Messick Validation on the Simulation Test of National Exam Using Rasch Model

Endro Suseno[1], Purwo Susongko[2], Dewi Apriani[3]
Pedagogy Masters Study Program,
Pancasakti University Tegal, Indonesia[1,2,3]
{endrosuseno.skom@gmail.com[1]}

Abstract. This study aims to validate the test items of the computer-based national exam simulation test subjects of the vocational theory of computer engineering competence and vocational high school networks in Tegal City using the Rasch model with the Messick validation model which includes validity aspects: (1) content, (2) substantive, (3) Structural, (4) External and (5) Consequential. To achieve this goal, this study reveals the quality of items including item suitability, Person-item Map, Person / Item Map, Test Information Functions, Person fit statistics, Collapsed Deviance, Casewise Deviance -Hosmer-Lemeshow, accuracy, sensitivity, specificity, unidimensional, invariance, separation and DIF. This research is a quantitative study, namely the evaluation of learning outcomes using quantitative descriptive methods with data collection techniques using documentation. Respondent data of 151 students were taken from two schools in Tegal City with 40 items. Data analysis using software R Program Studio version 4.02. Validation of the construct with Rasch modeling gave the following results: (1) The difficulty level of the items was in the range -2 to 2, (2) There were 40 items that matched the modeling, (3) There were 90.06% of student responses that matched the modeling, (4) There are 3 items that contain DIF. Based on the consideration of all aspects of validity, there are 37 items out of 40 that are suitable for use as test items in the test. The conclusion is that the analysis of the items using the Rasch model in the computer-based national exam simulation test, vocational theory exam subjects are declared valid with good item difficulty level.

Keywords: Messick Validity, National Exam Simulation Test, Rasch Model

1 Introduction

The National Examination is a system of evaluating standards for primary and secondary education nationally and the quality equality of education levels between regions carried out by the Education Assessment Center (Puspendik). The evaluation results will be used as the basis for changing the exam system for the better, and this will automatically change the learning method for the better as well. Computer-Based National Examination (UNBK) or also known as Computer Based Test (CBT) is a system for implementing the national exam using a computer as a test medium. Prior to the implementation of UNBK, Puspendik held an exam simulation which aims to prepare students for the exam so that they are accustomed to

ICONS 2020, October 13-14, Tegal, Indonesia

DOI 10.4108/eai.13-10-2020.2303711

operating equipment and are accustomed to doing computer-based exam questions. In addition, the exam simulation also aims to prepare students to face exam questions according to the national exam question grid.

Validity is the extent to which the test measures what it is intended to measure. In general, there are three approaches in examining the validity of a measuring instrument, namely 1) content validity, 2) construct validity, and 3) criterion validity (Suryabrata, 2005). Content validity is validity that focuses on what elements are in the measurement (Coaley, 2010), so that rational analysis is the main process carried out in content validity analysis (Azwar, 2005). Construct validity is a picture that shows the extent to which the measuring instrument shows results in accordance with the theory (Azwar, 2005). The process of testing construct validity is to connect the measuring instrument with other measuring tools that have the same concept or with other measuring tools that are theoretically related to it (Murphy & Davidshofer, 1991). The validity of the criteria is to link the measuring instrument with other measuring instruments as a criterion, whether the measuring instrument can be explained by its correlation with the criteria based on existing theories (Devellis, 2003). The validity to be measured in this study is the construct validity according to Messick. There are six items of construct validity concept according to Messick (1995).

a. Consequential- What are the potential risks if the scores are, in actuality, invalid or inappropriately interpreted? Is the test still worthwhile given the risks?
b. Content- Do test items appear to be measuring the construct of interest?
c. Substantive- Is the theoretical foundation underlying the construct of interest sound?
d. Structural- Do the interrelationships of dimensions measured by the test correlate with the construct of interest and test scores?
e. External- Does the test have convergent, discriminant, and predictive qualities?
f. Generalizability- Does the test generalize across different groups, settings and tasks?

To analyze the items in the simulation of the national exam, the TKJ vocational theory exam in this study uses the item response theory (IRT). The model that is often used in the IRT (Item Response Theory) is a logistic model. There are three logistic models, namely the 1 parameter model (1P), the 2 parameter model (2P) and the 3 parameter model (3P). The 1P model only uses the item difficulty level parameter, the 2P model uses the item difficulty level and the item distinguishing power, while the 3P model uses the coincidence parameter to answer correctly (pseudo guessing). One type of 1P model that is widely used is the Rasch Model. This is due to the existing logistic model, the Rasch Model is the simplest model with only one item parameter and uses a scale factor constant (D) of 1. The Rasch model relates the probability of answering each item correctly (P (θ)) as a function of ability (θ) with a constant item difficulty level (b) through the relationship as in equation 1.

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

The advantages of the Rasch model include being able to predict missing data, which is based on a systematic response pattern; able to produce standard error measurement values for the instruments used which can improve the accuracy of calculations; and the calibration is carried out simultaneously in three ways, namely the measurement scale, respondents, and items (Sumintono & Widhiarso, 2015). The superiority of the Rasch model is very suitable for use in this study because it is to evaluate the ability of the test items for the simulation of the

national exam in vocational theory examinations for students. The use of the Rasch model is more effective than classical analysis (Fisher, 1993).

The results of the observation show that the vocational theory test questions used in SMK Negeri 2 and 3 Tegal use multiple choice questions compiled nationally by the Puspendik and are not validated by the teacher and these questions have never been worked on before by students so that these questions cannot known the level of quality. Based on the above background, the problem to be examined in this study is how the validity of the national exam simulation questions using the Rasch model in terms of the validity of the Messick construct on the subject of the Vocational Theory of Computer Engineering and Networking Vocational Schools in Tegal City.

2 Research Methods

2.1 Data source

The data used in this study are secondary data from the simulation results of the national exam for the Vocational Theory Computer and Network Engineering Vocational School 2 and 3 Tegal City. The sample was 151 students with 40 multiple choice questions.

2.2 Method of Analysis

This research method uses a quantitative method with an exploratory descriptive approach that is used to describe, explain, or summarize various conditions, situations, phenomena, or various research variables according to events as they are which can be photographed, interviewed, observed, and which can be expressed through materials. documentary.

Quantitative data analysis with the help of the R version 4.02 program was used to obtain item parameters fit with the Rasch model. The determination of reliability is seen from the amount of Item Reliability and the overall item reliability value shown by the large Cronbach alpha value, while the item limit is declared fit with the model if it has MNSQ Outfit between 0.5 to 1.5; The ZSTD outfit is between -2.0 to 2.0; as well as item correlation values with total scores (point measure correlation) ranging from 0.4 to 0.85 (Sumintono & Widhiarso, 2015)

In Susongko (2016) the validation model used to reveal the quality of test item items includes the following aspects: (1) item fit test, (2) Person-item Map, (3) Person / Item Map, (4)) Test Information Function, (5) Person fit statistics, (6) Collapsed Deviance / Casewise Deviance / Hosmer-Lemeshow values, (7) accuracy, sensitivity, and specificity values, (8) Unidimensional test, (9) Invariance test (LRtest)), (10) separation Person strata value and (11) DIF based on the sex of the testees.

Table 1. Valid Test Criteria Seen From Various Aspects of Validity and Criteria

Construct Validity Aspect	Indicator	Criteria
Content	Itemfit	P > 0.05 0,5 <MNSQ<1,5 -2,0 < ZSTD<2,0
	Person-item Map	All item difficulty levels are in the testee ability domain
	Person/Item Map	The testee's ability is the same or close to the item difficulty level

Construct Validity Aspect	Indicator	Criteria
	Information Test Function	The test information function has the maximum value in the testee ability domain
Substantif	*Person fit statistic*	P > 0.05 0,5 <MNSQ<1,5 -2,0 < ZSTD<2,0
	Collapsed Deviance / Casewise Deviance /Hosmer-Lemeshow accuracy, sensitivity, dan specificity	P<0,05 approaching 1,0
Structural	Unidimension Tes	There is one main factor that is described through the Scree Plot of the factor analysis results
	LRTest	P> 0,05
External	Separation Reliability	approaching 1,0
Consekuential	DIF	there is no significant DIF

(Susongko, 2016)

3 Results and Discussion

3.1 Measurement of Content Validity (Content)

Content validity indicates whether all test items or tasks that involve cognitive processes in answering them are truly appropriate and are representative of the construct area being measured. There are 4 indicators in the aspect of content validity, namely: item fit test, Person item-map, Person / Item Map, and test information function. Item fit basically explains whether an item is functioning to take measurements normally or not. Quantitatively the test items are declared fit or can function properly if the MSQ Outfit value is between 0.5 to 1.5 while the outfit t value is between -2 to 2.0 and the chance of acceptance of model fit is greater than 0.05 (p> 0.05). The results of the Item Fit Test can be seen in table 2.

Table 2. Item Fit Test (Itemfit) on The Simulation Tes Using Rasch Model

	Chisq	df	p-value	Outfit MSQ	Infit MSQ	Outfit-t	Infit-t
V1	133.438	150	0.830	0.884	0.940	-0.927	-0.604
V2	110.534	150	0.993	0.732	0.933	-0.916	-0.203
V3	148.100	150	0.529	0.981	0.964	-0.369	-0.817
V4	170.463	150	0.121	1.129	1.097	1.463	1.462
V5	164.507	150	0.198	1.089	1.039	0.982	0.583
V6	137.664	150	0.756	0.912	0.974	-0.709	-0.246
V7	188.268	150	0.019	1.247	1.185	4.487	4.325
V8	119.341	150	0.969	0.790	0.900	-1.022	-0.562
V9	154.505	150	0.384	1.023	1.035	0.283	0.523
V10	154.079	150	0.393	1.020	1.030	0.431	0.770
V11	123.756	150	0.942	0.820	0.849	-3.436	-3.696
V12	132.785	150	0.840	0.879	0.974	-0.386	-0.049
V13	117.689	150	0.976	0.779	0.951	-0.399	-0.013
V14	104.102	150	0.998	0.689	0.911	-1.231	-0.348
V15	162.520	150	0.229	1.076	0.970	0.315	0.061
V16	136.243	150	0.783	0.902	0.919	-1.919	-2.009

	Chisq	df	p-value	Outfit MSQ	Infit MSQ	Outfit-t	Infit-t
V17	106.301	150	0.997	0.704	0.872	-1.284	-0.614
V18	90.261	150	1.000	0.598	0.883	-1.243	-0.327
V19	115.056	150	0.985	0.762	0.867	-1.975	-1.360
V20	126.510	150	0.919	0.838	0.897	-1.797	-1.463
V21	93.168	150	1.000	0.617	0.894	-0.975	-0.214
V22	113.907	150	0.987	0.754	0.904	-1.328	-0.586
V23	104.449	150	0.998	0.692	0.885	-1.284	-0.510
V24	137.155	150	0.766	0.908	0.949	-1.204	-0.874
V25	140.653	150	0.696	0.931	0.964	-0.968	-0.663
V26	171.026	150	0.115	1.133	1.025	0.854	0.242
V27	134.043	150	0.821	0.888	0.898	-0.986	-1.070
V28	161.063	150	0.254	1.067	1.065	0.764	0.962
V29	175.112	150	0.079	1.160	1.057	0.912	0.429
V30	141.396	150	0.680	0.936	0.984	-0.187	-0.018
V31	163.383	150	0.215	1.082	1.058	0.680	0.618
V32	157.164	150	0.328	1.041	1.029	0.862	0.738
V33	203.563	150	0.002	1.348	1.135	2.338	1.167
V34	187.169	150	0.021	1.240	1.200	3.260	3.636
V35	181.376	150	0.041	1.201	1.098	1.830	1.204
V36	150.242	150	0.479	0.995	0.990	-0.082	-0.201
V37	145.405	150	0.591	0.963	0.973	-0.555	-0.463
V38	128.057	150	0.902	0.848	0.870	-2.869	-2.747
V39	166.401	150	0.170	1.102	1.087	2.103	2.137
V40	157.483	150	0.322	1.043	1.041	0.907	1.011

From the data in table 2, it is known that all items can be accepted as good questions because the three criteria show that they are appropriate. The appropriate items must have a p-value> 0.05 and an MSQ Outfit of 0.5 <MNSQ <1.5 and an Outfit-t of -2.0 <ZSTD <2.0. Thus all items can be accepted 100%. Even though the items number 7, 34, and 35 have a p-value <0.05, the other two criteria can still be entered so that they are considered fit. Item Characteristic Curva (ICC) in the Rasch model only connects two variables, namely the ability of the testee (latent dimension parameter) and the probability of answering correctly. Difficulty level (b) is the ability where the testee has half the chance to answer correctly (0.5). All vocational theory test items can be well described as a logistical function, the ICC for most numbers can be seen in Figure 1.

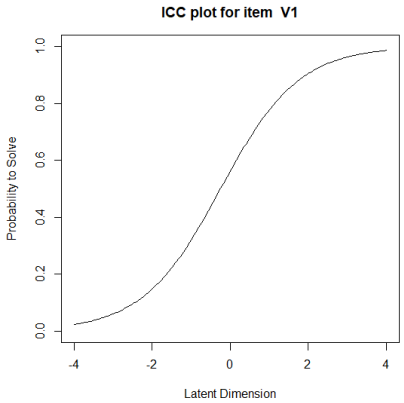

Fig. 1. ICC Plot for Item V1 1

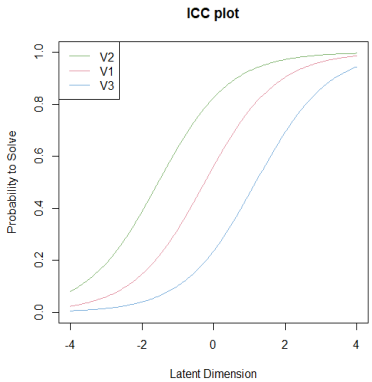

Fig. 2. ICC Plot fot Item V2

From Figure 1 and Figure 2, it can be seen that the higher the respondent's ability, the higher the chance of answering the correct answer.

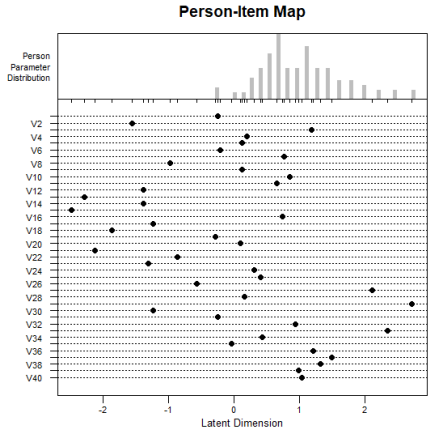

Fig. 3. Person-Item Map on the Simulation Tes

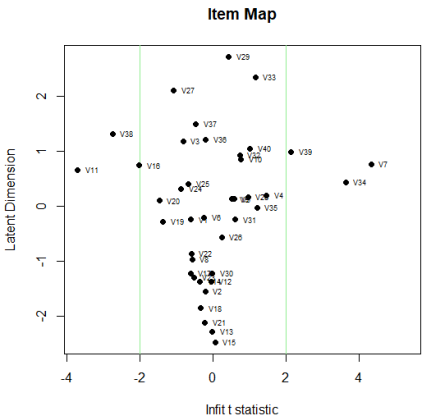

Fig. 4. Item Map on the Simulation Tes

Figure 3 which describes the item map and Figure 4 which describes the person-item map where all the difficulty levels of the items are at predetermined intervals. Figure 5 relates the test taker's ability and the item difficulty level.

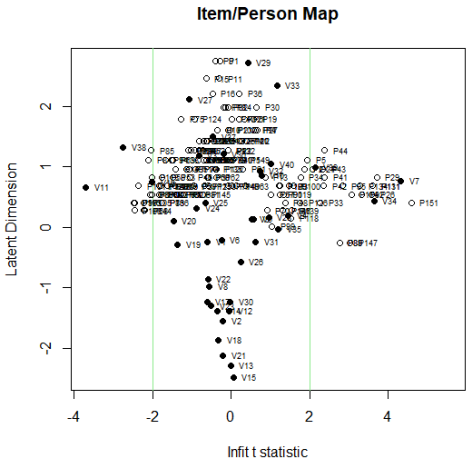

Fig. 5. Item/Person Map on the Simulation Tes

From Figure 5, it can be seen that all item difficulty levels are in the domain of the testee's ability except item no. 7, 11, 34, 38, and 39. Figure 6 also shows that the test will provide high information values in the ability range -2 to + 2.

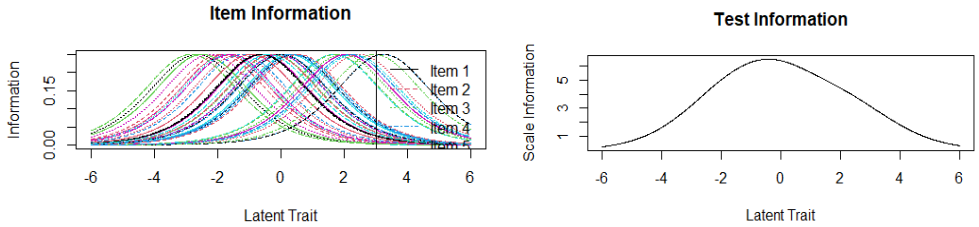

Fig. 6. Item Information and Test Information on the Simulation Tes

3.2 Measurement of Substantive Validity

To see the quality of the construct validity from the substantive aspects, the test taker's ability to fit the model was used. This test is basically testing the consistency of responses or different response patterns of participants to test items based on their level of difficulty.

Table 3. Person Fit Test on the Simulation Tes Using Rasch Model

	Chisq	df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t
P24	77.073	39	0.000	1.927	1.376	2.36	2.07
P26	73.541	39	0.001	1.839	1.635	2.64	3.58
P29	82.384	39	0.000	2.060	1.709	2.81	3.71
P33	63.194	39	0.008	1.500	1.370	2.05	2.27
P34	68.196	39	0.003	1.705	1.300	2.03	1.76
P42	72.520	39	0.001	1.813	1.401	2.43	2.36
P43	87.173	39	0.000	2.179	1.432	2.83	2.33
P44	105.479	39	0.000	2.637	1.503	3.09	2.38
P88	68.091	39	0.003	1.702	1.493	2.51	2.76
P89	68.091	39	0.003	1.702	1.493	2.51	2.76
P114	71.305	39	0.001	1.783	1.529	2.50	3.06
P131	74.906	39	0.000	1.873	1.660	2.57	3.61
P134	75.370	39	0.000	1.884	1.588	2.60	3.28
P147	71.320	39	0.001	1.783	1.549	2.74	3.03
P151	109.109	39	0.000	2.728	1.844	4.83	4.59

This deviant response can be caused by inaccuracy, cheating or even misconceptions. A person's response test has deviations or is not called person fit. The criteria for acceptance of a test taker's response are considered to have deviated or not the same as the fit item criteria. Quantitatively, the response of test takers who are declared fit or have no deviation is if the MSQ Outfit value is between 0.5 to 1.5 while the Outfit t value is between -2 to 2.0 and the chance of Ho acceptance (model fit) is greater than 0.05 ($p > 0.05$). Of the 151 test participants there were 15 test takers or 9.93% who experienced responses that deviated from the model and 136 or 90.06% accordingly. This can be seen from the 15 test participants who do not meet as many as two (p value and outfit MSQ) of the three person fit criteria.

3.3 Structural Validity Measurement

The test indicator has two structural aspects of construct validity, namely the test is unidimensional and has stability in estimating item parameters and test participants. Tests built in a one-dimensional paradigm must really have one dimension so that the measurement results they get can have meaning. The results of the unidimensional test analysis with the R

program using the ltm package can be seen in Table 4, while the results of the curve analysis can be seen in Figure 7. The results of the analysis of the invariance test with the R program using the ltm package can be seen in Table 5.

Table 4. Unidimensional Test Results for National Exam Simulation Test Items

Second eigenvalue in the observed data: 7.6373

Average of second eigenvalues in Monte Carlo samples: 4.6292

Monte Carlo samples: 100

p-value: 0.0099

From Table 4 it is known that the resulting unidimensional test probability is 0.0099, a value smaller than 0.05 so that it can be stated that the assumption is rejected. This condition can be stated that the test contains not only one dimension. And it can be concluded that the simulation test of the TKJ vocational theory national exam can be stated as multidimensional.

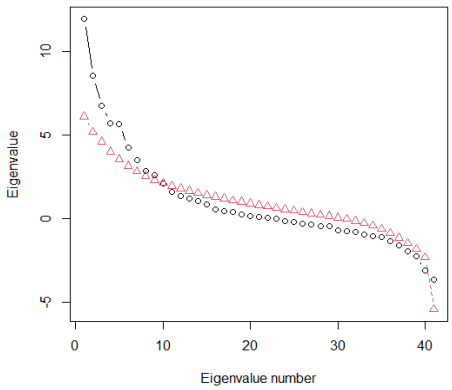

Fig. 7. Unidimensional Test Results for National Exam Simulation Question**s**

From Figure 7 it can be seen that the second eigenvalue (7.6373) of the observed data is substantially greater than the second eigenvalue (4.629) of the data under the assumed IRT model so that the unidimensional test results are rejected.

Table 5. The Results of the Test Items in the Simulation of the National Exam

The following items were excluded due to inappropriate response patterns within subgroups:

V14 V18 V21

Full and subgroup models are estimated without these items!

Andersen LR-test:

LR-value: 171.069

Chi-square df: 36

p-value: 0

Furthermore, to perform the measurement invariance test using the Anderson LR test. This test is used to determine the consistency of the Rasch modeling parameter estimates. The ideal condition in Rasch modeling occurs when the parameter estimates of the item difficulty level are consistent (invariant) even though they are obtained from a sample consisting of any population subgroup during the application of Rasch modeling. From the results of the analysis, the p value is 0, meaning that it does not accept the assumption, so it can be concluded that the parameter estimation is not invariant.

3.4 External Validity Measurement

The validity of the external aspect construct is used to determine the extent to which the test results are supported by other measurements (which measure the same or similar domains) so that it can be seen whether they have a strong relationship or not. One approach to determine the validity of the external aspect construct in this first year research is to use information on Person Separation reliability or Person Separation. Person separation is used to classify people based on information obtained from the test. The low separation of people (less than 2) from the sample of relevant people implies that the instrument may not be sensitive enough to distinguish between high and low performers. This means that more items are needed to measure it. The results of the Person separation analysis using the eRm package can be seen in Table 6.

Table 6. Test of Person Separation Reliability on Question

```
Separation Reliability: 0.5005

> summary(Z)
   Separation Reliability: 0.5005
   Observed Variance: 0.3084 (Squared Standard Deviation)
   Mean Square Measurement Error: 0.1541 (Model Error Variance)
```

From Table 6, it can be seen that the Person Separation reliability value is 0.50. Thus the value of the person separation for the test is 0.66. From the value of the person separation, it can be seen that the classification of the test takers obtained is close to one. This means that the question instrument can differentiate test participants into two categories, namely high and low. The consequence is that the test results only differentiate test participants into two groups, namely test takers who already have high exam results and those who have low test results.

3.5 Consequential Validity Measurement

The consequential aspect in the validity of the constructs implies the value interpretation of the score as a source of action. Evidence regarding the consequential validity aspect also addresses the actual and potential consequences of testing and using scores, particularly in terms of sources of invalidity such as bias, fairness and distributive justice. In Rasch modeling with the eRm package, the detection of grain bias can be approached by determining items that have a differential item functioning (DIF) using the Waldt Test. DIF deals with the

estimation of different item parameters in different subpopulations, in which test takers are differentiated by gender. If a male test taker deems an item more difficult or easier than a female or vice versa, then the item contains DIF. DIF or also known as item external bias is not a justification for grain bias because to find out whether there is a bias, an in-depth qualitative study must be carried out again regarding the causes of the emergence of DIF. However, the emergence of DIF can be an indication of the possibility of bias. The list of test items detected by DIF can be seen in Table 7, while the description of DIF can be seen in Figure 5. Statistical criteria with the Wald test, items that experience DIF are those that have a p-value less than 0.05 (significance level 0.05). From Table 7, it is known that 3 items are indicated to have DIF, namely items 16, 18, and 33.

Table 7. List of DIF-Indicated Test Items by Sex

Significance Level 0.05			
	Butir	**z-statistic**	**p-value**
beta	16	-2,596	0,009
beta	18	2,416	0,016
beta	33	-3,115	0,002

When using the 0.01 significance level, only no. 3 and 33 only experienced DIF. In accordance with the test taker's data, where the proportion of men is only 33%, far from the ideal proportion, of course the researchers are more careful in determining the level of significance when testing the presence of DIF on items caused by gender. If at the significance level of 0.05, it means that the probability of rejecting the correct assumption is 0.05, then at the significance level of 0.01 it means that the chance of rejecting the correct assumption is 0.01. The assumption here states that student responses to the test do not experience DIF. In connection with this in determining DIF, the researcher chose a significance level of 0.01 so that two items were considered detected by DIF. Meanwhile, other points with validity analysis that include content, psychometrics and constructs (content, substantive, structural, external, consequences) meet the requirements as good items.

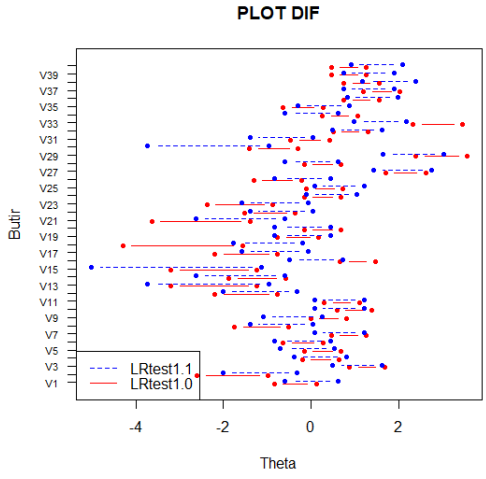

Fig. 8. Description of dif in Test Items for the Simulation of the National Examination

4 Conclusion

The results of measuring the validity of the TKJ vocational theory national examination simulation test have met the validity of the content aspect. All test items have met the validity of the psychometric aspects. Constructive validation with Rasch modeling gave the following results: (1) Content validity resulted in 100% acceptable items, (2) Substantive validity resulted in 9.93% or 15 students deviating and 90.06% or 136 students could be accepted, (3) Structural validity in the unidimensional test is rejected because this simulation test contains multidimensional, (4) External validity results in a reliability separation value of 0.66 which means that this simulation test can distinguish 2 groups of test participants, and (5) Consequential Validity is obtained from the DIF calculation produces 3 items containing DIF, namely items number 16, 18 and 33. So that this simulation test can be said to be accepted from gender bias because 37 items are appropriate.

Acknowledgements

From the results of this study it can be suggested that the implementation of simulation tests should be well prepared by students even though the results do not affect the results of the national exam but students should be more serious in working on simulation questions. The author is grateful to Pancasakti Tegal University for providing the opportunity to develop research. Likewise the authors would like to thank all parties concerned, especially the principal of SMK 2 Tegal and SMK 3 Tegal who have supported and granted research permission.

References

[1] Azwar, S. (2005). *Dasar-Dasar Psikometri*. Yogyakarta: Pustaka Pelajar.

[2] Coaley, K. (2010). *An Introduction to Psychological Assessment and Psychmetrics*. London: Sage.

[3] Devellis, R. F. (2003). *Scale Development*. London: Sage Publications.

[4] Ebel, R. L. (1991). *Essentials of educational measurement* (5th ed.) , . Englewood Cliffs New Jersey: : Prentice Hall

[5] Eleje, L., I & Onah, F., E. (2018). Comparative Study of Classical Test Theory and Item Response Theory Using Diagnostic Quantitative Economics Skill Test Item Analysis Results. *European Journal of Educational & Social Sciences*. 3(1), 71-89.

[6] Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *J*ournal of consumer research, 20(2), 303-315.

[7] Haynes, S. N., Richard, D. C., & Kubany, E.S. (1995). *Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods*. Psychological Assessment, 7, 238 - 247

[8] Indonesia, P. R. (2003). U*ndang-undang Republik Indonesia nomor 20 tahun 2003 tentang sistem pendidikan nasional*. Jakarta: Pemerintah Republik Indonesia.

[9] Murphy, K. R., & Davidshofer, C. O. (1991). *Psychological Testing: Principles and Applications*. New Jersey: Prentice Hall

[10] Samuel Messick (1995) "*Validity of Psychological Assessment*," American Psychologist 50, no. 9 : 741–49, doi:10.1037//0003-066X.50.9.741

[11] Siswanto, S. (2008). *Validitas Sebagai Alat Penentu Kehandalan Tes Hasil Belajar.* Jurnal Pendidikan Akuntansi Indonesia, *6*(1).
[12] Sudaryono, S. (2011). *Implementasi Teori Responsi Butir (Item Response Theory) Pada Penilaian Hasil Belajar Akhir di Sekolah.* Jurnal Pendidikan dan Kebudayaan, 17(6), 719-732.
[13] Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan*. Trim komunikata.
[14] Sugiyono. 2010. *Metode Penelitian Pendidikan Pendekatan Kuantitatif Kualitatif dan R&D*. Bandung: Alfabeta
[15] Susongko, P. (2016). *Validation of science achievement test with the rasch model*. Jurnal Pendidikan IPA Indonesia, *5*(2), 268-277.
[16] Susongko, P. (2019). *Aplikasi Model Rasch dalam Pengukuran Pendidikan Berbasis Program R.* Tegal : Badan Penerbitan Universitas Pancasakti Tegal
[17] Suryabrata, S. (2005). *Pengembangan Alat Ukur Psikologis*. Yogyakarta: Penerbit Andi.