International Journal of Instruction
e-ISSN: 1308-1470 ● www.e-iji.net

July 2020 ● Vol.13, No.3
p-ISSN: 1694-609X
pp. 617-632

Received: 06/07/2019
Revision: 19/02/2020
Accepted: 24/02/2020
OnlineFirst:08/05/2020

Empirical Analysis of Diagramatic Representation Test Instruments Using Partial Credit Model in Realizing Learning Outcomes

Warsono
Universitas Negeri Yogyakarta, Indonesia, *warsono@uny.ac.id*

Puji Iman Nursuhud
Universitas Negeri Yogyakarta, Indonesia, *nursuhudpujiiman_007.2017@student.uny.ac.id*

Rio Sandhika Darma
Universitas Negeri Yogyakarta, Indonesia, *riodarma58.2017@student.uny.ac.id*

Supahar
Universitas Negeri Yogyakarta, Indonesia, *supahar@uny.ac.id*

The study was conducted to analyze the items about the ability of high school students diagram representation and obtain Item Curve Characteristic. Grid test instruments are compiled based on competencies and indicators of diagram representation which are then used to compile items. The test instrument consisted of five items and was validated by measurement experts, physics education experts, material experts and practitioners. The validated instrument was piloted in 296 students from three schools in Batang Regency, Central Java. Politomous data were analyzed using the QUEST program for classical analysis and Item Response Theory for modern analysis based on Partial Credit Model using PARSCALE program. The results of item test data analysis showed that the whole item was fit with Partial Credit Model. Reliability of the test instrument is based on internal consistency of 0.66 and the level of difficulty of the items is in the range 0.84 to 1.10. The information function and Standard Error Measurement show that test items are developed reliably to measure the ability of students' diagrams with the medium category, namely $-1.2 < \theta < +2.2$ logit.

Keywords: diagrammatic, instrument, polytomous, partial credit model, representation

INTRODUCTION

Learning is a process that occurs repeatedly where students are able to explore, explore, discover and transform experiences that occur in the form of knowledge or a body knowledge (Suyono & Hariyanto, 2014). The Republic of Indonesia Minister of Education and Culture Regulation Number 22 Year 2016 concerning Basic and

Citation: Warsono, Nursuhud, P. I., Darma, R. S., & Supahar. (2020). Empirical Analysis of Diagramatic Representation Test Instruments Using Partial Credit Model in Realizing Learning Outcomes. *International Journal of Instruction*, 13(3), 617-632. https://doi.org/10.29333/iji.2020.13342a

Secondary Education Process Standards states that the learning process must be conducted interactively, inspiring, interesting, fun and challenging and can motivate students to actively participate. Learning is said to be successful if students can understand the material concepts that are taught very well (Georgiou & Sharma, 2015). Learning activities require a systematic curriculum structure. The learning environment with structured pedagogical concepts, good curriculum design and a comfortable learning atmosphere can make students transform knowledge effectively (Guney & Al, 2012).

Physics is a learning activity that has the purpose of developing logical abilities and inductive and deductive analysis of students using physics concepts so that they can solve problems (BSNP, 2006). Physics focuses on qualitative or quantitative measurements in finding and discovering basic laws relating to phenomena and using them to develop theories (Serway, R. & Jewett, J., 2004). Baran (2016) states that physics learning provides the ability for someone to solve problems in learning.

Problem solving is the most important basic element in physics learning (Docktor & Mestre, 2014). Merriënboer (2013) suggested four stages of problem solving namely (1) studying the problems raised, (2) exploring and interpreting information with appropriate procedures, (3) looking for references that support to solve problems, (4) the process of trying to solve problems. Whereas according to Dostál (2015), analyzing problem solving must consider several things such as the ability to see problems, perception of problems, ability to solve problems and problem solving strategies.

Problem solving strategies are very useful for solving problems in physics learning. Schoenfeld (2013) states that the process of finding a solution to a problem depends on the problem solving strategies used. The use of problem solving strategies must consider several things such as (1) identifying fundamental principles, (2) solving, and (3) checking (checking) (Gok, 2010). Docktor & Mestre (2014) explains that problem solving can be solved by applying representation as a solution strategy.

Diagram is a form of representation that plays an important role in solving problems (Docktor & Mestre, 2014). Chu, et al (2017) explained that diagrams as a form of representation can be used by students as a way of interpreting, representing and finding solutions to problems faced. Problems related to diagram representation that are often faced include (1) the process of describing diagrams and their components, and (2) using diagrams to translate mathematical equations (Rosengrant, et al., 2009; Samkoff, et al., 2016; Savinainen, et al ., 2013, 2017). Diagrams display various kinds of information so that it is easy to interpret and solve problems that are difficult to analyze (Pantziara, et al., 2009). Jian, et al., (2014) states that diagrams are more effectively used in the learning process. The advantages of using diagrams in the learning process include (1) it can be used easily to understand scientific phenomena, (2) presenting easy means and methods of analysis related to scientific phenomena that occur, (3) can be used to identify cognitive abilities (Sheredos, et. al., 2013). In addition, Savinainen, et al., (2017) states that diagrams as part of multiple representations have several advantages such as diagrams showing explicit scientific phenomena, diagrams acting as bridge representations between real and abstract events, and diagrams capable of growing scientific intuition.

Understanding of diagrams is determined by monitoring the process, progress and continuous improvement of learning outcomes so that an assessment is needed to measure the level of understanding of students' diagrams. Educational assessment is the process of collecting and processing information to determine student learning outcomes (Minister of Education Regulation No. 20, 2007). Assessment in the world of education can use two kinds of measurement theories, namely classical theory and modern theory. The use of classical measurement theory in Indonesia to analyze and estimate students' abilities is more desirable than modern measurement theory (Fajrianthi, et al., 2016). However, classical measurement theory has a weakness in its use. The disadvantages of classical measurement theory include the characteristics of test items such as the level of difficulty and the power of differences that depend on students (Persichitte, 2016). Problems with classical measurement theory will have an impact on the level of ability of students that is difficult to know (Awopeju & Afolabi, 2016). Problems that arise in classical measurement theory can be solved by applying modern measurement theory, namely the item response theory approach (Baker, 2001).

Item Response Theory (IRT) is a modern measurement theory that has the advantage of being able to find out the abilities and scores of students and have a more complex measurement model (Persichitte, 2016). DeMars (2010) explains that item response theory shows the relationship of ability or level trait measured using instruments and response points with a dichotomous or polytomus scoring model. The scoring model for dichotomous grains consists of: a) 1-PL model (Logistic Parameters) which emphasizes one parameter, namely the level of difficulty of the item, b) the 2-PL model which emphasizes two parameters, namely the level of grain difficulty and power difference, and c) the 3-PL model emphasizes three parameters, namely the level of difficulty of the item, different power and pseudo guessing (Mardapi, 2012). Scoring models for politomus items that are often used include the Graded Response Model (GRM), Modified Graded Response Model (MGRM), and Partial Credit Model (PCM) (Aybek & Demirtasli, 2017).

Partial Credit Model (PCM) is the development of a one-parameter logistic IRT model (1-PL) and is included in the Rasch model (Bacci et al., 2014). PCM is a politomus scoring model that uses several categories to analyze responses to an instrument (Master, 1999). For example, in a diagrammatic test instrument developed where the process of answering requires several steps of completion. The Master in Linden (2016) explained that PCM is the easiest and most widely applied polytomus item scoring model to analyze tests and assessments such as measuring critical thinking skills, computer adaptive tests (CAT), measuring conceptual understanding in science and diagnosing mathematical errors. Grunert et al., (2013) state that the PCM model is useful for knowing the level of conceptual understanding of students. The Partial Credit Model is an IRT analysis model that was developed with the aim of knowing the relationship of grain characteristics to the natural responses of students (ability or level trait). Bond & Fox (2015) states that PCM specifically combines the number of different response levels for different items on the same test which can combine dikotomous and politomous items. Muraki & Bock (1997) have mentioned the formula of PCM given in Eq. (1)

$$P_{ig}(\theta) = \frac{exp\left[\sum_{g=0}^{i}(\theta - b_{ig})\right]}{\sum_{h=0}^{m} exp\left[\sum_{g=0}^{h}(\theta - b_{ig})\right]} \quad , g = 1,2,3,\ldots,m+1 \qquad (1)$$

with $P_{ig}(\theta)$ is the probability of student with ability θ to answer i item correctly, θ is the student's ability, $m+1$ is the amount of i item category, and b_{ig} is the threshold index of i item category.

Parscale is a program for the analysis and scoring of rating-scale data. Parscale program was developed by Eiji Muraki and Darrell Bock. The interface of the Parscale program is as follows in Figure 1.

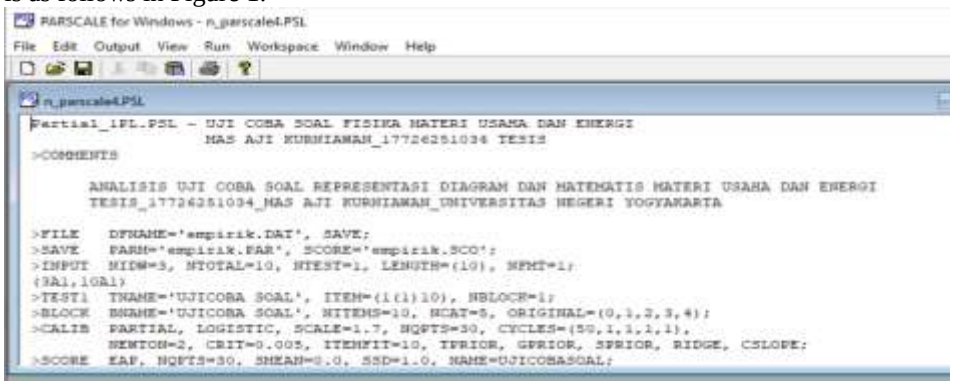

Figure 1
The Interface of the Parscale 4 Program

There are several standard menus that can be used for data analysis such as File, Edit, Output, View, Run, Workspace, Window and Help (du Toit, 2003).

METHOD

This research includes the type of development research with a quantitative approach. This development research uses a 4-D development model (Define, Design, Develop and Disseminate). The study began in November 2018 until February 2019. The development and preparation of diagram representation test instruments was carried out in November 2018 until January 2019. The trial was conducted in February 2019. The study was conducted in SMA N 1 Batang, SMA N 2 Batang and SMA N 1 Bandar, Batang regency, Central Java.

The steps for developing a test instrument follow the 4-D model which consists of: 1) Define (Defining Phase), 2) Design (Design Phase), 3) Develop (Development Phase), and 4) Disseminate (Deployment Stage). The defining stage includes: 1) determining the competency tested, 2) determining the material being tested, and 3) determining the indicator diagram representation. The design phase includes: 1) compiling the test grid and 2) arranging items according to the diagram representation indicator. The development phase includes: 1) validation of test items, 2) improvement of test items, and 3) preparation of scoring guidelines. The deployment phase includes: 1) the

determination of the trial subject, 2) the implementation of the trial, and 3) the analysis of the results of the trial data. The stages of test development are presented in Figure 2.

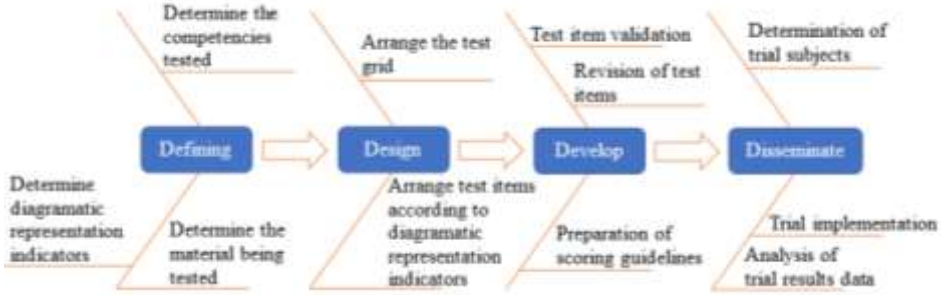

Figure 2
Steps for Developing Test Instrument using 4-D Model

The sample size used was 296 students. Bond & Fox (2015) stated that for analysis using the item response theory (IRT), a sample of between 30 and 300 people was used. While Reckase (2000) argues that the sample needed for analysis using 3-PL IRT which includes the level of difficulty, power difference and pseudoguessing is 300 people (Haladyna, 2004). The sample of this study was all students of class XI of the Science specialization program consisting of 3 classes in each school selected using the random sampling method. The number of research subjects was 296 students. So that by using the PCM model, 296 students were sufficient as subjects for empirical trials. The empirical trial was carried out by determining schools in Batang Regency based on national exam scores in physics in 2018 which were in the low, medium and high categories. The schools used for the trial include: SMA N 1 Batang, SMA N 2 Batang and SMA N 1 Bandar.

Analysis of the data from empirical trial results using PCM for fit test items for the ability of diagram representation. PCM is used to analyze test items that implement a number of completion steps. Consideration of the analysis using PCM, namely 1) can use a sample that is not too large compared to the calibration using the 2-PL logistic model and 3-PL, 2) the response characteristics of the items follow PCM. The analysis carried out included: 1) Compatibility of presentation diagram items, 2) reliability, 3) item characteristic curves, 4) item difficulty level, 5) item parameter estimation, 6) estimation of learners' abilities, 7) information function and standard error measurement (SEM). Goodness of fit analysis is carried out to determine item compatibility with the partial credit model (PCM). Goodness of fit is analyzed by interpreting the average MNSQ INFIT value along with the standard deviation or the average INFIT value t along with the standard deviation (Adams & Khoo, 1996). If the average INFIT MNSQ approaches 1.0 and the default deviation is 0.0 or the average INFIT t approaches 0.0 and the default deviation is 1.0 then the item is said to be fit with the model. The compatibility of the items with the model is known based on the INFIT MNSQ values in the range of values from 0.77 to 1.30 (Adams & Khoo, 1996). This value if converted using a standard value of t is in the range of -2 to +2 (rounding from -1.96 to +1.96) with an error rate of 5% (Bond & Fox, 2015). The item is said to be good if it has a

level of difficulty from -2 to +2 units of logit (Hambleton & Swaminathan, 1985). The QUEST and PARSCALE programs are used to analyze the results of the trial data. Scores obtained by students were analyzed using the QUEST program to determine the suitability of items for the PCM model and reliability. PARSCALE program is used to analyze data to show parameters of item characteristics such as: 1) item characteristic curve, 2) item parameter estimation, 3) estimation of student ability, 4) information function and standard error measurement (SEM).

The validation process involves six physics education experts (experts judgement). The results of experts' validation were analyzed using Aiken's formula. The Aiken formula used is as follows:

$$V = \frac{s}{n(c-1)} \qquad (2)$$

with r is the experts remarks, n is the number of points, c is the biggest scale for evaluation, and I_0 is the smallest scale of evaluation, that is:

$$s = r - I_0 \qquad (3)$$

The results of the analysis of the material aspects using aiken's v show that the test instruments developed are in the valid category. This is in accordance with the validation criteria according to Aiken's v (Aiken, l., 1985) which states that for 8 validators (expert judgements), item items are declared valid if they obtain Aiken's v score ≥ 0.75.

FINDINGS

Development Results of The Diagramatic Representation Ability Test

The test instruments developed amounted to 5 items which were in accordance with the material momentum and impulses. The test instrument is prepared and assembled in accordance with the diagram representation indicator which includes 1) describing the diagram and its components, 2) performing mathematical calculations according to the description of the diagram. The test instrument was assessed for its feasibility by 8 experts judgment before being used for the trial phase. Tables 1 and 2 show the distribution of test items the ability of diagram representation and the results of validation by expert judgment.

Table 1
Distribution of Diagramatic Representation Ability Test Items

Indicator of Diagramatic Representation	Item Number	Material
Describe the diagram and its components	1,2,3	Momentum and Impulse
Perform mathematical calculations according to the description of the diagram	4,5	

Table 2
Test Item Validation Results based on Aiken's V

Representation	Item Number	Score of Aiken's V	Criteria
Diagram	1,2,3	0,92	Valid
	4,5	0,88	Valid

The test instrument developed refers to diagramatic representation indicators which are part of problem solving. Based on the aiken's v analysis, the test instruments is qualified as valid instrument. The results of the analysis of the material aspects using aiken's v show that the test instruments developed are in the valid category. This is in accordance with the validation criteria according to aiken's v (Aiken, l., 1985) which states that for 8 validators, item items are declared valid if they obtain Aiken's v score ≥ 0.75.

Goodness of Fit of Diagramatic Representation Item Tests for the PCM Model

Overall testing of goodness of fit is done by analyzing the results of the trial test questions using the Quest program. Goodness of fit is tested according to the rules developed by Adams & Khoo (1996) by looking at the average INFIT MNSQ value and the standard deviation or by observing the average value of INFIT t and the standard deviation. The test instrument is said to be fit with the PCM model if the average INFIT MNSQ value is around 1.0 and the standard deviation is 0.0 or the INFIT average value is around 0.0 and the default deviation is 1.0. Table 3 shows item and test estimates from the ability test instrument diagram representation.

Table 3
Item Estimation and Test of the Test Instrument

Description	Item Estimation	Test Estimates
Average value and standard deviation	$0{,}01 \pm 0{,}29$	$-0{,}45 \pm 0{,}71$
Reliability	0,32	0,61
Average value and standard deviation of INFIT MNSQ	$0{,}97 \pm 012$	$0{,}94 \pm 0{,}69$
Average value and standard deviation of OUTFIT MNSQ	$0{,}95 \pm 0{,}12$	$0{,}95 \pm 0{,}69$
Average value and standard deviation of INFIT t	$-0{,}23 \pm 1{,}37$	$-0{,}08 \pm 1{,}10$
Average value and standard deviation of OUTFIT t	$-0{,}34 \pm 0{,}96$	$0{,}05 \pm 0{,}84$

Testing of goodness of fit for each item follows the rules developed by Adams & Khoo (1996) by looking at the INFIT MNSQ value of each item based on the output of the QUEST program. The item is declared fit or suitable for the model if the MNSQ INFIT value ranges from 0.77 to 1.30. In addition, items are also declared fit to the model if the INFIT t value is in the range of -2 to +2. Table 4 shows the INFIT MNSQ and INFIT values for each item. Table 4 shows that the diagram representation ability ability test items developed have MNSQ INFIT value ranges from 0.84 to 1.10.

Table 4
Distribution of INFIT MNSQ and INFIT t Each Test Item

Item	INFIT MNSQ	OUTFIT MNSQ	INFIT t	OUTFIT t
1	0,84	0,80	-1,50	-1,40
2	1,07	1,01	0,90	0,10
3	1,00	0,94	0,10	-0,50
4	0,86	0,88	-1,80	-1,00
5	1,10	1,11	1,20	1,00
Average	0,97	0,95	-0,20	-0,30

Reliability

The reliability value indicates that the diagram representation ability test instrument developed is qualified as good instrument. Reliability is obtained based on analysis

using the QUEST program is 0.66. The reliability value obtained has a medium category. (Lima, et al., 2018).

Item Characteristic Curve

Item characteristics are indicated by item characteristic curve (ICC). Analysis to find out the ICC was used by the PARSCALE 4 program. The analysis carried out was obtained as many as 5 items characteristic curves. Figure 3 presents ICC item number 4. The ICC chart in Figure 3 shows the opportunity for students to answer item test number 4 based on their ability. Opportunities for students working on item number 4 are as follows: 1) category 1 is $\beta_1 = 0{,}80$, 2) category 2 is $\beta_2 = 0{,}29$, 3) category 3 is $\beta_3 = 0{,}19$, 4) category 4 is $\beta_4 = 0{,}24$, 5) category 5 is $\beta_5 = 0{,}60$.

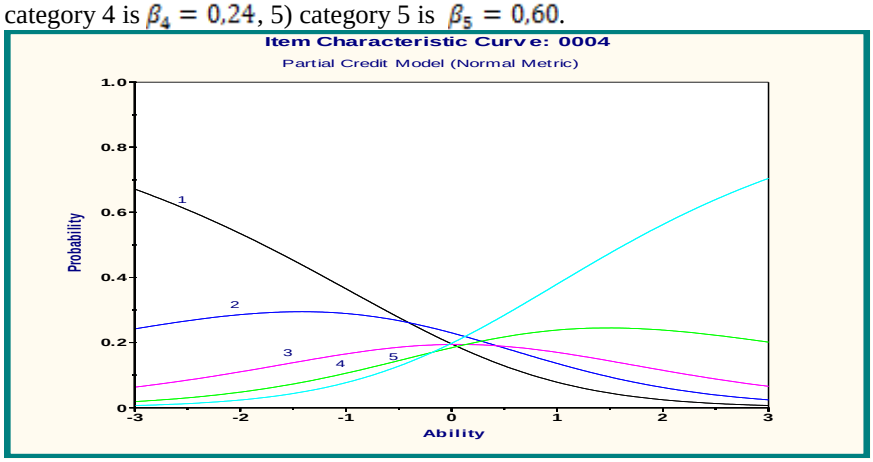

Figure 3
Item Characteristic Curve for Item Number 4

ICC for item number 4 contains information as follows: 1) category 1 is mostly obtained by students who have ability -4.0 scale logit. 2) category 2 mostly obtained by students who have ability -0.3 logit scale. 3) category 3 is mostly obtained by students who have the ability of 1.1 scale logit. 4) category 4 is mostly obtained by students who have the ability of 2.5 logit scale. 5) category 5 is mostly obtained by students who have a 4.0 scale logit ability.

Item Parameter Estimation

The estimated parameter of the diagram representation ability according to the PCM model is indicated by different difficulty levels for each item. Table 5 shows a summary of parameter estimates analyzed using the PARSCALE 4 program.

Table 5
Test Item Parameter Estimation

PARAMETER	MEAN	STN DEV	N
SLOPE	0,394	0,000	5
LOG (SLOPE)	-0,932	0,000	5
THRESHOLD	0,556	0,395	5
GUESSING	0,000	0,000	0

The power estimation of different items is indicated by the SLOPE parameter which has an average value of 0.394. The level of difficulty of the item is indicated by the THRESHOLD parameter which has an average value of 0.556. The pseudo guessing parameter is shown by the GUESSING parameter which has a value of 0,000. Partial Credit Model (PCM) refers to one parameter, namely the difficulty level of an item. Table 6 shows the difficulty level of each item diagram representation capability for each score category in PCM.

Table 6
Level of Difficulty Test Item Diagramatic Representation Ability

Item Number	Difficulty	Stage Difficulty				
		Category 1	Category 2	Category 3	Category 4	Category 5
1	0,31	-0,63	0,50	0,88	-0,68	-0,07
2	0,06	-0,85	0,28	-0,69	0,13	1,13
3	0,23	-0,89	-0,10	0,35	0,24	0,40
4	-0,32	-1,27	1,17	0,28	-0,12	-0,07
5	-0,28	-0,88	-0,32	0,99	0,25	-0,03

Table 6 shows that PCM measures the ability of students to work on test items based on the steps taken in the form of categories. Each category has different difficulty levels for each item.

Estimated Ability of Learners

Estimating the level of ability of students is shown by the histogram. Figure 4 shows a histogram of students' diagrammatic representation abilities. Figure 4 shows that students' diagrammatic abilities follow a normal curve.

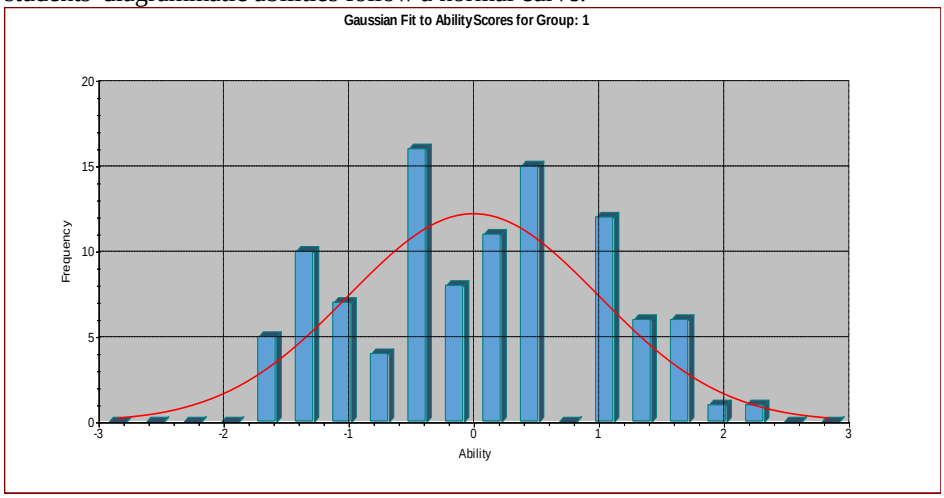

Figure 4
Histogram Estimated Diagrammatic Representation Capability

The histogram in Figure 4 can be interpreted by tabulating the frequency into the table. Table 7 shows the interpretation of the ability of diagrammatic representation of students on a logit scale.

Table7
Student Diagrammatic Representation Capability Category

Sample	Ability (Logit Scale)	Interpretation
1	+2,00 up to +3,00	Very High
25	+1,00 up to +2,00	High
54	-1,00 up to +1,00	Medium
22	-2,00 up to -1,00	Low
0	-3,00 up to -2,00	Very Low

The results of data interpretation based on Table 7 show that there are no students who have very low diagram representation capabilities. Table 7 shows that 0.98% has a very high diagrammatic representation ability, 24.50% has a high diagrammatic representation ability, 52.94% has medium diagrammatic representation capabilities and 21.56% has a low diagrammatic representation capability. The results of the study in Table 7 show that the ability of students to represent their diagrams is in the moderate to very high category. These results indicate that the ability of good students is in the range of -2 up to +2 (Nursuhud, et al., 2019). This value if converted using a standard value of t is in the range of -2 to +2 (rounding from -1.96 to +1.96) with an error rate of 5% (Bond & Fox, 2015).

Information Function and Standard Error Measurement (SEM)

Information functions and standard error measurement (SEM) were obtained based on analysis using the PARSCALE 4 program. Figure 5 shows a graph of total functions and SEM. The analysis results obtained intersection of information function lines and SEM lines at point -1.2 up to +2.2 logit scale.

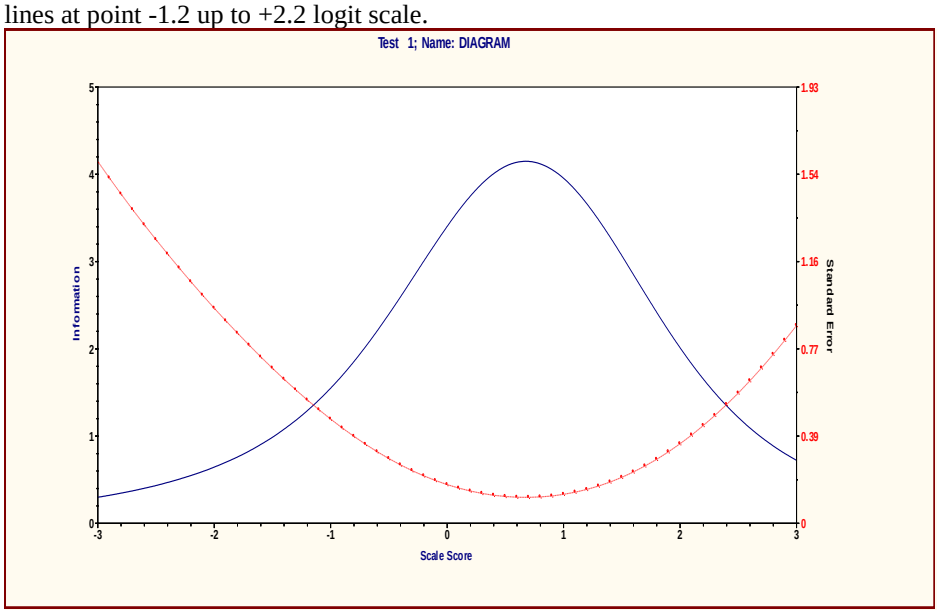

Figure 5
Information Function and Standard Error Measurement (SEM)

This value shows that the diagram representation ability test instrument is developed reliably when tested on students with moderate ability (θ) which is -1.2 < θ < +2.2 logit scale with SEM $\pm$ 0,48 . These results indicate tat the ability of diagrammatic representation of students is classified as medium.

DISCUSSION

The test instruments developed amounted to 5 items which were in accordance with the material momentum and impulses. The test instrument is prepared and assembled in accordance with the diagram representation indicator which includes 1) describing the diagram and its components, 2) performing mathematical calculations according to the description of the diagram. The test instrument developed refers to the indicator diagram representation which is part of problem solving. Docktor & Mestre (2014) states that problem solving is an important strategy in solving physics problems. This is supported by the results of the Merriënboer (2013) study which suggested that there are four stages of problem solving namely (1) studying the problems raised, (2) exploring and interpreting information with appropriate procedures, (3) looking for references that support solving problems, (4) the process of trying to solve a problem. Indicator diagram representation is developed with the aim of students being able to understand the concept as a whole so as to be able to apply it to solve physical problems, especially material momentum and impulses. The diagram representation indicators developed include (1) drawing diagrams and their components and (2) using diagrams to do mathematical calculations (Rosengrant et al., 2009; Samkoff et al., 2016; Savinainen et al., 2013, 2017). The test instrument was assessed for its feasibility by expert judgment before being used for the trial phase. The feasibility of the test instrument was assessed based on material aspects and empirical tests (Yadiannur & Supahar, 2017). Goodness of fit is tested according to the rules developed by Adams & Khoo (1996) by looking at the average INFIT MNSQ value and the standard deviation or by observing the average value of INFIT t and the standard deviation. The test instrument is said to be fit with the PCM model if the average INFIT MNSQ value is around 1.0 and the standard deviation is 0.0 or the INFIT average value is around 0.0 and the default deviation is 1.0. The item is declared fit or suitable for the model if the MNSQ INFIT value ranges from 0.77 to 1.30. In addition, items are also declared fit to the model if the INFIT t value is in the range of -2 to +2. Table 4 shows that the diagram representation ability ability test items developed have MNSQ INFIT value ranges from 0.84 to 1.10. This value indicates that all items have MNSQ INFIT values located in the range of acceptance of goodness of fit, so it is concluded that all test items are suitable and suitable for the partial credit model (PCM). This result is the rule of Adams & Khoo (1996) which states that items are declared fit to the model if they have MNSQ INFIT values in the range 0.77 to 1.30. This opinion is supported by Bond & Fox (2015) that good test items have a range of INFIT values from -2 to +2. Reliability is obtained based on the internal consistency value of the QUEST program analysis output. The internal consistency value obtained is 0.66 which indicates the magnitude of Cronbach's Alpha. The reliability value obtained has a medium category and shows that the ability of the diagram representation ability

test instrument to be developed can be used to make decisions about students (Suryabrata, 2002).

Findings of the study shows that the ttem characteristics are indicated by item characteristic curve (ICC) in figure 3. Figure 3 presents ICC item number 4. The ICC chart in Figure 3 shows the opportunity for students to answer item test number 4 based on their ability. Opportunities for students working on item number 4 are as follows: 1) category 1 is $\beta_1 = 0{,}80$, 2) category 2 is $\beta_2 = 0{,}29$, 3) category 3 is $\beta_3 = 0{,}19$,, 4) category 4 is $\beta_4 = 0{,}24$, 5) category 5 is $\beta_5 = 0{,}60$. ICC for item number 4 contains information as follows: 1) category 1 is mostly obtained by students who have ability -4.0 scale logit. 2) category 2 mostly obtained by students who have ability -0.3 logit scale. 3) category 3 is mostly obtained by students who have the ability of 1.1 scale logit. 4) category 4 is mostly obtained by students who have the ability of 2.5 logit scale. 5) category 5 is mostly obtained by students who have a 4.0 scale logit ability. The scoring of the politomus data model using PCM produces characteristic curves such as Figure 3. This is as explained by Master (1999) that PCM is a politomus scoring model that uses several categories to analyze the response to an instrument. Politomus newspapers use partial credit divided into several categories. Categories are sorted from easy to difficult categories, namely 1,2,3,4,5. This result agrees with the research of Grunert, et al. (2013) which states that the use of partial credit which is divided into several categories gives a significant impact on the item being tested. The categories on PCM specifically combine the number of different response levels for different items on the same test which can combine dichotomous and politomous items (Bond & Fox, 2015).

The estimated parameter of the ability to test diagram representation according to the PCM model is indicated by different difficulty levels for each item. Partial Credit Model (PCM) refers to one parameter, namely the difficulty level of an item. Findings of the study shows that the difficulty level of each item diagram representation capability divide for each score category in PCM. PCM measures the ability of students to work on test items based on the steps taken in the form of categories. Each category has different difficulty levels for each item. This result agrees with the research of Grunert, et al. (2013) which states that the use of partial credit which is divided into several categories gives a significant impact on the item being tested. The results of the research in Table 6 on the difficulty column show the difficulty level of each item. The difficulty value or the difficulty of the item is in the range of -2 to +2. This value is in accordance with the opinion of Bond & Fox (2015) which states that the level of difficulty for items with good categories is in the range of -2 to +2 (rounding from -1.96 to +1.96) with an error rate of 5%. Bond & Fox's opinion is supported by Hambleton & Swaminathan (1985) which shows that the item is said to be good if it has a level of difficulty from -2 to +2 units of logit.

Estimating the level of ability of students is shown by Figure 4. Figure 4 shows a histogram of students' diagrammatic representation abilities. Table 7 shows the interpretation of the ability of diagrammatic representation of students on a logit scale

based on histogram in Figure 4. Table 7 show that there are no students who have very low diagram representation capabilities. Table 7 shows that 0.98% has a very high diagrammatic representation ability, 24.50% has a high diagrammatic representation ability, 52.94% has medium diagrammatic representation capabilities and 21.56% has a low diagrammatic representation capability. The results of the study in Table 7 show that the ability of students to represent their diagrams is in the moderate to very high category. These results indicate that the ability of good students is in the range of -2 up to +2 (Nursuhud, et al., 2019). This value if converted using a standard value of t is in the range of -2 to +2 (rounding from -1.96 to +1.96) with an error rate of 5% (Bond & Fox, 2015). This proves that the test item diagram representation ability developed is able to measure the level of ability of students. The results of this study agree with the Master's statement in Linden (2016) which explains that PCM is the easiest and most widely applied polytomus item scoring model to analyze tests and assessments such as measuring critical thinking skills, computer adaptive test (CAT), measuring conceptual understanding in science and diagnose mathematical errors. In addition, DeMars (2010) explains that the use of item response theory in assessment can show the relationship between ability or level trait measured using instruments and response items with dichotomous or polytomus scoring models.

Findings of the study about information functions and standard error measurement (SEM) were obtained based on analysis using the PARSCALE program. Figure 5 shows a graph of total functions and SEM. The analysis results obtained intersection of information function lines and SEM lines at point -1.2 up to +2.2 logit scale. This value shows that the diagram representation ability test instrument is developed reliably when tested on students with moderate ability (θ) which is $-1.2 < \theta < +2.2$ logit. This is in agreement with Istiyono, et at. (2019) which states that the ideal ability to answer questions lies between the intersection of two curves. These results indicate that students' diagrammatic representation abilities are still relatively moderate. This agrees with the research of Jian et al. (2014) which states that diagrams can be used effectively in teaching and learning activities. Whereas, Sheredos et al., (2013) explained that diagrams can be used to identify cognitive abilities. The results of the analysis in Figure 5 prove that the test instrument developed has a medium category to measure the ability of diagrammatic representation of students.

CONCLUSION

This research can be concluded: (1) The diagram representation ability test instrument is developed in the form of description questions for indicators describing diagrams and their components and performing mathematical calculations according to the explanation of diagrams (2) Instrument representation ability diagram has fulfilled content validity based on expert judgment and obtained evidence construct validity fit with the Partial Credit Model (PCM) model based on five categories of polytomus data; (3) All items test the ability of diagram representation developed has good criteria.

Suggestions for further research can be developed on other learning material items to improve the ability of physics representation and high order thinking skills of high

school students. In addition, the items developed can be integrated with information technology in testing such as computer based adaptive tests.

ACKNOWLEDGEMENTS

Thank you to the Republic of Indonesia Ministry of Research and Higher Education for helping with research funding.

REFERENCES

Adams, R. J., & Khoo, S. T. (1996). *Quest: the interactive test analysis system version 2.1*. Victoria: The Australian Council for Educational Research.

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurements, 45*, 131-142.

Awopeju, & Afolabi. (2016). Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal*, *12*(28), 263–284. https://doi.org/10.19044/esj.2016.v12n28p263.

Aybek, E. C., Demirtasli, R. N., & Computerized, R. N. (2017). Computerized adaptive test (CAT) applications and item response theory models for polytomous items. *Int J of Res in Edu and Sci (IJRES), 3*(2), 475-487. https://doi.org/10.21890/ijres.327907.

Bacci, S., Bartolucci, F., & Gnaldi, M. (2014). A class of multidimensional latent class IRT models for ordinal polytomous item responses. *Communications in Statistics-Theory and Methods, 43*(4), 787-800. https://doi.org/10.1080/03610926.2013.827718.

Baker, F.B. (2001). *The basic of item response theory*. USA: ERIC Clearinghouse on Assessment and Evaluation.

Baran, M. (2016). An analysis on high school students' perceptions of physics courses in terms of gender (A Sample from Turkey). *Journal of Education and Training Studies, 4*(3), 150–160. https://doi.org/10.11114/jets.v4i3.1243.

Bond, T. G., & Fox, C. M. (2015). *Applying the rasch model. Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

BSNP. (2006*). Standar kompetensi dan kompetensi dasar mata pelajaran fisika untuk sma dan ma*. Jakarta: BSNP-Depdiknas.

Chu, J., Rittle-Johnson, B., & Fyfe, E. R. (2017). Diagrams benefit symbolic problem-solving. *British J of Edu Psyc, 87*(2), 273–287. https://doi.org/10.1111/bjep.12149.

DeMars, C. (2010). *Item response theory*. Oxford: Oxford University Press.

Depdiknas. (2007). *Peraturan menteri pendidikan nasional no 20 tahun 2007 tentang standar penilaian*.

Docktor, J. L., & Mestre, J. P. (2014). Synthesis of discipline-based education research in physics. *Physical Review Special Topics – Physics Education Research*, *10*(*020119*), 1–58. https://doi.org/10.1103/PhysRevSTPER.10.020119.

Dostál, J. (2015). Theory of problem solving. *Procedia - Social and Behavioral Sciences*, *174*, 2798–2805. https://doi.org/10.1016/j.sbspro.2015.01.970.

Du Toit, & Mathilda. (2003). *IRT from SSI: BILOG-MG MULTILOG PARSCALE TESTFACT*. USA: Scientific Software International, Inc.

Fajrianthi., Hendriani, W., & Septarini, B. G. (2016). Pengembangan tes berpikir kritis dengan pendekatan item response theory. *Jurnal Penelitian dan Evaluasi Pendidikan, 20*(1), 45–55.

Georgiou, H., & Sharma, M. D. (2015). Does using active learning in thermodynamics lectures improve students' conceptual understanding and learning experiences. *European J of Physics, 36*(1), 015020. https://doi.org/10.1088/0143-0807/36/1/015020.

Gok, T. (2010). The general assessment of problem solving processes and metacognition in physics education. *Eurasian Journla of Physics and Chemistry Education, 2*(2), 110–122. https://doi.org/10.1007/s11409-008-9026-0.

Grunert, M. L., Raker, R., Murphy, K. L., & Holme, T. A. (2013). Polytomous versus dichotomous scoring on multiple-choice examinations: Development of a rubric for rating partial credit. *Journal of Chemical Education, 90*(10), 1310-1315. https://doi.org/10.1021/ed400247d.

Guney, A., & Al, S. (2012). Effective learning environments in relation to different learning theories, *Procedia - Social and Behavioral Sciences, 46*, 2334–2338. https://doi.org/10.1016/j.sbspro.2012.05.480.

Haladyna, T. M. (2004). *Developing and validating multiple choice test items*. New Jersey: Lawrence Erlbaum Associates, Inc.

Hambleton, R. K, & Swaminathan, H. (1985). *Item response theory*. Boston: Kluer.

Istiyono, E., Dwandaru, W. S. B., Lede, Y. A., Rahayu, F., & Nadapdap, A. (2019). Developing IRT-based physics critical thinking skill test: A CAT to answer 21st century challenge. *Int J of Instruction, 12*(4), 267-280. https://doi.org/10.29333/iji.2019.12417a

Jian, Y. C., Wu, C. J., & Su, J. H. (2014). Learners' eye movements during construction of mechanical kinematic representations from static diagrams. *Learning and Instruction, 32*, 51–62. https://doi.org/10.1016/j.learninstruc.2014.01.005.

Lima, E., Teixeira-Salmela, L. F. Magalhaes, L. C., Laurentino, G. E., Simoes, L. C., Moretti, E., … & Lemos, A. (2018). Measurement properties of the Brazilian version of the motor assessment scale, based on rasch analysis. *Disability and Rehabilitation*, 1-6.

Mardapi, D. (2012). *Pengukuran penilaian dan evaluasi pendidikan.* Yogyakarta: Nuha Medika.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. Van Der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp.101-121). New York: Springer.

Merriënboer, J. J. G. Van. (2013). Computers & education perspectives on problem solving and instruction. *Computers & Education, 64*, 153–160. https://doi.org/10.1016/j.compedu.2012.11.025.

Muraki, E., & Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for ratingscale data [Computer software]*. Chicago: Scientific Software.

Nursuhud, P. I., Warsono, Supahar, & Darma, R. S. (2019). Development of vector representation test instrument for senior high school students in realizing learning outcomes. *International Journal of Educational Research Review*, *4*(4), 543-554. https://doi.org/10.24331/ijere.628328.

Pantziara, M., Gagatsis, A., & Elia, I. (2009). Using diagrams as tools for the solution of non-routine mathematical problems. *Educational Studies in Mathematics*, *72*(1), 39–60. https://doi.org/10.1007/s10649-009-9181-5.

Persichitte, K. A. (2016). *Educational Technology to improve quality and access on a global scale*. New York: Springer.

Rosengrant, D., Heuvelen, A. Van, & Etkina, E. (2009). Do students use and understand free-body diagrams ? *Physical Review Special Topics – Physics Education Research*, *5*(010108), 1–13. https://doi.org/10.1103/PhysRevSTPER.5.010108

Samkoff, A., Lai, Y., & Weber, K. (2016). Research in mathematics education on the different ways that mathematicians use diagrams in proof construction. *Research in Mathematics Education, 14(*1), 49-67. https://doi.org/10.1080/14794802.2012.657438.

Savinainen, A., Mäkynen, A., Nieminen, P., & Viiri, J. (2013). Does using a visual-representation tool foster students' ability to identify forces and construct free-body diagrams? *Physical Review Special Topics - Physics Education Research*, *9*(1), 1–11. https://doi.org/10.1103/PhysRevSTPER.9.010104.

Savinainen, A., Mäkynen, A., Nieminen, P., & Viiri, J. (2017). The effect of using a visual representation tool in a teaching-learning sequence for teaching Newton's third law. *Research in Science Education*, *47*(1), 119–135. https://doi.org/10.1007/s11165-015-9492-8.

Schoenfeld, A. H. (2013). Reflections on problem solving theory and practice. *The Mathematics Enthusiast*, *10*(1), 9-34.

Serway, R. A., & Jewett, J. W. (2012). *Physics for scientists and engineers.* United States: Brooks-Cole Publishing.

Sheredos, B., Burnston, D., Abrahamsen, A., & Bechtel, W. (2013). Why do biologists use so many diagrams. *Philosophy of Science*, *80*(5), 931–944. https://www.jstor.org/stable/10.1086/674047.

Suryabrata, S. (2002). *Pengembangan alat ukur psikologis*. Yogyakarta: Andi Offset.

Suyono, & Hariyanto. (2014). *Belajar dan Pembelajaran*. Bandung: PT Remaja Rosdakarya Offset.

Yadiannur, M., & Supahar. (2017). Mobile learning based worked example in electric circuit (WEIEC) application to improve the high school students' electric circuits interpretation ability. *International Journal of Environmental and Science Education, 12*(3), 539–558. https://doi.org/10.12973/ijese.2017.1246p.