlon Fisher methods on math test scores of VII grade of junior high school students (Widyastuti, 2014). Another study conducted by Rudner states that the Rasch model unweighted index fits the statistic (U1) and the bi-serial correlation (br) are not accurate in detecting participants with false high test score. Unweighted Birnbaum model fit statistic (U3) is also not accurate in detecting participants with low false test. Weighted index model fit statistic (W3) relative can falsely identify the test participants. Norm conformity index (NCI) and the modified caution index (C1) tend to be more statistically stable both at low and high scores compared with other indices (Rudner, 1983).

Innapropriate index methods based on clasical theory is in accordance with respondanced capability and index difficulty level. Asumption difference in order to sort index group between easy and difficult make capability difference possible in order to sort wether the test paticipant is appropriate or not. We can say that this study will differetiate accuracy from some methods based on innapropriate index that has already been efective in order to show the real data or standarized condition in confution matrix. Birenbaum stated that efective innapropriate index are $ECI2_z$, $ECI2_z$, and L_z (Birenbaum, 1985). This study used L_z innapropriate index as a compatibility refference in confution matrix in order to get accuracy proportion. Methods that have higher accuracy proportion will have higher accuracy.

The purpose of this study was to determine whether or not Donlon Fischer method is more accurate than Jacob's, to determine whether the SHL method is more accurate than Jacob, and to determine whether SHL method is more accurate than Donlon Fischer's or not.

METHOD

The design of the study used is a comparative quantitative studies which compares the proportion of accuracy based in confusion matrix of effective inappropriate index namely inappropriate index L_z.

This study uses physics test of National Exam instrument of 40 test items. The data is in the form of secondary data of National Exam in 2012 in West Kalimantan with the total of 1545 participants were obtained from Puspendik.

The first step is taking data samples from a population of 500. The analysis of the 40 items is done with the Rasch model to determine the item that matches the model of Rasch in order to acquire 40 items that fit the Rasch models. Further, an analysis of the 500 participants is done to determine the test participants who did not fit with the model in order to obtain as many as 481 participants tests that fit the Rasch model. This is evident from the value of OUTFIT MNSQ in which none of them exceeds 2.00. To test the ability of 500 participants, there are 19 test participants who have OUTFIT MNSQ score of more than 2.00 so it can be said

that the 19 participants of the tests do not match with the model, and therefore the 19 test participants should be excluded from the analysis.

The second step is one-dimensional inspection requirements by looking at the table of Standardized Residual variance (in Eigenvalue units) on the Rasch model analysis with winsteps program. Raw value of variance explained by measures of 30% which shows one-dimensional requirements are met and included in the category of good since it has minimum value of 20%. It can also be determined based on the value of unexplained variance none of which exceeds 15%, which shows the one-dimensional requirements are met. Furthermore, the calculation of the Lz inappropriate index is used to determine which test participants turned out to be appropriate and inappropriate.

$$L_z = \frac{L_0 - \mu_{L_0}}{\sigma_{L_0}},$$

$$L_0 = \ln L\,(X|\theta) = \prod_{i=1}^{N}[X_i \ln P_i(\theta) + (1 - X_i) \ln Q_i(\theta)]$$

$$\mu_{L_0} = \frac{\sum_{i=1}^{N} L_{0i}}{N} = \sum_{i=1}^{N} m_i(\theta) = P_i(\theta) \ln P_i(\theta) + Q_i(\theta) \ln Q_i(\theta)$$

$$\sigma_{L_0} = \sqrt{\frac{N \sum_{i=1}^{N} L_{0_i}^2 - \left(\sum_{i=1}^{N} L_{0_i}\right)^2}{N^2}} = \sqrt{\sum_{i=1}^{N}\left[P_i(\theta)Q_i(\theta)\left(\ln \frac{P_i(\theta)}{Q_i(\theta)}\right)^2\right]}$$

The third step is to estimate inappropriateness index of as many as 481 test participants by using Jacob, Donlon Fischer method, and thus the appropriate and inappropriate test participants based on each method can be determined.

The fourth step is to make confusion matrix by comparing the inappropriateness index of Jacob, Donlon Fischer, and SHL with an L_z inappropriateness index to acquire the number of participants test with the true positive, false negative, true negatives , and false negatives results in order to obtain the proportion of the accuracy of each method of inappropriate index.

Predicted		p	n
	Y	True Postive	False Postive
	N	False Negative	True Negative

Accuracy of the method is sought bt finding the proportion of (Fawcett, 2006):

$$p_{accuracy} = \frac{TP + TN}{P + N}$$

The higher the TP and TN values $p_{accuracy}$ values will be greater, hence making such method more accurat.

The fifth step is to analyze data by comparing the proportion of accuracy of confution matrix by using a difference test in proportion with the Z test.

Research Findings

The results of the accuracy proportion calculation of the inapropriate score of Jacob's method is obtained from the cconfusion matrix on the Jacob methods innapropriate index which showed the number of true positives and false positives of 257 and 224 samples, but there are no false negatives and true negatives from the total sample of 481. Then, the accuracy proportion of 0.5343 Jacob's method's inappropriate index may be calculated.

The results of the accuracy proportion calculation of Donlon Fischer method's inappropriate index is obtained from the Donlon Fischer's method inappropriate index confusion matrix that showed the number of true positives and false positives of 257 and 224 samples, but no false negatives and true negatives from the total sample of 481. Then, the accuracy proportion of 0.5343 Donlon Fischer's inappropriate index may be calculated.

The results of the accuracy proportion calculation of SHL method's innapropriate index is obtained from the SHL method inappropriate index confusion matrix that sowed the number of true positive and false positives of 221 and 135 samples the number of false negative is 36, and the number of true negative is 89 from the total samples of 481. 0,6445 may be calculated.

The first hypothesis test calculation is done by using the Z test to test the difference in the accuracy proportion between Dohlon Fischer innapropriate index and Jacob innapropriate index. The result of the hypothesis testing shows that Z_{count}= 0 is less than Z_{table} =1.645 hence the H_0 is acceptable. The results of this statistical test showed no difference in the accuracy or it can be said that the Donlon Fischer innapropriate index score is as accurate as Jacob inappropriate index score. However, the same results have showed that there are some things that need to be further studied.

The second hypothesis testing the calculation is done by using the Z test for testing the accuracy proportion differences between SHL innapropriate index score and Jacob's innapropriate index score. The result of Z_{count} = 3.38 is greater than Z_{table} = 1.645, thus the H_0 is rejected. The statistical test result shows that the SHL innapropriate index score is more accurate than Jacob's.

The third hypothesis testing the calculation is done by using the Z test for testing the accuracy proportion differences between SHL's innapropriate index score and Donlon Fischer innapropriate index score. Since the result of Z_{count} = 3.38 Z_{table} = 1.645) the H_0 is rejected. The statistical test result shows that the SHL innapropriate index score is more accurate than Donlon Fisher's.

The accuracy of the innapropriate score detection method is seen from the increasing number of true positive and true negatives that appearing in innapropriate method during participants test. The method would be accurate if the number of true positive and true negative, which compared with the number of samples, yield the proportions accuracy. The higher the accuracy proportion, the higher the method's accuracy. It can be interpreted that if the method is accurate then these methods have the ability to distinguish between appropriate and inappropriate score. It can also be said that if the innapropriate score detection method between Jacob, Donlon, and SHL will have a high accuracy views from their ability to separate natural and unnatural score. Score that indicated fair will be completely reasonable in effective methods and score that indicated unnatural will be indicated as well. This is similar to Naga's opinion who stated that effective appropriate index is able to correctly separate the participant's appropriate score against the inappropriate scores of another participant.

It is also predicted since the time of the analysis, the item and the person that would fit with Rasch model. In the Donlon and Jacob score inappropriate index calculation data that has been analyzed by the Rasch model is the number of test takers that is similarly estimated by the L_z inappropriate index. From the 500 samples of test participants taken, there are 19 test participants who do not fit with Rasch models that should be excluded from the level of difficulty estimation and the ability of the test taker estimation's analysys. It is assumed that the 19 test results are from participants who had high inappropriate score which means that some scores according to the Jacob and Donlon indexes had inappropriate scores which varies between the two methods. Finally, after 19 test participants that being removed from the analysis, the outcome of Donlon Jacob and Fischer index has the number of appriopriate and inappropriate score. This resulted in a proportion which fails to describe the accuracy proportion.

The second hypothesis testing showed that the SHL inappropriate index is more accurate than Jacob's. SHL inappropriate index shows true positive of 221 and true negative of 89. Although the inappropriate index of SHL shows false negatives and false positives, but the proportion of accuracy reached 0.64. SHL accuracy of this proportion is higher than the accuracy proportion of Jacob's Index. The higher the proportion of appropriate and inappropriate l matches, the more accurate the SHL index will be.

The third hypothesis testing also showed that SHL is also more accurate than Donlon Fisher. It is in accordance with the initial hypothesis which states that SHL index score is more stable in indicating an appropriate and inappropriate scores at high, low and avarage ability. This is consistent with Harahap (2007) research that stated bi-serial person correlation will effectively used if the score is the reference which includes all territories of Indonesia with the expectations that the reference group has normal distribution. In addition, Widyastuti (2014) stated that innapropriate score detection of test participants in a fewer number by using

SHL methods will be better. As for the large sample Donlon Fisher methods will be more suitable. Furthermore Rudner also says modified caution index (Ci) or also known as SHL index tends to be more stable statistically at low or high scores (Rudner, 1983).

CONCLUSION

SHL inappropriate index score is more accurate than the Donlon Fischer and Jacob's inappropriate index score. The detection of participants consistency, who are suspected to have appropriate and inappropriate test scores, using an accurate method on classical test theory such as the SHL's index. Detection using SHL will be easily done if it is made in the application form. In order to detect inappropriate score to be more stable than the difficulty level, the items difficulty level should be represent using a test scale.

REFERENCES

Anastasi, Anne. (1998). *Psycological Testing*. New York: Macmillan Publishing Company.

Barrow, Harold M. dan Rosemary. (1979). *A Practical Approach to Measurement in Physical Education*. Philadelphia: Lee & Febiger.

Birenbaum, Menucha. (1985). "Comparing the effectiveness of several IRT Based Appropriateness Measures in detecting unusual Response patterns". *Educational and Psychological Measurement*, No. 22. Israel: Tel-Aviv University.

Brown, Fredrick. G. (1976). *Principles of Educational and Psychological Testing*. New York: Holt, Rinehart & Winston.

Ebel, Robert L. dan David A. Frisbie. (1991). *Essentials of Educational Measurement*. New Jersey: Prentice-Hall International, Inc.

Ebel, Robert L. (1979). *Essensials of Education Measurement*. New Jersey: Prentice-Hall, Inc.

Fawcett, Tom. (2006). *"An Introduction to ROC Analysis", Pattern Recognition Letters*. United States of America: Palo Alto.

Harnisch, D. L. & Linn, R. L, "Analysis of Item Response Patterns: questionable tes data and dissimilar curriculum practices". *The Journal of Educational Measurement*, Volume 18, No. 3.

Hulin, Charles I., Pritsz Drasgrow, and Charles K. (1993). Peersons, *Item Respon Teory Application to Psychological Measurement*. Ilionis: Dow Jones-Irwin.

Levine dan Rubin. (1976). "Measuring The Appropriateness of Multiple-choice test scores". *Journal of Educational Statistics*, Vol. 4, No. 4.

Mahyuddin HarahaNugaan Yulia Wardani. (2007). "Detection of Inappropriateness score of National Examination Outcome Mathematic in Medans' Senior High School Student Batch 2006/2007" Indonesian University.

Mehrens, William A. dan Irrvin J. Jehmann. (1972). *Measurement and Evaluation in Education and Psychology*. Michigan: Holt, Rinehart and Winston, Inc.

Miller, M. David, Robert L. Linn, dan Norman E. Gronlund. (2009). *Measurement and Assessment In Teaching*. New Jersey: Pearson Education, Inc.

Naga, Dali S. (1992). *Introduction of Score Theory of Education Measurement*. Jakarta: Gunadarma.

Nitko, Anthony J. (2001). *Educational Assessment of Student*. New Jersey: Prentice-Hall, Inc.

Nunnally, Jum C. (1970). *Introduction to Psycological Measurement*. New York: Mcgraw-Hill, Inc.

Roderick P, McDonald. (1999). *Test Theory*. New Jersey: Lawrence Erlbaum Associates.

Rudner. (1983). "Individual assessment Accuracy". *Journal of Education Measurement".* Vol. 20. No. 3.

Sumintono, Bambang dan Wahyu Widhiarso. (2014). *Rasch Model Aplication for Social Knowledge Study*. Cimahi: Trim komunikata Publishing House.

_____. (2014). *Rasch Model Aplication for Educaton Assessment.* Cimahi: Trim komunikata Publishing House.

Tyler, Leona E. (1971). *Test and Measurement*. New Jersey: Prentice Hall, Inc.

Widyastuti, Suciati Rahayu. (2013). *"The Efectivity of SHL dan Donlon Fisher' method to detect mathematic outcomes score".* curriculum proceciding seminar.