**SE -1**

# USE OF COMPUTER MANAGEMANT INSTRUCTION FOR DEVELOPMENT STANDARDIZED TEST FOR EQUIVALENCY QUALITY ASSESSMENT AS DETERMINANTS OF SCHOOL GRADUATION IN THE NATIONAL EXAM SYSTEM FAIR

**Dadan Rosana[1)], Sukardiyono[2)]**

*[1), 2)] Science Education Study Program, Faculty of Mathematics and Science Yogyakarta State University, email: danrosana.uny@gmail.com*

**Abstract**

Issues around final school exams is still the main problem in education that spawned a lot of controversy, one of which is about the method of determining graduation. The final value for the determination of graduation obtained from the combined value of school subjects tested nationally and value the UN, which is weighted 40% of the value of school subjects tested nationally and 60% of the value UN (Regulation of the Minister of Education and Culture of the Republic of Indonesia Number 3 in 2013). The problem that then arises in this regard is the lack of equality of quality assessment used for assessment in school, so it can not guarantee the quality of the justice due to differences in a given test. It is very urgent to find a solution, because the value of the school is used also in the new admissions system (SNMPTN) invitation. The problem is very urgent to find a solution is to produce a standardized assessment system through school equivalency exam quality using equiting process and question bank. In most large-scale testing programs, the preparation of similar tests which were extremely important. This should be done for the rapid treatment in the event of a leak test and to compare the results of the test participants using different tests such. This activity can be done using the response theory item (item response theory). Due to the widespread use of computer technology, the utilization of virtualization as computer management instruction has provided opportunities for schools, teachers and students to interact with the server to access facilities, virtual desktop and applications without having to invest and maintenance independently. It is becoming an increasingly easy opportunity to do as the development of data networks increasingly varied and widespread.

(Times New Roman 10 pt, single space, right-left indent 1,5, justified)

**Key words**: standardized tests, graduation determination, equalization problem

## INTRODUCTION

Polemic developed in the community that there is no viable standardized assessment used to equalize the quality of the test in determining the final school exams should be overcome with a good system and ensure fairness for all Indonesian citizens. In the test 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 a certain level of equality some test devices can be implemented at the time of developing the test itself, but usually varies between a test device with other test devices, especially in terms of level of difficulty. This can be overcome by conducting equivalency between the test scores in a way that is appropriate and correct. Often found in schools, different test participants must be measured by different tests

even though the tests are not necessarily equivalent and are expected to measure the nature and demands of achieving the results that can be compared (Tumilisar, 2006: 3).

Although to a certain extent equality of some tests may be pursued at the time of preparing the tests itself, but in general the level of variation between tests difficult persists (Swediati, 1997: 1). In addition, equating tests necessary to remember that compose test truly parallel is not easy. So empirically make two tests are the same, never completely parallel, reliable or unidimensional, so that the resulting scores-scores can not be compared (Grounlund, 1985: 169). If the test results are used to determine the increase in class or program majors, of course, it becomes unfair because it does not do the equivalence of scores for the different tests. Therefore, it is important to do the adjustment of the test scores so that participants of different tests, using different tests can be compared.

These problems can be overcome by doing equivalency scores obtained from the participants who took the tests. Statistical process known as equating method (equiting), has been developed to address this problem. In other words, equating is a process to determine the relationship between the scale scores of two or more tests that test-scores scores are treated fairly. Activity equivalency test can be done by developing a system conversion unit test system to another test unit so that once converted scores from the two test devices become equal and interchangeable. This activity can be done by using Classical test theory and the theory of grain responsiveness. In this article the discussion is focused on the application of the response theory item (item response theory) using Quest program. Application of the theory of the response grains in equalizing the test is very useful especially for the development of a question bank. For that in this study developed a standardized assessment models based CMI (Computer Management Instructional) to ensure equality of quality assessment as a graduation in the determination of the data base system that is equitable School Final Examination.

## RESEARCH METHOD

Methods of Research and Development (R & D) used in developing the model assessment-based CMI (Computer Management Instructional), using the five phases of design activity spiral model adapted from 'Five phases of instructional design'. In the process of vertical equating use common-item nonequivalent groups design and determination of equating coefficients with the QUEST program, and in the quality of the tests used equating EXEL Program. The trial results equating, based on the results of the linear equating equation equating the third package was found that daily about Physics (The topics Quantities and Units and Motion).

CMI-SIPSMA applications used in the final school equivalency exam is a system based on client-server where the client computer machine only integrated with the end-user CMI-SIPSMA and client requirements. While the machine can be integrated with a server computer system database (database) and server requirements. CMI-SIPSMA Applications can also be applied to a machine that has a computer wrote a whole section of the system: the system end-user CMI-SIPSMA, server and client requirements, along with the base system database (database).

At CMI-SIPSMA applications, security and access rights are developed with user-level security (User) and User Roles (User Role). Each user is based on each individual teacher at each school. Only the user "admin" who act as Super User, and Administrator user role as the user "default" by not based on the individual teacher.

## RESULT AND DISCUSSION

Creating a test equivalent to two packs or more, of course, is not easy or even impossible, because there must be a difference. This is because almost not possible to organize a multi pack test that truly parallel (Petersen, Kolen, & Hoover, 1989). Although the authors tests using the

same test specifications in writing an item-item and just change the numbers, there is no guarantee that the level of difficulty of these items will be the same. Especially if that is different is the key word and the contents of the answer choices. According Angoff (1971) and Kolen (1988) as cited in Hambleton (1991), the equating method is divided into two categories, namely: 1) equatingequipercentile, and 2) linear equating (linear equating). The first category is an improvement scores by making a comparison between the test scores of X and Y be equivalent if the order of percent rank of each group is the same.

Furthermore, to equalize the score in two different tests, then a second test proficiency level should be given to examine the same group. Later in the second category, it is assumed that the test scores x x and y on test scores Y has a unidirectional relationship / line (linearly related). According Tumilisar (2006), equating methods are ways to find the relationship equating two test scores from two different research instruments using certain statistical and data collection specific to the design of data collection. Equipercentileequating method is divided into two, namely:

1. Equating method equipercentile chain is how to find equivalence equipercentile two test scores from two different research instruments, data collection is done with anchor test design and test nonekivalen anchor is an internal anchor tests using certain statistics. Equipercentile equivalence is calculated by the method of direct equipercentileequating separately on the test scores of both instruments, each of the test anchors, without the use of synthetic populations.
2. The method of frequency estimation equipercentileequating is how to find equivalence percentile two test scores from two different research instruments using certain statistical, and data collection is done by design unequivalence test anchors and anchor test is a test of the internal anchor. Equipercentile equivalence is calculated by estimating the cumulative distribution of two test scores of each of the anchor tests, using synthetic populations. The process of equating of multiple device test (equating) can be done in two ways, namely equating horizontally and vertically. Equating process obtained from two different test devices but measuring the same thing called horizontal equating. The process of equating of the two groups of participants of different tests in the levels / levels of education, but given the same problem called vertical equating (Crocker &Algina, 1986).

Basically equating aims to level the scores by comparing the scores obtained from working on a test device with scores obtained from other test devices that work is done through the process of equalizing the scores on the test device (Hambleton&Swaminthan, 1991). According to Zhu (1998), Silverback-scores on test A and test B can be synchronized if they meet four conditions, namely: 1) measures the ability or the same characteristics. So the tests are composed of different lattice can not be compared; 2) after equating, frequency distribution of scores on a test should be the same as the frequency distribution of scores on tests of B, so that scores on the test A and test B are interchangeable after equating; 3) equivalency test should be free of data or job candidates in the process of equating, and conversion from equating should apply to all similar situations; and 4) the transformation should be the same regardless of which test is used as a base or reference conversion, which means that the interpretation should be equally good scores equating of test A test to B or from B test to test A.

Lord (1980) put forward the notion or idea of equality in a number of implications, namely: 1) measurement tests with different properties can not be compared; 2) raw scores on the same test is not consistent, it can not be done equating process; 3) raw scores on tests with varying difficulty can not be compared because the test would not be consistent at the same level of difficulty; 4) mistakes or errors on test scores or package A and B can not be compared unless the tests are actually parallel; and 5) a perfect test reliability can be done equating.

Equating is done by converting one package to another package, which measures the ability of the package the same. Equivalency test device is the creation of a number of decisions of the scores obtained from a packet to be adjusted to different forms of the difficulty level. If

there is a package X is more difficult than the package Y, then X to Y equating package produce higher values of X package or valuable if equated to package Y (Crocker and Algina, 1986). There are three basic in designing the data to be retrieved and analyzed in doing equivalency test (Kolen& Brennan, 2004), namely: 1) the design of the data collected from the two groups were tested in different packages with the same grating, wherein the second division of the package are random or random; 2) for the equating process, one test group was given a package after it tested again with the package B, and another group was given first package B then rework package A; and 3) the instrument test given to different examinees. But in the second package contained the test anchor (anchor test) were given to all participants of the test. Anchor test that is used as a benchmark to perform equating. Participants test in this case does not need to be divided at random or random although the random division also will not affect this model.

The first test equating method is a method of regression. Determination of conversion constants a and b are regression method performed by observing the response of the test participants on both the X and Y. Estimation test item parameters and parameters of the ability of participants meet the following linear regression equation:

y = ax + b + a with a = r xyxy / Sx and b = y, - ax

Description:

y: estimation of ability or item parameter estimates on the test device Y

x: estimation of ability or item parameter estimates on the test device X

rxy: the correlation coefficient between X and Y

y, x: mean of y and x

Sy, Sx: standard deviation of x and y

E: error in estimating the regression error

The use of this method is not reciprocal (asymmetric) so inadequate for determining the conversion constants especially considering that the equivalency test two or more devices are in need of invariance requirements and the reciprocal of the test device synchronized. The second test equating method is the average sigma method. In this method, the determination of the conversion constants α and β according to the mean and sigma method is done by taking into account the value of the parameter estimate the difficulty level on the second test item test devices that bx and by. According Hambleton&Swaminathan (1985: 26), the relationship between the estimated parameter or parameters of the test item in the second participant's ability to be synchronized test devices and determination of the conversion constants satisfy the following equation:

y = ax + b with a = Sy/Sxand b = $\hat{Y}$ – ax

Mean and sigma method is reciprocal so that the same way the relationship of y to x can be determined. However, according to Hambleton&Swaminathan (1991: 26) argues that the mean and sigma equating method does not consider the variation of the parameter estimation error standard item.

The third test equating method called the method of mean and sigma tough. Hambleton and Swaminathan (1991: 26), states that the mean and sigma equating method is not mempertimbagkan grain variation parameter estimation. Equating method mean and sigma tough considering the variation of the standard error of the parameter estimate grain. The steps in the determination of the conversion constants for equivalency test using this method are as follows (Sukirno, 2007: 312):

1. Determination of the weight of item parameters (wi) in each pair (bxi and BYI), namely: wi = [max {v (xi), v (yi)}] - 1 where: i = 1,2,3,4 ... .k, v (xi) and v (yi) is a variant of the test difficulty level parameter estimates X and Y.
2. Determination of the scaling weights wi scale using the formula: wi` = k = number of anchor point on the test device X and Y.

3. Calculation of the estimated weighted test X and Y, using the formula: xi` = wi`xi and yi` = wi`xi
4. Determination of the mean and standard deviation of the estimated weighted test X and Y, ie x, y, Sx`, Sy`.
5. Determination of the conversion constants α and β by using the mean and standard deviation of the weighted estimation is done by substituting the mean and standard deviation of the estimated weight of the equation equating scale.

According Stocking and Lord (Hambleton, 1985) in mean and sigma equating method, the process of determining the conversion constants do not pay attention to the possibility of extreme group scores, whereas the mean and sigma equating method can toughen scores improved by observing extreme groups.

While all four methods that can be used in the test is a method equating characteristic curve. Determination of conversion constants α and β with characteristic curve method, carried out with due regard to the value of the second test item parameter estimates about the devices that x and y. Mean and sigma equating method and the method of mean and sigma rigid in determining the conversion constants only take into account the existing relationship between item difficulty parameters on which the test device to test other devices. The relationship between parameters of different power on both the tests have not been considered. Rahayu (2008), states that the characteristic curve method considers information from different power parameters of grain and grain in determining the level of difficulty of the conversion constants. Therefore, the characteristic curve equating method considered the relationship between the parameters of different power and relationship difficulties between item difficulty parameter tests to be synchronized. In addition, also in the method of original scores observed characteristic curve (true score) candidates in the second test device.

There are three basic in designing the data to be retrieved or analyzed by equating (Crocker and Algina, 1986), (Yi, Kim and Brennan, 2007), namely;

1. Design the data collected from two groups or groups that differ in the test package with the same grating, wherein the second division of the package at random or random.
2. For the equating process, one test group was given a package after it in the test came back with the package B, and another group was given first package B then rework package A.
3. The difference in the test instrument given to examinees different. But in the second package are common items or anchor test given to all participants of the test. Anchor that is used as a benchmark to perform equating. Participants test in this case does not need to be divided at random or random although the random division also will not affect this model. (Crocker and Algina, 1986).

Illustration of equating the third draft of the above description, it can be seen as shown in the following table.

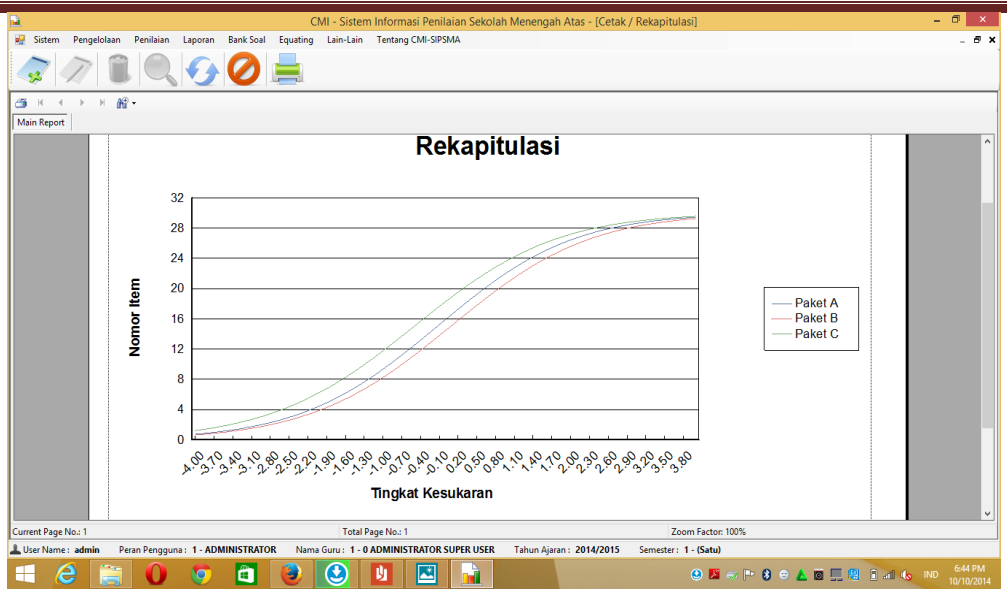


**Figure 1.**
**Summary of Results of equating Package A, B, and C**

A is a group 1were given a packet of X here in after given package Y, B is a group 1 were given a packet of 1 and there is an anchor (packet Z).

Thus it can be said that equating an empirical procedure performed to compare the scores of the test package with a package of other tests. By equating the right, then allow the direct conversion of the results of the exam candidates who take a different package. From the analysis of item response theory to the QUEST program, the obtained statistical information to third matter Package (equating be gradual process; Package A and Package B, Package B Package C, and Package C Package A), it can be concluded in the illustration the following:

**Table1**
**Results of Quest Problem Analysis Package A**

Item Estimates (Thresholds) In input Order 9/ 9/ 14 12:56
all on all (N =**** L = 30 Probability Level= .50)

| | ITEM NAME | SCORE | MAXSCR | THRSH | INFT MNSQ | OUTFT MNSQ | INFT t | OUTFT t |
|---|---|---|---|---|---|---|---|---|
| | | | 1 | | | | | |
| 1 | item 1 | 409316631 | .52 | 1.03 .02 | 1.07 | 3.0 | 4.7 | |
| 2 | item 2 | 491216553 | .25 | 1.00 .02 | 1.02 | -.1 | 1.4 | |
| 3 | item 3 | 472616561 | .31 | 1.00 .02 | 1.01 | -.4 | .5 | |
| 4 | item 4 | 994016645 | -1.09 | .98 | .98 | -3.4 | -2.1 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | .02 | | | | |
| 5 | item 5 | 769916624 | -.52 | .93 | .93 | -17.1 | -8.5 |
| | | | .02 | | | | |
| 6 | item 6 | 1122216627 | -1.45 | .94 | .91 | -10.2 | -7.4 |
| | | | .02 | | | | |
| 7 | item 7 | 739616635 | -.44 | 1.01 | 1.01 | 2.6 | 1.5 |
| | | | .02 | | | | |
| 8 | item 8 | 542216452 | .08 | 1.01 | 1.03 | 1.8 | 2.5 |
| | | | .02 | | | | |
| 9 | item 9 | 365216550 | .67 | 1.02 | 1.06 | 2.0 | 3.7 |
| | | | .02 | | | | |
| 10 | item 10 | 450316511 | .37 | 1.01 | 1.01 | 1.7 | .9 |
| | | | .02 | | | | |
| 11 | item 11 | 424116605 | .47 | .96 | .97 | -4.6 | -2.4 |
| | | | .02 | | | | |
| 12 | item 12 | 449016605 | .39 | 1.00 | 1.01 | .3 | 1.1 |
| | | | .02 | | | | |
| 13 | item 13 | 411216599 | .51 | 1.01 | 1.03 | .9 | 2.0 |
| | | | .02 | | | | |
| 14 | item 14 | 589416510 | -.05 | 1.05 | 1.07 | 8.0 | 6.0 |
| | | | .02 | | | | |
| 15 | item 15 | 1068616636 | -1.29 | .98 | 1.02 | -3.2 | 2.2 |
| | | | .02 | | | | |
| 16 | item 16 | 941416649 | -.95 | .95 | .94 | -13.0 | -6.0 |
| | | | .02 | | | | |
| 17 | item 17 | 749816596 | -.47 | 1.08 | 1.10 | 15.8 | 9.4 |
| | | | .02 | | | | |
| 18 | item 18 | 576416629 | .00 | .95 | .94 | -7.6 | -6.0 |
| | | | .02 | | | | |

*****Output Continues****

From the picture above, it was shown that the results of the linear equating line A package to package the same benchmark values B average, that's indeed the basis of the linear formula

equating. But the results of the linear equating to a low score is below the benchmark value, while a higher score will be above the benchmark value of it is because the process of equating performed a difficult package to package easily. When the equating process of the package easily kepaket difficult then the line would otherwise linear equating results.

## CONCLUSION AND SUGGESTION

In most large-scale testing program, the preparation of the tests are equivalent is a very important activity. This should be done for the rapid handling in the event of a leak test and to compare the results of the test participants using different tests such. This activity can be done by using the response theory item (item response theory). Because it is used in large scale utilization of computer technology management system (CMI) has provided opportunities for schools, teachers and students to interact with the facility to access servers, virtual desktops and applications without having to make an investment and maintenance independently. This becomes a more convenient opportunity to do with the development of data networks increasingly varied and widespread.

Standardized Assessment Model Based CMI (Computer Management Instructional) can only be developed to the level of high school, so the development still requires review and better test, given the still very heterogeneous quality of schools in the territory of the Republic of Indonesia. We hope slightest contribution that can still provide benefits for the next research. Do not forget to thank DITLITABMAS Higher Education for funding this research through grant schemes Competence so this research done.

## REFERENCES

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

Angoff, W. H. 1982. Uses of Difficulty and Discrimination Indices for Detecting Item Bias In RA Berk. *Handbook of Methods for Detecting Item Bias.* Baltimore: John Hopkins University Press.

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.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement,* 28 (4),227-246.

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.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.),*Journal of Educational measurement* (4th ed., pp. 187{220). Westport, CT: Greenwood.

Kim, S., von Davier, A. A., & Haberman, S. (2008). Small-sample equating using a synthetic linking function. *Journal of Educational Measurement,* 45, 325{342}

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

Kolen, Michael J., dan Robert L. Brennan. 1995. *Test Equating.* New York: Springer Verlag.

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.

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement*, 46, 330{343}

Lord, F. M. (2009). The standard error of equipercentile equating. *Journal of Educational Statistics*,7, 165{174}

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.

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.

Rahayu, Wardani. 2008. Pengaruh Metode Linking Terhadap Banyak Butir False Positive pada Pendeteksian DIF Berdasarkan Teori Responsi Butir. *Disertasi.* Jakarta: Universitas Negeri Jakarta.

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.