JPII 5 (2) (2016) 268-277

Jurnal Pendidikan IPA Indonesia

http://journal.unnes.ac.id/index.php/jpii

# VALIDATION OF SCIENCE ACHIEVEMENT TEST WITH THE RASCH MODEL

**P. Susongko***

Pancasakti University, Tegal, Indonesia

**DOI: 10.15294/jpii.v5i2.7690**

Accepted: August 25th 2016. Approved: September 4th 2016. Published: October 2016

**ABSTRACT**

The purpose of this study was to validate the test items of science achievement, which were used as a test at Pancasakti Science Competition, in order to obtain enough valid test items with the Rasch model application. Validation model used in this research is Messick validity covering aspects such as (1) content, (2) substantive, (3) structural, (4) external and (5) consequential. To achieve these objectives, this study investigates the quality of the items that include matching items, Person-Items Folder, Person/Item Folder, Information Function Tests, Person Fit Statistic, Collapsed Deviance, Casewise Deviance, Hosmer-Lemeshow, accuracy, sensitivity, specificity, unidimensional, invariance, separation and DIF. The test was given in the form of multiple choice as many as 40 items consisting of 15 items of physics, 10 items of chemistry, and 15 items of biology. The participants were 85 biology students who were given 60 minutes to do the test. Item analysis was conducted by using R 3:12 Program, eRm package version 0.15-6. The study results showed that the test items of science achievement were proven valid by the application of Rasch Model. The test items have met construct validity according to Messick (1996) which includes such aspects: (1) content, (2) substantive, (3) Structural, (4) External and (5) consequential.



**Keywords:** validation, Messick, Rasch model, science achievement test

## INTRODUCTION

Measurement is one of the initial steps in the evaluation program, which is a process for determining the characteristics of a number of attributes of the students, especially their learning achievement. In relation to student achievement, measurement is the process of giving the numbers that are expected to demonstrate the ability of students about a subject. To perform measurements, a measuring instrument is needed. Measuring tool that gives information about a person's position in the measured attribute.A good measurement tool would ensure valid and reliable results in order to accurately quantify the ability of students. Validity indicates the degree of evidence and theory that supports the interpretation of test scores as the intended use of the test (APA, 1999). Thus, validation is basically a collection of evidence to provide the scientific basis of test scores interpretation. The validation process used by the association of education and psychology in the United States (AERA and APA) can be divided into five types: (1) evidence based on the content of the test, (2) evidence based on the response, (3) evidence based on the internal structure, (4 ) evidence based on the linkages with other variables, and (5) evidence based on the consequences of the test (APA, 1999). Reliability is the coefficient which shows the degree of regularity or consistency of the measurement results of a test (Mardapi, 2012). Therefore, the reliability is not related to the test, but rather to the level of errors in the results of a test in the form of a score.

Messick (1996) argues that validity is a

***Alamat korespondensi:**
Email: kusumatirto@gmail.com

single concept which is expressed as the validity of a construct composed of six elements each: (1) content, (2) substantive, (3) structural, (4) generalizability, (5) external and (6) consequential. The validity of the suitability contents answers to all test items or tasks that involve cognitive processes and areas of the construct being measured. The content aspect of construct validity is at least associated with the suitability of the content, the representation, and the technical quality. The technical quality of test items is related to the issue of the level of readingaccuracy, language ambiguous and the answer keyaccuracy. The substantive aspectrefers to the substance of the content. This is achieved by investigating empirically in order to ensure that the test taker really involves the measured ability in answering the test items. For example, in the multiple-choice test, the test taker who chooses the wrong answer (distractor) has a low ability.

Structural aspect is linked to scoring, which is due prior to the priority of knowing the test structure before continue to scoring. Scores on multidimensional tests should be reported separately for each dimension. Generalization aspect assesses the extent of scores obtained that depict the actual capabilities of the test taker. External aspect reviews the extent of the scores obtained from the test correlated with other appropriate tests. Consequences aspect is merely associated with the meaning of the score obtained in the test or the implications of the score.

Classical test theory uses mathematical models that are very simple in showing the relationship among the score of observation, the actual score, and the score of error. This model was followed by a number of assumptions to simplify the formula in order to estimate the index of reliability and validity of an instrument. Although it has grown rapidly, classical test theory actually has several drawbacks. The weaknesses are: (1) the estimation of the test taker'sability depends on the characteristics of the test used; (2) the item parameter estimation depends on the ability of the test taker; and (3) the measurement error can only be sought for the group, not the individual (Mardapi, 2012). Moreover, the assumption of parallel tests are commonly used to search the index reliability tests, which are very hard statistically.

Some weaknesses in classical test theory are addressed by developing the Item Response Theory (IRT). There are three assumptions underlying the IRT. According to Hambleton et al (1991), the three assumptions are unidimensional, local independence and invariance parameter. Unidimensional means tests only measure one's ability. Unidimensional assumption can be proven by using either explanatory factor analysis and confirmatory. Local independence means in response to an item will not affect the other grains. Local independence assumption can be tested through the formula of probability and Chi-square dependence test. Parameter invariance means that the itemcharacteristicsare not dependent on the ability of participants and of the test parameter. This assumption is evidenced by estimating the parameters of ach item in the participantgroup of different tests, such as the group based on gender, place of residence, socioeconomic status, and others. With the implementation of IRT, the weakness of the application of classical test theory can be resolved, namely: (1) the estimation of the test taker's ability does not depend on the characteristics of the tests used; (2) the item parameter estimation does not depend on the ability of the test taker; and (3) the measurement error can be searched for each individual.

A modelthat is often used in IRT (Item Response Theory) is the logistic model. There are three logistic models namely one parameter model (1P), 2 parameter model (2P) and 3 parameters model (3P). 1P models only use the parameter level of difficulty of the grain. 2P model covers thelevel of difficulty and distinguishes grain, while 3P model covers1P's and 2P's parameters summed up with the parameter for pseudo guessing. One of 1P models widely used is the Rasch model. Due to the existing logistics model, Rasch is the simplest one, which only uses one point and constant parameter scale (D) of 1. Rasch links the opportunities of correct answer to each item (P (θ)) as a function of ability (θ) with a constant level of difficulty (b) through a relationship as in equation 1.

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

Rasch model has been developed separately from the IRT and even more widely to politomosscoring. With only involving one item, parameter estimation grain or participants in Rasch model needs less than the estimated data in other models. Application of the Rasch model in learning achievement, since being introduced by Georg Rasch in 1960, now extends not only in the world of education but also in the world of medicine and public health. The Rasch model application include: (1) the science and math education and attitudes towards science (Zain et

al, 2010), (2) the students' understanding of the philosophy of science (Neumann et al, 2010), (3) the opinion of math teachers towards the practice of teaching (Grimbeek & Nisbet, 2006), (4) the model's influence towards the law of science (Sjaastad, 2012), (5) and physics education (Oon & Subramaniam, 2012). Rasch model is also applied to formulate educational measurement as well as the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS) and other comparative tests (Schulz & Fraillon 2011; Stubbe, 2011; Wendt et al, 2011, Glynn, 2012). Rasch model is also used in the validation of a scale of attitudes toward science among countries compiled by TIMMS (Sabah et al, 2013). Likewise, reported by Long et al (2014) and Bansilal (2015), Rasch model can be applied either at the middle school math literacy tests in South Africa and measurement capabilities of science inquiry (Lou et al, 2015). The application of the Rasch model in the health sector includes: (1) Measurement of Oswestry Disability Index (ODI) (Lu et al, 2013), (2) Measurement of the General Health Questionnaire (GHQ) (Smith et al, 2010), (3) Measurement of the Malaria Diagnostic Test (Ayele et al, 2014).

Rasch analysis is basically aimed to meet the construct validity as described by Messick (1996). Compatibility of grain is used to check the item contents and the construct being measured. Meanwhile, fit test item is an item indicating that the contents have been relevant to the construct being measured. Person items folder and Person / Item Folder are two indicators the extent to which the test can measure the construct region to be measured (Baghai, 2008).

To examine the substantive type, testee and model compatibility test may be applied. Testee compatibilitytest (Person fit statistic) reveals the extent to which the testee's consistency in answering the questions. Person fit statistics empirically produces guides to what extent the pattern of testee'sanswers / responses can be predicted accurately by the model (Smith, 2001: 296). Consistent or fit testee answers all of the items that have the level of difficulty under their capabilities, and does not answer the items that have the level of difficulty above theirability. Inconsistent testee (Person Misfit) may be caused by carelessness, guessing, or dishonesty (cheating) (Wolfe & Smith, 2007: 209). Model capability test included the value of Collapsed Deviance, Casewise Deviance, Rost Deviance, Hosmer-Lemeshow, accuracy, sensitivity, and specificity (Mair et al., 2008). The validity of the structural type can be achieved by two methods, namely with unidimensional test and item parameter invariance-test. Person separation or stratum is a measure of the construct validity of external type (Wright and Stone, 1999). DIF identification (differential item functioning) is used to test the consequential validity (Baghaei & Amrahi, 2011). Table 1 is a summary of the criteria for a valid viewed from various aspects and criteria.

The purpose of this study was to validate the test items of science achievement, which were used as a test at Pancasakti Science Competition, in order to obtain enough valid test items with the Rasch model application. Validation model used in this research is Messick validity covering aspects such as (1) Person-Items Folder, (2) Person/ Item Folder, (3) Information Function Tests, (3) Person Fit Statistic, (4) Collapsed Deviance, (5) Casewise Deviance, (6) Hosmer-Lemeshow, (7) the value of accuracy, sensitivity and specificity, (8) the unidimensional test, (9) the Invariance test (LRtest), (10) the value of Person strata separation and (11) DIF according to the testee's gender.

## METHOD

This research was carried out on the test items of science learning achievement that were used for the first stage of the junior high school competition in Pekalongan on March 20,2016 at the Pancasakti University. The test was given in the form of multiple choice as many as 40 items consisting of 15 items of physics (numbers 1-15), 10 items of chemistry (numbers 16-25) and 15 items of biology (numbers 26-40). The test was attended by 85 junior high school students with the time allocation of 60 minutes. To achieve the objectives of the study, item analysis was conducted by using the Rasch model application withR 3:12 software program, eRm package, version 0.15-6. The program was selected as it is open source and has more than 3,000 packages for a specific purpose (Nettekoven, & Ledermüller, 2012). Extended Rasch model (ERM) is the application of Rasch model analysis with parameter estimation techniques using Conditional Maximum Likelihood (CML). R Program eRm-Package is more widely used because it has a lot of facilities and with the Maximum Likelihood estimation techniques, it does not require a prerequisite normal distribution, which results more consistent items (Mair, & Hatzinger, 2007). While for the unidimensional test, explanatory factor analysis with SPSS version 17.0 was used (Reise, et al, 2000).

**Table 1**. Criteria for a valid test viewed from various aspects and criteria

| Construct Validity Aspect | Indicator | Criteria |
|---|---|---|
| Content | Fit item test | P < 0.05<br>0,5 <MNSQ<1,5<br>-2,0 < ZSTD<2,0 |
| | Person-item Map | All level of item difficulties are in the testee's domain capability |
| | Person/Item Map | The level of item difficulties are is at or near tester's ability |
| | Information function test | Information function test has maximal values on the testee's domain capability |
| Substantive | Person fit statistic | P < 0.05<br>0,5 <MNSQ<1,5<br>-2,0 < ZSTD<2,0 |
| | Collapsed Deviance / Casewise Deviance /Hosmer-Lemeshow | P<0,05 |
| | Accuracy, sensitivity, and specificity | Close to 1,0 |
| Structural | Unidimensional test | There is one main factor that is pictures through Screen Plot's factor analysis result |
| | Invariation test (LRtest) | P< 0,05 |
| External | Person strata separation test | Close to 1,0 |
| Consequential | DIF | No significant DIF found |

## RESULT AND DISCUSSION

Item Characteristic Curves (ICC) on the Rasch model simply connects two variables: the testee'sability (latent dimension parameter) and the probability of answering correctly. The difficulty level (b) is the capability where the testee has a chance to answer correctly by half (0.5). All items of science learning achievement test can be described well as logistics functions ICC as can be seen in Figure 1.

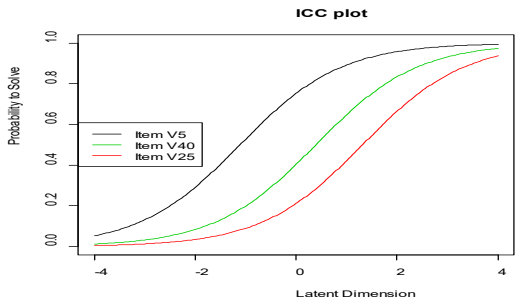


**Figure 1**. Item Characteristic Curva (ICC) untuk nomor 5, 25 dan 40

The analysis results of the difficulty level of the science learning achievement test with the Rasch model application are shown in Table 2. Test match items in Table 3, Person-Item folder in Figure 2 and Person / Item Folder in Figure 3, whereas function test and item informationin Figure 4. Of the 40 test items, all have difficulty values between -2 level up to +2. It is something that is ideal due to the learning achievement test based on the norm reference test, just like in a test competition, thedifficulty level should be in accordance with the testee'sability and away from an extreme value of difficulty level. It is seen that items measuring the ability of physics are more difficult compared to what of chemistry and biology. The most difficult item is number 7 (physics), while the most convenient is number 27 (biology).

Table 3 shows, of all the 40 test items, all qualifies as suitable items for measuring science learning achievement test, as seen from the MNSQ and ZSTD. When viewed from a significance test with Kai Square, it appears items number 7 and number 34 still have a chance (P) below 0.05 that therefore both of these items are declared as less suitable to be a measuring tool. However, if we take a significance of 1% (0.01), then the two points are still appropriate to describe the testee's response. From Figure 2 and 3, it is shown all levels of itemdifficulties are in the domain of testee'scapabilities except item number 7

**Table 2**. The difficulty level of the science learning achievement test with the Rasch model

| No | Difficulty level | Deviation Standard | No | Difficulty lever | Standard Deviation |
|---|---|---|---|---|---|
| 1 | 0.228 | 0.228 | 21 | -0.633 | 0.224 |
| 2 | -0.081 | 0.223 | 22 | 0.175 | 0.227 |
| 3 | 0.020 | 0.224 | 23 | 0.742 | 0.247 |
| 4 | -0.633 | 0.224 | 24 | 0.336 | 0.231 |
| 5 | -1.113 | 0.236 | 25 | 1.316 | 0.286 |
| 6 | 0.020 | 0.224 | 26 | -1.289 | 0.244 |
| 7 | 2.382 | 0.419 | 27 | -1.350 | 0.247 |
| 8 | 2.212 | 0.391 | 28 | -1.171 | 0.238 |
| 9 | 1.928 | 0.351 | 29 | 0.391 | 0.233 |
| 10 | 0.175 | 0.227 | 30 | 0.228 | 0.228 |
| 11 | 0.282 | 0.230 | 31 | 0.336 | 0.231 |
| 12 | -0.031 | 0.223 | 32 | -0.432 | 0.222 |
| 13 | 0.620 | 0.242 | 33 | -1.229 | 0.241 |
| 14 | 1.493 | 0.302 | 34 | 0.871 | 0.254 |
| 15 | 0.806 | 0.251 | 35 | -0.482 | 0.222 |
| 16 | -1.413 | 0.250 | 36 | -0.132 | 0.222 |
| 17 | -0.031 | 0.223 | 37 | -0.947 | 0.231 |
| 18 | -0.282 | 0.222 | 38 | -2.806 | 0.390 |
| 19 | 0.071 | 0.225 | 39 | -0.232 | 0.222 |
| 20 | -0.736 | 0.225 | 40 | 0.391 | 0.233 |

and number 38. Figure 4 also shows the test will provide information of high value in the range of -2 to + 2 capability.

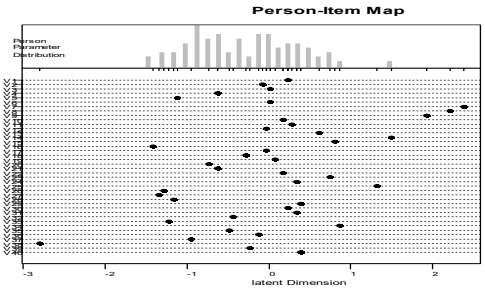


**Figure 2**. Person-Item Map test of science learning achievement test with the Rasch model.

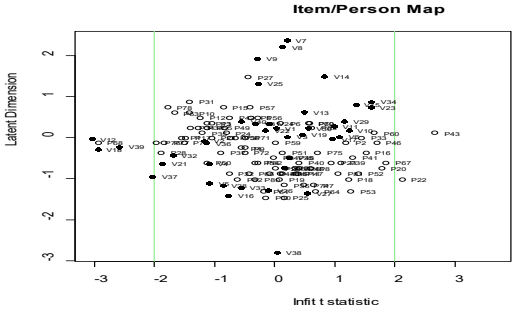


**Figure 3**. Person/Item Map test of science learning achievement test with the Rasch model.

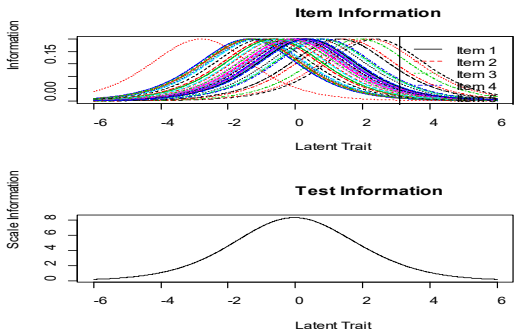


**Figure 4**. Information function of science learning achievement test with the Rasch model.

The analysis result of Good-of-Fit with the help of eRm model version 0.15-6 shows different parameters as shown in Table 4.

According to the 6 parameter values of Good-of-Fit Tests as described in Table 4, except the value of Collapsed Deviance, it is indicated that science learning achievement tests meet Substantive validity. Nevertheless, Deviance Collapsed parameter is significantly affected by the sample. By taking a 10% error, these tests still fulfill the prerequisites value of Collapsed Devian-

**Table 3**. Compatibility item of science learning achievement test with the Rasch model

| No Item | chisq | df | p-value | Outfit MSQ | Infit MSQ | Outfit t | Infit t |
|---|---|---|---|---|---|---|---|
| 1 | 84.955 | 84 | 0.450 | 0.999 | 1.000 | 0.03 | 0.02 |
| 2 | 76.969 | 84 | 0.694 | 0.906 | 0.924 | -1.29 | -1.17 |
| 3 | 88.310 | 84 | 0.353 | 1.039 | 1.013 | 0.51 | 0.21 |
| 4 | 77.485 | 84 | 0.679 | 0.912 | 0.927 | -1.11 | -1.10 |
| 5 | 73.152 | 84 | 0.795 | 0.861 | 0.899 | -1.16 | -1.09 |
| 6 | 93.686 | 84 | 0.220 | 1.102 | 1.074 | 1.27 | 1.07 |
| 7 | 120.042 | 84 | 0.006 | 1.412 | 1.031 | 0.99 | 0.20 |
| 8 | 91.508 | 84 | 0.270 | 1.077 | 1.009 | 0.32 | 0.13 |
| 9 | 62.437 | 84 | 0.962 | 0.735 | 0.903 | -0.77 | -0.30 |
| 10 | 94.298 | 84 | 0.207 | 1.109 | 1.095 | 1.20 | 1.23 |
| 11 | 93.118 | 84 | 0.233 | 1.096 | 1.080 | 0.97 | 0.98 |
| 12 | 66.584 | 84 | 0.919 | 0.783 | 0.807 | -3.03 | -3.05 |
| 13 | 93.421 | 84 | 0.226 | 1.099 | 1.049 | 0.76 | 0.49 |
| 14 | 120.795 | 84 | 0.005 | 1.421 | 1.158 | 1.56 | 0.83 |
| 15 | 115.222 | 84 | 0.013 | 1.356 | 1.167 | 2.12 | 1.34 |
| 16 | 72.338 | 84 | 0.814 | 0.851 | 0.909 | -0.95 | -0.76 |
| 17 | 90.716 | 84 | 0.289 | 1.067 | 1.064 | 0.88 | 0.96 |
| 18 | 68.710 | 84 | 0.886 | 0.808 | 0.826 | -2.92 | -2.96 |
| 19 | 89.568 | 84 | 0.319 | 1.054 | 1.031 | 0.67 | 0.45 |
| 20 | 84.753 | 84 | 0.456 | 0.997 | 1.009 | 0.00 | 0.15 |
| 21 | 72.475 | 84 | 0.811 | 0.853 | 0.879 | -1.90 | -1.87 |
| 22 | 81.841 | 84 | 0.546 | 0.963 | 0.985 | -0.39 | -0.17 |
| 23 | 104.432 | 84 | 0.065 | 1.229 | 1.192 | 1.50 | 1.60 |
| 24 | 83.187 | 84 | 0.505 | 0.979 | 0.989 | -0.17 | -0.10 |
| 25 | 79.209 | 84 | 0.627 | 0.932 | 0.942 | -0.23 | -0.28 |
| 26 | 85.370 | 84 | 0.438 | 1.004 | 0.984 | 0.08 | -0.12 |
| 27 | 93.546 | 84 | 0.223 | 1.101 | 1.057 | 0.70 | 0.54 |
| 28 | 74.277 | 84 | 0.767 | 0.874 | 0.915 | -0.99 | -0.87 |
| 29 | 98.243 | 84 | 0.137 | 1.156 | 1.104 | 1.40 | 1.16 |
| 30 | 89.014 | 84 | 0.333 | 1.047 | 1.041 | 0.52 | 0.54 |
| 31 | 80.694 | 84 | 0.582 | 0.949 | 0.970 | -0.46 | -0.32 |
| 32 | 75.892 | 84 | 0.724 | 0.893 | 0.896 | -1.53 | -1.70 |
| 33 | 75.293 | 84 | 0.740 | 0.886 | 0.941 | -0.84 | -0.56 |
| 34 | 119.844 | 84 | 0.006 | 1.410 | 1.214 | 2.28 | 1.61 |
| 35 | 85.641 | 84 | 0.430 | 1.008 | 1.013 | 0.13 | 0.23 |
| 36 | 77.071 | 84 | 0.691 | 0.907 | 0.927 | -1.30 | -1.14 |
| 37 | 68.361 | 84 | 0.892 | 0.804 | 0.840 | -1.96 | -2.04 |
| 38 | 88.647 | 84 | 0.343 | 1.043 | 0.981 | 0.24 | 0.04 |
| 39 | 70.784 | 84 | 0.848 | 0.833 | 0.845 | -2.50 | -2.59 |
| 40 | 80.438 | 84 | 0.590 | 0.946 | 0.949 | -0.46 | -0.55 |

ce. Value Accuracy, Sensitivity and specificity are also close to 1 or more than 0.6 so that science learning achievement test is considered suitable in describing the testee's response. Table 5 contains the majority of the test Person's match. Atthe significant of 5 percent, there are 5 testees who are inconsistent and rejected by the model, and there are 80 testees who are considered consistent. As a conlcusion, 94 percent of the population meets the substantive validity. The inconsistent five testees are testee number 22, 41, 43, 60, and 67.

**Table 4**. Good-of-Fit TestsParameter

| Parameter | df | Value | P Value |
|---|---|---|---|
| Collapsed Deviance | 878,187 | 820 | 0,078 |
| Hosmer Lemeshow | 30,027 | 800 | 0,00 |
| Casewise Deviance | 3958,932 | 3425 | 0,00 |
| Accuracy | | 0,714 | |
| Sensitivity | | 0,638 | |
| Specificity | | 0,776 | |

Figure 5 shows scree plot, which is the analysis results of unidimensional with the help of SPSS version 17,00.

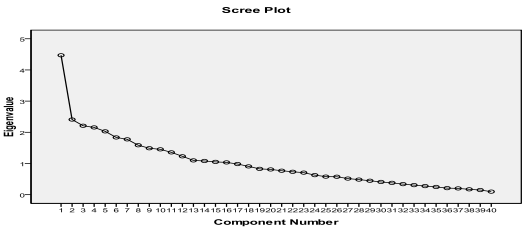


**Figure 5**. Scree plot of the respond factor analysis of science learning achievement.

Scree plot analysis shows that there is only one line while the others are sharply sloping that they cannot be considered as a factor. Therefore, figure 5 shows that the learning achievement tests can be expressed only consist of one factor or one-dimensional.

Invariance parameters test with Anderson LR-test shows a value of 123.193 with chi-square df by 39 with an opportunity by 0,000. At the 5% significance level can be concluded that there has been invariance parameter estimation. Person se-

**Table 5**. Person-matching test of science learning achievement test with the Rasch model

| NoPerson | chisq | df | p-value | Outfit MSQ | Infit MSQ | Outfit t | Infit t |
|---|---|---|---|---|---|---|---|
| 7 | 47.462 | 39 | 0.166 | 1.187 | 1.112 | 0.66 | 0.65 |
| 8 | 53.887 | 39 | 0.057 | 1.347 | 1.082 | 1.35 | 0.60 |
| 9 | 34.146 | 39 | 0.691 | 0.854 | 0.937 | -0.72 | -0.50 |
| 20 | 52.524 | 39 | 0.073 | 1.313 | 1.274 | 1.23 | 1.78 |
| 21 | 37.668 | 39 | 0.531 | 0.942 | 0.877 | -0.25 | -1.05 |
| 22 | 74.578 | 39 | 0.001 | 1.864 | 1.389 | 2.48 | 2.12 |
| 23 | 42.451 | 39 | 0.325 | 1.061 | 1.128 | 0.34 | 0.94 |
| 40 | 40.378 | 39 | 0.409 | 1.009 | 1.050 | 0.12 | 0.41 |
| 41 | 58.429 | 39 | 0.023 | 1.461 | 1.171 | 1.94 | 1.30 |
| 42 | 35.267 | 39 | 0.641 | 0.882 | 0.897 | -0.50 | -0.75 |
| 43 | 72.579 | 39 | 0.001 | 1.814 | 1.345 | 3.51 | 2.67 |
| 44 | 36.750 | 39 | 0.573 | 0.919 | 0.980 | -0.22 | -0.08 |
| 59 | 38.414 | 39 | 0.496 | 0.960 | 0.997 | -0.15 | 0.01 |
| 60 | 67.324 | 39 | 0.003 | 1.683 | 1.205 | 3.04 | 1.66 |
| 61 | 34.903 | 39 | 0.657 | 0.873 | 0.951 | -0.27 | -0.18 |
| 62 | 36.172 | 39 | 0.600 | 0.904 | 0.955 | -0.35 | -0.29 |
| 63 | 26.831 | 39 | 0.930 | 0.671 | 0.775 | -1.56 | -1.67 |
| 64 | 54.512 | 39 | 0.051 | 1.363 | 1.126 | 1.05 | 0.67 |
| 65 | 42.689 | 39 | 0.316 | 1.067 | 1.019 | 0.35 | 0.18 |
| 66 | 40.744 | 39 | 0.394 | 1.019 | 0.928 | 0.16 | -0.42 |
| 67 | 57.210 | 39 | 0.030 | 1.430 | 1.267 | 1.73 | 1.85 |
| 68 | 24.736 | 39 | 0.963 | 0.618 | 0.690 | -2.25 | -2.94 |

**Table 6**. Wald test according to gender of science learning achievement test.

| No item | Z statistics | P Value | No | Z statistics | P Value |
|---|---|---|---|---|---|
| 1 | 0.127 | 0.899 | 21 | 0.628 | 0.530 |
| 2 | -0.260 | 0.795 | 22 | 0.482 | 0.629 |
| 3 | -0.467 | 0.641 | 23 | 1.061 | 0.289 |
| 4 | 1.125 | 0.261 | 24 | 0.948 | 0.343 |
| 5 | 1.888 | 0.059 | 25 | 0.216 | 0.829 |
| 6 | -0.965 | 0.335 | 26 | -1.212 | 0.225 |
| 7 | 0.096 | 0.923 | 27 | 0.827 | 0.408 |
| 8 | 0.688 | 0.491 | 28 | -1.408 | 0.159 |
| 9 | 0.911 | 0.362 | 29 | 1.107 | 0.268 |
| 10 | -1.034 | 0.301 | 30 | -1.883 | 0.060 |
| 11 | -1.252 | 0.211 | 31 | -0.597 | 0.550 |
| 12 | 0.881 | 0.378 | 32 | -1.759 | 0.079 |
| 13 | -0.361 | 0.718 | 33 | 0.578 | 0.563 |
| 14 | 0.637 | 0.524 | 34 | -0.269 | 0.788 |
| 15 | 0.686 | 0.492 | 35 | -0.927 | 0.354 |
| 16 | 0.115 | 0.908 | 36 | 0.586 | 0.558 |
| 17 | 0.881 | 0.378 | 37 | 0.264 | 0.791 |
| 18 | -0.346 | 0.729 | 38 | -1.745 | 0.081 |
| 19 | -0.821 | 0.411 | 39 | 0.294 | 0.768 |
| 20 | -0.166 | 0.868 | 40 | 0.072 | 0.942 |

paration value reliability or strata, with the help of software version 0.15-6 eRm program, shows the value of 0.6552. Although this value is not ideal but still considered to conform to external validity (Chyi Lo et al, 2014). DIF analyzes by sex is used to determine the validity of the consequential aspects.

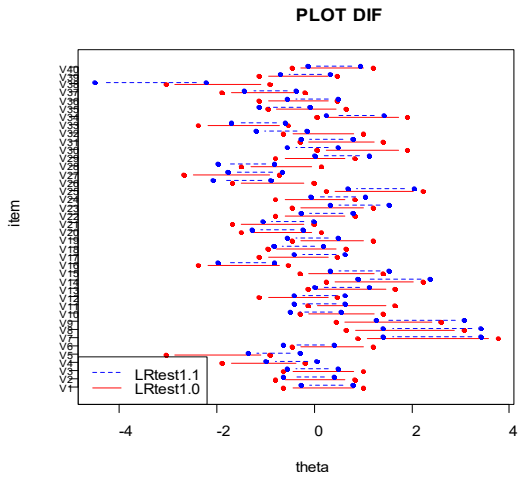


**Figure 6**. DIF plot ofscience learning achievement test items.

From Table 6, it is known as much as 40 test items do not contain visible from DIF Value of Test Wald and everything gets above 0.05. Therefore, it can be concluded that the science learning achievement test used is not gender biased. This is further explained in Figure 6 that describes the LR test for each item between men and women, where all the items show different prices and no coincident. This research result shows science learning achievement test used by Pancasakti Science Competition is consistent without bias.

## CONCLUSION

According to the results of this study, it can be concluded that science learning achievement tests used in Pancasakti Science Competition is proven to be valid by the application of the Rasch model. The science learning achievement test items have met construct validity according to Messick (1996) which includes: (1) content, (2) substantive, (3) structural, (4) external and (5) the consequential. The science learning achievement test items can then be further developed as a standardized test for implementation of Pancasakti Science Competition.

## REFERENCES


APA. 1999. Standards for educational and psychological test and manuals .Washington DC

Ayele, Dawit G, Zewotir, Temesgen, Mwamb., & Henry. 2014. Using Rasch modeling to re-evaluate rapid malaria diagnosis test analyses. *International Journal of Environmental Research and Public Health,* 11(7 ): 23-32.

Bansilal, Sarah. 2015. A Rasch analysis of a grade 12 test written by mathematics teachers. *South African Journal of Science* , 111.5(6) : 1-9. Retrieved from: http://search.proquest.com/docview/1733144353/fulltext/C1F8BDAEDD3D4127PQ/2?accountid=62691

Baghaei,P. 2008. The Rasch model as a construck validation tool. Rasch Measurement Transaction , 22(1): 1145-1146. Retrieved from: http://www.rasch.org/rmt/rmt221a.htm.

Baghaei,P & Amrahi, N. 2011. Validation of a Multiple Choice English Vocabulary Test with the Rasch Model. *Journal of Languange Teaching and Research* , 2(5) : 1052-1060

Chyi Lo,Wen-Miin Liang, Liang-Wen Hang , Tai-Chin Wu , Yu-Jun Chang & Chih-Hung Chang. 2015. A psychometric assessment of the St. George's respiratory questionnaire in patients with COPD using rasch model analysis, *Health and Quality of Life Outcomes,* 13 (1): 45-60 , August 2015. Retrieved from:http://hqlo.biomedcentral.com/articles/10.1186/s12955-015-0320-7

Glynn,S.M. 2012. International assessment: A Rasch model and teachers' evaluation of TIMSS science achievement items**.***Journal of Research in science teaching (JRST)* , 49 (10): 1321–1344

Grimbeek, P., & Nisbet, S. 2006. Surveying primary teachers about compulsory numeracy testing: Combining factor analysis with Rasch analysis. *Mathematics Education Research Journal, 18*(2): 27-39.

Hambleton,R.K, Swaminathan,H & Rogers,H.J. 1991. Fundamentals of item response theory . Newbury Park London New Delhi: Sage Publication

Long, Caroline., Bansilal, Sarah., Debba., & Rajan. 2014. An investigation of Mathematical Literacy assessment supported by an *application of Rasch measurement , Pythagoras 35 (1)* : 1-17. Retrieved from:http://search.proquest.com/docview/1636192258/fulltext/C1F8BDAEDD3D4127PQ/10?accountid=62691

Lou, Yiping, Blanchard, Pamela, Kennedy., & Eugene. 2015. Development and validation of a science Inquiry skills assessment. *Journal of Geoscience Education* , 63 (1 ): 73-85. Retrieved from::http://search.proquest.com/docview/1700950287/fulltext/1573E9C5292A4E15PQ/19?accountid=62691

Lu, Yen-Mou, Wu, Yuh-Yih, Hsieh, Ching-Lin, Lin, Chih-Lung. 2013. Measurement precision of the disability for back pain scale-by applying Rasch analysis**.** *Journal of Health and Quality of Life Outcomes* ,11, Page 119 Retrieved from : http://search.proquest.com/docview/1412368063?accountid=62691

Mair, P & Hatzinger , R . 2007. CML based estimation of extended Rasch models with the eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20(9): 1-20.

Mair, P, Reise, S.P & Bentler, P,M. 2008. IRT Goodness og fit Using Approaches from Logistic Regression. UCLA.California, US. Retrieved from:http://escholarship.org/uc/item/1m46j62q#page-1

Mardapi,D. 2012. Pengukuran Penilaian Dan Evaluasi Pendidikan .Yogyakarta: Nuha Medika

Messick. 1996.Validaty and washback in language testing, Languange Testing, 13(3):241-256.

Nettekoven, Michaela& Ledermüller, Karl**.** 2012**.** Assess the Assessment: An Automated Analysis of Multiple Choice Exams and Test Items.European Conference on e-Learning**:** 397-XV. Kidmore End: Academic Conferences International Limited**.** AustriaRetrieved from:.(http://search.proquest.com/docview/1328342405/fulltext/761A0EBA4252446BPQ/2?accountid=62691

Neumann, I., Neumann, K., & Nehm, R. 2010. Evaluating instrument quality in science education: Rasch-based analyses of a Nature of Science test. *International Journal of Science Education, 33*(10): 1373-1405.

Oon, P.T., & Subramaniam, R. 2012. Factors influencing Singapore students' choice of physics as a tertiary field of study: A Rasch analysis. *International Journal of Science Education, 35*(1), 86- 118. doi: 10.1080/09500693.2012.718098.

Reise, SP, Waller, NG & Comrey, AL. 2000. Factor analysis and scale revision. Psychological Assesment, 12 (3): 287-297 Retrieved from : http://psych.colorado.edu/~willcutt/pdfs/Reise_2000.pdf

Sabah, Saed., Hammouri, Hind., & Akour, Mutasem. 2013. Validation of a scale of attitudes toward science across countries using rasch model :findings from TIMSS. *Journal of Baltic Science Education* ,12 (5): 692-702. 11p.Retrieved from :http://oaji.net/articles/2015/987-1425810820.pdf

Schulz, W., & Fraillon, J. 2011. The analysis of measurement equivalence in international studies using the Rasch model. *Educational Research and Evaluation, 17*(6): 447-464.

Sjaastad, J. 2012. Measuring the ways significant persons influence attitudes towards science and mathematics. *International Journal of Science Education*, 35(2): 192-212.

Smith, A.B, Fallowfield, L. J, Stark, D. P, & Velikova, G. 2010. A Rasch and confirmatory factor analysis of the General Health Questionnaire (GHQ) – 12. *Journal Health and Quality of Life Outcomes*, 8(1): 45-56. Retrieved from :http://search.proquest.com/docview/902252382?acc

ountid=62691

Smith , E.V.Jr. 2001. Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective . *Journal of Applied Measurement* ,2(3): 281-311

Stubbe, T. C. 2011. How do different versions of a test instrument function in a single language? A DIF analysis of the PIRLS 2006 German assessments. *Educational Research and Evaluation, 17*(6): 465-481.

Wendt, H., Bos, W., & Goy, M. 2011. On applications of Rasch models in international comparative large-scale assessments: A historical review. *Educational Research and Evaluation, 17*(6): 419-446.

Wolfe,E.W & Smith, E.V.Jr. 2007. Instrument development tools and activities for measure validation using Rasch models; Part II-validation activities. *Journal of Applied Measurement*, 8(2): 204-234.

Wright, B.D & Stone, M.H. 1999. Measurement essentials. Wide Range Inc, Wilmington.Retrieved from:(http://www.rasch.org/measess/me-all.pdf 28Mb)

Zain, A. N. M., Samsudin, M. A., Rohandi, & Jusoh, A. 2010. Using the Rasch model to measure students' attitudes toward science in 'low performing' secondary schools in Malaysia. *International Education Studies, 3*(2): 56-63.