Validation School Exam Evaluation Study Subject in Chemistry SMA Negeri 3 Tegal Using Rasch Model

E. Y. Wijayanti[1], Basukiyatno[2], P. Susongko[3]
Master of Pedagogy, Universitas Pancasakti, Tegal, Indonesia[1,2,3]

{liyatliyut@gmail.com[1], basukiyatnofkip@gmail.com[2], purwosusongko@upstegal.ac.id[3]}

Abstract. This study aims to analyze the construct validity of chemistry test istrument on the school examination at SMAN 3 Tegal by using Rasch modeling in term of contents, subtantive, and consequential aspects. The participants of this study consisted of 193 students of XII grade. The test consists of 34 items presented in multiple choice form which covered the scope of basic chemistry, analytical chemistry, physical chemistry, organic chemistry, and inorganic. The construct validation using Rasch modeling on the school examination of chemistry subject through the R Program Studio application 4.02 version gave the following results: there were 31 items that matched the modeling; (2) 91.71% of the students' responses matched the modeling; (3) There was 1 item that contained DIF. It means that 31 of 34 items on the school examination of chemistry subject at SMAN 3 Tegal meet the costruct validity of contents, subtantive, and consequential aspects.

Keywords: School Exams, Chemistry, Rasch Model

1 Introduction

Assessment is the process of gathering and processing information to measure the achievement of student learning outcomes (Kemendikbud, 2016). Assessment or evaluation of learning outcomes can also be defined as the process of giving a value achieved by students with certain criteria. One of the evaluations on the educational unit scale that is used as one of several components of the graduation criteria is the School Examination (US). Permendikbud No.23 of 2016 concerning with Assessment Standards also defines that school / madrasah exams are activities carried out to measure the achievement of student competencies as recognition of learning achievement and / or completion of an educational unit. One of the characteristic subjects tested in school exams in high school is chemistry

Subagia described that chemistry is one of the science lessons that most high school students are less interested in. This is inseparable from the way the book presents the material, the way the teacher teaches chemistry, the public information received by students, and the students' goals for learning chemistry (Subagia, 2014). In the same statement, Subagia added that to construct chemistry subjects, at least three main ideas can be recommended which are used to strengthen students' interests in chemistry subjects in high school. First, high school chemistry learning needs to begin by building new ways of thinking of students about chemistry subjects. This can be done by explaining that chemistry is important, prosperous,

ICONS 2020, November 30, Tegal, Indonesia

DOI 10.4108/eai.30-11-2020.2303725

fun, healthy, and beneficial for everyone. Second, every high school chemistry learning must be linked back to the existence of chemistry in everyday life. Third, students are trained to think critically and creatively on every aspect of the chemical material being studied. Thus, students are able to see the role of chemistry in explaining or solving daily problems and not only seen as mere knowledge. Through these three methods, it is hoped that the awareness, interest, and motivation of students to learn chemistry can be increased so that it is possible for chemistry subjects to become increasingly desirable while obtaining optimal learning outcomes.

Improving the quality of education can be done in various ways. One way that can help is to conduct a comprehensive education assessment and evaluation. For this reason, rasch modeling is very effective to use. This implies that racsh modeling converts the raw score data into data at the same interval so as to produce a measurement scale that is linear, precise, and has units. Rasch modeling can be used for the analysis of the quality of the questions, knowing the level of student ability and the difficulty level of the questions, to the detection of misconceptions, the existence of bias in the questions, as well as the possibility of knowing that students are cheating (Sumintono & Widhiarso, 2015).

Bond Trevor argues that the Rasch model minimizes invariant failure, or what is often referred to as DIF, and reminds researchers to modify the assessment procedures or subsistence theory under investigation (Bond, 2003). In modern measurement test theory, the Rasch model is seen as the most objective measurement model. The type of data generated from the learning achievement test and from the attitude scale is ordinal not interval so that the analytical tools that can be used are limited (Mari et al., 2012). The Rasch model has other advantages such as connecting the probability of answering each item correctly (P (θ)) as a function of ability (θ) with the difficulty level of item (b). This relationship can be shown in Equation 1.

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

The Rasch model can also be used for dichotomous or two-category responses as well as multiple choice questions. Whereas for responses that are politomos or more than two categories, the Rasch Model is developed more broadly as the Partial Credit Model (PCM) or the partial credit model. The general odds in PCM are given by Equation 2.

$$P(X_{ni} = x) = \frac{exp \sum_{k=0}^{x} (\theta_n - \delta_n)}{\sum_{h=0}^{mi} exp \sum_{k=0}^{h} (\theta_n - \delta_n)} \tag{2}$$

The Rasch model was further developed separately from the IRT. Even the Rasch model was also developed more broadly in the politomos scoring. Since its introduction by its founder Georg Rasch in 1960, the application of the Rasch model of learning achievement is widespread not only in education but also in medicine and public health (Smith et al., 2010; Lu et al., 2013; Ayele et al, 2014). To analyze these instruments, several questions need to be answered as follows: (1) How is the construction of the school examination instrument for chemistry subjects ?; (2) How is the validity of the content aspects of the school examination test instrument in chemistry subject ?; (3) How is the validity of the substantive aspects of the school examination test instrument for chemistry subjects? And: (4) How is the validity of the consequential aspects of the test instrument for the school examination in chemistry subjects?

Thus, the aim of the study to analyze the construct validity of the school examination test instruments in chemistry subjects using Rasch modeling can be achieved.

2 Method

To build objectivity, this instrument was validated with the Rasch modeling. This is an effort to evaluate the school test instrument for chemistry subjects which is carried out through the test of construct validity for the content, substantive, and consequential aspects.

2.1 Participants

Participants in this study were 193 students of class XII MIPA SMAN 3 Tegal Tegal as many as 70 male students and 123 female students. Their age range is 16-18 years. All students come from the city of Tegal and its surroundings. The initial abilities and family backgrounds of students vary widely as a result of the implementation of zoning policies in the admission of new high school students.

2.2 Instrument

The school test instrument for chemistry is presented in the form of a multiple choice test consisting of 34 question items. This instrument consists of the scope of basic chemistry, analytical chemistry, physical chemistry, organic chemistry, and inorganic chemistry which refers to the BSNP Regulation Number 0053 / P / BSNP / I / 2020 concerning Standard Operational Procedures for the Implementation of the National Examination for the 2019/2020 Academic Year with cognitive level composition of 10% -15% for reasoning, 50% -60% for application, and 25% -30% for knowledge and understanding.

2.3 Data analysis

The data analysis was carried out by dividing the validity of the construct into three aspects, namely the validation of the constructs of content aspects, substantive aspects, and consequential aspects with Rasch modeling. Validity in this study refers to the concept of construct validity (Messick, 1996) which is divided into six aspects, namely content, substantive, structural, external, consequential and generalization (Baghaei & Amrahi, 2011). Susongko provides quantitative criteria related to the validity indicator of the construct according to the Rasch modeling as described in Table 1.

Table 1. Valid test criteria seen from various aspects of validity and criteria with the application of the Rasch Model (Susongko, 2016)

Construct Validity Aspek	Indicator	Criteria
Content	Item fit test (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 Aspek	Indicator	Criteria
	Function of Information Test	The test information function has a maximum value in the testee ability domain
Substantive	Person fit statistic	P > 0.05 0,5 <MNSQ< 1,5 -2,0 < ZSTD< 2,0
	Collapsed Deviance / Casewise Deviance /Hosmer-Lemeshow	P< 0,05
	accuracy, sensitivity, and specificity	approching 1,0
Structural	Unidimensional test	There is one main factor that is described through the Scree Plot of the factor analysis results
	Invariance Test (LRtest)	P> 0,05
External	value of separation person strata	approching 1,0
Consequential	DIF	there is no significant DIF

In this study, the software used in analyzing Rasch modeling used the R program version 4.0.2 with the eRm package version 0.16-2. This software is used because it is open source so it is easy to access and is developed for observers of educational assessment research.

3 Results and Discussion

The analysis of the validity of the test instruments for the school examination in chemistry subject at SMAN 3 Tegal aims to meet the suitability of measuring instruments that can ensure that student competencies are in accordance with predetermined competency standards. The National Examination is insufficient to measure competency standards that have been determined due to several things, including: (1) National Examination is no longer used as a determinant of graduation so that it is no longer a guarantee of compliance with graduate competency standards, (2) students only choose one of the special characteristic subjects MIPA. (3) there is a discourse on the elimination of the UN in the following year. This test instrument is expected to facilitate graduate competency expectations by paying attention to three aspects including content, chemistry learning outcomes, and measurement models.

The grid for the school examination test instrument for chemistry subjects consists of the scope of material (1) basic chemistry, (2) analytical chemistry, (3) physical chemistry, (4) organic chemistry, and (5) inorganic chemistry.

Table 2. Indicators of Chemistry Subject School Examination Instruments

Achievement of School Examinations	Indicators used
Chemistry Subject	Basic chemistry Analytical Chemistry Physical Chemistry Organic Chemistry Inorganic Chemistry

Item validation uses IRT modeling (Rasch for dichotomous). Apart from paying attention to the success of school examination results data, this instrument also pays attention to basic chemical, analytical, physical, organic and inorganic aspects. These five aspects integratedly form comprehensive information on the competence of chemistry subjects. The scoring of each item in one item is dichotomous (1 or 0)

Table 3. Test Scoring Model

Score	Criteria
0	False
1	True

In accordance with the explanation of Table 1 regarding to the criteria for construct validity on the content aspect, then, it will explain some of the data from the analysis results with Rasch modeling for dichotomous data (IRT). Table 4 contains the results of the item fit analysis on the model (Item Fit). Fit items basically explain whether an item is functioning to take measurements normally or not. Quantitatively, the test items that are declared fit or can function properly are 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). Outfit is an outlier-sensitive fit, which is a measure of the sensitivity of the response pattern to items with a certain difficulty level from the respondents (students) or vice versa. Outfit t is the t test for the hypothesis of the suitability of the data to the model. While the MSQ Outfit value is calculated from the chi square value divided by the degrees of freedom (df). From Table 4, it can be seen that all items are generally accepted as good items except items 4, 6, and 28. These items have an outfit t of less than -2 and more than 2.0 and p value <0.05. This means that the item is seen from the out fit t of more than 2.0 which contains the meaning of the data appearing to be unpredictable while the probability of model fit is also less than 0.05. Two criteria reject these items so that it can be concluded that at the 0.05 significance level the three numbers cannot be accepted by the model.

Table 4. Results of Item Fit Analysis of Chemistry Subject School Test Measurement Instruments at SMAN 3 Tegal

No	Chisq	Df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t	Discrim
1	202,862	192	0,282	1,051	1,056	0,803	1,098	0,157
2	173,477	192	0,827	0,899	0,889	-1,906	-2,696	0,412
3	169,827	192	0,874	0,880	0,883	-1,726	-2,302	0,443
4	285,965	192	0,000	1,482	1,039	1,668	0,261	-0,102
5	186,333	192	0,602	0,965	0,978	-0,635	-0,497	0,314
6	252,714	192	0,002	1,309	1,108	2,848	1,465	-0,012
7	189,423	192	0,539	0,981	0,980	-0,324	-0,460	0,304
8	239,401	192	0,011	1,240	1,099	1,822	1,055	-0,028
9	164,723	192	0,924	0,853	0,971	-0,601	-0,119	0,172
10	208,184	192	0,201	1,079	0,948	0,333	-0,105	0,082
11	179,19	192	0,737	0,928	0,964	-0,764	-0,517	0,302
12	165,983	192	0,913	0,860	0,898	-2,477	-2,335	0,428
13	142,138	192	0,997	0,736	0,902	-1,592	-0,743	0,343
14	121,278	192	1,000	0,628	0,844	-1,492	-0,745	0,426
15	193,079	192	0,465	1,000	1,018	0,026	0,425	0,239
16	237,427	192	0,014	1,230	0,988	0,697	0,052	-0,040
17	208,913	192	0,191	1,082	1,045	1,167	0,831	0,138
18	170,705	192	0,863	0,884	0,960	-0,822	-0,374	0,273
19	173,392	192	0,828	0,898	0,911	-0,258	-0,340	0,200
20	198,617	192	0,357	1,029	1,028	0,321	0,412	0,159
21	165,781	192	0,915	0,859	0,869	-2,428	-2,983	0,483

No	Chisq	Df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t	Discrim
22	191,494	192	0,497	0,992	0,981	-0,098	-0,350	0,244
23	195,544	192	0,415	1,013	1,033	0,162	0,472	0,197
24	197,262	192	0,382	1,022	1,038	0,286	0,632	0,196
25	187,773	192	0,573	0,973	1,010	-0,328	0,188	0,248
26	135,212	192	0,999	0,701	0,874	-1,546	-0,809	0,375
27	154,735	192	0,978	0,802	0,913	-1,492	-0,862	0,391
28	247,226	192	0,004	1,281	1,049	1,341	0,368	-0,037
29	178,863	192	0,743	0,927	0,899	-1,026	-1,855	0,373
30	221,746	192	0,069	1,149	1,102	1,938	1,829	0,061
31	190,239	192	0,522	0,986	0,962	-0,016	-0,242	0,195
32	197,409	192	0,379	1,023	1,031	0,335	0,559	0,198
33	195,104	192	0,424	1,011	0,987	0,166	-0,208	0,266
34	188,451	192	0,559	0,976	0,991	-0,185	-0,090	0,215

This outfit value illustrates the deviation of the test taker's response from the ideal model. By an outfit value that exceeds the reasonable limit, it can be stated that the item has a significant deviation from the Rasch model. The deviation in this case is that some test takers who have an ability lower than the item difficulty level successfully answer the item correctly or some test takers who have the ability above the difficulty level but do not succeed in answering the item correctly. The incompatibility of responses to the model can be caused by many factors, such as carelessness, misconceptions or success in guessing (Sumintono & Widhiarso, 2015). Thus the Rasch model can be used to identify the occurrence of misconceptions.

A lot of studies have shown that the Rasch Model can be used to identify the occurrence of misconceptions on large scale tests. It happens on the tests of mastery of physics, chemistry and science. (Herrmann-Abell & DeBoer, 2011); (Wind & Gale, 2015); (Romine et al., 2015); (Morris et al., 2012); (Edwards & Alcock, 2010); (Planinic et al., 2010). Item number 4 contains the ratio of N2 gas to H2 in NH3 compounds; item 6 contains an analysis of why HF is the weakest acid solution but HF has the highest boiling point; and item 28 which contains testing of the pH of 3 types of indicators. These three points are very susceptible to student misconceptions.

Table 5. Value of Difficulty Levels for Items of School Test Instruments for Chemistry Subjects at SMAN 3 Tegal

Item	p-value	Item	p-value
1	0,868	18	-2,689
2	-0,189	19	-5,044
3	-1,046	20	-1,849
4	4,765	21	-0,610
5	0,020	22	0,913
6	-1,849	23	-1,899
7	-0,189	24	-1,456
8	2,505	25	-1,226
9	-4,181	26	-3,884
10	-5,732	27	-2,686
11	-1,649	28	3,886
12	-0,526	29	1,135
13	-3,378	30	-1,136
14	-4,763	31	-3,537
15	0,312	32	1,179
16	5,948	33	-1,226
17	1,089	34	2,110

From Table 5, it can be seen that the lowest difficulty level value is in item number 10 of -5,732 while the highest difficulty level is in item number 16 of 5,948. The difficulty level of 5,948 means that participants are expected to do the items correctly if they have at least 5,948 skills. However, the difficulty level that is more than 3 and less than -3 is only for a few question items. Most are within a reasonable difficulty range. It is possible that there are students who feel that the results of the school exam scores will always be made safe or converted so that all students must meet the minimum completeness. So, in the process it is careless. The item difficulty level is a location parameter that shows the position of the item characteristic curve in relation to the ability scale. The item difficulty level parameter is described by a point on the ability scale where the probability of answering correctly is 0.5. The greater the difficulty level parameter value, the greater the ability needed by the respondent to get the opportunity to answer the questions correctly as much as 0.5. For more details, Figure 1 and Figure 2 explain the characteristic curves of items 1 and 4.

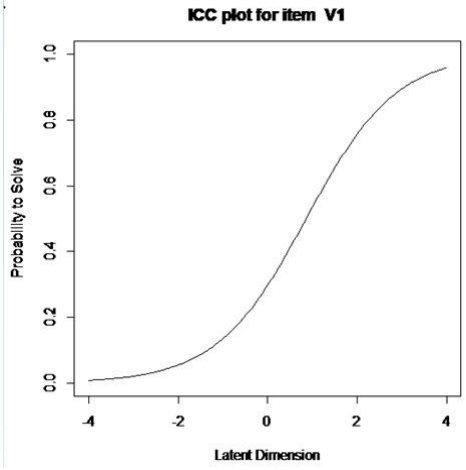

Fig 1. Characteristics Curve Number 1

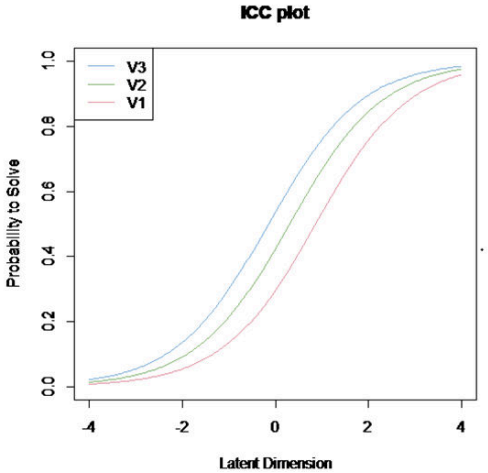

Fig 2. Characteristics Curve Number 2

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. From Table 5, it can be seen that the item difficulty level moves from -5,732 to 5,948. The test, which is effective, has item difficulty level between -2.00 and 2.00 (Stenbeck et al., 1992). However, tests that are built to measure competence as well as instruments for measuring chemistry learning outcomes should be able to measure the basic competencies of all test takers so that the distribution of material is measured accurately. If it is assumed as developed by the item response theory / normal distribution, the item difficulty level for measuring competency can start from -3.00 to 3.00, because at that interval it can measure around 99.98% of participants. Thus, from the results of the analysis of all the test items for the Chemistry Subject School Examination Test at SMAN 3 Tegal that have been compiled, most of them are at intervals of -3.00 to 3.00 so they are quite effective as an instrument for measuring learning outcomes. This is clarified by 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.

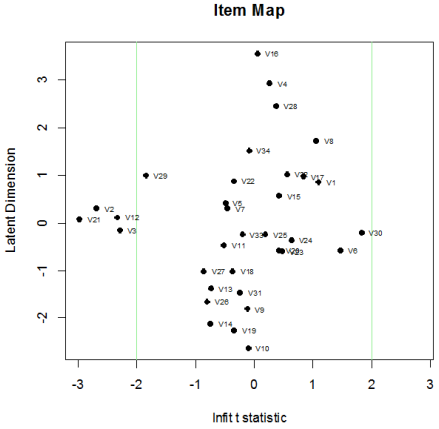

Fig 3. *Item Map* of Items for the School Exam Test Instruments for Chemistry at SMAN 3 Tegal

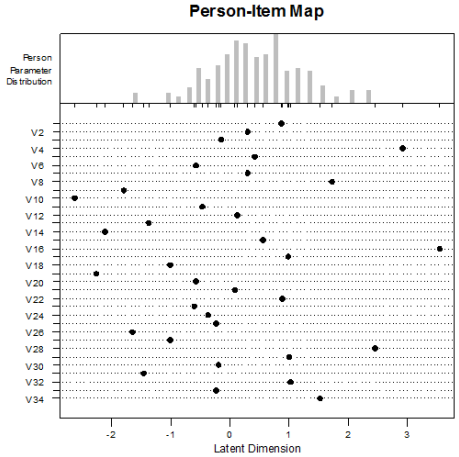

Fig 4. *Person- Item Map* Items of Chemistry Subject School Examination Test Instruments at SMAN 3 Tegal

The evidence that most of the items on the School Chemistry Subject School Examination Test at SMAN 3 Tegal are effective for the test taker's ability between -3.00 to 3.00 is explained by the item and test information function (Figure 6). The figure explains that the information function will be maximal at the student ability interval between 0 to 1.0 and effective between -3.0 to 3.00.

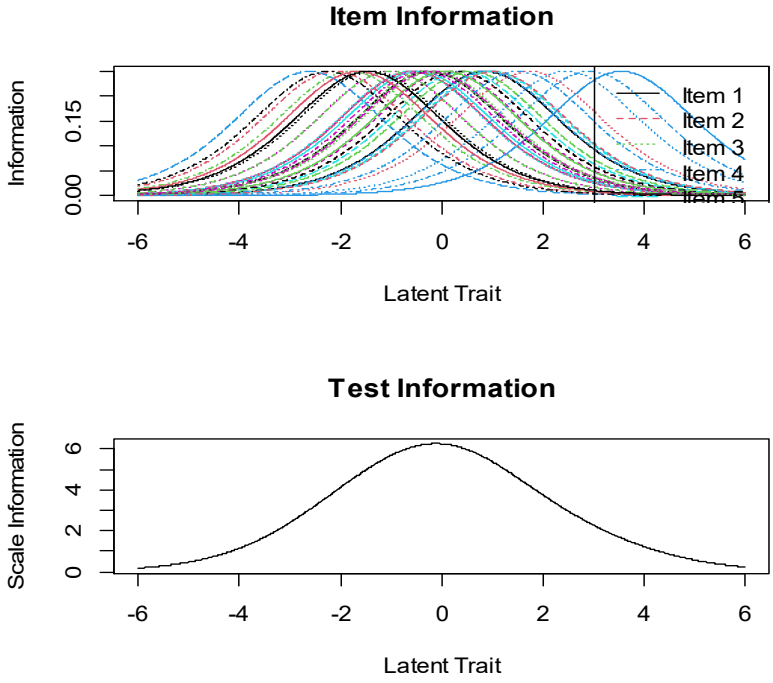

Fig 5. Information Function Items for Chemistry Subject School Examination Instruments at SMAN 3 Tegal

3.1 Validity of the Constructive Substantive Aspects

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 from participants to test items based on their level of difficulty. A different response pattern is the mismatch of responses given based on their abilities compared to the ideal model. A test taker who has the ability (Ø) of 1.5 should be able to answer all the questions that have a difficulty level below 1.5, but in the field of course there are some students who are inconsistent or cause an abberant response. How many students experienced this abstract response is a measure of the validity of the substantive type constructs.

This deviant response can be caused by inaccuracy, cheating or even misconceptions. Person fit is whether someone deviates from the response test. The item fit criterion is a criterion for acceptance of responses from test takers who experience deviations or not. 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, 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 273 participants, 16 test

participants experienced responses that deviated from the model. This can be seen from the 16 participants who do not meet as many as two (p value and outfit MSQ) of the three person fit criteria. Even one participant (P33) did not meet all the person fit criteria. The list of test participants is presented in Table 6.

Table 6. Test takers who have aberrant responses (*aberrant response*)

Test Takers	Chisq	Df	p-value	Outfit MSQ	Infit MSQ	Outfit t	Infit t
30	52,852	33	0,016	1,554	1,357	1,56	2,01
46	75,619	33	0,000	2,224	1,032	2,78	0,25
79	67,024	33	0,000	1,971	1,145	2,09	0,74
92	77,795	33	0,000	2,288	1,168	2,72	0,99
95	62,592	33	0,001	1,841	1,186	2,08	1,11
103	60,322	33	0,003	1,774	1,250	2,13	1,46
117	70,614	33	0,000	2,077	1,387	2,38	2,08
118	95,875	33	0,000	2,820	1,719	4,06	3,56
122	60,827	33	0,002	1,789	1,277	1,64	1,44
128	55,089	33	0,009	1,620	1,231	1,67	1,23
133	62,491	33	0,001	1,838	1,273	2,21	1,49
138	61,302	33	0,002	1,803	1,554	2,17	2,95
139	63,386	33	0,001	1,864	1,213	2,33	1,26
158	69,469	33	0,000	2,043	1,086	2,04	0,45
168	70,713	33	0,000	2,080	1,389	2,77	2,16
169	73,038	33	0,000	2,148	1,113	2,34	0,68

From this explanation, it can be concluded that there were 91.71% of test takers' responses that were reasonable according to the model or did not experience deviations, while 8.29% of the responses experienced deviations. The large percentage of test takers who have a reasonable response in accordance with this model can be the basis that the test is sufficient to meet substantive validity. Students' responses that deviate from the Rasch model indicate an indication of students doing careless or lucky guess or even cheating (Sumintono & Widhiarso, 2015). Several studies have shown that person fit can be used as initial data on the existence of cheating, careless or lucky guess when taking tests by students (Shu et al., 2013); (Meyer & Zhu, 2013); (Hohensinn & Kubinger, 2011); (Magis et al., 2012); (Lamprianou, 2010); (Liu & Yu, 2011)

3.2 Validity of Structural Aspects Constructions

There are two indicators that the test has the validity of the structural aspects of the construct, namely the test is unidimensional and has stability in estimating the parameters of the items and the test participants. Tests built in a one-dimensional paradigm need to have one dimension so that the measurement results obtained have meaning. The principle of unidimensional testing is first stated by the null hypothesis which states that the second eigenvalue is not greater than the first eigenvalue with the alternative hypothesis that the second eigenvalue is greater than the first eigenvalue. The results of the unidimensional test analysis with the R program using the ltm package can be seen in Table 8, while the results of the curve analysis can be seen in Figure 6.

Table 7. Unidimensional Test Results Instrument Items for School Examination in Chemistry Subjects at SMAN 3 Tegal

Alternative hypothesis: the second eigenvalue of the observed data is substantially larger than the second eigenvalue of data under the assumed IRT model
Second eigenvalue in the observed data: 3.9983
Average of second eigenvalues in Monte Carlo samples: 3.5344
Monte Carlo samples: 100
p-value: 0.1089

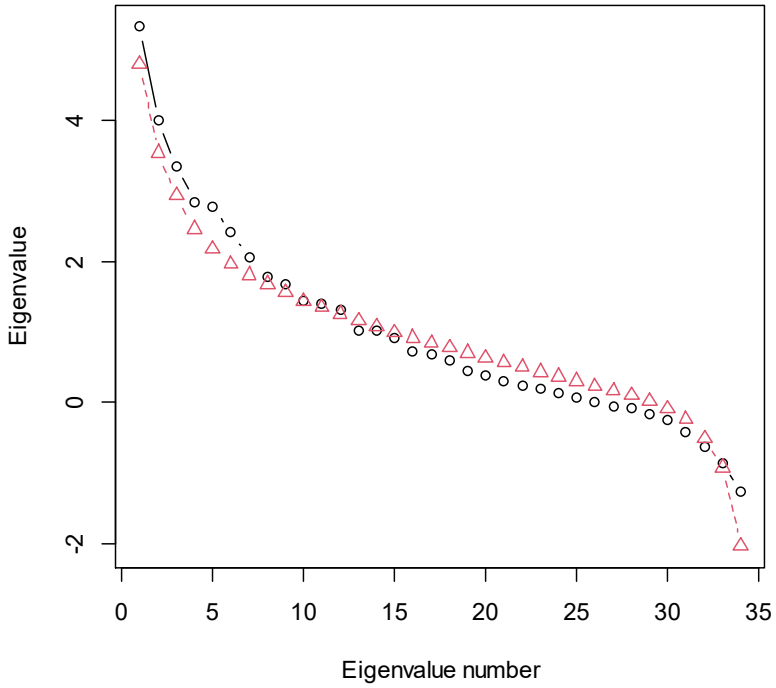

Fig 6. Graph of Analysis of Dimensionality Test Instruments Items of School Test Instruments for Chemistry Subjects at SMAN 3 Tegal

From Table 7, it can be seen that the resulting unidimensional test opportunity is 0.1089, a value greater than 0.05 so that it can be stated that Ho is accepted. If Ho is accepted, it means that the second eigen value and so on is smaller than the first eigen value. This condition can be stated that the test contains only one dimension. Thus it can be concluded that the test instrument for the Chemistry Subject School Examination at SMAN 3 Tegal can be declared to be unidimensional.

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 for Rasch modeling occurs when the estimation of the item difficulty level parameter is consistent (invariant) even though it is obtained from a sample consisting of any population subgroup during the application of Rasch modeling, in this case using PCM. The results of the Anderson LR test analysis can be seen in Table 9. From the results of the

analysis, the p value is 0 which means rejecting Ho, so it can be concluded that the parameter estimation is invariant.

Table 8. Invariance Test Measurement Using *Anderson LR test*

Andersen LR-test:
LR-value: 115.47
Chi-square df: 33
p-value: 0

3.3 Validity of External Aspects Extract

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. Ideally, researchers have data on other tests that are more accurate, such as scientific literacy tests, general intelligence tests, or special aptitude tests that support science. It can be interpreted that the external construct validity test is basically an evaluation of an instrument that has been developed.

One approach to determine the validity of the external aspect construct 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 9.

Table 9. Test of Person Separation reliability on Instrument Items for School Test in Chemistry Subjects at SMAN 3 Tegal

Separation Reliability: 0.6338
Observed Variance: 0.4823 (Squared Standard Deviation)
Mean Square Measurement Error: 0.1766 (Model Error Variance)

From Table 9, it can be seen that the Person Separation reliability value is 0.6338. Thus the person separation value for the test is 1.178. From the value of the person separation, it can be seen that the classification of the test participants obtained exceeds one. It means that the instrument that has been made can differentiate test participants in more than one category, namely reaching the KKM and not reaching the KKM. The consequence is that the results of this test only differentiate test participants into two groups, namely test participants who already have a minimum adequacy of school examination results for chemistry subjects and who do not have sufficient minimum results for school exams in chemistry subjects. This information can be followed up in determining the limits for the completeness of the school examination results in chemistry subjects at SMAN 3 Tegal.

3.4 Validity of Consequence Aspects Constructiveness

The consequential aspect in the validity of the construct lies in the implication of the score interpretation value as a source of action. The consequential validity aspect also discusses the actual and potential consequences of testing and using the score, especially in terms of sources of invalidity such as bias, fairness and distributive justice. In this regard, the measurement of school exams for chemistry subjects at SMAN 3 Tegal needs to detect a bias test.

In Rasch modeling with the eRm package, the detection of bias item 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 an item is considered more difficult or easier by male test takers than women or vice versa, then the item contains DIF. DIF or also known as item external bias is not a justification for bias item 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 10, while the description of DIF can be seen in Figure 8. Statistical criteria with the Wald test, items experiencing DIF are those that have a p-value of less than 0.05 (when using the 0.05 significance level). From Table 10, it is known that there is 1 item indicated to have DIF, namely item number 20

Table 10. List of DIF-Indicated Test Items by Gender
Significance Level 0.05

	Item	z-statistic	p-value
beta	20	2,620	0,009

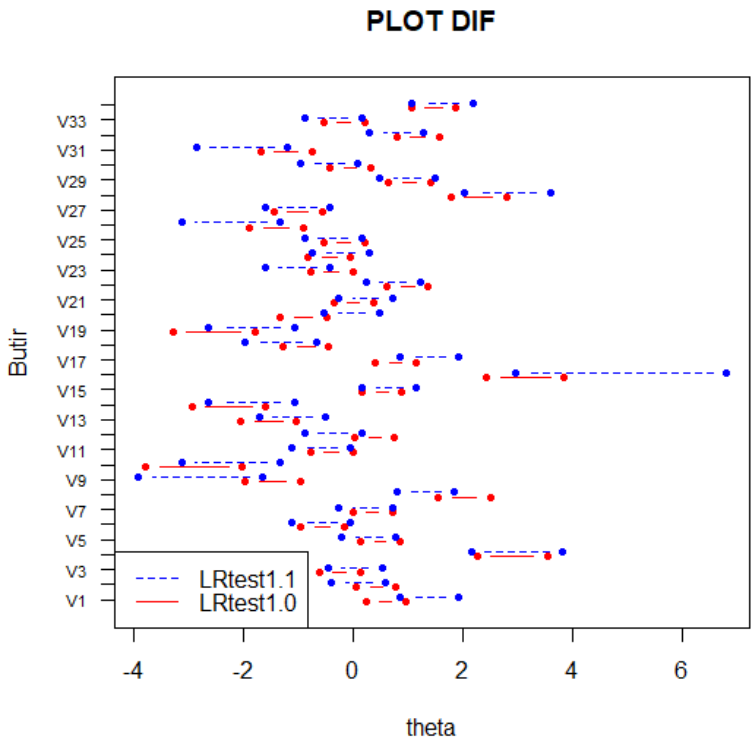

Fig 7. DIF Description on the Items for the School Examination for Chemistry Subjects at SMAN 3 Tegal

From Table 10, it can be seen that there is 1 item where the opportunity to answer correctly on each item in one questionnaire is DIF. When using a significance level of 0.05, items

number 1, 9, 12, and 20 experience DIF. When using the 0.01 significance level, only number 20 experienced DIF. In accordance with the test taker data, where the proportion of men is only 36.3%, far from the ideal proportion, of course, researchers must be 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 Ho is 0.05, then at the significance level of 0.01 it means that the chance of rejecting the correct Ho is 0.01. Ho here stated that the student's response to the test did not experience DIF. In connection with this in determining DIF, the researcher chose a significance level of 0.01 so that one item was considered detected by DIF.

Item number 20 contains material about the equilibrium reaction if the pressure system is enlarged. This material discusses many concrete things. For item number 20, the proportion of female students who answered correctly was 0.492 while male was 0.207. This point benefits women and significantly contains DIF that benefits women. This phenomenon supports several previous studies where it was found that women are easier to think abstractly while men have advantages in concrete thinking (Dietz et al., 2012; Bates et al., 2013; Madsen et al., 2013; Wilson et al., 2016)

From the research results it is known that there are three items that are not suitable for use as an instrument for measuring chemistry learning outcomes, namely items that do not fit the model (items number 4, 6, and 28) and one item detected by DIF at the 0.01 significance level, namely item number 20. Meanwhile, the other points by analyzing the construct validity of the content, substantive, structural, and external aspects of the consequences fulfill the requirements as good items.

The weakness of this study is that it has not tested the validity of the criteria for the test instrument. The criterion validity test is needed in order to ensure that the test results are in line with other standard tests that have the same construct. The validity test of this criterion can be done by comparing the results of the school exam tests for this chemistry subject with the results of other tests such as intelligence tests, aptitude tests or national exam results.

4 Conclusion

The school examination instrument for chemistry subjects consists of 34 statement items regarding the scope of basic chemistry, analytical chemistry, physical chemistry, organic chemistry, and inorganic chemistry which refers to BSNP Regulation Number 0053 / P / BSNP / I / 2020 concerning Standard Operating Procedures Implementation of the 2019/2020 Academic Year National Examination. All test items have met the construct validity. Construct validation with Rasch modeling gave the following results: (1) most of the item difficulty levels were in the range -3 to 3, (2) There were 31 items that matched the modeling, (3) There were 91.71% of student responses that matched modeling, (4) There are as many as 1 item that contains DIF. Based on the consideration of all aspects of validity, there are 31 items out of 34 that are suitable for use as test items for the chemistry subject at SMAN 3 Tegal.

Acknowledgements

The author is grateful to Pancasakti Tegal University for providing the opportunity to develop research. Likewise the authors would like to thank all parties involved, especially the principal of SMA 3 Tegal who has supported and granted research permission.

References

[1] Ayele, D. G., Zewotir, T., & Mwambi, H. (2014). Using rasch modeling to re-evaluate rapid malaria diagnosis test analyses. International Journal of Environmental Research and Public Health. https://doi.org/10.3390/ijerph110706681

[2] Baghaei, P., & Amrahi, N. (2011). Validation of a Multiple Choice English Vocabulary Test with the Rasch Model. Journal of Language Teaching and Research. https://doi.org/10.4304/jltr.2.5.1052-1060

[3] Bates, S., Donnelly, R., Macphee, C., Sands, D., Birch, M., & Walet, N. R. (2013). Gender differences in conceptual understanding of Newtonian mechanics: A UK cross-institution comparison. European Journal of Physics. https://doi.org/10.1088/0143-0807/34/2/421

[4] Bond, T. G. (2003). Validity and assessment : a Rasch measurement perspective. Metodologia de Las Ciencias Del Comportamiento.

[5] Dietz, R. D., Pearson, R. H., Semak, M. R., & Willis, C. W. (2012). Gender bias in the force concept inventory? AIP Conference Proceedings. https://doi.org/10.1063/1.3680022

[6] Edwards, A., & Alcock, A. (2010). Using rasch analysis to identify uncharacteristic responses to undergraduate assessments. Teaching Mathematics and Its Applications. https://doi.org/10.1093/teamat/hrq008

[7] Herrmann-Abell, C. F., & DeBoer, G. E. (2011). Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. Chemistry Education Research and Practice. https://doi.org/10.1039/c1rp90023d

[8] Hohensinn, C., & Kubinger, K. D. (2011). On the impact of missing values on item fit and the model validness of the Rasch model. Psychological Test and Assessment Modeling.

[9] Kemendikbud. (2016). Salinan Permendikbud Nomor 23 tahun 2016 Tentang Standar Penilaian Pendidikan. 2016.

[10] Lamprianou, I. (2010). The practical application of Optimal Appropriateness Measurement on empirical data using rasch models. Journal of Applied Measurement.

[11] Liu, M. T., & Yu, P. T. (2011). Aberrant learning achievement detection based on person-fit statistics in personalized e-learning systems. Educational Technology and Society.

[12] Lu, Y. M., Wu, Y. Y., Hsieh, C. L., Lin, C. L., Hwang, S. L., Cheng, K. I., & Lue, Y. J. (2013). Measurement precision of the disability for back pain scale-by applying Rasch analysis. Health and Quality of Life Outcomes. https://doi.org/10.1186/1477-7525-11-119

[13] Madsen, A., McKagan, S. B., & Sayre, E. C. (2013). Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? Physical Review Special Topics - Physics Education Research. https://doi.org/10.1103/PhysRevSTPER.9.020121

[14] Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of snijders's l z*

index of person fit with emphasis on response model selection and ability estimation. In Journal of Educational and Behavioral Statistics. https://doi.org/10.3102/1076998610396894

[15] Mari, L., Carbone, P., & Petri, D. (2012). Measurement fundamentals: A pragmatic view. IEEE Transactions on Instrumentation and Measurement. https://doi.org/10.1109/TIM.2012.2193693

[16] Messick, S. (1996). Validity and washback in language testing. Language Testing. https://doi.org/10.1177/026553229601300302

[17] Meyer, J. ., & Zhu, S. (2013). Fair and Equitable Measurement of Student Learning in MOOCs: An Introduction to Item Response Theory, Scale Linking, and Score Equating. Research & Practice in Assessment.

[18] Morris, G. A., Harshman, N., Branum-Martin, L., Mazur, E., Mzoughi, T., & Baker, S. D. (2012). An item response curves analysis of the Force Concept Inventory. American Journal of Physics. https://doi.org/10.1119/1.4731618

[19] Planinic, M., Ivanjek, L., & Susac, A. (2010). Rasch model based analysis of the Force Concept Inventory. Physical Review Special Topics - Physics Education Research. https://doi.org/10.1103/PhysRevSTPER.6.010103

[20] Romine, W. L., Schaffer, D. L., & Barrow, L. (2015). Development and Application of a Novel Rasch-based Methodology for Evaluating Multi-Tiered Assessment Instruments: Validation and utilization of an undergraduate diagnostic test of the water cycle. International Journal of Science Education. https://doi.org/10.1080/09500693.2015.1105398

[21] Shu, Z., Henson, R., & Luecht, R. (2013). Using Deterministic, Gated Item Response Theory Model to Detect Test Cheating due to Item Compromise. Psychometrika. https://doi.org/10.1007/s11336-012-9311-3

[22] Smith, A. B., Fallowfield, L. J., Stark, D. P., Velikova, G., & Jenkins, V. (2010). A Rasch and confirmatory factor analysis of the General Health Questionnaire (GHQ) - 12. Health and Quality of Life Outcomes. https://doi.org/10.1186/1477-7525-8-45

[23] Stenbeck, M., Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1992). Fundamentals of Item Response Theory. Contemporary Sociology. https://doi.org/10.2307/2075521

[24] Subagia, I. W. (2014). Paradigma Baru Pembelajaran Kimia SMA. Prosiding Seminar Nasional MIPA UNDIKSHA.

[25] Sumintono, B., & Widhiarso, W. (2015). Aplikasi Permodelan Rasch Pada Assessment Pendidikan. In Aplikasi Permodelan Rasch Pada Assesment Pendidikan.

[26] Susongko, P. (2016). Validation of science achievement test with the Rasch model. Jurnal Pendidikan IPA Indonesia. https://doi.org/10.15294/jpii.v5i2.7690

[27] Wilson, K., Low, D., Verdon, M., & Verdon, A. (2016). Differences in gender performance on competitive physics selection tests. Physical Review Physics Education Research. https://doi.org/10.1103/PhysRevPhysEducRes.12.020111

[28] Wind, S. A., & Gale, J. D. (2015). Diagnostic Opportunities Using Rasch Measurement in the Context of a Misconceptions-Based Physical Science Assessment. Science Education. https://doi.org/10.1002/sce.21172