JIPF (JURNAL ILMU PENDIDIKAN FISIKA)
p-ISSN: 2477-5959 | e-ISSN: 2477-8451 *Vol. 7 No. 3, September 2022, Page 276-287*

Four-Tier Test Model to Measure Prospective Physics Teacher Students' Multiple Representation Ability on Electricity Topic

Rahmawati [1*)], Widiasih [2], Nasrah [3], A. Muafiah Nur [4], Dewi Hikmah Marisda [5]
Universitas Muhammadiyah Makassar, Indonesia[1,3,4,5], Universitas Terbuka, Indonesia[2],
***Corresponding email: rahmawatisyam@unismuh.ac.id**

Received: January 1st, 2022. Revised: March 11th, 2022. Accepted: March 15th, 2022

Keywords :
Multiple Representation; Item Response Theory, Rasch Model; Electricity Topic; Winstep Version 3.68.2

ABSTRACT
The advantage of four-tier model test is that it has four levels of questions that require complex reasoning abilities from students to express the reasons for their answers to the problems given. This study aimed to develop a multiple representation test with a four-tier model to measure students' multiple representation in electricity topics. The stages of developing this test instrument used a stage model of The Design Based Research which consists of five stages, namely developing an assessment framework, designing items, developing rubrics, conducting trials, and applying the Racsh Model analysis with the Item Response approach. Theory (IRT) program assisted Winsteps version 3.68.2. The research method used was descriptive-exploratory method to describe the results of the development and validation of four-tier model test. This test consisted to 20 items were developed based on four types of representation, namely verbal representations, diagrammatic representations (pictures), mathematical representations, and graphic representations with each indicator. The research subjects involved were 30 prospective physics students at the pilot test stage and 79 prospective physics students at the field test stage from different universities in Makassar city. The results of the development of the four-tier test model test overall items are valid with a high level of reliability (Cronbach's Alpha value = 0.80). Based on the results of expert judgment validation and testing, it can be concluded that the multiple representation test with four-tier model on electricity topic is feasible to use.

INTRODUCTION

Multiple representation is a fundamental aspect of scientific knowledge and reasoning [1]. This multi-representation can be used as a tool for thinking, making speculations, finding possible explanations of a problem, and re-checking the explanation of the results obtained [2]. Multi-representation is also very useful because it can act as a visual aid and foster a better understanding of students' physics

problems [3]. Based on the description, it can be concluded that multiple representations are needed in helping students understand physics concepts that tend to be abstract.

Various kinds of difficulties faced by students in solving physics problems are mostly caused by the lack of familiarity with the use of multiple representations in learning and application in forms of assessment [4]. Previous studies related to students' difficulties in understanding and solving physics problems in Basic Physics course, especially related to electrical material, showed that the form of questions tested in both the mid-semester and final semester exams only contained aspects of mathematical representation and some verbal aspects. Meanwhile, the representation of diagrams and graphs is still very limited. Furthermore, the form of assessment used is in the form of an open description that has not been specifically designed to be able to dig up information on a number of forms of representation [5].

Multi-representation ability can be measured using assessment techniques [6]. Several relevant studies have used various types of test to measure the multi-representation of students, for example using multiple choice test techniques [7] [8] [9] [10], *two tier test* [11] [12] and *three-tier test* [6] [13] [14] [15] [16]. Several weaknesses found in each of the test instrument models became the fundamental basis for the development of a four-tier test model.

Electrical material in the Basic Physics course is one of the dominant materials that is abstract. This abstract nature provides opportunities for students to experience difficulties in understanding various concepts. There are several research results that confirm the nature of abstraction and the level of difficulty of electrical material to be understood by students. The abstract nature of electrical material makes this material difficult to understand starting at the elementary level [17] [18] [19] 20] [21], intermediate level [22] [23] [24] [25] [26] [27] [28], high education [29] [30] [31], until teacher level [22] [32] [33].

Difficulties in understanding concepts have the potential to cause misconceptions among students. To help students' difficulties in understanding concepts, it is necessary to construct an assessment tool that represents several representations (verbal, diagrams, mathematical, and graphs) so that they are able to assist students in constructing their thoughts on the concept. One of the assessment tools that can be used is a reasoned objective test. Therefore, this study focuses on measuring students' multi-representation abilities in the electrical material of the four-tier test instrument.

Obtaining a test instrument that is valid, reliable and effective in its use, it is necessary to have an analytical model for the testing process. There are two types of theories that can be used in analyzing test, namely classical theory and modern theory or known by other terms Item Response Theory (IRT) or item response theory. The classical theory has a number of fundamental weaknesses, namely: (1) the classical test theory model uses several statistics such as the level of difficulty and discriminating power of items depending on the respondents tested in the analysis; (2) classical test theory is more oriented to the test results obtained than to the items of the test instrument itself; (3) the concept of test reliability in the context of classical theory based on the parallelism of test sets is very difficult to fulfill. In practice, it is very difficult to get two sets of tests that are completely parallel; (4) classical test theory does not provide a basis for determining how the test takers respond when given certain items; (5) the standard error index of measurement is assumed to be the same for each test taker; (6) the procedure for testing item bias and test equivalence is not practical and difficult to do. The same is true for vertical equivalence [34] [35]. For this reason, psychometricians offer an alternative measurement theory and model called item response theory (IRT).

The item response theory (IRT) model shows the relationship between the ability or trait measured by the instrument and an item response [36] [37]. There are three assumptions underlying IRT, namely unidimensionality, local independence, and parameter invariance. Unidimensional means that each test item only measures one abilitie. In practice, the unidimensional assumption cannot be met strictly because of the external factors that are not constant. Local independence is a condition in which external factors affecting performance are constant which causes the subject's response to any item to

be statistically independent of each other. The assumption of local independence will be fulfilled if the participant's answer to one item does not affect the participant's answer to the other items. Parameter invariance means that the characteristics of the items do not depend on the distribution of the test taker's ability parameters and the parameters that characterize the test takers do not depend on the characteristics of the items [35] [36] [37].

A popular model in the use of item response theory (IRT) is known as the logistic model. There are three types of logistics models, namely one-parameter logistics model, two-parameter logistics model, and three-parameter logistics model. The one-parameter logistics model is one of the most widely used IRT models. This model is also known as the Rasch Model. Based on the advantages of the IRT theory with the Rasch model, it was decided in this study to use the Rasch model as a model approach assisted by the Winsteps software used in analyzing the feasibility of the developed test instrument. As for the formulation of the research problem, how is the development and eligibility of four-tier test to measure prospective physics teacher students' multi-representation ability on electrical topic?

METHOD

Research Design

This research is a type of research and development using design-based research (The Design Based Research) which was adapted from Kuo, Wu, Jen, & Hsu [38]. The research design includes five steps, namely (1) developing an assessment framework; (2) designing items; (3) developing a scoring rubric; (4) conducting trials; and (5) applying the Rasch Model analysis. The systematic steps of developing this test instrument can be seen in Figure 1.

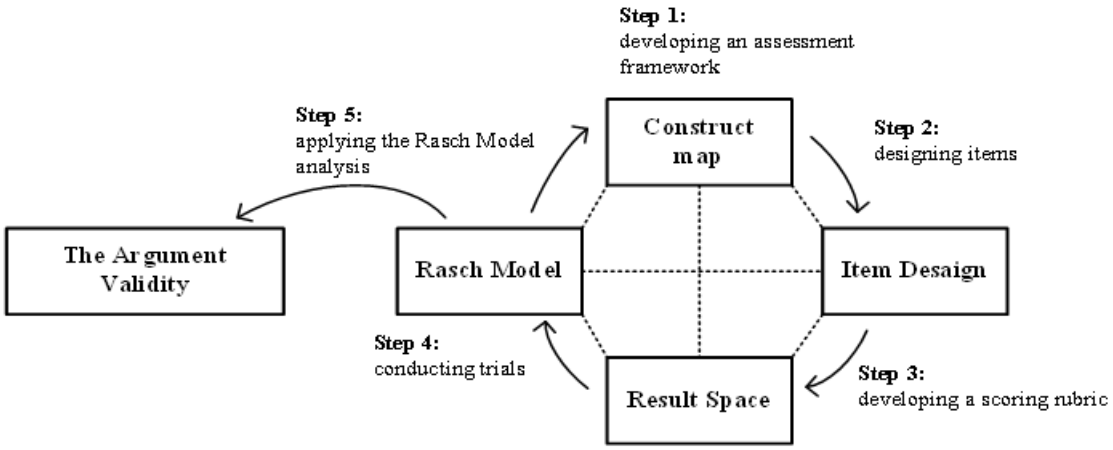

Fig 1. *Development Design of Four Tier Model Multiple Representation Test*

Participants

This study involved student physics teacher candidates from two different universities in the city of Makassar, namely institution A and institution B. The number of participants involved in this study was different at each stage of the study. The summary of the participants in this study is shown in Table 1.

Table 1. Participants in the Instrument Trial Stage

Stage	Total	Institution
Pilot Test	30	A
Field Test	79	B

Data Collection Techniques, Data Analysis Techniques, and Research Instruments
Data collection in this study was carried out by several techniques. The description of the instruments and techniques of data analysis as well as the division of labor in the research can be shown in Table 2.

Table 2. Data Collection Techniques and Procedures

Measured Aspect	Data Resources	Data collection techniques	Data Analysis technique	Instrument
Four tier test design	Researcher	Document analysis	Qualitative and quantitative description	Basic physics course material
Validation of instrument assessment developed	Expert Judgement	Validation questionnaire	CVR and I-CVI agreement index analysis	Validation questionnaire, four tier test
Pilot test	Students	Dissemination of test	Rasch Model analysis with PCM (Partial Credit Model) type	Four tier test
Field test	Students	Administration of test	Rasch Model analysis with PCM (Partial Credit Model) type	Four tier test

The feasibility analysis of the resulting test instrument product is carried out based on agreement between expert judgments by determining the value of the agreement coefficient using the CVR and I-CVI equations formulated by Lawshe [39]. CVR content validity analysis uses the formulation:

$$\mathrm{CVR} = \frac{\mathrm{N_e} - \frac{\mathrm{N}}{2}}{\frac{\mathrm{N}}{2}} \quad (1)$$

Note:
CVR = Content Validity Ratio
N_e = Number of experts who declared relevant
N = Number of expert judgments

Furthermore, the I-CVI analysis is calculated using the formula:

$$\mathrm{I-CVI} = \frac{\mathrm{N_e}}{\mathrm{N}} \quad (2)$$

Note:
I- CVI = *Item Content Validity Index*
Ne = Number of experts who declared relevant
N = Number of expert judgments

An item is said to be feasible if the CVR coefficient is in the range 0-1. However, the determination of whether the items are accepted or rejected is done by comparing the calculated CVR value with the critical value of CVR [40]. The critical value of CVR depends on the number of reviewers. This study uses as many as five reviewers so that the critical CVR coefficient value is 0.99. Furthermore, the value of the item content validity index (I-CVI) is interpreted with a number of categories. The item content validity index is in the range of 1-0 with the category level divided into three categories as shown in Table 3.

Table 3. Item Content Validity Index Category

Interval Index*I-CVI*	Category
$I\text{-}CVI_{account} \geq 0.79$	Relevant
$0 \leq I\text{-}CVI_{account} < 0.79$	Revision
$I\text{-}CVI_{account} < 0$	Elimination

The test instrument that has been tested for feasibility is then tested on several samples. The data from the field trials were analyzed using the Rasch PCM (Partial Credit Model) model with the help of the Winstep version 3. 68 program. 2. There are several criteria that need to be observed in determining the quality of the test instrument through statistical summary analysis, including the following:

a. Comparing the value of a person measure with an item measure (in logic) (the value of an item measure is always 0.0 logic). If the logic person measure value is higher than the item measure, this indicates that the respondent's ability (ability) tendency is higher than the item's difficulty level.
b. Cronbach's Alpha coefficient value data. This value measures the reliability of the instrument, namely the interaction between the person and the item as a whole. Furthermore, Cronbach's Alpha coefficient values are categorized based on the range of coefficient values. Cronbach's Alpha reliability level category [41] is shown in Table 4.

Table 4. Cronbach's Alpha Reliability Categorization

Coefficient reliability of Cronbach's Alpha	Category
$0,8 \leq \alpha$	Very good
$0,7 \leq \alpha < 0,8$	Good
$0,6 \leq \alpha < 0,7$	Middle good
$0,5 \leq \alpha < 0,6$	Poor
$\alpha < 0,5$	Very bad

c. Data value of person and item reliability. The level of quality of person reliability and item reliability can be divided into several categories based on the reliability coefficient value. The categorization of the level of reliability of persons and items is presented in Table 5. showing the categorization of the level of reliability of persons and items.

Table 5. Categorization of Person and Item Reliability Level

Person and Item Reliability Coefficient Value Interval (*r*)	Category
$0,94 \leq r$	Excellent
$0,91 \leq r < 0,94$	Very good
$0,80 \leq r < 0,91$	Good
$0,67 \leq r < 0,80$	Midle good
$r < 0,67$	Weaknes

d. Data person and item separation. Person and item separation aims to group people (respondents) based on their level of ability to items. Meanwhile, item separation is used to verify the item hierarchy. The greater the value of separation, the quality of the instrument in terms of overall respondents and items is better because it can identify groups of respondents and groups of items [42] [43].

Data on the value of person fit and item fit in the infitmnsq and outfit mnsq columns, as well as the value of infitZstd and outfit Zstd which follow the ideal value of the Rasch model (ie 1.00). Meanwhile, the standard Z value (Zstd) on infitZstd and outfit Zstd, both on person fit and item fit refers to the ideal value of 0.0. To see the level of suitability of the items, there are several criteria that must be met, namely 1) the value of the outfit mean square (mnsq) received: 0.5 <mnsq< 1.5; 2) the value of outfit Z-standard (Zstd) received: -2.0 <zstd< +2.0; and 3) point measure correlation value (Pt

mean corr): 0.4 <Pt measure corr. < 0.85 [42] [43] [44].

RESULTS AND DISCUSSIONS

Developing Assessment Framework

The assessment framework was developed as a reference for the next steps. The development of this assessment framework refers to the theoretical analysis related to the learning taxonomy by Revised Bloom's Taxonomy [45] and Marzano's Taxonomy [46].

There are three criteria used to build the right framework for this test model. First, the learning taxonomy covers the cognitive domains of behavior and knowledge in one model. Second, this taxonomy of learning clearly distinguishes between thought processes and knowledge. Third, the learning taxonomy can predict behavior related to essential concepts in Electrical material. The empirical analysis was carried out on several aspects related to the analysis of students' initial concepts on the Dynamic Electricity material in the Basic Physics course using a two-tier test instrument.

Designing Items

The item design follows the assessment framework that has been prepared previously. Items are developed in a paper-and-pen assessment format in the form of an objective four-tier model.Several aspects are taken into consideration in designing a four-tier model multi-representation instrument item. First, the context aspect, the context aspect considers the criteria that the items developed must (a) be in accordance with the real life of undergraduate students (aged 18-25 years); (b) the context is authentic; (c) includes content that should be mastered by prospective physics teacher students; (d) in accordance with the competencies formulated. Second, the sensitivity aspect, namely the items developed must (a) be used nationally, free from the cultural context and knowledge of certain cultural groups; (b) not gender biased. Third, the technical aspects include (a) the assessment can be used in the classroom both online and offline; (b) easy scoring and interpretation of the results; and (c) the test can be answered within a maximum of 90 minutes so as not to cause boredom to the tester which can result in bias in the test results.

The items of this test instrument are 20 items spread over a number of materials and electrical concepts. The distribution of the material, the concept of electricity, and the number of items developed are presented in Table 6.

Table 6. *Blue Print* of the Test on Electricity Topic

Topics	Sub Topics	Item number
Electricity and Resistance	Electric current	1, 2
	Resistance	3, 4
	Resistivity	5, 6
	Ohm's Law	7
	Electric motion voltage	8, 9
	Energy and Electric current	10, 11, 12
Direct current	Resistors in series, parallel, and mixed circuits	13, 14, 15
	Kirchhoff's Law	16, 17, 18
	RC circuit capacitor charging and discharging	19, 20

Developing Scoring Rubric

The development of the scoring rubric is related to the construct modeling approach, namely item design and result space (Figure 1). The results space consists of a set of different qualitative categories for identifying, evaluating, and scoring student answers [47]. The development of the scoring rubric refers to the scoring model scheme developed by Gurcay & Gulbas [6] adapted according to the four-tier multi-representation test instrument model. Students' multi-representation abilities are divided into three categories, namely understanding, not understanding, and misconceptions. The categorization of

multi-representation abilities based on the pattern of interpretation of answers is shown in Table 7.

Table 7. Categorization of Understanding Levels based on Interpretation of Answer Patterns

Answer	Confidence level of answer	Reason	Confidence level of reason	Criteria
Correct	High	Correct	High	Understand
Correct	High	Correct	Low	Not understand
Correct	Low	Correct	High	
Correct	Low	Correct	Low	
Correct	High	Wrong	Low	
Correct	Low	Wrong	Low	
Wrong	Low	Correct	High	
Wrong	Low	Correct	Low	
Wrong	Low	Wrong	Low	
Correct	High	Wrong	High	
Correct	Low	Wrong	High	Misconception
Wrong	High	Correct	High	
Wrong	High	Correct	Low	
Wrong	High	Wrong	Low	
Wrong	High	Wrong	High	
Wrong	Low	Wrong	High	

Pilot Testing Instrument

The assessments that have been developed include frameworks, items, and scoring rubrics which are validated by experts first as a validation process for expert judgment. The expert judgment process aims to see the quality of the assessment developed including the quality of the items (simple/uncomplicated language, and clear), the suitability of the content of the construct being measured, and the alignment between the items developed and the construct [48] [49] [50] [51]. The expert validation process was given to five experts each in the fields of assessment, learning, and physicists.

The score given for each aspect assessed is 0 and 1. If the item is in accordance with the aspect being assessed, then it is given a score of 1 by putting a check mark (√) in the column provided and giving a score of 0 if it does not match the aspect of the assessment by placing a mark times (X) in the column provided on the judgment sheet. The results of expert judgment were analyzed using the CVR and I-CVI equations. The results of the calculation analysis show that a CVR value of 0.99 is accepted for the number of SME (Subject Matter Expert) as many as 5 expert judgments based on the provisions of the allowed critical CVR value. Meanwhile, the I-CVI coefficient value was obtained at 0.99 with the appropriate category. From the results of the validation of the contents of the CVR and I-CVI, it can be concluded that the four-tier test model multi-representation ability test instrument has appropriate content validity for all items.

The next step is to test the test instrument. The trial process is carried out through the pilot testing and field testing stages. The pilot testing stage involved 30 prospective physics teacher students from one of the public universities in the city of Makassar. Sampling for pilot test needs does not have to go through strict procedures [38]. Linacre explained that the use of a sample with a range of 16-36 respondents for pilot test purposes was feasible to obtain stable estimation results with a range of ±1 logic and a 95% confidence level [51]. The criteria for at least 50% of the answers to all items tested have been met by 30 respondents.

The results of the pilot test analysis showed that the test instrument developed was feasible to use.

There are only two questions (S10 and S19) that require a little revision in terms of language in the questions. Furthermore, the trial was continued with a field test involving 79 prospective physics teacher students from different universities from the pilot test sample in the city of Makassar. The following is a description of each stage of data analysis of field test results from the four-tier test instrument developed with the application of the Rasch Model.

Applying the Rasch Model Analysis

The Rasch model analysis was applied to the data obtained from the test results. All Rasch analyzes were performed using Winsteps software *version 3.68.2* [52]. Because the item answer score model is in the form of a polytomy and also the maximum score between items is not the same, the Rasch analysis used is PCM (Partial Credit Model).

Rasch Modeling Analysis of the Reliability and Separation of Items and Persons

Analysis of reliability level, item separation and test person were obtained from the output data of WinstepsMinistep program version 3.68.2. Analysis of test reliability was reviewed on three aspects, namely the reliability value of Alpha Cronbach (KR-20), the value of person reliability, and the value of item reliability. For the observation of the separation variable, it is possible to observe the separation of items and persons. The results of the analysis of several aspects of reliability and separation observations are shown in Table 8.

Table 8. Summary of Analysis on Cronbach's Alpha, Person and Item Reliability, and Person and Item Separation

Statistic	Statistic aspect	Value
Reliability	Cronbach’s Alpha	0,80
	Person reliability	0,76
	Item reliability	0,88
Separation	Person separation	1,77
	Item separation	3,6

Table 8 shows that the reliability value of Cronbach's Alpha (KR-20) is 0,80. The reliability value of Alpha Cronbach (KR-20) indicates that this four-tier test instrument has internal consistency reliability in a good category [53]. Bond & Fox confirmed that the Cronbach's Alpha coefficient obtained through the Rasch analysis approach is in the range of 0,70 to 0,99 which is the allowable value with the best acceptance category [53].

The results of Rasch's analysis on person reliability and person separation are 0,76 and 1,77 respectively [41]. The person reliability value obtained is in the fairly good category which indicates that the responses from the respondents are quite good and consistent [53]. For the aspect of person separation, the coefficient value obtained is 1,50. Krishnan & Idris stipulated that the person separation value must be greater than 1,00 to ensure that the respondents being measured are spread throughout. Person separation of 1,50 (< 3.0) is included in the acceptable category although this value indicates the test instrument is less sensitive to distinguish between high-skilled and low-skilled persons [42].

The value of item reliability and item separation obtained from the results of the analysis respectively were 0,88 and 3,6 (> 3,0). The value of this reliability item is in the good category [41]. The item separation coefficient value obtained is in the good category.Linacre (2002)confirmed that the item separation value greater than 2,00 is interpreted as good. This implies that the person sample is sufficient to confirm the item difficulty hierarchy [42] [54].

Rasch modeling analysis on item fit

Determination of item fit is based on three criteria, namely the outfit means-square (MNSQ), the outfit z-standard (ZSTD), and the point measure correlation (PT-MEASURE CORR). If one of these three criteria is not met, it can be ascertained that the item is not good enough so that it needs to be revised

or discarded [41] [55] [56].

The results of the analysis show that item number 10 (S10) has a tendency not to fit because it does not meet the requirements for Outfit Zstd (-2,1), but meets the criteria for Outfit Mnsq and Pt. measure corr. S10 item is still within the allowed limit so that S10 item can be maintained. There are 10 items (S1, S2, S3, S4, S6, S7, S12, S15, S19, and S20) that do not meet the Pt criteria. Measure corr. but the other two criteria are met. This shows that S10item is still within the allowable limits so they do not need to be omitted. Meanwhile, nine items (S5, S8, S9, S11, S13, S14, S16, S17, and S18) have met the three criteria, so they can be accepted well. Thus, it can be concluded that there are no items that need to be changed or discarded.

Rasch Modeling Analysis on person fit
Information that can be used to observe items that do not fit the model (misfit) are: 1) the value of the outfit mean square (mnsq) received: 0.5 <mnsq< 1.5; 2) the value of outfit Z-standard (Zstd) received: -2,0 <zstd< +2,0; and 3) point measure correlation value (Pt mean corr): 0,4 <Pt measure corr< 0,85 [42] [43] [44]. By using the three criteria for observing person fit, there are several respondents (persons) who experience misfits and are summarized in Table 9.

Table 9. *Misfit* Order ofRespondens (Person) in test

Person	Total score	Measure	Outfit MNSQ (0,5 – 1,5)	Outfit ZSTD (-2,0 – 2,0)	PT-Measure Corr. (0,4 – 0,85)
R01	30	0,00	1,51	2,2	0,18
R17	26	-0,16	1,53	2,1	0,28
R37	33	0,12	1,61	2,5	0,16
R41	26	-0,16	1,61	2,4	0,32

Table 9 shows the respondents whose responses to items were misfit based on the Rasch modeling analysis. In other words, the respondent (person) gives an answer that is not in accordance with his ability compared to the ideal model. All persons (responses) are given the initial initial code "R". Based on Table 9, there are four respondents (R01, R17, R37, and R41) who do not meet the three criteria for determining the suitability of items that do not fit (misfit), namely outside the value limits of MNSQ outfit, ZSTD outfit, and PT-Measure Corr. Meanwhile, other respondents have a pattern of answers with the value of the three criteria meeting the requirements. Based on the misfit order person fit analysis on the criteria for the MNSQ outfit value, the ZSTD outfit, and the PT-Measure Corr, it can be concluded that there are seventeen of the seventy-nine respondents (persons) who experience an unusual response pattern.

CONCLUSION AND SUGGESTION

Based on the results of the development and feasibility test process through the validation stage, pilot test, and field test, it can be concluded that the four-tier test model has met the requirements of content validity (expert judgment), construct validity (empirical validity), reliability test, and test the level of suitability of items through the analysis of the Rasch model with Item Response Theory (IRT) approach. Thus, this test consists of 20 questions and their scoring rubric was declared suitable to be used to measure the multi-representation ability of prospective physics teacher students on the topic of electricity.

ACKNOWLEDGMENTS

This study is supported by Department of Physics Education, Faculty of Teacher Training and Education, Muhammadiyah University of Makassar. We thank to the research institute LP3M Muhammadiyah University of Makassar which has provided research funding assistance in the PUPT research grant scheme. We also are grateful to the participants who have been contributed in this study.

REFERENCES

[1] Hubber, P., & Tytler, R. (2017). Enacting a representation construction approach to teaching and learning astronomy. In *Multiple representations in physics education* (pp. 139-161). Springer, Cham.
[2] Waldrip, B., Prain, V., & Carolan, J. (2010). Using multi-modal representations to improve learning in junior secondary science. *Research in science education*, *40*(1): 65-80.
[3] Nieminen, P., Savinainen, A., & Viiri, J. (2017). Learning about forces using multiple representations. In *Multiple Representations in Physics Education* (pp. 163-182). Springer, Cham.
[4] Ornek, F., Robinson, W. R., & Haugan, M. P. (2008). What Makes Physics Difficult?. *International Journal of Environmental and Science Education*, *3*(1): 30-34.
[5] Rahmawati, R., Rustaman, N. Y., Hamidah, I., & Rusdiana, D. (2018). The Development and Validation of Conceptual Knowledge Test to Evaluate Conceptual Knowledge of Physics Prospective Teachers on Electricity and Magnetism Topic. *Jurnal Pendidikan IPA Indonesia*, *7*(4): 283-490.
[6] Gurcay, D., & Gulbas, E. (2015). Development of three-tier heat, temperature and internal energy diagnostic test. *Research in Science & Technological Education*, *33*(2): 197-217.
[7] Beichner, R. J. (1994). Testing student interpretation of kinematics graphs. *American journal of Physics*, *62*(8): 750-762.
[8] Wiser, M. (1986). The Differentiation of Heat and Temperature: An Evaluation of the Effect of Microcomputer Teaching on Students' Misconceptions. Technical Report 87-5.
[9] Baser, M., & Geban, Ö. (2007). Effectiveness of conceptual change instruction on understanding of heat and temperature concepts. *Research in science & technological education*, *25*(1): 115-133.
[10] Başer, M., & Geban, Ö. (2007). Effect of instruction based on conceptual change activities on students' understanding of static electricity concepts. *Research in Science & Technological Education*, *25*(2): 243-267.
[11] Tan, K. C. D., Goh, N. K., Chia, L. S., & Treagust, D. F. (2002). Development and application of a two-tier multiple choice diagnostic instrument to assess high school students' understanding of inorganic chemistry qualitative analysis. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *39*(4): 283-301.
[12] Tan, K. C. D., Taber, K. S., Goh, N. K., & Chia, L. S. (2005). The ionisation energy diagnostic instrument: a two-tier multiple-choice instrument to determine high school students' understanding of ionisation energy. *Chemistry Education Research and Practice*, *6*(4): 180-197.
[13] Arslan, H. O., Cigdemoglu, C., & Moseley, C. (2012). A three-tier diagnostic test to assess pre-service teachers' misconceptions about global warming, greenhouse effect, ozone layer depletion, and acid rain. *International journal of science education*, *34*(11): 1667-1686.
[14] Caleon, I., & Subramaniam, R. (2010). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International journal of science education*, *32*(7): 939-961.
[15] Romine, W. L., Schaffer, D. L., & Barrow, L. (2015). Development and application of a novel Rasch-based methodology for evaluating multi-tiered assessment instruments: Validation and utilization of an undergraduate diagnostic test of the water cycle. *International Journal of Science Education*, *37*(16): 2740-2768.
[16] Peşman, H., & Eryılmaz, A. (2010). Development of a three-tier test to assess misconceptions

about simple electric circuits. *The Journal of educational research*, *103*(3): 208-222.
[17] Shipstone, D. (1988). Pupils' understanding of simple electrical circuits. Some implications for instruction. *Physics education*, *23*(2): 92.
[18] Shipstone, D. M. (1984). A study of children's understanding of electricity in simple DC circuits. *European journal of science education*, *6*(2): 185-198.
[19] Arnold, M., & Millar, R. (1987). Being constructive: An alternative approach to the teaching of introductory ideas in electricity. *International Journal of Science Education*, *9*(5): 553-563.
[20] Osborne, R. J., & Cosgrove, M. M. (1983). Children's conceptions of the changes of state of water. *Journal of research in Science Teaching*, *20*(9): 825-838.
[21] Osborne, R. (1983). Towards modifying children's ideas about electric current. *Research in Science & Technological Education*, *1*(1): 73-82.
[22] Hekkenberg, A., Lemmer, M., & Dekkers, P. (2015). An analysis of teachers' concept confusion concerning electric and magnetic fields. *African Journal of Research in Mathematics, Science and Technology Education*, *19*(1): 34-44.
[23] Tarciso Borges, A., & Gilbert, J. K. (1999). Mental models of electricity. *International journal of science education*, *21*(1): 95-117.
[24] Cosgrove, M. (1995). A study of science-in-the-making as students generate an analogy for electricity. *International journal of science education*, *17*(3): 295-310.
[25] Cohen, R., Eylon, B., & Ganiel, U. (1983). Potential difference and current in simple electric circuits: A study of students' concepts. *American Journal of Physics*, *51*(5): 407-412.
[26] Paatz, R., Ryder, J., Schwedes, H., & Scott, P. (2004). A case study analysing the process of analogy-based learning in a teaching unit about simple electric circuits. *International Journal of Science Education*, *26*(9): 1065-1081.
[27] Psillos, D., Koumaras, P., & Valassiades, O. (1987). Pupils' representations of electric current before, during and after instruction on DC circuits. *Research in Science & Technological Education*, *5*(2): 185-199.
[28] Engelhardt, P. V., & Beichner, R. J. (2004). Students' understanding of direct current resistive electrical circuits. *American journal of physics*, *72*(1): 98-115.
[29] Finkelstein, N. (2005). Learning physics in context: A study of student learning about electricity and magnetism. *International Journal of Science Education*, *27*(10): 1187-1209.
[30] Zacharia, Z. C., & De Jong, T. (2014). The effects on students' conceptual understanding of electric circuits of introducing virtual manipulatives within a physical manipulatives-oriented curriculum. *Cognition and instruction*, *32*(2): 101-158.
[31] Stocklmayer, S. M., & Treagust, D. F. (1996). Images of electricity: How do novices and experts model electric current?. *International Journal of Science Education*, *18*(2): 163-178.
[32] Heller, P. M., & Finley, F. N. (1992). Variable uses of alternative conceptions: A case study in current electricity. *Journal of Research in Science Teaching*, *29*(3): 259-275.
[33] Heywood, D., & Parker, J. (1997). Confronting the analogy: primary teachers exploring the usefulness of analogies in the teaching and learning of electricity. *International Journal of Science Education*, *19*(8): 869-885.
[34] Retnawati, H. (2016). *Validitas, Reliabilitas, & Karakteristik Butir (Panduan untuk Peneliti, Mahasiswa, dan Psikometrian)*. Yogyakarta: Parama Publishing.
[35] Hambleton, R. K., Swaminatan, H. & Rogers, H. J. (1991). *Fundamentals of Item Respons Theory*. United States of America: Sage Publication, Inc.
[36] Demars, C. (2010). *Item Respons Theory: Understanding Statistics Measurement*. United States of America: Oxford University Press.
[37] Hambleton, R. K., & Swaminatan, H. (1985). *Item Response Theory: Principles and Applications*. New York: Springer Science+Business Media, LLC.
[38] Kuo, C. Y., Wu, H. K., Jen, T. H., & Hsu, Y. S. (2015). Development and validation of a multimedia-based assessment of scientific inquiry abilities. *International Journal of Science Education*, *37*(14): 2326-2357.
[39] Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology*, *28*(4): 563-575.
[40] Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for

Lawshe's content validity ratio. *Measurement and evaluation in counseling and development*, *45*(3): 197-210.
[41] Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Cimahi: Penerbit Trim Komunikasi.
[42] Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. New York, London: Springer.
[43] Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, *90*(2): 253-269.
[44] Davidowitz, B., & Potgieter, M. (2016). Use of the Rasch measurement model to explore the relationship between content knowledge and topic-specific pedagogical content knowledge for organic chemistry. *International Journal of Science Education*, *38*(9): 1483-1503.
[45] Anderson, L. W., & Krathwohl, D. R. (2001) *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: David McKay Company, Inc.
[46] Marzano, R. J. (2006). *Classroom Assessment & Grading that Work*. Alexandria, Virginia: Association for Supervision and Curriculum Development.
[47] Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, New Jersey, London: Lawrence Erlbaum Associates, Publishers.
[48] Bansilal, S. (2015). A Rasch analysis of a Grade 12 test written by mathematics teachers. *South African Journal of Science*, *111*(5-6): 1-9.
[49] Gronlund, N. E. (1985). Measurement and evaluation in teaching. In *Measurement and evaluation in teaching* (pp. xv-540).
[50] Thorndike, R. L. (1971). *Educational Measurement, Second Edi*. United State of America: American Council on Education.
[51] Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of applied measurement*, *3*(1): 85-106.
[52] Linacre, J. M. (2010). Predicting responses from Rasch measures. *Journal of Applied Measurement*, *11*(1): 1.
[53] Bond, T. G., & Fox, C. M. (2-15). *Applying the Rasch Model: Fundamental Measurement in the Human Science, Third Edit*. New York and London: Routledge Taylor & Francis Group.
[54] Krishnan, S., & Idris, N. (2014). Investigating reliability and validity for the construct of inferential statistics. *International Journal of Learning, Teaching and Educational Research*, *4*(1): 51-60.
[55] Sumintono, B., &Widhiarso, W. (2013). *Aplikasi Model Rasch untuk Penelitian Ilmu-ilmu Sosial*. Cimahi: Trim Komunikasi Publishing House.
[56] Sumintono, B. (2018). *Rasch Model Measurement as Tools in Assessment for*, ResearchGate, Kuala Lumpur.