JIPF (JURNAL ILMU PENDIDIKAN FISIKA)
p-ISSN: 2477-5959 | e-ISSN: 2477-8451 *Vol. 9 No. 1, January 2024, Page 74-87*

The Design and Validation of the Four Tier Test Instrument for Energy Literacy Using the Rasch Model Analysis

Tri Hastiti Fiskawarni [1], Rahmawati [2*)], Widiasih [3], Rezkawati Saad [4]
Universitas Muhammadiyah Makassar, Indonesia[1,2,4], Universitas Terbuka, Indonesia[3]
[*)]Corresponding E-mail: rahmawatisyam@unismuh.ac.id

Received: August 25th, 2023. Revised: October 11th, 2023. Accepted: November 3rd, 2023

Keywords :
Energy literacy; Item response theory, Rasch model; Energy topic; Winstep version 3. 68. 2

ABSTRACT
The science education curriculum needs to contain content on environmental issues including energy and its use as an effort to equip prospective physics teacher students with knowledge about the importance of reducing the impact of energy use. For this reason, it is important to measure energy literacy knowledge in prospective physics teacher students. This study aims to design and validate a four-tier test instrument to measure the energy literacy knowledge of prospective physics teacher students. The test instrument format used is a four-tier test. This test model has the advantage of being able to capture more accurate information with various answer patterns. The stages of developing this test used the design-based research model which consisted of five stages, namely developing an assessment framework, designing items, developing rubrics, conducting tests, and applying the Racsh Model analysis. The application of the Racsh Model analysis aims to obtain a valid and reliable test instrument with the Item Response Theory (IRT) approach assisted by the Winsteps program. The research method used is a descriptive-exploratory method to describe the results of the development and validation of the Four Tier test to measure Energy Literacy for prospective physics teacher student. The validation of the test was carried out through an assessment by five experts to assess the construct and content of the test instrument. The results of the item validation showed that the questions were acceptable in all aspects. The conclusion is that the test with four-tier format is suitable for identifying the knowledge of prospective physics teacher students about Energy Literacy. The four tier test model in exploring the energy literacy abilities of prospective teacher students can basically also be applied to students at the elementary school, middle school and high school levels. However, the complexity of the content tested needs to be adjusted to the existing curriculum at each level.

INTRODUCTION

Energy is an important concept in physics so that citizens can make the right decisions regarding important social issues such as energy production and use and climate change [1] [2]. Energy is a key issue for sustainable development which is also the responsibility of science education [3] [4] [5] [6] [7]. It because education that has the potential to change the behavior of young adults to use energy rationally and increase energy literacy [3] [7] [8]. Therefore, science education has an important role in preparing young adults from an early age to become future decision makers regarding energy [9] [10] [11].

Physics teacher candidate students are today's citizens who have personal responsibility in terms of energy use. In addition, in the future they are teachers who have the responsibility to teach students the concept of energy. Teachers play a key role in improving students' conditions [12] [13] [14]. It is important for prospective physics teacher students to have energy literacy so that in the future they can grow and develop their students' energy literacy. With their energy literacy, teachers were expected to be the main agents who can reorient education so that they can bring change towards a sustainable world [10] [15] [16].

Measuring energy literacy for prospective physics teacher students is important for at least two reasons. First, the measurement results will inform the state of the energy literacy of the respondents being measured. Second, measurement results provide data to make the right decisions [9] [11] [17]. To obtain data about the energy literacy of prospective physics teachers, measurement instruments are needed. Research on the development of instruments and measurement of energy literacy has been carried out a lot. In terms of the age group of the participants, the studies that have been carried out vary from elementary school age children [10,18,19] to junior high and high school age groups [11] [18] [20] [21] [22]. Research on energy literacy at the student teacher level has so far not been conducted.

An instrument is said to be good if it has three characteristics; valid, reliable, and usable [23] [24] [25]. There are several approaches that can be used to perform instrument validity, namely content validity, construction validity, and criterion validity [23] [26]. To obtain a test instrument that is valid, reliable and effective in its use, an analytical model is needed. for the testing process. There are two types of theory that can be used in analyzing test instruments, namely classical theory and modern theory, also known as Item Response Theory (IRT) or item response theory. Classical theory has a number of fundamental weaknesses, one of which is the classical test theory model using some statistics such as the level of difficulty and the discriminating power of the items depending on the respondents tested in the analysis [27] [28]. For this reason, psychometricians offer an alternative measurement theory and model called item response theory (IRT).

A popular model in the use of item response theory (IRT) is known as the logistic model. There are three types of logistic models, namely one-parameter logistic models, two-parameter logistic models, and three-parameter logistic models. The one-parameter logistic model is one of the most widely used IRT models. This model is also named as Rasch Model [27] [29] [30]. Based on the advantages of IRT theory and the Rasch model, it was decided in this study to use the Rasch model as a modeling approach assisted by the Winsteps software. Based on the background previously described and to meet these needs, it is necessary to develop a Four Tier Test instrument with the Rasch Model analysis assisted by the Winsteps program to measure the energy literacy of prospective physics teacher students. The form of the research problem is how to design and validate the four tier test instrument on the topic of energy literacy using the Rasch model analysis?

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 [32]. 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.

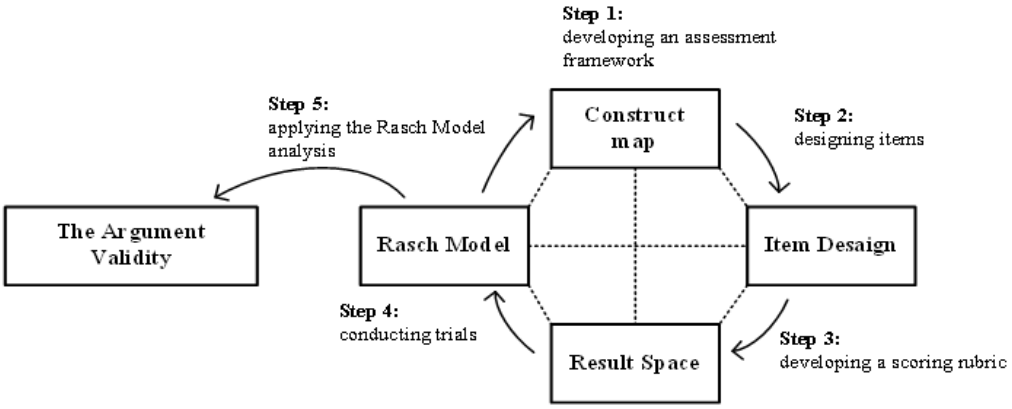


Fig 1. *Development Design of Four Tier Model Multiple Representation Test*

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	35	A
Field Test	63	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	Item of Energy Literacy Test
Validation of instrument assessment developed	*Expert Judgement*	Validation questionnaire	CVR and I-CVI agreement index analysis	Validation questionnaire, Energy literacy four tier test
Pilot test	Students	Dissemination of test	Rasch Model analysis with PCM (Partial Credit Model) type	Energy literacy four tier test

Measured Aspect	Data Resources	Data collection techniques	Data Analysis technique	Instrument
Field test	Students	Administration of test	Rasch Model analysis with PCM (Partial Credit Model) type	Energy literacy 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 [33]. CVR content validity analysis uses the formulation:

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

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:

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

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 [34]. 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 category [35] is shown in Table 4.

Table 4. Cronbach's Alpha Reliability Categorization

Coefficient reliability of Cronbach's Alpha	Category
$0,8 \leq \alpha$	Special
$0,7 \leq \alpha < 0,8$	Very Good
$0,6 \leq \alpha < 0,7$	Very nice
$0,5 \leq \alpha < 0,6$	Enough
$\alpha < 0,5$	Weak

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. Interpretation of Item Reliability

Value of Item reliability (*r*)	Category
$0,94 \leq r$	Special
$0,91 \leq r < 0,94$	Very good
$0,80 \leq r < 0,91$	Very nice
$0,67 \leq r < 0,80$	Enough
$r < 0,67$	Weak

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 [36] [37].

Data on the value of person fit and item fit in the infit mnsq and outfit mnsq columns, as well as the value of infit Zstd and outfit Zstd which follow the ideal value of the Rasch model (ie 1.00). Meanwhile, the standard Z value (Zstd) on infit Zstd 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 [36] [37] [38].

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 [39] and Marzano's Taxonomy [40].

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 format. Several

aspects were 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 40 items spread over a number of materials about energy. The distribution of the material, the concept of energy, and the number of items developed were presented in Table 6.

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

No.	Learning outcome	Item number
1	Identifying non-renewable and renewable energy resources	1,2,3,4
2	Explaining the relationship between energy consumption and the result of emissions	5,6,7,8
3	Calculating the cost of electricity consumption	9,10,11,12
4	Calculating the energy produced from an energy source	13,14,15,16
5	Identifying energy-related misconceptions	17,18,19
6	Making conclusions from information about energy presented in the form of tables or graphs	20,21,22,23
7	Analyzing the impact of using certain energy sources on the environment	24,25,26,27
8	Using information to make decisions regarding energy consumption and purchases	28,29,30,31,32
9	Proposing alternative solutions to energy-related problems	33,34,35,36
10	Showing a tendency to behave energy-saving	37,38,39,40
	Total of number item	40

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 [41]. The development of the scoring rubric refers to the scoring model scheme developed by Gurcay & Gulbas [42] adapted according to the four-tier energy literacy test. Students' energy literacy abilities were divided into three categories, namely understanding, not understanding, and misconceptions. The categorization of energy literacy abilities based on the pattern of interpretation of answers was 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	

Answer	Confidence level of answer	Reason	Confidence level of reason	Criteria
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 [43] [44] [45]. 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. 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 [47]. 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* [48]. 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).

Figure 2 show a problem measurement report that displays the results of expert validation for the

problem categories with Rasch Model analysis through *Winstep* software.

INPUT: 63 PERSONS 40 ITEMS MEASURED: 63 PERSONS 40 ITEMS 2 CATS 3.68.2

PERSON: REAL SEP.: 1.50 REL.: .69 ... ITEM: REAL SEP.: 3.58 REL.: .93

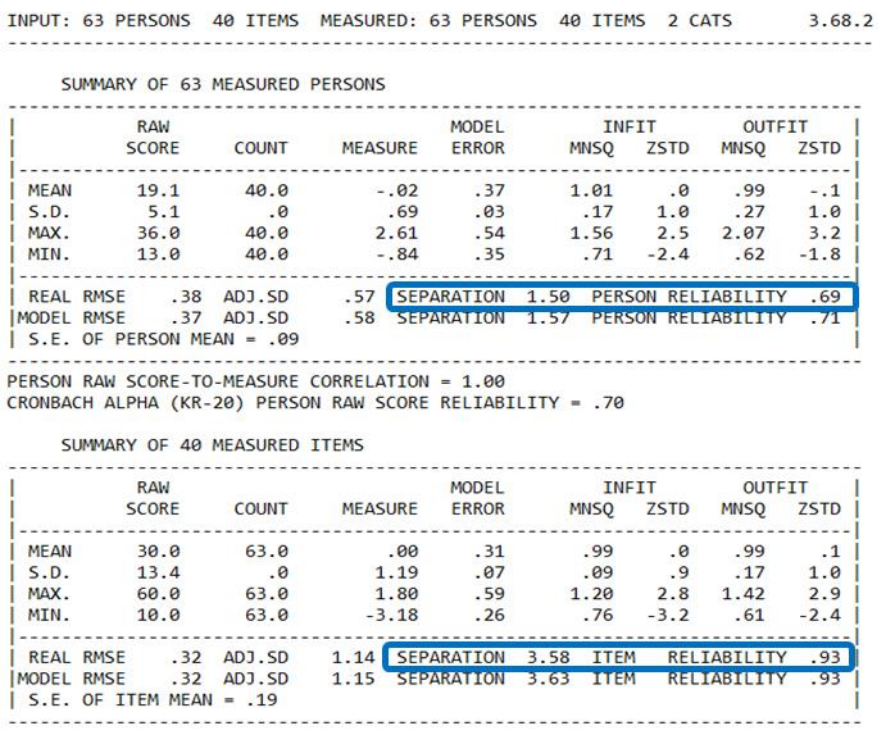
ITEM STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
27	10	63	1.80	.36	.76	-1.0	.66	-1.2	.63	.30	85.7	84.7	S27
33	11	63	1.67	.35	.88	-.5	.93	-.2	.43	.31	84.1	83.2	S33
22	13	63	1.44	.33	1.12	.6	1.33	1.3	.10	.31	81.0	80.4	S22
25	16	63	1.14	.31	.89	-.7	.89	-.6	.47	.32	81.0	76.6	S25
29	16	63	1.14	.31	1.01	.1	1.11	.6	.27	.32	77.8	76.6	S29
8	17	63	1.05	.30	1.03	.3	.99	.0	.29	.32	71.4	75.5	S8
9	17	63	1.05	.30	.93	-.4	.93	-.4	.41	.32	77.8	75.5	S9
11	17	63	1.05	.30	1.03	.3	1.09	.5	.26	.32	74.6	75.5	S11
39	17	63	1.05	.30	1.03	.2	1.03	.3	.28	.32	74.6	75.5	S39
37	19	63	.88	.29	1.03	.3	1.05	.3	.27	.32	74.6	73.2	S37
16	21	63	.71	.28	.89	-.9	.91	-.6	.45	.32	73.0	70.9	S16
38	21	63	.71	.28	1.11	1.0	1.11	.8	.17	.32	66.7	70.9	S38
31	22	63	.64	.28	1.02	.2	1.00	.0	.30	.32	68.3	69.7	S31
36	23	63	.56	.28	1.05	.5	1.07	.6	.24	.31	68.3	68.6	S36
10	24	63	.48	.27	.99	.0	1.00	.0	.32	.31	69.8	67.6	S10
28	25	63	.41	.27	1.10	1.0	1.19	1.7	.15	.31	63.5	66.6	S28
40	25	63	.41	.27	.99	.0	.98	-.1	.32	.31	63.5	66.6	S40
15	26	63	.34	.27	.96	-.5	.94	-.6	.37	.31	66.7	65.6	S15
32	26	63	.34	.27	1.02	.3	1.02	.2	.28	.31	60.3	65.6	S32
2	27	63	.27	.27	1.04	.5	1.04	.4	.25	.31	65.1	64.6	S2
21	27	63	.27	.27	.96	-.5	.94	-.5	.37	.31	65.1	64.6	S21
23	27	63	.27	.27	.99	-.1	.98	-.1	.33	.31	68.3	64.6	S23
34	27	63	.27	.27	.99	.0	.99	-.1	.31	.31	68.3	64.6	S34
12	31	63	-.02	.26	1.20	2.8	1.24	2.4	.00	.29	50.8	61.5	S12
19	33	63	-.15	.26	1.01	.1	1.01	.2	.27	.29	58.7	61.0	S19
26	34	63	-.22	.26	.97	-.4	.95	-.4	.33	.28	63.5	61.1	S26
6	35	63	-.29	.26	.80	-3.2	.76	-2.4	.57	.28	82.5	61.2	S6
35	35	63	-.29	.26	1.05	.7	1.00	.0	.23	.28	50.8	61.2	S35
4	36	63	-.36	.27	.87	-2.0	.82	-1.7	.47	.28	71.4	61.6	S4
5	36	63	-.36	.27	1.08	1.1	1.09	.8	.16	.28	55.6	61.6	S5
18	36	63	-.36	.27	1.01	.2	1.02	.2	.26	.28	58.7	61.6	S18
20	36	63	-.36	.27	1.10	1.5	1.37	2.9	.07	.28	61.9	61.6	S20
30	36	63	-.36	.27	1.00	.0	1.00	.1	.28	.28	61.9	61.6	S30
17	46	63	-1.12	.29	1.09	.7	1.42	1.9	.01	.22	73.0	73.0	S17
14	48	63	-1.30	.30	1.01	.1	.94	-.2	.21	.21	76.2	76.2	S14
24	52	63	-1.71	.34	.91	-.3	.78	-.7	.33	.18	82.5	82.5	S24
1	55	63	-2.09	.38	.93	-.2	.73	-.7	.30	.16	87.3	87.3	S1
3	59	63	-2.87	.52	.95	.0	.66	-.5	.24	.11	93.7	93.6	S3
13	59	63	-2.87	.52	.94	.0	.61	-.6	.27	.11	93.7	93.6	S13
7	60	63	-3.18	.59	1.00	.2	.94	.1	.11	.10	95.2	95.2	S7
MEAN	30.0	63.0	.00	.31	.99	.0	.99	.1			71.7	71.6	
S.D.	13.4	.0	1.19	.07	.09	.9	.17	1.0			11.0	9.8	

Fig 2. *The results of expert validation for the problem categories with Rasch Model analysis through Winstep software*

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 Winsteps Ministep 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. Figure 3 showed the results of the test reliability analysis using the Winstep software.



```
INPUT: 63 PERSONS  40 ITEMS  MEASURED: 63 PERSONS  40 ITEMS  2 CATS       3.68.2
--------------------------------------------------------------------------------

    SUMMARY OF 63 MEASURED PERSONS
--------------------------------------------------------------------------------
|          RAW                          MODEL         INFIT        OUTFIT     |
|         SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|------------------------------------------------------------------------------|
| MEAN     19.1      40.0        -.02     .37      1.01     .0    .99    -.1 |
| S.D.      5.1        .0         .69     .03       .17    1.0    .27    1.0 |
| MAX.     36.0      40.0        2.61     .54      1.56    2.5   2.07    3.2 |
| MIN.     13.0      40.0        -.84     .35       .71   -2.4    .62   -1.8 |
|------------------------------------------------------------------------------|
| REAL RMSE    .38  ADJ.SD     .57  SEPARATION  1.50  PERSON RELIABILITY  .69 |
|MODEL RMSE    .37  ADJ.SD     .58  SEPARATION  1.57  PERSON RELIABILITY  .71 |
| S.E. OF PERSON MEAN = .09                                                    |
--------------------------------------------------------------------------------
PERSON RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) PERSON RAW SCORE RELIABILITY = .70

    SUMMARY OF 40 MEASURED ITEMS
--------------------------------------------------------------------------------
|          RAW                          MODEL         INFIT        OUTFIT     |
|         SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|------------------------------------------------------------------------------|
| MEAN     30.0      63.0         .00     .31       .99     .0    .99     .1 |
| S.D.     13.4        .0        1.19     .07       .09     .9    .17    1.0 |
| MAX.     60.0      63.0        1.80     .59      1.20    2.8   1.42    2.9 |
| MIN.     10.0      63.0       -3.18     .26       .76   -3.2    .61   -2.4 |
|------------------------------------------------------------------------------|
| REAL RMSE    .32  ADJ.SD    1.14  SEPARATION  3.58  ITEM   RELIABILITY  .93 |
|MODEL RMSE    .32  ADJ.SD    1.15  SEPARATION  3.63  ITEM   RELIABILITY  .93 |
| S.E. OF ITEM MEAN = .19                                                      |
--------------------------------------------------------------------------------
```

***Fig 3**. The results of the test reliability analysis using the Winstep software*

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	0,70
	Person reliability	0,69
	Item reliability	0,93
Separation	Person separation	1,50
	Item separation	3,58

Table 8 shows that the reliability value of Cronbach's Alpha (KR-20) is 0,70. The reliability value of Alpha Cronbach (KR-20) indicates that this four-tier test instrument has internal consistency reliability in a good category [49]. 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 [50].

The results of Rasch's analysis on person reliability and person separation are 0,69 and 1,50 respectively [35]. The person reliability value obtained is in the fairly good category which indicates that the responses from the respondents are quite good and consistent [51]. For the aspect of person separation, the coefficient value obtained is 1,50. Krishnan & Idris [52] 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 [36].

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 [35]. The item

separation coefficient value obtained is in the good category. Linacre 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 [36] [52].

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 [35] [53] [54].

The following was the result of the item fit analysis presented in Table 9.

Table 9. Summary of Fit Item Statistics Results of the Energy Literacy Test for Pilot Test

No.	Item	Infit		Outfit		Pt-Measure	
		Mnzq	Zstd	Mnzq	Zstd	Corr.	Exp.
1.	S1	0,93	-0,2	0,73	-0,7	0,30	0,16
2.	S2	1,04	0,5	1,04	0,4	0,25	0,31
3.	S3	0,95	0,0	0,66	-0,5	0,24	0,11
4.	S4	0,87	-2,0	0,82	-1,7	0,47	0,28
5.	S5	1,08	1,1	1,09	0,8	0,16	0,28
6.	S6	0,80	-3,2	0,76	-2,4	0,57	0,28
7.	S7	1,00	0,2	0,94	0,1	0,11	0,10
8.	S8	1,03	0,3	0,99	0,00	0,29	0,32
9.	S9	0,93	-0,4	0,93	-0,4	0,41	0,32
10.	S10	0,99	0,0	1,00	0,0	0,32	0,31
11.	S11	1,03	0,3	1,09	0,51	0,26	0,32
12.	S12	1,20	2,8	1,24	2, 4	0,00	0,29
13.	S13	0,94	0,0	0,61	-0,6	0,27	0,11
14.	S14	1,01	0,1	0,94	-0,2	0,21	0,21
15.	S15	0,96	-0,5	0,94	-0,6	1, 37	0,31
16.	S16	0,89	-0,9	0,91	-0,6	0,45	0,32
17.	S17	1,09	0,7	1,42	1,9	0,01	0,22
18.	S18	1,01	0,2	1,02	0,2	0,26	0,28
19.	S19	1,01	0,1	1,01	0,2	0,27	0,29
20.	S20	1,10	1,5	1,37	2, 9	0,07	0,28
21.	S21	0,96	-0,5	0,94	-0,5	0,37	0,31
22.	S22	1,12	0,6	1,33	1,3	0,10	0,31
23.	S23	0,99	-0,1	0,98	-0,1	0,33	0,31
24.	S24	0,91	-0,3	0,78	-0,7	0,33	0,18
25.	S25	0,89	-0,7	0,89	-0,6	0,47	0,32
26.	S26	0,97	-0,4	0,95	-0,4	0,33	0,28
27.	S27	0,76	-1,0	0,66	-1,2	0,63	0,30

No.	Item	Infit		Outfit		Pt-Measure	
		Mnzq	Zstd	Mnzq	Zstd	Corr.	Exp.
28.	S28	1,10	1,0	1,19	1,7	0,15	0,31
29.	S29	1,01	0,1	1,11	0,6	0,27	0,32
30.	S30	1,00	0,0	1,00	0,1	0,28	0,28
31.	S31	1,02	0,2	1,00	0,00	0,30	0,32
32.	S32	1,02	0,3	1,02	0,2	0,28	0,31
33.	S33	0,88	-0,5	0,93	-0,2	0,43	0,31
34.	S34	0,99	0,0	0,99	-0,1	0,31	0,31
35.	S35	1,05	0,7	1,00	0,0	0,23	0,28
36.	S36	1,05	0,5	1,07	0,6	0,24	0,31
37.	S37	1,03	0,3	1,05	0,3	0, 27	0,32
38.	S38	1,11	1,0	1,11	0,8	0,17	0,32
39.	S39	1,03	0,2	1,03	0,3	0,28	0,32
40.	S40	0,99	0,0	0,98	-0,1	0,32	0,31

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 were some items that do not meet the Pt criteria. Measure corr. but the other two criteria are met. This showed that S10 item was still within the allowable limits so they do not need to be omitted. Meanwhile, nine items 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 [36] [37] [38]. By using the three criteria for observing person fit, none of the respondents (people) experienced misfit. The results of this person fit analysis can be a reference for further research. To obtain better data, the research sample used should be enlarged so that the data distribution can be more comprehensive.

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 energy literacy ability of prospective physics teacher students on the topic of energy.

ACKNOWLEDGMENTS

This study is supported by Kementerian Pendidikan Kebudayaan Riset dan Teknologi (Kemendikbudristek) in Hibah Dosen Pemula DRTPM 2023. We also are grateful to the participants who have been contributed in this study.

REFERENCES

[1] Duit, R., Schecker, H., Höttecke, D., & Niedderer, H. (2014). Teaching physics. In *Handbook of Research on Science Education, Volume II* (pp. 448-470). Routledge.
[2] Tajudin, N. A. M., & Chinnappan, M. (2015). Exploring Relationship between Scientific Reasoning Skills and Mathematics Problem Solving. *Mathematics Education Research Group of Australasia*.
[3] Wellington, J. (2003). Science education for citizenship and a sustainable future. *Pastoral Care in Education*, *21*(3), 13-18.
[4] Hoque, F., Yasin, R. M., & Sopian, K. (2022). Revisiting education for sustainable development: Methods to inspire secondary school students toward renewable energy. *Sustainability*, *14*(14), 8296.
[5] Khushik, F., & Diemer, A. (2018). Critical analysis of education policies in Pakistan: A sustainable development perspective. *Social Science Learning Education Journal*, *3*(09), 01-16.
[6] Eilks, I. (2015). Science education and education for sustainable development–justifications, models, practices and perspectives. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(1), 149-158.
[7] Glavič, P. (2020). Identifying key issues of education for sustainable development. *Sustainability*, *12*(16), 6500.
[8] Zografakis, N., Menegaki, A. N., & Tsagarakis, K. P. (2008). Effective education for energy efficiency. *Energy Policy*, *36*(8), 3226-3232.
[9] Martín-Gámez, C., & Erduran, S. (2018). Understanding argumentation about socio-scientific issues on energy: a quantitative study with primary pre-service teachers in Spain. *Research in Science & Technological Education*, *36*(4), 463-483.
[10] Aguirre-Bielschowsky, I., Lawson, R., Stephenson, J., & Todd, S. (2017). Energy literacy and agency of New Zealand children. *Environmental Education Research*, *23*(6), 832-854.
[11] Chen, K. L., Liu, S. Y., & Chen, P. H. (2015). Assessing multidimensional energy literacy of secondary students using contextualized assessment. *International Journal of Environmental and Science Education*, *10*(2), 201-218.
[12] Kaltakci-Gurel, D., Eryilmaz, A., & McDermott, L. C. (2017). Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. *ReseaRch in science & Technological educaTion*, *35*(2), 238-260.
[13] Fraenkel, J., Wallen, N., & Hyun, H. (2018). *How to design and evaluate research in education (10th) ed.)*. McGraw-Hill.
[14] Khaeruddin, K., Rahmawati, R., Nurfazlina, N., Salwa, R., & Nurhayati, N. (2022). The Development of Students' Worksheets Face to Face Online Based on Hypercontent on Temperature and Heat Topic. *Jurnal Penelitian Pendidikan IPA (JPPIPA)*, *8*(6), 3011-3019.
[15] Stephens, J. C., Hernandez, M. E., Román, M., Graham, A. C., & Scholz, R. W. (2008). Higher education as a change agent for sustainability in different cultures and contexts. *International journal of sustainability in higher education*, *9*(3), 317-338.
[16] Gänswein, W. (2011). *Effectiveness of information use for strategic decision making*. Wiesbaden: Gabler.
[17] Yusup, M., Setiawan, A., Rustaman, N. Y., & Kaniawati, I. (2017, July). Developing a framework for the assessment of pre-service physics teachers' energy literacy. In *Journal of Physics: Conference Series* (Vol. 877, No. 1, p. 012014). IOP Publishing.
[18] Chen, S. J., Chou, Y. C., Yen, H. Y., & Chao, Y. L. (2015). Investigating and structural modeling energy literacy of high school students in Taiwan. *Energy Efficiency*, *8*, 791-808.

[19] Fell, M. J., & Chiu, L. F. (2014). Children, parents and home energy use: Exploring motivations and limits to energy demand reduction. *Energy Policy*, *65*, 351-358.
[20] Davis, P. (1985). The attitude and knowledge of tasmanian secondary students towards energy conservation and the environment. *Research in Science Education*, *15*(1), 68-75.
[21] DeWaters, J., & Powers, S. (2013). Establishing measurement criteria for an energy literacy questionnaire. *The Journal of Environmental Education*, *44*(1), 38-55.
[22] Halder, P., Pietarinen, J., Havu-Nuutinen, S., Pöllänen, S., & Pelkonen, P. (2016). The Theory of Planned Behavior model and students' intentions to use bioenergy: A cross-cultural perspective. *Renewable Energy*, *89*, 627-635.
[23] E Gronlunds, N. (2021). Measurement and assessment in teaching. *Pakistan Journal of Educational Research and Evaluation (PJERE)*, *5*(2).
[24] Doran, R. L. (1980). *Basic Measurement and Evaluation of Science Instruction*. National Science Teachers Association, 1742 Connecticut Ave., NW, Washington, DC 20009 (Stock No. 471-14764; no price quoted)..
[25] Engeström, Y., Virkkunen, J., Helle, M., Pihlaja, J., & Poikela, R. (1996). The change laboratory as a tool for transforming work. *Lifelong learning in Europe*, *1*(2), 10-17.
[26] Gronlund, N. E. (1998). *Assessment of student achievement*. Allyn & Bacon Publishing, Longwood Division, 160 Gould Street, Needham Heights, MA 02194-2310; tele.
[27] Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.
[28] Retnawati, H. (2016). *Validitas Reliabilitas & Karakteristik Butir (Panduan untuk Peneliti, Mahasiswa, dan Psikometrian) berbasis software*. Nuha Medika.
[29] DeMars, C. (2010). *Item response theory*. Oxford University Press.
[30] Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.
[31] 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.
[32] 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.
[33] Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology*, *28*(4), 563-575.
[34] 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.
[35] Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan*. Trim komunikata.
[36] Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences*. Springer Science & Business Media.
[37] 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.
[38] 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.
[39] Bloom, B. S. (2010). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
[40] Marzano, R. J. (2006). *Classroom assessment & grading that work*. ASCD.
[41] Wilson, M. (2023). *Constructing measures: An item response modeling approach*. Taylor & Francis.
[42] 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.
[43] 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.

[44] E Gronlunds, N. (2021). Measurement and assessment in teaching. *Pakistan Journal of Educational Research and Evaluation (PJERE)*, *5*(2).
[45] Thorndike, R. L. (1982). Educational measurement: Theory and practice. *The improvement of measurement in education and psychology*, 3-13.
[46] Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of applied measurement*, *3*(1), 85-106.
[47] Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of applied measurement*, *3*(1), 85-106.
[48] Linacre, J. M. (2010). Predicting responses from Rasch measures. *Journal of Applied Measurement*, *11*(1), 1.
[49] Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.
[50] Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.
[51] Bond, T. G., & Fox, C. M. (2013). Applying the Rasch model: Fundamental measurement in the human sciences. Psychology Press.
[52] 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.
[53] Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial (edisi revisi)*. Trim Komunikata Publishing House.
[54] Sumintono, B. (2017). Rasch Model Measurement as Tools in Assessment for Learning.