Analysis of Validity and Reliability of Diagnostic Test of Picture Representation Ability in High School Physics Learning

M F T Nirmala[1,*] A M R Tumanggor[2,] Supahar[3]

[1,2,3] *Physics Education, Graduate School, Universitas Negeri Yogyakarta, Sleman, Indonesia*
[*]Corresponding author. Email: maria0020pasca.2018@student.uny.ac.id

ABSTRACT

The purpose of this research was to determine the feasibility of a diagnostic test of the picture representation ability of high school students in physics learning. This research was a developmental research of test instrument. Analysis of test instruments includes qualitative and quantitative analysis. Qualitative data analysis was data of content validation result by 7 raters with four scale rating. Quantitative data analysis was the result of the response of 256 high school students to the test instrument. The number of items tested was 36, divided into 2 test sets; set A and set B, where anchor items were 4. Analysis of content validity using Aiken V proved to be valid. Valid based empirically data because it matches the MNSQ INFIT value of the Rasch model. Based on the TIC and SEM charts, the diagnostic test instrument of picture representation ability was reliable if applied to students with moderate to very high category ability. Therefore, the diagnostic test instrument of picture representation ability developed is feasible to use.

*Keywords: Validity, Reliability, Diagnostic test, Picture representation ability.*

1. INTRODUCTION

Assessment is an important part of the learning process. In Permendikbud No. 23 of 2016 concerning the assessment standards, explained that assessment is a process of gathering and processing information to measure the achievement of student learning outcomes [1]. Assessment is not just limited to providing a number of test questions, then shows the true and false scores of students, but provides information about the achievement patterns of mastery ability of students concerned [2]. Assessment is needed in terms of monitoring the process, progress and improvement of student learning outcomes on an ongoing basis. Information in the form of assessment results can help teachers and students in planning follow-up activities, which in turn will improve learning outcomes. Assessment helps teachers to understand what students know and able to do after learning [3]. A challenge for teachers is how to design assessments that can support learning for students. Assessment is an inseparable part of learning, and in an intentionally done context.

Physics is one area of learning that can contribute to the advancement of science and technology. In the

contents of Permendikbud Number 21 Year 2016 it is explained that physics is related to a way of finding out about natural phenomena or phenomena systematically [4]. To understand this phenomenon, one ability is needed, that is representation. The way of presenting something in other different or another situation called representing. Forms of representation called images are images, diagrams, tables and graphs [5].

The challenge for teachers in learning is to detect students' difficulties and weaknesses because students have different characteristics [6]. One form of assessment that can detect students' difficulties and weaknesses is a diagnostic assessment [7]. With the diagnosis of abilities and weaknesses, it is expected to provide appropriate assistance to students according to their needs so that no students are left behind in learning. If a test item is used for diagnostic use, the focus is on analyzing students' wrong answers and the analysis is usually done at the individual level [8]. However, the results of observations and interviews with several Physics subject teachers showed that teachers did not have enough time to develop an instrument that meets valid and reliable criteria. Aside from the teacher being preoccupied with various
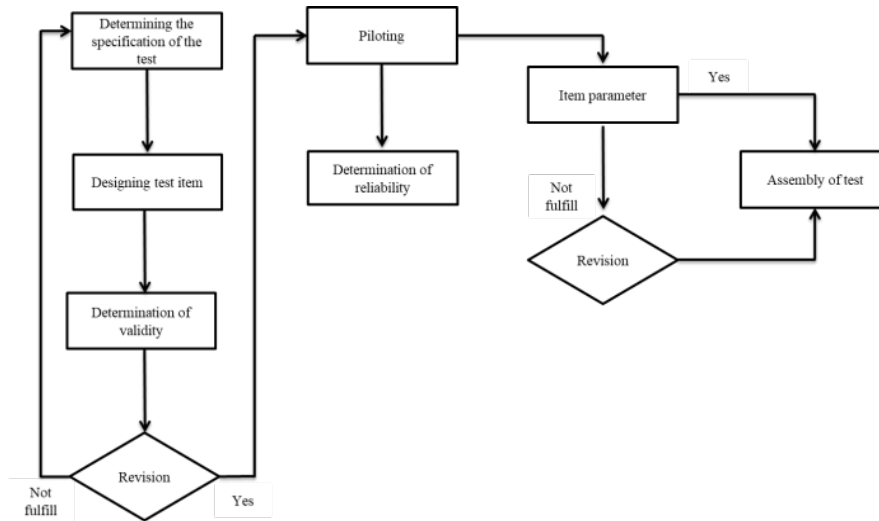
Picture 1. Research procedure scheme

administrations in the school, the teacher also needed time to prepare for learning activity. Therefore, it is necessary to develop a diagnostic test instrument for the picture representation ability of high school students in physics that meets valid and reliable criteria.

2. RESEARCH METHOD

This research is part of developmental research of diagnostic test instruments in high school physics learning. The aim was to produce a diagnostic test instrument that is suitable for use. The stages of the study as shown in Figure 1, refer to the stages of making the test instruments that have been made by Supahar and Prasetyo [9].

Diagnostic test of picture representation ability was developed based on aspects and indicators of picture representation ability. The ability of picture representation is divided into four aspects, namely (1) collecting data in the form of picture; (2) interpreting picture; (3) connecting variables; and (4) provide conclusions.

Analysis of the validity and reliability of the instrument was determined based on qualitative analysis and quantitative analysis. Data on qualitative analysis obtained by giving sheet study of diagnostic test instruments to one assessment expert lecturer, one material expert lecturer and three practitioners or high school physics teachers. Furthermore, readability test, by giving the diagnostic test instrument to 30 students.

Quantitative analysis was done by analyzing empirical data. Empirical data obtained through trials on 256 students of class XI MIPA SMAN 5 Yogyakarta and SMAN 10 Yogyakarta with the criteria that students have studied Newton's law material in the previous year. Quantitative analysis includes determining empirical validity and alpha coefficients.

Data from the study by experts and practitioners were then tabulated, then analyzed. The validity standard of an instrument using Aiken V is only influenced by the number of rater and rating scale used [10]. This research used 7 raters with 4 rating categories. Based on the standards set by Aiken the minimum Aiken V for this study is 0.76 with a probability of 0.45. Content validity using Aiken's coefficient V [11] [12] is obtained through the formula:

$$V = \frac{\sum s}{[n(c-1)]} \tag{1}$$

Description: s = r – lo; n = number of panels of assessors; c = highest validity assessment; lo = lowest validity assessment; r = the number given by an assessor.

Empirical content validity was done by analyzing the test results of the diagnostic test instrument to 256 respondents. The instrument consisted of 36 items with 5 alternative choices. Empirical validity can be determined using Classical Test Theory (CTT) or Item Response Theory (IRT) [13]. The Rasch model is one part of IRT, with one parameter characteristic namely item difficulty level (b_j). test designs generally have difficulty levels from -2 to +2 [14]. The Rasch model is expressed by equations [15] [16]:

$$P_i(0) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}} \tag{2}$$

Note: $P_i(0)$ is the probability of a test taker who is able to answer point I correctly; θ is a ability parameter; b_i is the parameter of item difficulty level; D is a parameter value of 1.7.

Empirical content validity was done by analyzing the trial result data using the QUEST program. The aim is to see the suitability of items according to the Rasch model. The item is declared valid if INFIT MNSQ is in the range of 0.77 to 1.30 [17] [16], or the value of INFIT t is ± 2.0 with a probability of 0.5 [16]. Thus, an item becomes incompatible with the Rasch model if the

INFIT value is t < - 2.0 or > + 2.0, with a probability of 0.5.

One way to determine reliability is by determining Cronbach's alpha coefficient. Diagnostic test instrument developed by dichotomous scaling. Reliability estimation using Cronbach's alpha coefficient [18] is determined by the following formula:

$$\rho_{XX'} \geq \alpha = \left[\frac{k}{k-1}\right] \frac{\sigma_y^2 - \sum \sigma_X^2 g}{\sigma_y^2} \qquad (3)$$

Where k = the number of items; σ_y^2 is the total score variance; and $\sum \sigma_X^2 g$ is the number of item score variances. The instrument is said to be reliable if the previous research stage has an alpha coefficient of 0.7; basic research of 0.8; and medical research of 0.95 [19].

In addition to using the Cronbach alpha coefficient, the reliability testing of each test item using IIC and the reliability of the test instrument were obtained from the TIC (Total Info Curve) from the results of data processing using Parscale for Windows software [20]. ICC and TIC charts have a logic scale range of -3 to +3. According to [21] the higher the value of the test information function, the more reliable the test.

Table 1. Category of picture representation ability

| Range of Ability	Category
$+1.5\sigma < \theta$	Very high
$+0.5\sigma < \theta \leq +1.5\sigma$	High
$-0.5\sigma < \theta \leq +0.5\sigma$	Moderate
$-1.5\sigma < \theta \leq -0.5\sigma$	Low
$\theta \leq -1.5\sigma$	Very low

3. RESULT AND ANALYSIS

The results of the analysis using Aiken V to 36 items of the diagnostic test of picture representation ability can be seen in figure 2.

The result of the analysis shown that the Aiken V coefficient of the test is in the range 0.76 to 0.95. Based on the standards set by Aiken [10] for 7 raters with 4

rating categories, the minimum aiken value was 0.76 with a probability of 0.45. Item with 7 raters using 4 rating categories said to be valid if the Aiken coefficient V ≥ 0.76 [23]. Therefore, all items are declared valid and fit to be used for further research.

Question items that have been valid, then carried out a trial testing. The items were divided into two sets namely set A and set B with 20% anchor items (4 items). So that there were the same 4 items that got two sets of the test. The trial limitation was conducted on 30 students. The purpose of the readability test was to avoid the use of language and images that are difficult for students to understand. The instrument was subsequently revised.

The second trial was conducted on 256 students of class XI MIPA of SMAN 5 Yogyakarta and SMAN 10 Yogyakarta who had studied Newton's law material in the previous year. Student responses to subsequent tests were analyzed using IRT, assisted by QUEST and PARSCALE software.
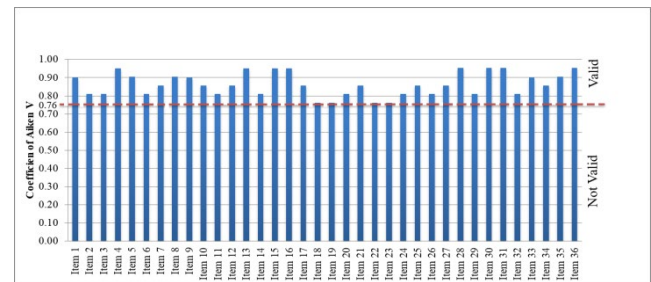
Figure 2. Analysis result of diagnostic test picture representation using V Aiken

The Goodness of Fit test aims to determine the suitability of the test with the Rasch model. Analysis using the QUEST program for 36 items. The MNSQ INFIT value for each item was shown in figure 3. The items with the highest MNSQ INFIT value were items 9 and 26. That is, students with the ability close to the difficulty level of an item have not responded consistently according to the model. Items with the highest MNSQ OUTFIT score were item 8 (1.24)

Table 2. Result of item and testi estimation diagnostic test of picture representation ability

No	Aspect	Item Estimate	Testi Estimate
1	Mean & standard deviations	0.00 ± 0.51	-0.96 ± 0.40
2	Adjusted mean & standard deviation	0.00 ± 0.49	-0.96 ± 0.11
3	Reliability	0.92	
4	Mean & standard deviation of INFIT MNSQ	0.08 ± 0.61	1.00 ± 0.09
5	Mean & standard deviation of OUTFIT MNSQ	-0.05 ± 0.95	0.99 ± 0.14
6	Mean & standard deviation of INFIT t	0.2 ± 1.40	0.80 ± 0.61
7	Mean & standard deviation of OUTFIT t	-0.10 ± 0.90	0.04 ± 0.54

```
--------------------------------------------------------------------------------
Item Fit                                                              6/ 7/20 15:26
all on all (N = 256 L = 36 Probability Level= .50)
--------------------------------------------------------------------------------
INFIT
MNSQ        .56      .63      .71      .83     1.00     1.20     1.40     1.60     1.80
-------------+--------+--------+--------+--------+--------+--------+--------+--------+--------
   1 item 1                            .               |*       .
   2 item 2                            .               | *      .
   3 item 3                            .               | *      .
   4 item 4                            .             * |        .
   5 item 5                            .            *  |        .
   6 item 6                            .               *        .
   7 item 7                            .               |    *   .
   8 item 8                            .               |   *    .
   9 item 9                            .               |    *   .
  10 item 10                           .              *|        .
  11 item 11                           .               |*       .
  12 item 12                           .               *        .
  13 item 13                           .             * |        .
  14 item 14                           .            *  |        .
  15 item 15                           .              *|        .
  16 item 16                           .            * |        .
  17 item 17                           .               | *      .
  18 item 18                           .            * |        .
  19 item 19                        *   .              |        .
  20 item 20                           .               |*       .
  21 item 21                           .            * |        .
  22 item 22                           .               | *      .
  23 item 23                           .               |  *     .
  24 item 24                           .              *|        .
  25 item 25                      *     .              |        .
  26 item 26                           .               |  *     .
  27 item 27                           .            * |        .
  28 item 28                           .            * |        .
  29 item 29                           .               |*       .
  30 item 30                           .               |*       .
  31 item 31                           .          *   |        .
  32 item 32                           .          *   |        .
  33 item 33                           .             * |        .
  34 item 34                           .          *   |        .
  35 item 35                           .             * |        .
  36 item 36                           .             * |        .
================================================================================
```
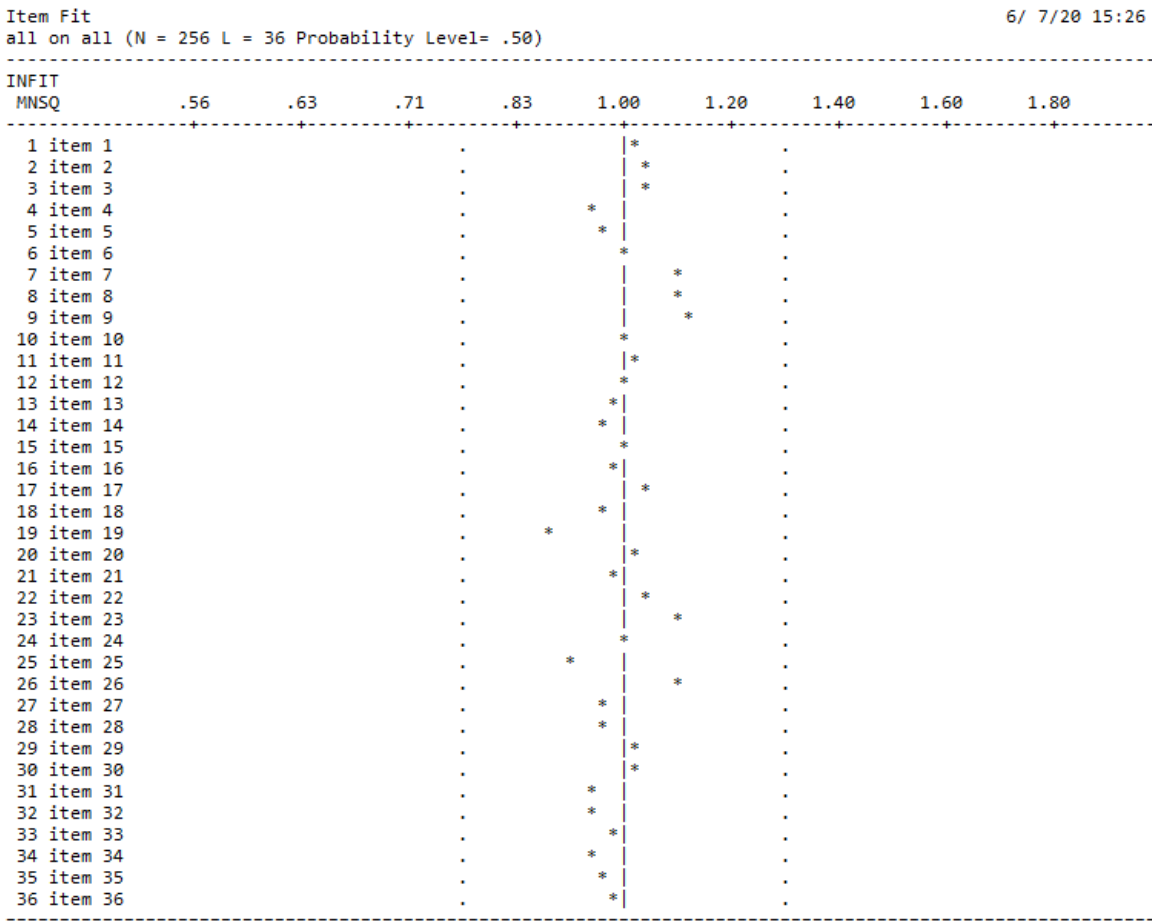
Figure 3. Map distribution of diagnostic test item of picture representation ability fit with the Rasch model

meaning that students with picture representation ability did not answer the item correctly. High MNSQ OUTFIT scores on items with low levels of difficulty indicated that students with high ability, responded to items incorrectly [24].

MNSQ INFIT value as a whole item was in the range of 0.77 to 1.30. Therefore, all test items were valid. This is consistent with the results of research by Bashooir and Supahar (2018), items that have an INFIT MNSQ value between 0.77 to 1.30 are declared valid. In addition, the INFIT value of item t is 0.2 with a standard deviation of 1.40. This shows the value of INFIT t is in the range of -2 to +2. An item becomes

incompatible with the Rasch model if the INFIt value is t < - 2.0 or > +2.0, with a probability of 0.5 [16]. Thus, the diagnostic test instrument developed for picture representation ability is valid.

The model used in this research was the Rasch Model, so the items parameters analyzed were only the level of difficulty. Item difficulty level is an opportunity to answer correctly on test items and at a certain ability level [25]. The level of difficulty for each item of the diagnostic test for picture representation ability can be seen in Figure 4. The item with the highest difficulty level was item 32 (1.08). This means that the minimum ability needed for students to answer the item correctly
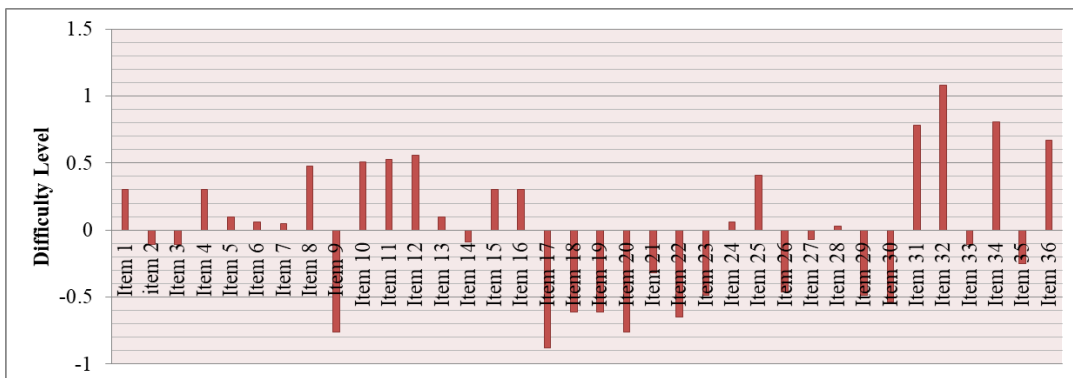

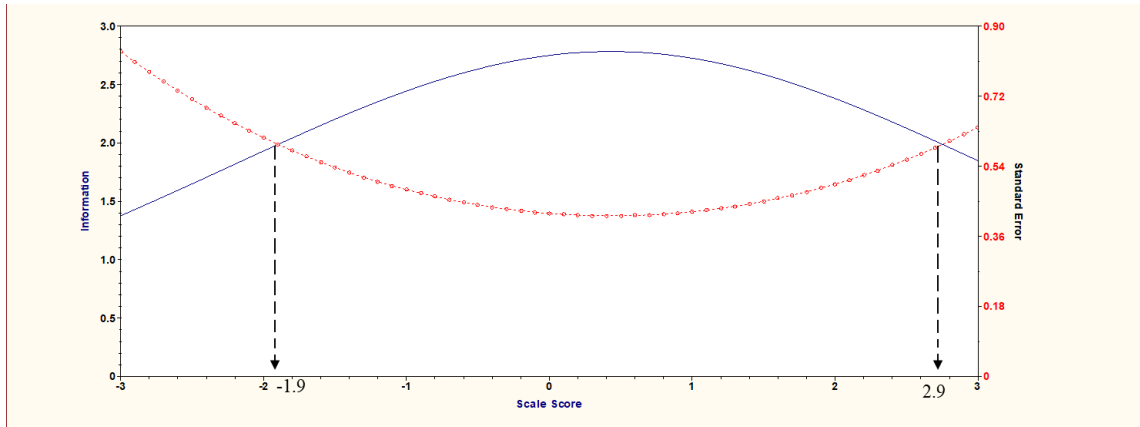**Figure 4.** Difficulty level each item of the diagnostic test of picture representation ability
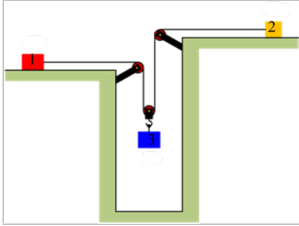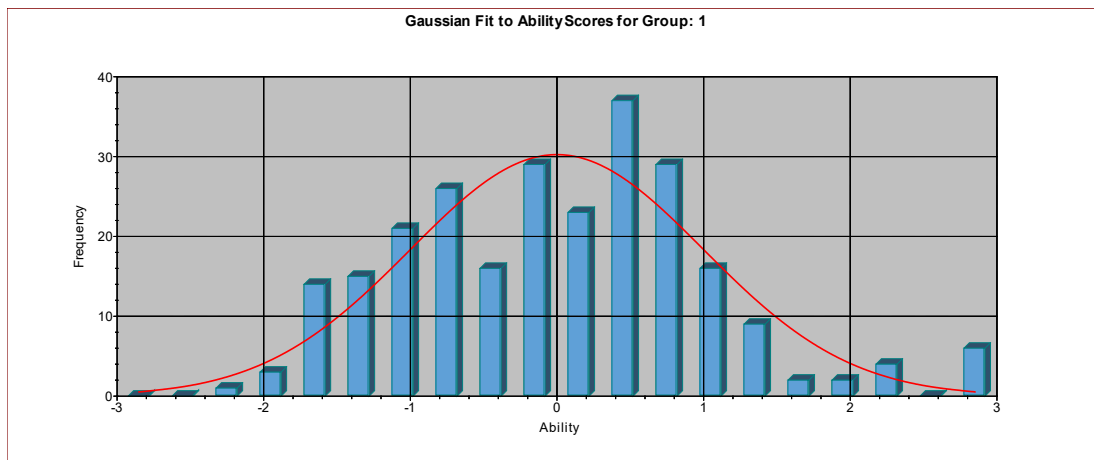
Figure 5. Total of information curve and SEM

Figure 6. Distribution of students' picture representation ability

was 1.08. The item with the lowest difficulty level was item 17 (-0.88). This showed that students' minimum ability to be able to answer correctly on these items was -0.88. On average, the item had a good difficulty level. Item difficulty level was in the range of -2 to +2 [26]. The difficulty level of the item is getting closer to zero, the better the item is [27].

MNSQ INFIT value as a whole item was in the range of 0.77 to 1.30. Therefore, all test items were valid. This is consistent with the results of research by Bashooir and Supahar (2018), items that have an INFIT MNSQ value between 0.77 to 1.30 are declared valid. In addition, the INFIT value of item t is 0.2 with a standard deviation of 1.40. This shows the value of INFIT t is in the range of -2 to +2. An item becomes incompatible with the Rasch model if the INFIT value is t < - 2.0 or > +2.0, with a probability of 0.5 [16]. Thus, the diagnostic test instrument developed for picture representation ability is valid.

The model used in this research was the Rasch Model, so the items parameters analyzed were only the level of difficulty. Item difficulty level is an opportunity to answer correctly on test items and at a certain ability level [25]. The level of difficulty for each item of the

diagnostic test for picture representation ability can be seen in Figure 4. The item with the highest difficulty level was item 32 (1.08). This means that the minimum ability needed for students to answer the item correctly was 1.08. The item with the lowest difficulty level was item 17 (-0.88). This showed that students' minimum ability to be able to answer correctly on these items was -0.88. On average, the item had a good difficulty level. Item difficulty level was in the range of -2 to +2 [26]. The difficulty level of the item is getting closer to zero, the better the item is [27].

Table 2 showed the estimated item reliability of 0.92, meaning that the order of items based on the level of difficulty will remain the same if retested to the same student or students with the same picture representation ability. Respondent reliability was very low (0.08). This showed that similar/parallel instruments are less reliable when tested on the same student . This is consistent with the results of Susac, Planinic, Klemencic, & Sipus [28], who stated that instruments with low respondent reliability would be less reliable if parallel instruments were retested to the same student.

Test reliability can be known from the Total Of Information Curve (TIF) and Error Measurement (SEM)

Three objects are connected by a rope through the combined pulley system as shown on the right. If the masses of objects 1, 2 and 3 are 2 kg, 3 kg, and 4 kg, respectively. All three pulleys are in a slippery state, and the mass of the rope and pulley is negligible. The amount of rope tension experienced by the system is ...

A.	8,9 N	C.	15 N
B.	10,9 N	D.	17,1 N
		E.	24 N

Answer option	Diagnostic	Advice
If you choose A is wrong	You have not been able to describe the rope tension force by a system connected with two fixed pulleys and one free pulley	You should review the application of Newton's second law concept to objects connected by free and fixed pulleys
If you choose B is wrong	You have been able to describe the rope tension force by a system connected with two fixed pulleys and one free pulley	You can continue to the next question
If you choose C is wrong	You have been able to describe the rope tension force by a system connected with two fixed pulleys and one free pulley, but misrepresented the direction of the rope tension force	You should review again how to draw a rope tension force diagram
If you choose D is right	You have been able to describe the rope tension force by the system connected with two fixed pulleys and one free pulley, but you have mistakenly determined the acceleration of the system motion	You should study again how to determine the speed of motion for a connected system with a fixed pulley and a free pulley
If you choose E is wrong	You have been able to describe the rope tension force by a system connected with two fixed pulleys and one free pulley, but have mistakenly drawn the tension force of the rope and determined the acceleration of the object on the pulley	You should review again how to draw a rope tension force diagram and how to determine the acceleration of motion for a system connected with a fixed pulley and a free pulley

Figure 7. Item test to diagnose picture representation ability

charts which are outputs of Parscale software. TIF and SEM graphs can be seen in Figure 5. Based on Figure 5, it was known that the test instrument is declared to be reliable if it is used on students with a minimum ability of -1.9 to 2.9. That is, the diagnostic test instrument of picture representation ability would be appropriate if it was tested on students with medium ability categories (-1.9) to very high ability categories (2.9).

The distribution of the picture representation ability of the trial students can be seen in Figure 6. The picture representation ability of the trial students can be categorized based on table 1. The ability categories became 5, namely very high (+2), high (+1 < +2), moderate (-1 < +1), low (-2 < -1), very low (-1). Figure 5 and 7 showed that the ability of the trial students was mostly in the moderate category. Or in other words,
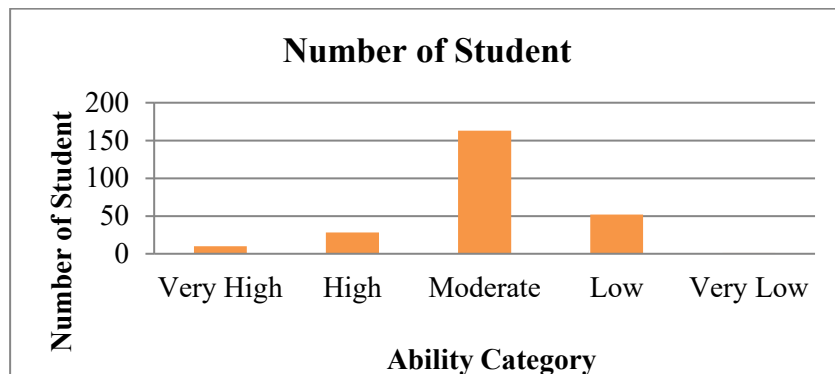

Figure 8. Level of representation ability of students

most students' abilities were in the range of -1 < +1. Figure 6 shown item of test instrument to diagnose student's picture representation ability.

4. CONCLUSION

The diagnostic test instrument for picture representation ability was proven to be valid from the content point of view using Aiken V. Valid empirically because it matches the MNSQ INFIT value of the Rasch model. Based on the TIC and SEM charts, the diagnostic test instrument of picture representation ability was reliable if applied to students with moderate to very high category ability (-1.9 to 2.9). Therefore, the diagnostic test instrument of picture representation ability developed is feasible to be used because it meets the requirements in content validity, empirical validity and reliability.

ACKNOWLEDGMENTS

REFERENCES

[1] Permendikbud Nomor 21 Tahun 2016. .

[2] M. Duskri, K. Kumaidi, and Suryanto, "Developing diagnostic test of mathematics learning difficulties in elementary schools," J. Penelit. dan Eval. Pendidik., vol. 18, no. 1, pp. 44–56, 2014.

[3] J. T. Laverty and M. D. Caballero, "Analysis of the most common concept inventories in physics: What are we assessing?," Phys. Rev. Phys. Educ. Res., vol. 14, no. 1, p. 10123, 2018.

[4] Permendikbud No 21 Tahun 2016 tentang standar isi pendidikan dasar dan menengah. .

[5] A. P. Bal, "Skills of Using and Transform Multiple Representations of the Prospective Teachers," Procedia - Soc. Behav. Sci., vol. 197, no. February, pp. 582–588, 2015.

[6] Y. Prihatni, Kumaidi, and Mundilarto, "Pengembangan Instrumen Diagnostik Kognitif pada Mata Pelajaran IPA di SMP," J. Penelit. dan Eval. Pendidik., vol. 20, no. 1, pp. 111–125, 2016.

[7] A. M. R. Tumanggor, Supahar, E. S. Ringo, and M. D. Harliadi, "Detecting Students ' Misconception in Simple Harmonic Motion Concepts Using Four-Tier Diagnostic Test Instruments," J. Ilm. Pendidik. Fis. Al-Biruni, vol. 9, no. 1, pp. 21–31, 2020.

[8] X. Liu, Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach. Iap, 2010.

[9] S. Supahar and Z. K. Prasetyo, "Pengembangan Instrumen Penilaian Kinerja Kemampuan Inkuiri Peserta Didik Pada Mata Pelajaran Fisika Sma," J. Penelit. dan Eval. Pendidik., vol. 19, no. 1, pp. 96–108, 2015.

[10] L. R. Aiken, "Three Coefficients For Analysing Reliability and Validity of Ratings," Educ. Psychol. Meas., vol. 45, pp. 131–142, 1985.

[11] K. Bashooir and Supahar, "Validitas dan reliabilitas instrumen asesmen kinerja literasi sains pelajaran Fisika berbasis STEM," J. Penelit. dan Eval. Pendidik., vol. 22, no. 2, pp. 168–181, 2018.

[12] F. Kurnia, D. Rosana, and Supahar, "Developing Instruments using CIPP Evaluation Model in the Implementation of Portfolio Assessment in Science Learning," Int. J. Environ. Sci. Educ., vol. 12, no. 8, pp. 1989–1998, 2017.

[13] B. Sumintono and W. Widhiarso, Aplikasi Pemodelan Rasch pada Assesment Pendidikan. Bandung: Trim komunikata, 2015.

[14] D. Mardapi, Pengukuran, Penilaian dan Evaluasi Pendidikan. Yogyakarta: Nuha Medika, 2012.

[15] V. Mešić et al., "Measuring students' conceptual understanding of wave optics: A Rasch modeling approach," Phys. Rev. Phys. Educ. Res., vol. 15, no. 1, pp. 1–20, 2019.

[16] M. Planinic, W. J. Boone, A. Susac, and L. Ivanjek, "Rasch analysis in physics education research: Why measurement matters," Phys. Rev. Phys. Educ. Res., vol. 15, no. 2, p. 20111, 2019.

[17] R. J. Adams and S.-T. Khoo, "Acer Quest: The interactive test analysis system," Australian Council for Educational Research. 1998.

[18] L. J. Cronbach, "Coefficient Alpha and The Internal Structure of Tests," Psychometrika, vol. 16, no. 3, pp. 297–334, 1951.

[19] S. D., "Starting at the beginning: An introduction to coefficient alpha and internal consistency," J. Pers. Assess., vol. 80, no. 1, pp. 99–103, 2003.

[20] M. du Toit, IRT from SSI: Bilog-MG, multilog, parscale, testfact. Scientific Software International, 2003.

[21] E. Istiyono, Pengembangan Instrumen: Penilaian dan Analisis Hasil Belajar Fisika. UNY Press, 2017.

[22] S. Azwar, Penyusunan Skala Psikologi. Yogyakarta: Pustaka pelajar, 2012.

[23] A. Abdulfattah and Supahar, "The development of high school physics problem solving skills test instruments based problem-based learning," J. Educ. Gift. Young Sci., vol. 7, no. 4, pp. 1037–1052, 2019.

[24] M. Planinic, L. Ivanjek, and A. Susac, "Rasch model based analysis of the Force Concept Inventory," Phys. Rev. Spec. Top. - Phys. Educ. Res., vol. 6, no. 1, pp. 1–11, 2010.

[25] Suwarto, Pengembangan Tes Diagnostik dalam Pembelajaran. Yogyakarta: Pustaka Pelajar, 2017.

[26] S. Wei, X. Liu, Z. Wang, and X. Wang, "Using Rasch Measurement To Develop a Computer Modeling-Based Instrument To Assess Students' Conceptual Understanding of Matter," J. Chem. Educ., vol. 89, no. 3, pp. 335–345, 2012.

[27] S. Wei, X. Liu, and Y. Jia, "Using Rasch Measurement to Validate the Intrument of Students' Understanding of Models in Science (SUMS)," Int. J. Sci. Math. Educ., vol. 12, no. 5, pp. 1067–1082, 2013.

[28] A. Susac, M. Planinic, D. Klemencic, and Z. Milin Sipus, "Using the Rasch model to analyze the test of understanding of vectors," Phys. Rev. Phys. Educ. Res., vol. 14, no. 2, p. 23101, 2018.

[1] Permendikbud Nomor 21 Tahun 2016. .

[2] M. Duskri, K. Kumaidi, and Suryanto, "Developing diagnostic test of mathematics learning difficulties in elementary schools," J. Penelit. dan Eval. Pendidik., vol. 18, no. 1, pp. 44–56, 2014.

[3] J. T. Laverty and M. D. Caballero, "Analysis of the most common concept inventories in physics: What are we assessing?," Phys. Rev. Phys. Educ. Res., vol. 14, no. 1, p. 10123, 2018.

[4] Permendikbud No 21 Tahun 2016 tentang standar isi pendidikan dasar dan menengah. .

[5] A. P. Bal, "Skills of Using and Transform Multiple Representations of the Prospective Teachers," Procedia - Soc. Behav. Sci., vol. 197, no. February, pp. 582–588, 2015.

[6] Y. Prihatni, Kumaidi, and Mundilarto, "Pengembangan Instrumen Diagnostik Kognitif pada Mata Pelajaran IPA di SMP," J. Penelit. dan Eval. Pendidik., vol. 20, no. 1, pp. 111–125, 2016.

[7] A. M. R. Tumanggor, Supahar, E. S. Ringo, and M. D. Harliadi, "Detecting Students ' Misconception in Simple Harmonic Motion Concepts Using Four-Tier Diagnostic Test Instruments," J. Ilm. Pendidik. Fis. Al-Biruni, vol. 9, no. 1, pp. 21–31, 2020.

[8] X. Liu, Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach. Iap, 2010.

[9] S. Supahar and Z. K. Prasetyo, "Pengembangan Instrumen Penilaian Kinerja Kemampuan Inkuiri Peserta Didik Pada Mata Pelajaran Fisika Sma," J. Penelit. dan Eval. Pendidik., vol. 19, no. 1, pp. 96–108, 2015.

[10] L. R. Aiken, "Three Coefficients For Analysing Reliability and Validity of Ratings," Educ. Psychol. Meas., vol. 45, pp. 131–142, 1985.

[11] K. Bashooir and Supahar, "Validitas dan reliabilitas instrumen asesmen kinerja literasi sains pelajaran Fisika berbasis STEM," J. Penelit. dan Eval. Pendidik., vol. 22, no. 2, pp. 168–181, 2018.

[12] F. Kurnia, D. Rosana, and Supahar, "Developing Instruments using CIPP Evaluation Model in the Implementation of Portfolio Assessment in Science Learning," Int. J. Environ. Sci. Educ., vol. 12, no. 8, pp. 1989–1998, 2017.

[13] B. Sumintono and W. Widhiarso, Aplikasi Pemodelan Rasch pada Assesment Pendidikan. Bandung: Trim komunikata, 2015.

[14] D. Mardapi, Pengukuran, Penilaian dan Evaluasi Pendidikan. Yogyakarta: Nuha Medika, 2012.

[15] V. Mešić et al., "Measuring students' conceptual understanding of wave optics: A Rasch modeling approach," Phys. Rev. Phys. Educ. Res., vol. 15, no. 1, pp. 1–20, 2019.

[16] M. Planinic, W. J. Boone, A. Susac, and L. Ivanjek, "Rasch analysis in physics education research: Why measurement matters," Phys. Rev. Phys. Educ. Res., vol. 15, no. 2, p. 20111, 2019.

[17] R. J. Adams and S.-T. Khoo, "Acer Quest: The interactive test analysis system," Australian Council for Educational Research. 1998.

[18] L. J. Cronbach, "Coefficient Alpha and The Internal Structure of Tests," Psychometrika, vol. 16, no. 3, pp. 297–334, 1951.

[19] S. D., "Starting at the beginning: An introduction to coefficient alpha and internal consistency," J. Pers. Assess., vol. 80, no. 1, pp. 99–103, 2003.

[20] M. du Toit, IRT from SSI: Bilog-MG, multilog, parscale, testfact. Scientific Software International, 2003.

[21] E. Istiyono, Pengembangan Instrumen: Penilaian dan Analisis Hasil Belajar Fisika. UNY Press, 2017.

[22] S. Azwar, Penyusunan Skala Psikologi. Yogyakarta: Pustaka pelajar, 2012.

[23] A. Abdulfattah and Supahar, "The development of high school physics problem solving skills test instruments based problem-based learning," J. Educ. Gift. Young Sci., vol. 7, no. 4, pp. 1037–1052, 2019.

[24] M. Planinic, L. Ivanjek, and A. Susac, "Rasch model based analysis of the Force Concept Inventory," Phys. Rev. Spec. Top. - Phys. Educ. Res., vol. 6, no. 1, pp. 1–11, 2010.

[25] Suwarto, Pengembangan Tes Diagnostik dalam Pembelajaran. Yogyakarta: Pustaka Pelajar, 2017.

[26] S. Wei, X. Liu, Z. Wang, and X. Wang, "Using Rasch Measurement To Develop a Computer Modeling-Based Instrument To Assess Students' Conceptual Understanding of Matter," J. Chem. Educ., vol. 89, no. 3, pp. 335–345, 2012.

[27] S. Wei, X. Liu, and Y. Jia, "Using Rasch Measurement to Validate the Intrument of Students' Understanding of Models in Science (SUMS)," Int. J. Sci. Math. Educ., vol. 12, no. 5, pp. 1067–1082, 2013.

[28] A. Susac, M. Planinic, D. Klemencic, and Z. Milin Sipus, "Using the Rasch model to analyze the test of understanding of vectors," Phys. Rev. Phys. Educ. Res., vol. 14, no. 2, p. 23101, 2018.