Cakrawala
Jurnal Pendidikan
Volume 16 No 2 (2022)
http://cakrawala.upstegal.ac.id/index.php/Cakrawala
email: cakrawala.upstegal@gmail.com

The Functioning of The Differential Items on The Physics Learning Result Test Device at MAN Tegal City

[1]Tri Setiya Utami [1]Purwo Susongko [1]Suriswo

[1]Universitas Pancasakti Tegal, Indonesia
Email: trisetiya04@gmail.com

History

Received 29 March 2022
Revised 24 August 2022
Accepted 30 October 2022
Published 31 November 20222
DOI:
https://10.24905/cakrawala.v16i2.292

Abstract

This study aims to (1) analyze the characteristics of the physics questions tested at MAN Tegal City, (2) find out which questions were detected with DIF (3) analyze the confidence plot on the test items for the end of the 2020/2021 school year assessment test. at MAN Tegal City. The population in this study were all students of MAN Tegal City. As for the sample, all the answer sheets of students who took the final assessment test for the 2020/2021 school year on physics subjects consisting of 337 answer sheets. Data collection is by using the documentation method. The analysis technique uses Rasch modeling with Wald test. From the results of this study, it was concluded that (1) there were characteristics in the form of differences in the level of difficulty of the final assessment items for the 2020/2021 school year in physics subjects tested on class XI students at MAN Tegal City, (2) the results of the DIF detection analysis using the Wald test. shows empirical facts that in the final assessment test kit for the 2020/2021 school year physics class XI at MAN Tegal City there are no questions containing DIF, (3) the confidence plot can be used properly on the test kit..

Keywords: Confidence Plot, Final Assesment, Rasch Model, Wald Test

Fungsi Soal Diferensial Pada Alat Tes Hasil Belajar Fisika di MAN Kota Tegal

Abstrak

Penelitian ini bertujuan untuk (1) menganalisis karakteristik soal-soal fisika yang diujikan di MAN Kota Tegal, (2) mengetahui soal-soal yang terdeteksi dengan DIF (3) menganalisis confidence plot pada butir-butir tes akhir tahun pelajaran 2020/2021. tes penilaian tahun ajaran. di MAN Kota Tegal. Populasi dalam penelitian ini adalah seluruh siswa MAN Kota Tegal. Adapun sampelnya adalah seluruh lembar jawaban siswa yang mengikuti tes penilaian akhir tahun pelajaran 2020/2021 pada mata pelajaran fisika yang terdiri dari 337 lembar jawaban. Pengumpulan data dengan menggunakan metode dokumentasi. Teknik analisis menggunakan pemodelan Rasch dengan uji Wald. Dari hasil penelitian ini disimpulkan bahwa (1) terdapat karakteristik berupa perbedaan tingkat kesukaran butir soal penilaian akhir tahun pelajaran 2020/2021 pada mata pelajaran fisika yang diujikan pada siswa kelas XI MAN Tegal. City, (2) hasil analisis deteksi DIF menggunakan uji Wald. menunjukkan fakta empiris bahwa pada test kit penilaian akhir fisika kelas XI tahun ajaran 2020/2021 di MAN Kota Tegal tidak terdapat soal yang mengandung DIF, (3) confidence plot dapat digunakan dengan baik pada test kit tersebut..

Kata Kunci: Confidence Plot, Penilaian Akhir, Model Rasch, Uji Wald

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)

INTRODUCTION

The state Madrasah aliyah of Tegal City is one of the madrasas that uses the 2013 curriculum which has been revised every year. The process carried out to achieve the success of the objectives of the 2013 curriculum or to determine the success of the learning process is one of them by using evaluation, with evaluation it can be seen to what extent the learning objectives can be achieved. Learning is part of the teaching and learning process in a social interaction in order to achieve learning objectives. Measuring an achievement of learning outcomes can be through daily assessments, assignments, mid-semester assessments, end-of-semester assessments, and year-end assessments. One of the mandatory components in the evaluation is the test instrument. After passing the content validation stage, the instrument will then go through the empirical validation stage by testing it on a number of students. A good and appropriate test instrument can be identified through the characteristics of the test itself by conducting an analysis of both the items and the test as a whole, so that the good test items and the bad test items will be known.(Andayani et al., 2019)

Year-end assessment is one of the activities carried out at the state madrasah aliyah of Tegal City in order to evaluate student learning outcomes for one semester, especially in the even semester or second semester in each period of the school year. This is carried out in order to measure the ability of students, especially cognitive abilities that have been carried out by the teacher during the learning process related to the materials being taught. Assessment is an important part of the learning process. By conducting an assessment, the teacher as the manager of learning activities can find out the abilities of the students, the accuracy of the teaching methods used, and the success of students in achieving the competencies that have been set.(Safihin, 2019).

The world of education itself requires educators to conduct an evaluation to find out the problem solving skills possessed by students. Evaluation in education, one of which can use tests that function to provide information about certain aspects (Retnawati, 2014), using tests, in addition to obtaining information about students' problem solving skills, can also improve problem solving skills. Hidayat et al. (2017) mentions that students' problem solving skills also need to be measured with the aim of knowing how prepared students are in facing the challenges of the 21st century(Mulyani et al., 2021).

Analysis The state Madrasah aliyah of Tegal City which has an Islamic background and is under the auspices of the Ministry of Religion has a curriculum that is not much different from senior high schools in general, it's just that there are some special characteristic lessons given at the State Madrasah Aliyah of Tegal City. The territory of the state madrasa aliyah of Tegal City is under the auspices of the Ministry of Religion of the City of Tegal, and within one residency scope, namely the Pekalongan residency area. In one Pekalongan district, which consists of Batang district, Pekalongan City, Pekalongan Regency, Pemalang Regency, Tegal Regency, Tegal City and Brebes Regency, there is a working group for Madrasah Heads called the Working Group for Heads of Madrasah Aliyah (K3MA) Pekalongan Residency. The working group of the head of Madrasah aliyah (K3MA) of the Pekalongan residency has the authority to regulate the distribution and manufacture of end-of-semester assessment questions and year-end assessments. The Madrasah Head Working Group (KKMA) gave the mandate to appoint teachers selected by the madrasa to make end-of-semester assessment questions (PAS) and year-end assessment questions (PAT).

Physics is one of the subjects tested in the year-end assessment (PAT) with questions made by teachers who are members of the Pekalongan residency's subject teacher deliberations (MGMP). The results of learning Physics for students at the State Madrasah Aliyah of Tegal City are used to measure abilities, especially students' cognitive abilities. The results of studying Physics are what researchers will use as research material with questions whether there is bias

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)

and a differential item function (DIF) occurs. Measuring children's abilities requires a quality test device so that students' cognitive abilities can be expressed. The quality of a test kit can be seen by conducting qualitative and quantitative analysis (Hutabarat, 2009).

The state Madrasah aliyah of Tegal City carried out a year-end assessment (PAT) for the 2020/2021 school year independently, during the Covid-19 Pandemic. Therefore, for the end-of-year assessment (PAT) for the 2020/2021 school year, the questions are made by the Tegal City State Madrasah Aliyah teacher, in this case the author. In order to obtain a high-quality instrument, in addition to theoretical analysis (a review of items based on content, construction, and language aspects), an empirical item analysis is also necessary. This empirical item analysis can be divided into two, namely the classical test theory approach and item response theory (IRT).(Retnawati, 2015). The results of studying physics at the state madrasah aliyah of Tegal City in the last two years obtained data thatthere are differences in learning outcomes obtained between male students and female students. The average physics learning outcome of female students is higher than that of male students in the joint daily assessment (PHB) in the 2019-2020 academic year for odd and even semesters, as well as for the 2020-2021 academic year for odd semesters, which are 82, respectively. 78, 78 for women, while for men 78, 75, and 70. This possibility is influenced by the occurrence of item bias in the learning tools made, or it could also be due to differences in the learning process between male and female students. The ability of each student has a tendency or what is commonly called a bias.

In the physics subject test conducted at the state madrasah aliyah of Tegal City according to the data above, there is one gender of students who are more advantaged in answering the test, where the female gender is easier to work on physics questions with literacy and memorization characteristics. While the male gender is easier to work on physics problems with numerical characteristics.

Based on the description above, the researcher intends to identify which items contain DIF, describe the characteristics of the statistical curve (maple curve) on the PAT test kit that the researcher has compiled, and describe the implications of the occurrence of DIF in physics learning.

METHODS

Rasch Model and Objective Measurement

The concept of objective measurement in the social sciences according to (Bond et al., 2020; Lamprianou, 2019) must have five criteria, namely:

1. give a measure that is linear with equal intervals;
2. carry out an appropriate estimation process;
3. finding items that are inappropriate (misfits) or uncommon (outliers);
4. resolve lost data;
5. produce measurements that are replicable (independent of the parameters studied)

Five requirements of Rasch model has been able to fulfill these five conditions. In other words, the quality of measurements in the social sciences carried out with the Rasch Model will have the same quality as measurements made in the field of physics. The improvement of students' critical thinking skills can be evaluated with the presence of relevant measuring tools or instruments. The instrument is said to be good if it is able to evaluate or assess something with results such as the conditions being evaluated, to get a good test instrument, an analysis of the instrument must be carried out.(Rosidah et al., 2018).

When viewed further, the logit scale (log odds unit) generated in the Rasch Model is a scale with equal and linear intervals derived from the ratio data (odds ratio) and not the raw

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)

score data obtained. Assessment data is collected through tests or exams given to students. Tests must represent all materials or materials that have been given. Multiple choice questions (multiple choice test) is a form of test that can be used to represent all learning materials. The main component in making questions in the form of multiple choice tests is that the questions are suitable for use. The question must go through an analytical process to prove that the question is suitable for use. The method used to determine the feasibility of the item is item analysis (Fernanda & Hidayah, 2020; Nafiati et al., 2022; Santosa et al., 2022).

Therefore, the process of estimating a person's ability or level of difficulty will have a more precise estimation value and can be compared with each other because they have the same unit (logit). Since the algorithm used will perform a structured sorting between respondents from high to low ability, which at the same time also sorts questions from easy to difficult, then any inaccuracy/consistency of answers from respondents (misfit) or unusual patterns (outliers) will result in easy to detect; Likewise for the pattern of responses received by a particular question. The order of the respondent's abilities and the difficulty of the questions in a structured manner also allows the Rasch Model to make predictions if there is missing data. The resulting logit scale will display a value that depends on the response pattern given, rather than at a predetermined initial score, so the Rasch model will always produce an independent measure. There are two types of tests, namely objective tests and description tests. Objective test is a test that has several answer choices and students answer by choosing one answer that is considered correct. While the description test is a test where students answer questions by analyzing, organizing and finding their own answers and then writing them on the answer sheet (Setiawan, 2020).

Summative assessment is an assessment carried out to find out what the student already knows or what he can do, at the end of the study period. The goal is to provide information, what has happened has been achieved; in popular terms is called the assessment of learning (assessment of learning). The results of tests carried out by students are usually presented in various ways. The score a student gets on a test can show how well he is doing in class, or a comparison of his previous achievements. Moreover, the results of these exams can be used by teachers to: (a) determine students' abilities relative to other students on the same test; (b) showing the development of students' knowledge and abilities over a certain period of time; (c) show evidence of understanding of a particular subject matter, knowledge or idea; and (d) can predict student performance for the future. In order for the test results to be reliable and feasible to use, it is very important to know and report the validity and reliability of the instrument aspects.(Purwo Susongko, 2019).

Classical test theory has several fundamental drawbacks. Most of the statistics used in the classical test model, such as the level of difficulty and discriminating power of the questions, are highly dependent on the sample used in the analysis. The average ability level, range, and distribution of students' abilities that are sampled in the analysis greatly affect the statistical value obtained.(Triyatno & Ngazizah, 2014). Classical test theory uses a very simple mathematical model in showing the relationship between observation scores, actual scores, and error scores. This model is followed by a number of assumptions to simplify the formula for estimating the reliability and validity index of an instrument. Although it has developed rapidly, classical test theory actually has several drawbacks. The weaknesses are: (1) the estimation of the test taker's ability depends on the characteristics of the test used; (2) the estimation of item parameters depends on the ability of the test takers; and (3) measurement errors can only be searched for groups, not individuals (Engelhard Jr. & Wind, 2017; Lamprianou, 2019). In addition, the assumption of parallel testing is usually used to find the index reliability test, which is statistically very difficult.(P Susongko, 2016)

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)

Rasch's measurement model fixes the weaknesses of classical test theory(Science & Journal, 2020). The analysis using the Rasch model produces a statistical fit analysis that provides information to researchers whether the data obtained ideally illustrates that people who have high abilities provide patterns of answers to items according to their level of difficulty. The parameters used are infit and outfit from the mean square and standardized values.

For the most part, the measurement models found in item response theory (IRT) can provide the information needed to develop and/or assess the quality of the desired measure. The desired measure is one that is simple and easy to use and is characterized by the quality of the information obtained—usually reported as reliability and validity.(Green & Frantom, 2002)

Differential Item Functioning (DIF) Concept

DIF describes the phenomenon of one or more items of the test "functioning" differently in the group of individuals to be compared. These individuals are distinguished based on characteristics such as age, gender, ethnicity, religion, country of origin (Chiang & Tzou, 2018)(P. Susongko et al., 2021). The results of research conducted by Liu & Jane Rogers (2022) in a daily test based on CBT in science learning at a junior high school in Tegal City showed that there were some items that benefited males but there were also some items that benefited females.

In the world of education, we are often unaware of the occurrence of gender bias. Understanding gender itself is a sociocultural dimension in the psychology of men and women. The term gender is distinguished from gender (sex). Sex relates to the dimensions of men and women. Gender roles are social expectations that define how men and women should think, feel and act. The determinants of gender inequality include: 1) old problems; 2) gender values held by the community; 3) values and gender roles contained in textbooks; 4) gender values instilled by teachers; and 5) gender-biased policies.

In the detection of DIF, the existing methods that have been developed by previous measurement experts/researchers still revolve around DIF analysis using unidimensional item response theory. Unidimensional, meaning that each test item only measures one ability. The assumption of unidimensionality can be demonstrated only if the test contains only one dominant component that measures the performance of a subject.(Retnawati, 2013)

There are several methods to detect DIF in a test kit. From detection with a classic approach to detection with a modern approach. The Wald test method is one of the methods used to detect DIF. To detect the presence of DIF in the test items, a population was divided into two groups, namely the focal group and the reference group. The focal group is a group that is investigated whether there are items containing DIF in that group. The reference group is a comparison group. Both groups were drawn from the population and worked on items on the same test set. The same test set has the same validity and reliability. Population grouping can be based on gender (gender), culture, language, and ethnicity(Setiawan, 2020).

Item bias is a threat to the validity of the measurement because the score is polluted by something that is not planned to be measured. If an item is relatively more difficult for a group that has a certain culture and experience background, it means that the item is biased. Grain bias in a measurement indicates a systemic error in the measurement. Grain bias has two characters, namely direction and magnitude. The amount of bias can be estimated statistically. An item is said to be biased if two groups that have the same ability get different results on the item. Mathematically, item bias can be expressed in terms of probability. That means people who have the same ability, but do not have the same opportunity to get the right answer. If an

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)

item is relatively more difficult for a group that has a certain culture and experience background, it means that the item is biased. Grain bias in a measurement indicates a systematic error in the measurement. The procedure for detecting item bias used will determine whether the items given will provide valid information.

It appears here that the bias of the item or item that is biased arises because:

1. The test items measure the characteristics of participants that they should not measure; and
2. The test items also measure the characteristics that should not be measured, so that the item scores between groups or subgroups of test takers that should not be different, are now different.

Wald Test

The Wald-Wolfowitz test can be applied if you want to test the null hypothesis that two independent samples come from the same population or not. This means that the samples are large enough. The Wald-Wolpowitz test in principle uses the number of circuits contained in two samples. The sample distribution is tested based on the number of Runs in both samples. The minimum data scale that can be used is the ordinal scale.(Lendert et al., 2019).

Item response theory is a solution to overcome the weaknesses that exist in classical test theory because item response theory has the concept of releasing the relationship between items and samples or test takers. The characteristics/ability of the examinees remained even though they worked on items with different characteristics, and conversely, the characteristics of the items remained even though they were worked on by examinees with different abilities. In addition, item response theory is based on items/items no longer on test kits.(Hartono et al., 2022)

In this study, documentation and observation methods were used to obtain a set of questions and answers, a list of the names of students in class XI MIPA Madrasah aliyah Negeri Tegal City and the standard content of physics subjects as well as a grid for writing questions. The data collection in this study was in the form of student answer sheets, question sheets, grids and answer keys for the year-end assessment of physics class XI MIPA at the state madrasah aliyah of Tegal City for the 2020/2021 academic year.

Meanwhile, the data analysis techniques used are:

1. Estimation of item difficulty level
2. Detection of grain difficulty
3. DIF detection

RESULTS AND DISCUSSION

The estimation results using R program version 3.2.4 eRm package in the form of difficulty level (bi) for male participants and difficulty level (bi) for female participants. The analysis data is used to answer the first problem formulation in the form of an estimation of the difficulty level of the Rasch model.

In the item difficulty level parameter, if the value of bi is located at bi < -2, it is an item that is too easy, it is categorized as an easy item, it is categorized as a difficult item and bi > 2 is categorized as an item that is too difficult. Based on the criteria, the results of the item difficulty parameter analysis are presented in Table 1 as follows:

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)

Table 1. The results of the parameter analysis of the item difficulty level

No	Est. Std	Information	No	Est. Std	Information
1	2.233	Very Difficult	18	-2.495	Very Easy
2	-0.840	Easy	19	1.042	Hard
3	1,682	Hard	20	-0.229	Easy
4	-2.440	Very Easy	21	-1,770	Easy
5	-2.024	Very Easy	22	0.798	Hard
6	1.345	Hard	23	1.256	Hard
7	0.960	Hard	24	1.422	Hard
8	2,402	Very Difficult	25	2,535	Very Difficult
9	2.058	Very Difficult	26	0.403	Hard
10	1,717	Hard	27	1,717	Hard
11	-1,240	Easy	28	-0.364	Easy
12	-1,737	Easy	29	-0.859	Easy
13	-2.065	Very Easy	30	-1.673	Easy
14	-1,389	Easy	31	1.098	Hard
15	-0.288	Easy	32	0.905	Hard
16	-2.553	Very Easy	33	-1,612	Easy
17	-0.840	Easy	34	-1.838	Easy
			35	2,681	Very Difficult

Table 2. The results of the parameter analysis of the item difficulty level between men and women number 1 – 16

Item difficulty level					
No	**Man**	**Information**	**No**	**Woman**	**Information**
1	2.213	Very Difficult	1	2,243	Very Difficult
2	-0.806	Easy	2	-0.853	Easy
3	1,402	Hard	3	1,799	Hard
4	-2,611	Very Easy	4	-2.358	Very Easy
5	-1.91	Easy	5	-2.082	Very Easy
6	1.346	Hard	6	1.348	Hard
7	1,238	Hard	7	0.86	Hard
8	2.052	Very Difficult	8	2,555	Very Difficult
9	2,392	Very Difficult	9	1,953	Hard
10	1,767	Hard	10	1,702	Hard
11	-0.992	Easy	11	-1.367	Easy

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)

Item difficulty level					
No	**Man**	**Information**	**No**	**Woman**	**Information**
12	-1.516	Easy	12	-1.854	Easy
13	-2.026	Very Easy	13	-2.082	Very Easy
14	-1.516	Easy	14	-1.33	Easy
15	-0.579	Easy	15	-0.166	Easy
16	-2,611	Very Easy	16	-2.522	Very Easy

Table 3. The results of the parameter analysis of the item difficulty level between men and women number 17-35

Item difficulty level					
No	**Man**	**Information**	**No**	**Woman**	**Information**
17	-0.866	Easy	17	-0.825	Easy
18	-2,611	Very easy	18	-2.438	Very easy
19	1.033	Hard	19	1.049	Hard
20	-0.078	Easy	20	-0.294	Easy
21	-1.605	Easy	21	-1.854	Easy
22	0.839	Hard	22	0.785	Hard
23	1,291	Hard	23	1,246	Hard
24	1.346	Hard	24	1.454	Hard
25	2.3	very difficult	25	2,634	very difficult
26	0.289	Hard	26	0.453	Hard
27	1,577	Hard	27	1,774	Hard
28	-0.69	Easy	28	-0.229	Easy
29	-0.634	Easy	29	-0.968	Easy
30	-1.516	Easy	30	-1.752	Easy
31	1,238	Hard	31	1.049	Hard
32	0.984	Hard	32	0.878	Hard
33	-1.431	Easy	33	-1,704	Easy
34	-1.7	Easy	34	-1.908	Easy
35	2,392	very difficult	35	2.806	very difficult

Based on tables 2 and 3 above, male students show that there are 4 items that are categorized as very easy, namely questions number 4, 13, 16, and 18 or 11%. And there are 14 items that are categorized as easy, namely questions number 2, 5, 11, 12, 14, 15, 17, 20, 21, 28, 29, 30, 33, and 34 or 40%. There are 12 items that are categorized as difficult, namely questions number 3, 6, 7, 10, 19, 22, 23, 24, 26, 27, 31, and 32 or 34%. And there are 5 items that are categorized as very difficult, namely questions number 1, 8, 9, 25, and 35 or 14%.

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)

Meanwhile, female students showed that there were 5 items in the very easy category, namely questions number 4, 5, 13, 16, and 18 or 14%. And there are 13 items that are categorized as easy, namely questions number 2, 11, 12, 14, 15, 17, 20, 21, 28, 29, 30, 33, and 34 or 37%. And there are 13 questions that are categorized as difficult, namely questions number 3, 6, 7, 9, 10, 19, 22, 23, 24, 26, 27, 31, and 32 or 37%. And there are 4 items that are categorized as very difficult, namely questions number 1, 8, 25, and 35 or 12%.

The difficulty level of male students ranged from -2.611 to 2.393. The test items with the highest level of difficulty were found in items number 9 and 35, namely 2.393 and the lowest level of difficulty was found in items number 4 and 18, namely -2.611. Meanwhile, for female students, the difficulty level ranges from -2.522 to 2.806. The test item with the highest difficulty level is in item number 35, which is 2.806 and the lowest difficulty level is in item number 16, which is -2.522.

Table 4. The results of the DIF analysis using the Wald test item number 1-15

	z-statistics	**p-value**	**DIF**
beta V1	-0.087	0.931	Non DIF
beta V2	0.161	0.872	Non DIF
beta V3	-1.402	0.161	Non DIF
beta V4	-0.507	0.612	Non DIF
beta v5	0.421	0.674	Non DIF
beta v6	-0.008	0.994	Non DIF
beta v7	1.418	0.156	Non DIF
beta v8	-1.494	0.135	Non DIF
beta v9	1,274	0.203	Non DIF
beta v10	0.218	0.827	Non DIF
beta v11	1.182	0.237	Non DIF
beta v12	0.918	0.359	Non DIF
beta V13	0.134	0.893	Non DIF
beta V14	-0.535	0.593	Non DIF
beta V15	-1.515	0.13	Non DIF

Table 5. The results of the DIF analysis using the Wald test item number 16-35

	z-statistics	**p-value**	**DIF**
beta v16	-0.175	0.861	Non DIF
beta v17	-0.139	0.89	Non DIF
beta v18	-0.345	0.73	Non DIF
beta v19	-0.062	0.951	Non DIF
beta v20	0.823	0.41	Non DIF

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)

	z-statistics	**p-value**	**DIF**
beta v21	0.663	0.507	Non DIF
beta v22	0.21	0.834	Non DIF
beta v23	0.168	0.867	Non DIF
beta v24	-0.391	0.695	Non DIF
beta V25	-0.937	0.695	Non DIF
beta v26	-0.65	0.516	Non DIF
beta v27	-0.681	0.496	Non DIF
beta v28	-1.661	0.097	Non DIF
beta v29	1.155	0.248	Non DIF
beta v30	0.651	0.515	Non DIF
beta v31	0.705	0.481	Non DIF
beta v32	0.405	0.685	Non DIF
beta v33	0.769	0.442	Non DIF
beta V34	0.539	0.59	Non DIF
beta V35	-1.12	0.263	Non DIF

Based on tables 6 and 7, the results of DIF detection using the Wald test show that there is no 0.01 significance level. The 35 items all have a significance level above 0.01 so they do not contain DIF.

Plot DIF

With the results of the analysis using R programming version 3.2.4, the DIF plot that can be seen is a Confidence plot. The confidence plot determines which items are detected by DIF and which are not detected by DIF, located at interval for each item. With a significance level of 0.01. The following are the results of the analysis using the DIF plot for the 35 items as follows: The results of the above analysis, the questions can be categorized as easy questions because there are 40% of the items in the easy category for male students and 37% items in the easy category for students woman.

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)

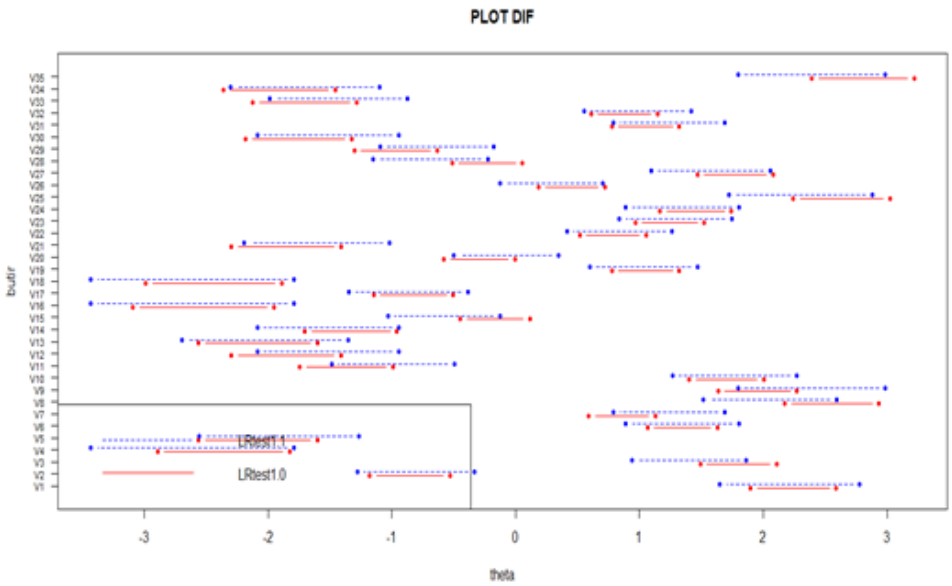

Figure 1. Plot DIF

Characteristic Curve

Detection of DIF using a characteristic curve is to compare the differences in the characteristic curves of the two groups studied. An item shows DIF if the item characteristic curves in two groups of students who have the same ability show that they do not coincide, and conversely an item does not show DIF if the item characteristic curves of two groups of students who have the same ability show that they coincide. The definition of overlap is that two groups have the same line pattern and are parallel to the ability of the students who answered the test items. To draw the item characteristic curve using the Maple 15 program, enter the output results of the equivalence test into the Rasch equation with the value of the student difficulty parameter ranging from .

The following is a figure of the gain characteristic curve on Maple 15 for items with a very difficult category. The sample is item number 9.

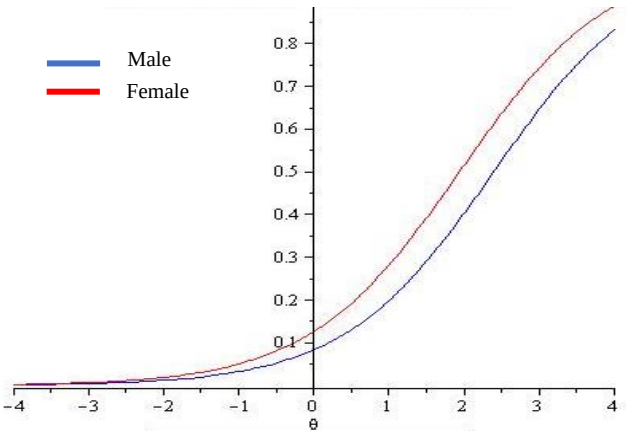

Figure 1. Gain Characteristic Curve

Detection of DIF using the Rasch model in this study did not find items containing DIF based on gender differences. The results of this study can be used as a reference, that the items

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)

used are ideal items because they do not indicate the existence of DIF based on gender, where students, both female students and male students, have the same opportunity to answer correctly.

Many factors support why this study did not find any DIF on each item, such as the ability of teachers, facilities and infrastructure, appropriate teaching methods. Different learning processes can have different effects on two groups that have the same ability. The implementation of the test can also affect whether or not DIF occurs(Hadi et al., 2021), especially with regard to test supervision and test administration procedures.

The limitation in this study is that it does not test the invariance assumption because a separate study is needed to test this assumption. As a result, if this test item is used again in a different population, the item estimation results are not always the same as the results of this study and it does not test for internal bias, namely the bias on the item itself. still requires further and in-depth research. The written discussion is attached to the data discussed. The discussion is endeavored not to be separated from the data discussed. So if not separate results and discussion.

However, in this study the authors tried to test 3 parameters to determine the gussing and discriminant values of each item that did not match the above. And the following are the results of the analysis of the 3 parameters as shown in table 10 below.

Table 10. Test results of 3 item parameters that do not match the model

No Question	Gussng	Dffclt	Dscrmn
1	0.161	2.070	28,200
3	0.200	1.379	31,828
8	0.136	2,048	51,561
10	0.229	1971	25.232
23	0.318	1,965	20,207
25	0.118	2.039	61.895
27	0.229	2006	45,330
35	0.089	1,943	3.465

After testing the 3 parameters, it turns out that the discriminant value is very high, namely for item number 1 the discriminant value reaches 28,200. Item number 3 has a discriminatory score of 31,828. Item number 8 has a discriminatory value that is quite high, namely 51.561. Meanwhile, item number 10 has a discriminative value of 25.232. Item number 23 is 20,207. Item number 25 has a very high discriminatory value, which is 61,895. Meanwhile, item number 27 has a discriminatory value of 45,330 and item number 35 has a discriminatory value of 3,465.

From the test results mentioned above, it shows a fairly high result, it can be said that for very difficult items, students tend to guess the answers they choose, so it is said that items that do not fit this model cannot measure their ability. learners.

These items cannot distinguish which students have high abilities and which students have less abilities, they only tend to guess the answers. This is because the items are very difficult

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)

for students. In this case, it is necessary to do further research on why students are less able to understand the character of the questions given by the teacher.

The possibility of the learning model delivered by the teacher in learning activities also has an influence on why the type of question is included in the category of very difficult and difficult questions. Different and creative learning models need to be done by teachers so that student learning outcomes are quite satisfactory.

During the 2020/2021 school year, learning takes place through PJJ (Distance Learning). The method used is a distance learning method using a variety of different learning models according to the abilities of teachers and students. Every learning model must have shortcomings, but a teacher must be able to choose an appropriate and appropriate learning model used to be able to convey the material being taught in the situations and conditions encountered.

The learning model that is often used by researchers is the online learning model using the zoom meeting application, and e learning. The online model is also accompanied by a home visit to find out the progress of student learning outcomes. The home visit schedule is carried out every 2 months. With the implementation of the home visit program, the hope is that teachers can find out and measure the extent to which the material is acceptable to students, as well as the obstacles faced by students in participating in distance learning methods.

CONCLUSION

In accordance with the results of this study and the discussion that the researchers put forward in the previous chapter, in this study the following conclusions can be drawn: (1) there are differences in the level of difficulty of the year-end assessment items for physics class XI in the 2020/2021 school year between male students -male and female students, (2) based on the results of the DIF detection analysis using the wald test, it shows that in the year-end assessment questions for physics class XI for the 2020/2021 academic year there are no items containing DIF, (3) confidence plot can be put to good use on each item of year-end assessment of physics subject Class XI for the 2020/2021 school year in detecting DIF which is located in the interval of students' ability levels,(4) from the model fit test, it can be said that there are several items that do not match the Rasch model, (5) the three-parameter test is carried out by the author in order to find out which items that do not match have high gussing and discriminant values.

Researchers need to convey the following suggestions (1) for teachers, especially physics teachers, in learning should adjust the conditions and environment of students and choose learning methods and models that are acceptable to students and in accordance with the circumstances. that are being faced such as the current covid 19 pandemic era, (2) for researchers, this research is a lesson in order to increase knowledge about learning models and methods so that physics subjects are no longer considered a difficult subject for students, (3) for the team MGMP physics madrasa aliyah se Ex Pekalongan Residency, in compiling the test device it would be better to be more thorough and pay attention to the conditions and learning situations of students,(4) for the problem-writing team, they should use questions that are in accordance with the level of ability of students and which are empirically proven to be of good quality and do not contain DIF, (5) for other research, further research is needed using different data and methods. to expand the study of DIF which is even better.

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)

REFERENCES

Andayani, A., Purwanto, & Ramalis, T. R. (2019). Kajian implementasi teori respon butir dalam menganalisis instrumen tes materi fisika. *Prosiding Seminar Nasional Fisika 5.0*, *1*(1), 37–42.

Bond, T. G., Yan, Z., & Heene, M. (2020). Applying the rasch model: Fundamental measurement in the human sciences. In *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. https://doi.org/10.4324/9780429030499

Engelhard Jr., G., & Wind, S. (2017). Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments. In *Invariant Measurement with Raters and Rating Scales: Rasch Models for Rater-Mediated Assessments*. https://doi.org/10.4324/9781315766829

Fernanda, J. W., & Hidayah, N. (2020). Analisis Kualitas Soal Ujian Statistika Menggunakan Classical Test Theory dan Rasch Model. *Square : Journal of Mathematics and Mathematics Education*, *2*(1), 49. https://doi.org/10.21580/square.2020.2.1.5363

Green, K. E., & Frantom, C. G. (2002). Survey Development and Validation With the Rasch Model. *International Conference on Questionnaire Development, Evaluation, and Testing*, *January 2002*, 42.

Hadi, S., Basukiyatno, B., & Susongko, P. (2021). *Differential Item Functioning National Examination on Device Test Mathematics High School in Central Java*. https://doi.org/10.4108/eai.30-11-2020.2303726

Hartono, W., Hadi, S., Rosnawati, R., & Retnawati, H. (2022). Uji Kecocokan Model Parameter Logistik Soal Diagnosa Kemampuan Matematika Dasar. *JNPM (Jurnal Nasional Pendidikan Matematika)*, *6*(1), 125–144.

Hutabarat, I. M. (2009). Analisis Butir Soal dengan Teori Tes KLasik (Classical Test Theory) dan Teori Respons Butir (Item Response Theory). *PYTHAGORAS: Jurnal Pendidikan Matematika*, *5*(2).

Lamprianou, I. (2019). Applying the Rasch Model in Social Sciences Using R. In *Applying the Rasch Model in Social Sciences Using R*. https://doi.org/10.4324/9781315146850

Lendert, R. M., Aulele, S. N., & Lesnussa, Y. A. (2019). Analisis Pengaruh Daerah Asal Sma Terhadap Nilai Ujian Mahasiswa Dengan Menggunakan Uji Wald-Wolfowitz. *VARIANCE : Journal of Statistics and Its Applications*, *1*(1), 11–15. https://doi.org/10.30598/variancevol1iss1page11-15

Liu, X., & Jane Rogers, H. (2022). Treatments of Differential Item Functioning: A Comparison of Four Methods. *Educational and Psychological Measurement*, *82*(2). https://doi.org/10.1177/00131644211012050

Mulyani, S., Efendi, R., & Ramalis, T. R. (2021). Karakterisasi Tes Keterampilan Pemecahan Masalah Fisika Berdasarkan Teori Respon Butir. *JURNAL Pendidikan Dan Ilmu Fisika*, *1*(1), 1. https://doi.org/10.52434/jpif.v1i1.1006

Nafiati, D. A., Sukirno, S., & Mulyani, E. (2022). Entrepreneurial Attitudes and Interpersonal Communication in Teaching Students. *Cakrawala: Jurnal Pendidikan*, *16*(2), 77–87. https://doi.org/10.24905/CAKRAWALA.V16I2.338

Retnawati, H. (2013). Pendeteksian Keberfungsian Butir Pembeda Dengan Indeks Volume Sederhana Berdasarkan Teori Respons Butir Multidimensi. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *17*(2), 275–286. https://doi.org/10.21831/pep.v17i2.1700

Retnawati, H. (2015). Karakteristik Butir Tes dan Analisisnya. *Uny*, *53*(5), 1–116.

Rosidah, N. A., Ramalis, T. R., & Suyana, I. (2018). Karakteristik Tes Keterampilan Berpikir Kritis (Kbk). *Jurnal Inovasi Dan Pembelajaran Fisika*, *March*, 54–63.

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)

Safihin, M. (2019). Pengembangan Tes Menggunakan Model Rasch Materi Gaya Untuk SMA. *Jurnal Pendidikan Dan Pembelajaran*, *8*(6), 1–11.

Santosa, A. D., Sartika, S. H., Surgawati, I., & Doi, : (2022). Increasing Student's Learning Motivation and Learning Outcomes through the Application of Participant Centered Learning (PCL) Cards. *Cakrawala: Jurnal Pendidikan*, *16*(2), 88–98. https://doi.org/10.24905/CAKRAWALA.V16I2.340

Science, P., & Journal, E. (2020). Pancasakti Science Education Journal. *Pancasakti Science Education Journal*, *5*, 4–11. https://doi.org/10.24905/psej.v6i2.127

Setiawan, A. (2020). *KEBERFUNGSIAN BUTIR DIFERENSIAL MENGGUNAKAN PERMODELAN RASCH PADA PENILAIAN AKHIR SEMESTER IPA DI SMP NEGERI KABUPATEN*.

Susongko, P., Arfiani, Y., & Kusuma, M. (2021). Determination of gender differential item functioning in tegal-students' scientific literacy skills with integrated science (Slisis) test using rasch model. *Jurnal Pendidikan IPA Indonesia*, *10*(2). https://doi.org/10.15294/jpii.v10i2.26775

Susongko, P. (2016). VALIDATION OF SCIENCE ACHIEVEMENT TEST WITH THE RASCH MODEL. *JPII*, *5*(2), 268–277. https://doi.org/10.15294/jpii.v5i2.7690

Susongko, Purwo. (2019). *Aplikasi Model Rasch Dalam Pengukuran Pendidikan Berbasis Program R*. Badan Penerbitan Universitas Pancasakti Tegal.

Triyatno, N., & Ngazizah, N. (2014). Bias Gender Ujian Akhir Semester Genap Fisika Kelas X SMA Negeri Kabupaten Purworejo Tahun Pelajaran 2012/2013. *Nur Triyatno*, *4*(1), 34–37.

ISSN: 2549-9300 (Online) | ISSN: 1858-4497 (Print)