REID (Research and Evaluation in Education), 11(1), 2025, 45-58
Available online at: http://journal.uny.ac.id/index.php/reid

Enhancing the five-tier diagnostic test on cell concepts through Rasch model analysis

Oky Rizkiana Silaban; Kusnadi*; Ana Ratna Wulan
Universitas Pendidikan Indonesia, Indonesia
*Corresponding Author. E-mail: okysilaban@upi.edu

ARTICLE INFO

Article History
Submitted:
June 18, 2025
Revised:
July 7, 2025
Accepted:
July 11, 2025

Keywords
Rasch model; five-tier diagnostic test; cell concept; misconception

Scan Me:

ABSTRACT

Misconceptions in biology can prevent students from gaining a deeper understanding of biological concepts. There is a five-tier diagnostic test that can explore the misconceptions experienced by students. This research aims to develop a five-tier diagnostic test that is feasible to use to identify student misconceptions with Rasch model analysis. The research method used was quantitative descriptive, and the sample was 103 people with a purposive sampling technique, which is a purposeful sampling technique, namely, the school that is the research location, experiencing misconceptions related to the concept of cells. Based on the results of the study, it was found that the five-tier diagnostic test developed was very feasible to use as an instrument to identify students' misconceptions on the concept of cells. Each indicator is represented by several items that have been tested for validity, reliability, difficulty level and differentiation using Rasch model analysis with the help of the Winsteps program. Based on the analysis with the Rasch model, out of 36 items that were externally validated, 23 items were obtained that met the eligibility criteria and were declared valid for implementation.

To cite this article (in APA style):
Silaban, O. R., Kusnadi, K., & Wulan, A. R. (2025). Enhancing the five-tier diagnostic test on cell concepts through Rasch model analysis. *REID (Research and Evaluation in Education)*, *11*(1), 45-58. https://doi.org/10.21831/reid.v11i1.87212

INTRODUCTION

Misconceptions in biology learning (Brown & Schwartz, 2009; Suwono *et al*., 2021; Tambo *et al*., 2003; Rahma *et al*., 2022) have been widely reported, including in cell learning (Suwono *et al*., 2021; Rahma *et al*., 2022). Students' misconceptions in the cell structure and function section are that most students cannot understand the special characteristics of prokaryotic cells. Students chose that prokaryotic cells have a nucleus enveloped by a nuclear membrane (Suwono *et al*., 2021). This can be explained by the fact that students do not imagine cells without a nucleus and only learn cell models with a clear nucleus at school (Rahma *et al*., 2022). Another misconception related to cell structure is the presence of a cell wall in prokaryotic cells, causing some students to mistakenly believe that it is a plant cell (Tambo *et al*., 2003). Regarding the function of cell organelles, students often misconceive the role of mitochondria, consider mitochondria as a place for food storage, and have misconceptions related to the function and location of chloroplasts (Suwono *et al*., 2021; Rahma *et al*., 2022).

The existence of misconceptions in biological material can hinder students in mastering biological material more deeply, because concepts in biology are interconnected and become the key to understanding other concepts. Thus, if misconceptions in certain concepts are not immediately addressed, it can cause misconceptions in other concepts (Tekkaya, 2002). In addition, misconceptions are also persistent and resistant to change (Wandersee *et al*., 1994). Therefore, efforts need to be made to reduce or prevent the emergence of misconceptions in further learning (Koray & Bal, 2002; Zulfianto & Abduh, 2023) with misconception detection.

Some diagnostic tests can be used to detect misconceptions such as, regular multiple-choice tests can be used with a large number of participants and can identify students' concepts, but multiple-choice tests have some disadvantages, such as students can guess the answers, which can reduce the reliability of the test, and the selected answer options do not provide deep insight into students' thinking or conceptual understanding (Zimmerman & Williams, 2003; Chang et al., 2010). Meanwhile, two-tier tests cannot distinguish between students who do not understand and misconceptions, so all incorrect answers are considered misconceptions (Peşman & Eryılmaz, 2010). Three-tier tests are still unable to fully distinguish belief choices for main answers (first tier) from belief choices for reasons (second tier) (Gurel *et al*., 2015). The five-tier diagnostic test has been able to distinguish between the tier of confidence of the answer and the level of confidence of the reason for the student's choice, so that it can better identify the misconceptions experienced by students (Rusilowati, 2015). However, the four-tier diagnostic test has not been able to deeply identify the misconceptions experienced by students, because the four-tier diagnostic test has provided answer choices for the first and third tiers, so students can choose answers only by guessing (Silaban, 2021).

The five-tier diagnostic test is considered one of the most effective instruments for providing a clear picture of students' misconceptions, as it not only requires students to choose an answer from multiple choices. In the fifth tier, students are asked to respond by drawing or explaining the concept being assessed. The five-tier diagnostic test consists of five tiers. The first tier is a standard multiple-choice question with five answer options. The second tier asks students to rate their confidence in the answer chosen in the first tier. The third tier requires students to select a reason supporting their answer from a list of options. The fourth tier assesses the students' confidence in the reason selected. Finally, the fifth tier involves a drawing task (Ermawati et al., 2019; Ramadhani & Ermawati, 2020; Anam et al., 2019), in which students illustrate their concept understanding.

Drawing is used to investigate understanding in science and has been applied in various contexts. Drawing activities have proven effective in exploring students' ideas about abstract concepts when combined with interviews. The five-tier diagnostic test is essentially a four-tier test with an additional tier that allows students to represent their reasoning through visual depiction (Köse, 2008). Moreover, the five-tier test can reveal what students think about a concept and can identify misconceptions more detailed and comprehensively (Anam et al., 2019).

Before implementing the developed test items, it is essential to conduct an analysis to ensure their quality. A good test item is the item that can accurately convey information in accordance with the intended measurement objectives (Friatma & Anhar, 2019). For an instrument to be considered high quality, it must meet several criteria, including validity (measuring what it is intended to measure), high reliability, a range of difficulty levels, and good item discrimination to distinguish between students' varying levels of ability (Quaigrain & Arhin, 2017).

There are two main approaches to item analysis: Classical Test Theory (CTT) and the modern Item Response Theory (IRT) (Sumintono & Widhiarso, 2015). One of the models within IRT is the Rasch model. Compared to the classical approach, IRT provides more effective and accurate analysis (Tavakol & Dennick, 2012). However, the use of the Rasch model remains relatively limited in the field of education, particularly in the design of concept inventories and in the analysis of diagnostic test development (Boone, 2017; Ibrahim et al., 2024).

The Rasch model is considered a more rigorous alternative to Classical Test Theory (CTT) because it provides more objective, accurate, and sample- and item-independent measurements (Sumintono & Widhiarso, 2015). Unlike CTT, which only provides raw scores and reliability estimates that are dependent on sample characteristics, the Rasch model offers linear-scale measurement, item fit analysis using statistics such as Infit and Outfit Mean Square (MNSQ), and the ability to detect the unidimensionality of an instrument through residual analysis (Bond & Fox, 2015). Additionally, Rasch is more flexible in handling incomplete data and allows for detailed mapping of student ability and item difficulty (Boone et al., 2014). Research by Erfan et al. (2020) indicates that Rasch is more effective than CTT in selecting valid and reliable items, making it a more appropriate approach for evaluating the quality of test instruments. Therefore, a thorough Rasch model analysis of the five-tier diagnostic test in this study is a critical step to ensure its validity in identifying persistent misconceptions about the cell concept among 11th-grade science students, paving the way for data-driven instructional interventions.

METHOD

Research Design

The research method used in this research is descriptive quantitative. This method was chosen because this research focuses on analyzing the results of diagnostic test trials using the Rasch model involving 103 students. This analysis aims to examine the characteristics of the items, including aspects of validity, reliability, difficulty level, and question differentiation. Data were obtained from student responses to the test trials, then analyzed using Rasch modelling software to provide an objective and detailed evaluation of the quality of the items.

Population and Sample

This study was conducted in one school that was specifically selected using a purposive sampling technique. The school was chosen based on the results of preliminary research, which showed that students in the school experienced misconceptions related to the concept of cells. The study population consisted of all Class XII students (both science and social studies tracks), totalling 216 students. However, because the concept of cells is studied by science specialization students, the research sample is limited to three science classes, namely XII IPA-1, XII IPA-2, and XII IPA-3, with a total of 103 students.

Instrument

The test instrument used in this study is a five-tier diagnostic test. The five-tier diagnostic test is a tiered multiple-choice assessment designed to explore students' conceptual understanding in greater depth. In the first tier, students are required to answer a multiple-choice question with four answer options. The second tier assesses the students' confidence level in the answer they selected in the first tier. In the third tier, students are asked to choose the most appropriate reason to support their answer. The fourth tier again measures students' confidence, this time in their selected reason. Finally, the fifth tier involves a drawing task related to the concept tested in either the first or third tier, aiming to explore students' visual representation of the concept (Table 1).

The test blueprint was developed based on the learning outcomes outlined in the *Kurikulum Merdeka* for Phase F, which emphasizes students' understanding of cells and the bioprocesses that occur within them. The test included seven indicators arranged progressively. It began with a basic understanding of the chemical components of cells (Indicator 1), followed by the structure and function of cell organelles (Indicators 2 and 3). Once these foundational concepts were mastered, students were expected to analyze various cellular bioprocesses, such as membrane transport (Indicator 4) and cell division (Indicators 5 and 6). Finally, students should be able to compare bioprocesses within the cell, specifically distinguishing between mitosis and meiosis (Indicator 7).

Table 1. Example of Five-Tier Diagnostic Test Question

Tier	Question
First Tier	**The problem and the multiple-choice answer** *i. Protein, lipid, karbohidrat, dan asam nukleat adalah biomolekul utama penyusun sel. Manakah dari pilihan berikut yang dominan sebagai penyusun utama membran sel?* *a. Protein.* *b. Fosfolipid.* *c. Karbohidrat.* *d. Asam Nukleat.* i. Proteins, lipids, carbohydrates and nucleic acids are the major biomolecules that make up the cell. Which of the following options is dominant as the main constituent of the cell membrane? a. Proteins. b. Phospholipids. c. Carbohydrates. d. Nucleic Acids.
Second Tier	**Confidence level in choosing the answer (i)** *ii. Apakah Anda yakin dengan pilihan jawaban di atas?* *a. Yakin* *b. Tidak Yakin* ii. Are you sure about the answer choices above? a. Sure b. Not sure
Third Tier	**Reason in choosing the answer (i)** *iii. Alasan memilih jawaban?* *a. Berperan membentuk lapisan ganda yang memungkinkan membran sel menjadi selektif permeabel.* *b. Penyusun minor membran sel yang berperan dalam membantu molekul melakukan transportasi sel.* *c. Berperan dalam mengatur keluar masuknya zat, tetapi tidak membentuk membran sel utama.* *d. Ikut terlibat dalam pembentukan membran sel, namun berperan dalam penyimpanan genetik.* iii. Reason for choosing the answer? a. It plays a role in forming a double layer that allows the cell membrane to be selectively permeable. b. A minor constituent of the cell membrane plays a role in helping molecules carry out cellular transportation. c. Plays a role in regulating the entry and exit of substances but does not form the main cell membrane. d. Involved in cell membrane formation but plays a role in genetic storage.
Fourth Tier	**Confidence level in choosing the reason (iii)** *iv. Apakah Anda yakin dengan alasan jawaban Anda?* *a. Yakin* *b. Tidak Yakin* i. Are you sure of the reason for your answer? a. Sure b. Not sure
Fifth Tier	**Drawing Task** *v. Gambarlah sketsa struktur yang Anda pilih dan yakini sebagai penyusun dominan membran sel tersebut!* ii. Draw a sketch of the structure you chose and believe to be the dominant constituent of the cell membrane!

The content scope includes the chemical components of cells, cell structure, membrane transport, and cell division, as taught at the senior high school level. This instrument was developed with reference to the Revised Bloom's Taxonomy, specifically targeting the cognitive levels of C2 (understanding) and C4 (analyzing) (Anderson & Krathwohl, 2014).

Procedure

The first stage involved internal validation by two experts, an assessment expert and a content expert, who reviewed the instrument based on content, construct, and language aspects. Based on the validation results, the instrument obtained an average score of 84.85%, which falls into the "highly valid" category.

After that, the revised question set was then tested externally on 103 students of class XII IPA who had studied the concept of cells, to be done manually with question-and-answer sheets printed on A4 paper. Before being done, the researchers explained the five-tier diagnostic test and how to work on the question, followed by explaining how to answer on the answer paper. Guidelines for working on the five-tier diagnostic test have also been provided on the front page of the question sheet.

The process of working on the questions was carried out using stationery in the form of pencils or pens. Researchers had provided pencils for students, but the number was insufficient, so some students used pens. When using a pencil, students could correct their answers with an eraser.

The time for working on the questions, which was originally allocated for 150 minutes (two hours and 30 minutes), was felt to be insufficient by students. Therefore, at the request of students and to ensure that the work was carried out optimally, the researchers gave an additional 30 minutes, so that the total work was 180 minutes (three hours). Students said that in working on one question number, the average time needed was around four to six minutes.

The next stage was to review and score the test items, then input them into an Excel worksheet using a scoring guide. The knowledge aspects in tiers one, three, and five were scored. Tiers one and three were each assigned a score of 1 for correct responses and 0 for incorrect responses (Fariyani et al., 2015). In addition, the scoring rubric for tier five can be found in Table 2.

Table 2. Categories of Student Responses at Tier Five
(Adapted from Dikmenli, 2010; Anam et al., 2019; Lailiyah & Ermawati, 2022; Kurnaz & Eksi, 2015)

No.	Categories	Description	Score
1.	SD (*scientific drawing)*	The student provides a drawing that is consistent with the scientific concept	**4**
2.	PD *(partial drawing)*	The student provides a drawing that is somewhat aligned with the cell concept, but contains minor errors	**3**
3.	MD (*misconception drawing)*	The student provides a drawing that reflects a misconception or deviates from the correct concept	**2**
4.	UD (*undefined drawing)*	The student provides a drawing that is not related to the concept of cells	**1**
5.	ND (*no drawing)*	The student does not provide a drawing.	**0**

Table 2 categorizes student responses on the fifth-tier drawing task based on their accuracy and relevance to the cell concept. Scores range from 4 for scientifically accurate drawings to 0 for no drawing provided. This scoring system helps identify the depth of students' conceptual understanding and misconceptions visually.

Data Analysis Techniques

The data analysis technique used in this research is Rasch Model analysis with the help of Winstep 4.7.0 software. Validity test analysis can be seen in the Item Measure menu, by considering the Outfit MNSQ, ZSTD, and PT Mean Corr values, where items are declared valid if they meet at least two of the following three criteria. First, the Outfit MNSQ value that is considered feasible is in the range of greater than 0.5 and less than 1.5. Second, the range of Outfit ZSTD values that are accepted according to the standard is -2.0 to 2.0. Third, the range of PT Mean Corr values that are accepted is between 0.4 and 0.85 (Sumintono & Widhiarso, 2015; Sari & Mahmudi, 2024).

Then, the results of instrument reliability based on the results of the Summary Statistic Output of the Winstep program were analyzed by considering the Cronbach Alpha coefficient value. The instrument uses the reliability coefficient criteria; 0.70 or more is acceptable as good reliability (Streiner, 2010).

After that, the Wright Map item analysis was conducted. The Wright Map is a map that describes the distribution of item difficulty levels, which highlights how difficult each item is compared to other items, to illustrate the strength of items that will be used as a measure of student ability (Sumintono & Widhiarso, 2015; Sari & Mahmudi, 2024).

Finally, the differentiation of questions was analyzed by looking at the results of item analysis in the SE (standard error) model value section. Determination of the differentiation of questions based on the average value (Mean) SE and Standard Deviation of the SE value. If the SD value is smaller than the average value (Mean), the question's differentiation is said to be good, and if the average value is smaller than the SD value, the question's differentiation cannot distinguish well (Ramadhan & Hidayatullah, 2023).

FINDINGS AND DISCUSSION

A total of 36 items were given to 103 students for external validation. The instrument was analyzed using the Rasch model through the Winstep version 4.7.0 program with the aim of evaluating the validity, reliability, difficulty level, and question differentiation.

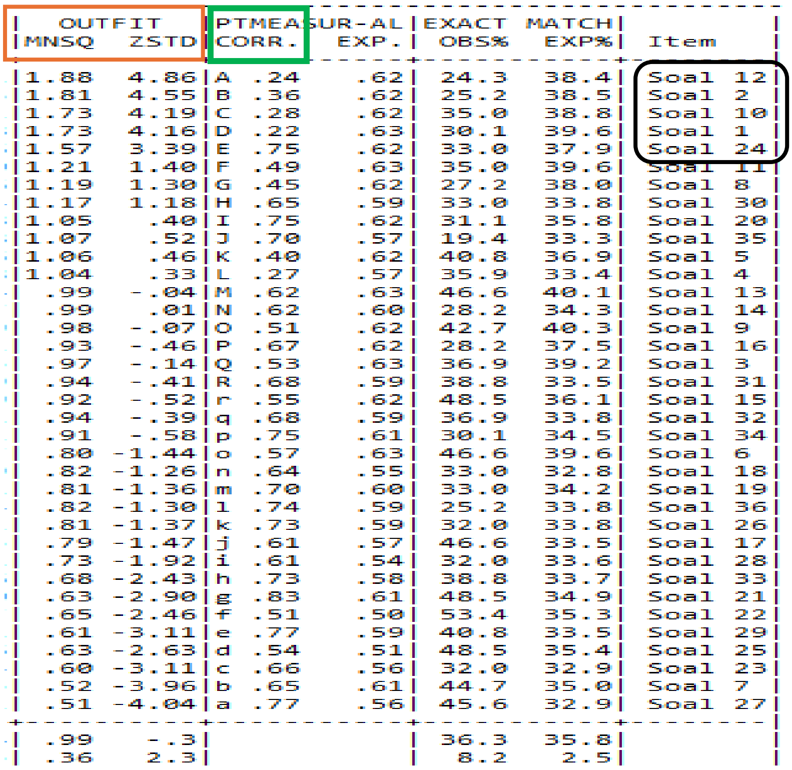
Validity

The first stage in the analysis is carried out by analyzing the MNSQ (Mean Square) outfit value to assess the level of item suitability. Outfit MNSQ values that are considered suitable fall within the range of greater than 0.5 and less than 1.5. The range indicates that the item is within the productive limit, which can measure students' abilities effectively without too much deviation from the expected model (Boone *et al.*, 2014). Based on the MNSQ outfit analysis (Figure 1), 31 items met the MNSQ Outfit value criteria, while five items did not meet the criteria, namely item numbers 1, 2, 10, 12 and 24.

The second stage is to analyze the ZSTD outfit value (Figure 1). The range of ZSTD outfit values accepted according to the standard is -2.0 to 2.0 (Sumintono & Widhiarso, 2015; Sari & Mahmudi, 2024). Values within this range indicate that the item has conformity with the Rasch model and can be considered a valid item. The items that did not meet were question numbers 1, 2, 7, 10, 12, 21, 22, 23, 24, and 25.

The third stage is to analyze the PT Mean Corr value by paying attention to the range of acceptable values, which is between 0.4 and 0.85 (Sumintono & Widhiarso, 2015; Sari & Mahmudi, 2024). Values within this range indicate that the items have a good correlation with the participants' abilities, so they are considered valid. From the results of this stage of analysis, six questions were obtained that did not meet the standards, namely questions number 1, 2, 4, 5, 10, and 12. The results of the item fit analysis for all items are presented in Figure 1, where the yellow color information is a value that does not meet the established validity criteria.

Question items are categorized as valid if they meet at least two of the three outfit value criteria set. Based on Figure 1, there are 31 items that are valid because they meet two of the three outfit value requirements, while five items, namely numbers 1, 2, 10, 12 and 24, are declared invalid because they do not meet the specified criteria.



```
                          -------------------------------------------
|   OUTFIT   |PTMEASUR-AL |EXACT  MATCH|              |
|MNSQ   ZSTD|CORR.   EXP.|  OBS%   EXP%|  Item        |
                          ------------+-------------+
|1.88   4.86|A  .24   .62|  24.3   38.4|  Soal  12|
|1.81   4.55|B  .36   .62|  25.2   38.5|  Soal  2 |
|1.73   4.19|C  .28   .62|  35.0   38.8|  Soal  10|
|1.73   4.16|D  .22   .63|  30.1   39.6|  Soal  1 |
|1.57   3.39|E  .75   .62|  33.0   37.9|  Soal  24|
|1.21   1.40|F  .49   .63|  35.0   39.6|  Soal  11|
|1.19   1.30|G  .45   .62|  27.2   38.0|  Soal  8 |
|1.17   1.18|H  .65   .59|  33.0   33.8|  Soal  30|
|1.05    .40|I  .75   .62|  31.1   35.8|  Soal  20|
|1.07    .52|J  .70   .57|  19.4   33.3|  Soal  35|
|1.06    .46|K  .40   .62|  40.8   36.9|  Soal  5 |
|1.04    .33|L  .27   .57|  35.9   33.4|  Soal  4 |
|  .99  -.04|M  .62   .63|  46.6   40.1|  Soal  13|
|  .99   .01|N  .62   .60|  28.2   34.3|  Soal  14|
|  .98  -.07|O  .51   .62|  42.7   40.3|  Soal  9 |
|  .93  -.46|P  .67   .62|  28.2   37.5|  Soal  16|
|  .97  -.14|Q  .53   .63|  36.9   39.2|  Soal  3 |
|  .94  -.41|R  .68   .59|  38.8   33.5|  Soal  31|
|  .92  -.52|r  .55   .62|  48.5   36.1|  Soal  15|
|  .94  -.39|q  .68   .59|  36.9   33.8|  Soal  32|
|  .91  -.58|p  .75   .61|  30.1   34.5|  Soal  34|
|  .80 -1.44|o  .57   .63|  46.6   39.6|  Soal  6 |
|  .82 -1.26|n  .64   .55|  33.0   32.8|  Soal  18|
|  .81 -1.36|m  .70   .60|  33.0   34.2|  Soal  19|
|  .82 -1.30|l  .74   .59|  25.2   33.8|  Soal  36|
|  .81 -1.37|k  .73   .59|  32.0   33.8|  Soal  26|
|  .79 -1.47|j  .61   .57|  46.6   33.5|  Soal  17|
|  .73 -1.92|i  .61   .54|  32.0   33.6|  Soal  28|
|  .68 -2.43|h  .73   .58|  38.8   33.7|  Soal  33|
|  .63 -2.90|g  .83   .61|  48.5   34.9|  Soal  21|
|  .65 -2.46|f  .51   .50|  53.4   35.3|  Soal  22|
|  .61 -3.11|e  .77   .59|  40.8   33.5|  Soal  29|
|  .63 -2.63|d  .54   .51|  48.5   35.4|  Soal  25|
|  .60 -3.11|c  .66   .56|  32.0   32.9|  Soal  23|
|  .52 -3.96|b  .65   .61|  44.7   35.0|  Soal  7 |
|  .51 -4.04|a  .77   .56|  45.6   32.9|  Soal  27|
+------------+-------------+-------------+----------
|  .99    -.3|             |  36.3   35.8|          |
|  .36    2.3|             |   8.2    2.5|          |
```

□ : MNSQ and ZSTD outfit value

□ : Point Measure Correlation (Pt. Measure Corr.)

□ : Invalid item

Figure 1. Output Misfit Order

Reliability

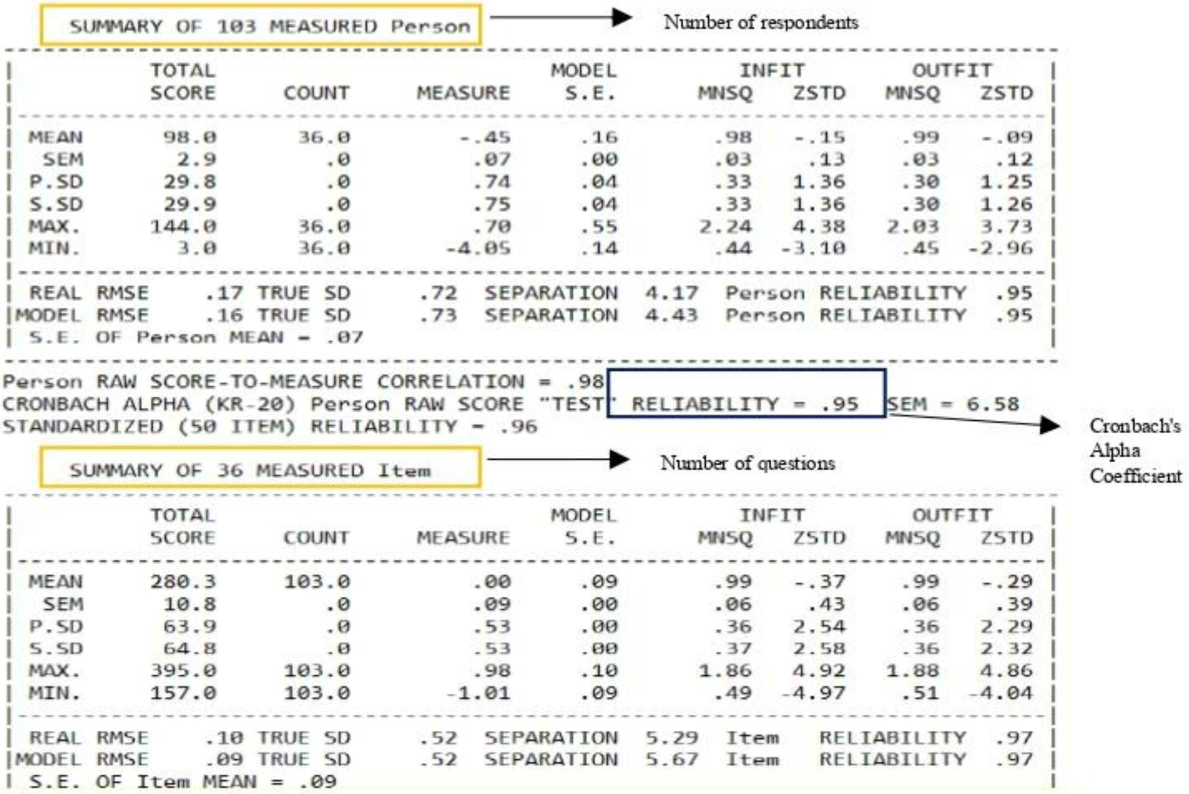
Item reliability can be seen from the Ouput Tables menu in the Winstep program by selecting the summary statistic menu. Reliability refers to the level of consistency of the results given by a test, so that it can be used as a reference to assess the reliability of the test according to the specified standards (Ramadhan & Hidayatullah, 2023).

The instrument is said to be reliable if the reliability value is greater than 0.7, which indicates that the instrument has a good level of reliability (Streiner, 2010). Based on the reliability analysis obtained, the reliability value of the respondents (students) is 0.95, and the reliability value of the 36 items is 0.97, and the Cronbach Alpha coefficient is 0.95 (Figure 2).

Difficulty Level

The level of item difficulty can be analyzed using Winstep Software from the output menu and selecting the item measure option. On the menu, information about the items will be presented in the form of a table that displays the logit value from highest to lowest (Figure 3).

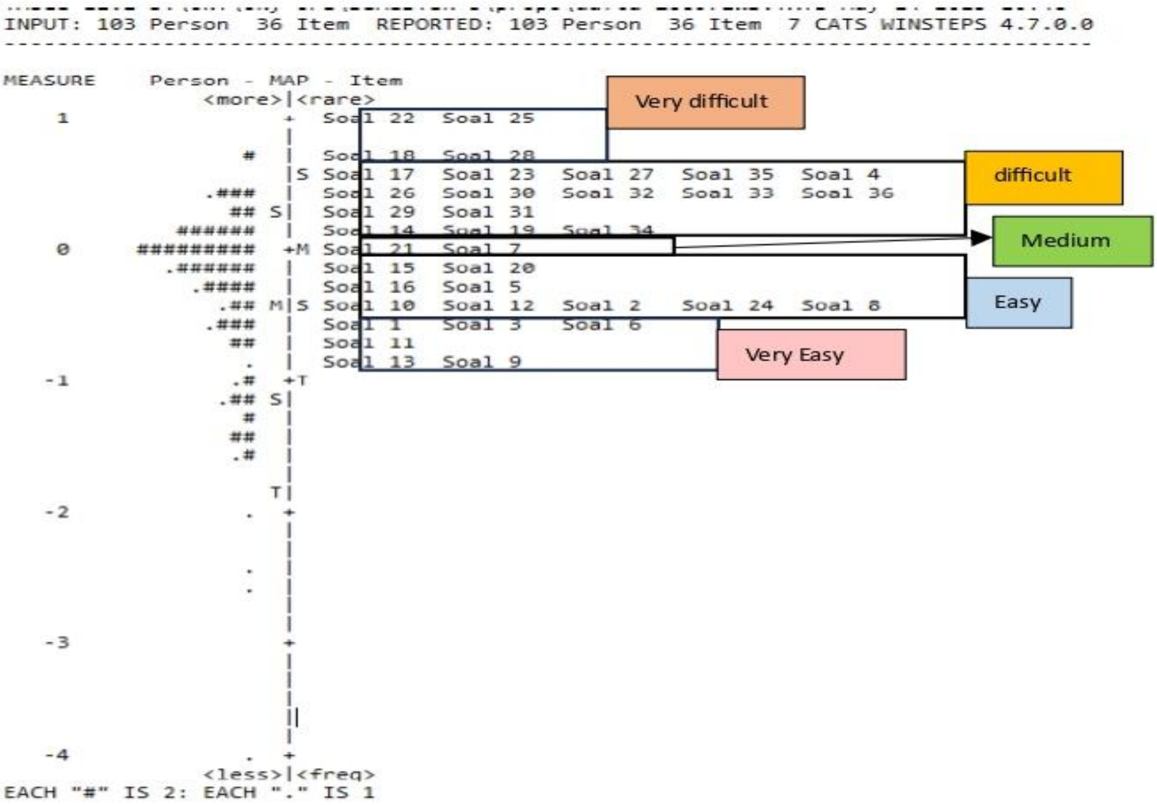
The category of question difficulty can be analyzed based on the Item Wright Map, which presents more clearly the category of question difficulty in a display, as in Figure 4. The difficulty level of the items is analyzed by combining the mean logit value with the Standard Deviation value (Sumintono & Widhiarso, 2015).



```
SUMMARY OF 103 MEASURED Person
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT         OUTFIT   |
|          SCORE      COUNT    MEASURE   S.E.      MNSQ   ZSTD    MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      98.0      36.0       -.45    .16        .98   -.15     .99   -.09 |
|  SEM       2.9        .0        .07    .00        .03    .13     .03    .12 |
| P.SD      29.8        .0        .74    .04        .33   1.36     .30   1.25 |
| S.SD      29.9        .0        .75    .04        .33   1.36     .30   1.26 |
| MAX.     144.0      36.0        .70    .55       2.24   4.38    2.03   3.73 |
| MIN.       3.0      36.0      -4.05    .14        .44  -3.10     .45  -2.96 |
|-----------------------------------------------------------------------------|
| REAL RMSE     .17 TRUE SD     .72  SEPARATION  4.17  Person RELIABILITY  .95 |
|MODEL RMSE     .16 TRUE SD     .73  SEPARATION  4.43  Person RELIABILITY  .95 |
| S.E. OF Person MEAN = .07                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .95  SEM = 6.58
STANDARDIZED (50 ITEM) RELIABILITY = .96

SUMMARY OF 36 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT         OUTFIT   |
|          SCORE      COUNT    MEASURE   S.E.      MNSQ   ZSTD    MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     280.3     103.0        .00    .09        .99   -.37     .99   -.29 |
|  SEM      10.8        .0        .09    .00        .06    .43     .06    .39 |
| P.SD      63.9        .0        .53    .00        .36   2.54     .36   2.29 |
| S.SD      64.8        .0        .53    .00        .37   2.58     .36   2.32 |
| MAX.     395.0     103.0        .98    .10       1.86   4.92    1.88   4.86 |
| MIN.     157.0     103.0      -1.01    .09        .49  -4.97     .51  -4.04 |
|-----------------------------------------------------------------------------|
| REAL RMSE     .10 TRUE SD     .52  SEPARATION  5.29  Item   RELIABILITY  .97 |
|MODEL RMSE     .09 TRUE SD     .52  SEPARATION  5.67  Item   RELIABILITY  .97 |
| S.E. OF Item MEAN = .09                                                     |
```

Figure 2. Picture of Summary Statistic Output Results



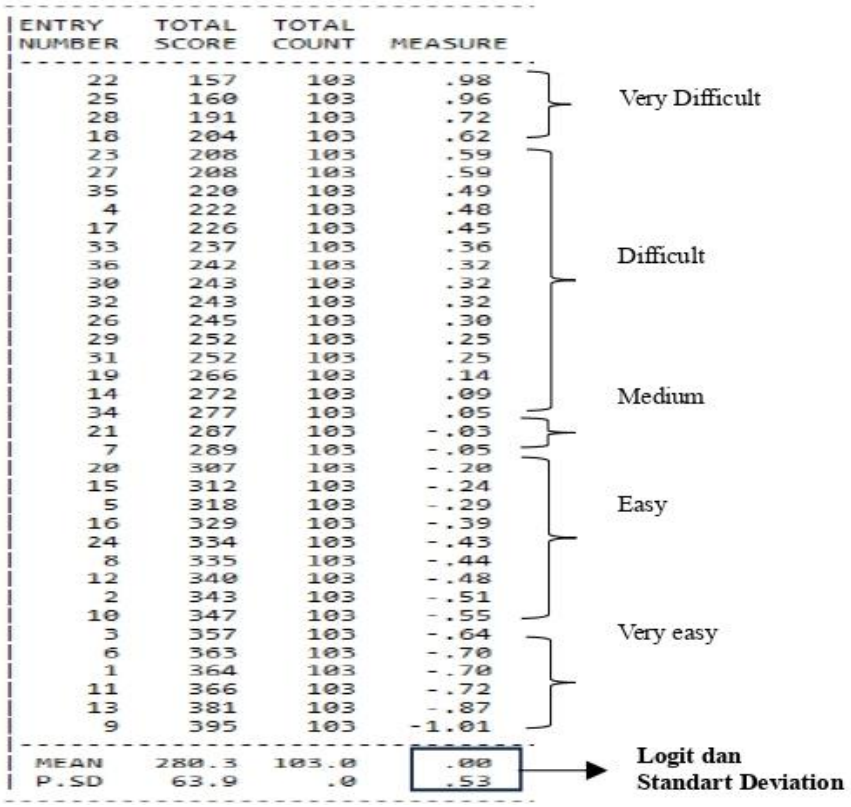
```
INPUT: 103 Person  36 Item  REPORTED: 103 Person  36 Item  7 CATS WINSTEPS 4.7.0.0
-----------------------------------------------------------------------------------

MEASURE     Person - MAP - Item
                   <more>|<rare>
    1                    +   Soal 22   Soal 25
                         |
                      #  |   Soal 18   Soal 28
                         |S  Soal 17   Soal 23   Soal 27   Soal 35   Soal 4
                   .###  |   Soal 26   Soal 30   Soal 32   Soal 33   Soal 36
                    ## S|    Soal 29   Soal 31
                 ######  |   Soal 14   Soal 19   Soal 34
    0         #########  +M  Soal 21   Soal 7
                .######  |   Soal 15   Soal 20
                 .####   |   Soal 16   Soal 5
                   .## M|S   Soal 10   Soal 12   Soal 2    Soal 24   Soal 8
                  .###   |   Soal 1    Soal 3    Soal 6
                    ##   |   Soal 11
                     .   |   Soal 13   Soal 9
   -1               .#   +T
                   .## S|
                     #   |
                    ##   |
                    .#   |
                         |
                        T|
   -2                .   +
                         |
                         |
                     .   |
                     .   |
                         |
                         |
   -3                    +
                         |
                         |
                         ||
                         |
   -4                .   +
                   <less>|<freq>
EACH "#" IS 2: EACH "." IS 1
```

Figure 3. Item Wright Map Result

The logit value and standard deviation information are at the bottom of the table. Based on Figure 4, the logit value of the Measure item is 0.0, and the standard deviation is 0.5. Grouping the level of difficulty of the question can be done by summing the mean logit value of 0.0 + and S.D 0.5, this category is a group of difficult questions, while values greater than 0.5 include questions with a very difficult category, in addition, the value range of 0.5 to 0.0 is an easy question, and the value range is less than-0.5 is a question with a very easy category (more detailed results can be seen in Table 3). Based on these groupings, each item's level of difficulty is presented in Figure 4.

According to Nuryanti *et al.* (2018), a good question is one that is neither too high nor too low, or in other words, in the medium, easy, or difficult category. Because the level of difficulty is a level that can be reached by the ability of students, and includes a good level of difficulty. Based on the grouping of the level of difficulty of the questions, four items were obtained in the too difficult category and five items in the too easy category. Questions that are too easy and too difficult are questions that have a difficulty level outside the limit of one Standard Deviation (Ramadhan & Hidayatullah, 2023), so they need to be replaced or revised. Of the nine questions that were outside the standard deviation, only one was revised, namely question number 3, because it was a representative question from indicator 1, while the other eight questions were not used.



```
-------------------------------------------------
|ENTRY    TOTAL   TOTAL
|NUMBER   SCORE   COUNT   MEASURE
|------------------------------------------------
|     22    157     103       .98
|     25    160     103       .96
|     28    191     103       .72
|     18    204     103       .62
|     23    208     103       .59
|     27    208     103       .59
|     35    220     103       .49
|      4    222     103       .48
|     17    226     103       .45
|     33    237     103       .36
|     36    242     103       .32
|     30    243     103       .32
|     32    243     103       .32
|     26    245     103       .30
|     29    252     103       .25
|     31    252     103       .25
|     19    266     103       .14
|     14    272     103       .09
|     34    277     103       .05
|     21    287     103      -.03
|      7    289     103      -.05
|     20    307     103      -.20
|     15    312     103      -.24
|      5    318     103      -.29
|     16    329     103      -.39
|     24    334     103      -.43
|      8    335     103      -.44
|     12    340     103      -.48
|      2    343     103      -.51
|     10    347     103      -.55
|      3    357     103      -.64
|      6    363     103      -.70
|      1    364     103      -.70
|     11    366     103      -.72
|     13    381     103      -.87
|      9    395     103     -1.01
|------------------------------------------------
|  MEAN   280.3   103.0       .00
|  P.SD    63.9      .0       .53
-------------------------------------------------
```

Figure 4. Item Measure Result

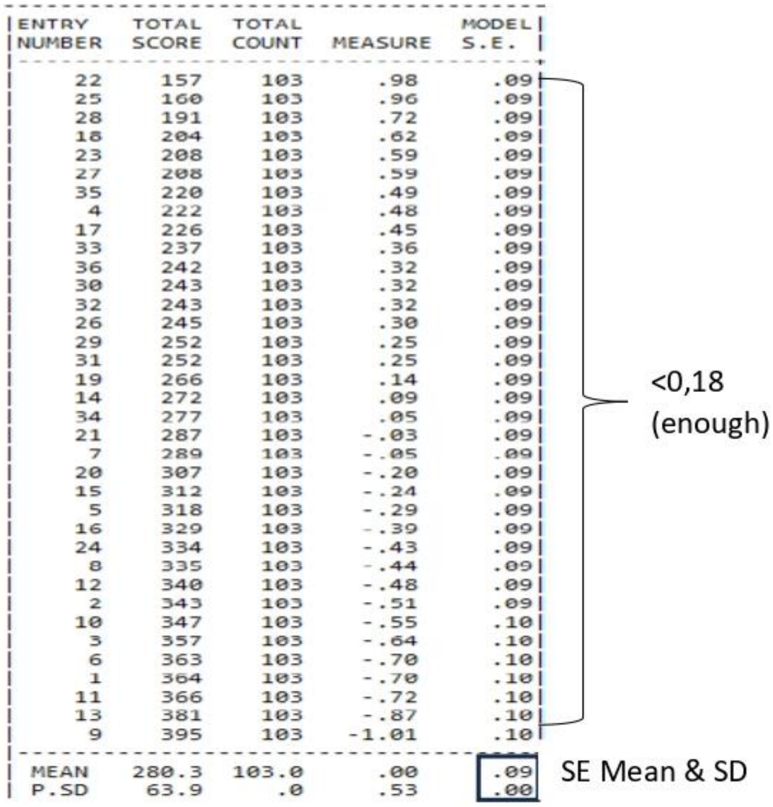
Table 3. Results of Problem Difficulty Analysis

Item Number	Total	Category	Percentage
18, 22, 25, 28	4	Very difficult	11.11%
4, 14, 17, 19, 23, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36	15	Difficult	41.67%
7, 21	2	Medium	5.55 %
2,5, 8, 10, 12, 15, 16, 20, 24	9	Easy	25%
1, 3, 6, 9, 11, 13	6	Very Easy	16.67%
Total	**36**		**100%**

Items classified as very difficult (18, 22, 25, 28) are the most difficult questions to do and are beyond the range of students' abilities, while very easy items (1, 6, 9, 11, 13) are the easiest questions to do and are less than students' abilities. Questions that are too easy and difficult cannot function properly as instruments (Sumintono & Widhiarso, 2015). Thus, based on the results of the analysis of the difficulty level of the questions, 23 questions were obtained that were suitable for use to identify student misconceptions, namely question numbers 3, 4, 5, 7, 8, 14, 15, 16, 17, 19, 20, 21, 23, 26, 27, 29, 30, 31, 32, 33, 34, 35, and 36.

Distinguishing Power

The differentiation of items can be analyzed using the Winstep program by looking at the item measure results in the Model S.E. (Standard Error) value section. The S.E. Model value can show the accuracy of the item in distinguishing the level of understanding and ability of the students on the question being tested.



```
-------------------------------------------------
|ENTRY    TOTAL   TOTAL                  MODEL|
|NUMBER   SCORE   COUNT    MEASURE       S.E. |
|---------------------------------------------+
|    22     157     103        .98        .09|
|    25     160     103        .96        .09|
|    28     191     103        .72        .09|
|    18     204     103        .62        .09|
|    23     208     103        .59        .09|
|    27     208     103        .59        .09|
|    35     220     103        .49        .09|
|     4     222     103        .48        .09|
|    17     226     103        .45        .09|
|    33     237     103        .36        .09|
|    36     242     103        .32        .09|
|    30     243     103        .32        .09|
|    32     243     103        .32        .09|
|    26     245     103        .30        .09|
|    29     252     103        .25        .09|
|    31     252     103        .25        .09|
|    19     266     103        .14        .09|
|    14     272     103        .09        .09|
|    34     277     103        .05        .09|
|    21     287     103       -.03        .09|
|     7     289     103       -.05        .09|
|    20     307     103       -.20        .09|
|    15     312     103       -.24        .09|
|     5     318     103       -.29        .09|
|    16     329     103       -.39        .09|
|    24     334     103       -.43        .09|
|     8     335     103       -.44        .09|
|    12     340     103       -.48        .09|
|     2     343     103       -.51        .09|
|    10     347     103       -.55        .10|
|     3     357     103       -.64        .10|
|     6     363     103       -.70        .10|
|     1     364     103       -.70        .10|
|    11     366     103       -.72        .10|
|    13     381     103       -.87        .10|
|     9     395     103      -1.01        .10|
|---------------------------------------------
| MEAN    280.3   103.0        .00        .09
| P.SD     63.9      .0        .53        .00
```

<0,18
(enough)

SE Mean & SD

Figure 5. Model S.E. Value Results

Determination of the question's differentiation power based on the average (Mean) S.E. value and the Standard Deviation of the S.E. value. If the S.D value is less than the average (Mean), then the differentiation power of the question is said to be good, and if the average value is less than the S.D value, then the differentiation power of the question cannot distinguish well (Ramadhan & Hidayatullah, 2023).

The results of the analysis of the average value (mean) of 0.09, while the S.D value is 0.0, which means that the SD value is less than (<) the average value (mean), and it can be said that the question is, on average, a good question differentiation. The S.E. value can then be grouped based on the quality of the question's differentiating power, namely, (<0.09) is categorized as good, the value (0.09-0.18) is sufficient, and (>0.18) is a bad category.

Based on Figure 5, from a total of 36 items analyzed, the instrument can be categorized as having sufficient discriminating power, with item discrimination indices ranging from 0.09 to

0.18. Although the values are relatively low, they still fall within an acceptable range, indicating that the instrument is feasible to use, especially with potential improvements (Ramadhan & Hidayatullah, 2023). Overall, the items can distinguish between students with higher and lower levels of understanding. Item discrimination is important because it shows the ability of each question to differentiate between high-achieving and low-achieving students, which is crucial for evaluating whether an assessment instrument effectively measures student performance and understanding.

Based on the results of the validity, reliability, difficulty level, and discrimination index analyses, 23 items were found to meet the criteria for a suitable instrument. Prior to the implementation, the test items were reviewed again for necessary revisions. The review revealed that two items, numbers 26 and 27, had similar content, both asking about anaphase division. Additionally, items 31 and 32 were also similar, both related to the topic of cancer cells. These items were revised and combined into a single item to avoid redundancy.

This study has several limitations that should be acknowledged. The research was conducted in only one purposively selected school, and the sample was limited to Grade XII science students, which restricts the generalizability of the findings to broader student populations. Therefore, future research is recommended to involve a more diverse range of schools and student groups from various regions to strengthen generalizability.

CONCLUSION

Based on the research results described in the research discussion, it can be concluded that the five-tier diagnostic test developed is very feasible to use as an instrument to identify student misconceptions on the concept of cells. The five-tier diagnostic test, developed based on the specification table, consists of seven indicators, namely identifying the chemical components that make up cells, explaining cell structure, explaining the function of cell parts, analysing various membrane transport mechanisms in cells, applying the mitotic division process, and comparing mitosis and meiosis. Each indicator is represented by several items that have been tested for validity, reliability, difficulty level and differentiation using Rasch model analysis with the help of the Winsteps program. Based on the analysis with the Rasch model, a total of 31 questions were declared valid with a very high reliability value of 0.97, indicating good instrument consistency in measuring students' abilities. The level of difficulty of the questions varied, with four very difficult questions, 15 difficult questions, two questions of medium difficulty, nine easy questions, and six very easy questions, thus reflecting adequate variation to accommodate the diverse ability levels of students. However, the differentiating power of the questions was moderate, with a value of less than 0.18, indicating that some questions still had a limited ability to distinguish between high and low ability. From a total of 36 externally validated items, 23 items were obtained that met the eligibility criteria and were declared valid for implementation.

The use of Rasch model analysis in developing this five-tier diagnostic instrument ensures strong validity, reliability, and appropriate item difficulty, enabling accurate measurement of student abilities, as well as effective identification of misconceptions about the cell concept. Implementing this instrument provides educators with valid insights into student understanding, helping them design targeted teaching strategies to address learning gaps. Furthermore, this study offers a practical guide for teachers on applying Rasch analysis, empowering them to create and refine their own assessment tools, thereby enhancing the quality and fairness of educational measurements.

DISCLOSURE STATEMENT

The authors acknowledge the financial support in the form of a scholarship from the Indonesia Endowment Fund for Education Agency (*Lembaga Pengelola Dana Pendidikan* or LPDP). However, they declare that they have no conflicts of interest related to the content of this study.

FUNDING STATEMENT

This research was supported by a scholarship from the Indonesia Endowment Fund for Education Agency (*Lembaga Pengelola Dana Pendidikan* or LPDP). The authors declare that no additional external funding was received for this study.

ETHICS APPROVAL

The research participants in this study were anonymized, all data collected during the study were used only for the purpose of research, and it is guaranteed that the results of the study will not cause any harm to the research participants.

REFERENCES

Anam, R. S., Widodo, A., Sopandi, W., & Wu, H. K. (2019). Developing a five-tier diagnostic test to identify students' misconceptions in science: An example of the heat transfer concepts. *Elementary Education Online*, *18*(03), 1014-1029.

Anderson, L. W. & Krathwohl, D. R. (2014). *Kerangka landasan untuk pembelajaran pengajaran dan asesmen* (A. Prihantoro, trans.). Pustaka Pelajar.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge. https://doi.org/10.4324/9781315814698

Boone, W. J. (2017). Rasch analysis for instrument development: Why, when, and how? *CBE-Life Sciences Education, 15*(4), 1-7. https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4

Brown, M. H., & Schwartz, R. S. (2009). Connecting photosynthesis and cellular respiration: Preservice teachers' conceptions. *Journal of Research in Science Teaching, 46*(7), 791 – 812. https://doi.org/10.1002/tea.20287

Chang, C. Y., Yeh, T. K., & Barufaldi, J. P. (2010). The positive and negative effects of science concept tests on student conceptual understanding. *International Journal of Science Education*, *32*(2), 265-282. https://doi.org/10.1080/09500690802650055

Dikmenli, M. (2010). Misconceptions of cell division held by student teachers in biology: A drawing analysis. *Scientific Research and Essays,* *5*(2), 235–247. https://academicjournals.org/article/article1380539915_Dikmenli.pdf

Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Analisis kualitas soal kemampuan membedakan rangkaian seri dan paralel melalui teori tes klasik dan model Rasch. *Indonesian Journal of Educational Research and Review,* *3*(1), 11–18. https://ejournal.undiksha.ac.id/index.php/IJERR/article/view/24080/pdf

Ermawati, F. U., Anggrayni, S., & Isfara, L. (2019). Misconception profile of students in senior high school IV Sidoarjo East Java in work and energy concepts and the causes evaluated using four-tier diagnostic test. *Journal of Physics*: *Conference Series*, *1387*, 012062. https://doi.org/10.1088/1742-6596/1387/1/012062

Fariyani, Q., Rusilowati, A., & Sugianto. (2015). Pengembangan four-tier diagnostic test untuk mengungkap miskonsepsi fisika siswa SMA kelas X. *Journal of Innovative Science Education*, *4*(2), 41–49.

Friatma, A., & Anhar, A. (2019). Analysis of validity, reliability, discrimination, difficulty and distraction effectiveness in learning assessment. *Journal of Physics: Conference Series, 1387*(1), 012063. https://doi.org/10.1088/1742-6596/1387/1/012063

Gurel, D. K., Eryilmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(5), 990–1008. https://doi.org/10.12973/eurasia.2015.1369a

Ibrahim, N. L., Laliyo, L. A. R., Hafid, R., Panigoro, M., & Bumulo, F. (2024). Applying diagnostic assessment with Rasch analysis to measure students' basic understanding of economics. *JETL (Journal of Education Teaching and Learning), 9*(2), 1-7. http://dx.doi.org/10.26737/jetl.v9i2.5376

Koray, C. Ö. & Bal, Ş. (2002). Misconceptions in science teaching and conceptual change strategy. *Gazi University Kastamonu Education Journal, 10*(1), 83-90. https://academicjournals.org/journal/ERR/article-full-text-pdf/ADB2D5B4634

Köse, S. (2008). Diagnosing student misconceptions: Using drawings as a research method. *World Applied Sciences Journal*, *3*(2), 283-293. https://idosi.org/wasj/wasj3(2)/20.pdf

Kurnaz, M. A. & Eksi, C. (2015). An analysis of high school students' mental models of solid friction in physics. *Educational Sciences: Theory and Practice, 15*(3), 787-795. https://jestp.com/article-detail/?id=675

Lailiyah, S. & Ermawati, F. U. (2020). Materi gelombang bunyi: Pengembangan tes diagnostik konsepsi berformat Five-Tier, uji validitas dan reliabilitas serta uji terbatas. *Jurnal Pendidikan Fisika Tadulako Online (JPFT), 8*(3), 104-119.

Nuryanti, S., Masykuri, M., & Susilowati, E. (2018). Analisis Iteman dan model Rasch pada pengembangan instrumen kemampuan berpikir kritis peserta didik sekolah menengah kejuruan. *Jurnal Inovasi Pendidikan IPA*, *4*(2), 224-233. https://doi.org/10.21831/jipi.v4i2.21442

Peşman, H. & Eryilmaz, A. (2010). Development of a three-tier test to assess misconceptions about simple electric circuits. *Journal of Educational Research*, *103*(3), 208-222. https://doi.org/10.1080/00220670903383002

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1301013. https://doi.org/10.1080/2331186X.2017.1301013

Rahma, B., Zaki, M., Boujemaa, A., Nadia, B., & Lhoussaine, M. (2022). University students' knowledge and misconceptions about cell structure and functions. *European Journal of Education Studies*, *9*(10), 121-138. http://dx.doi.org/10.46827/ejes.v9i10.4494

Ramadhani, N. N. & Ermawati, F. U. (2020). Five-tier diagnostic test instrument for uniform circular motion concepts: Development, validity, reliability and limited trials. *Jurnal Pendidikan Fisika*, *9*(1), 4763-4776.

Ramadhan, A. F., & Hidayatullah, R. S. (2023). Analisis kualitas butir soal ujian satuan pendidikan (USP) materi C2 teknik pemesinan kelas XII di SMK PGRI 1 Lamongan melalui model Rasch. *Jurnal Pendidikan Teknik Mesin (JPTM)*, *12*(3), 1–10. https://ejournal.unesa.ac.id/index.php/jurnal-pendidikan-teknik-mesin/article/view/56146

Rusilowati, A. (2015). Pengembangan tes diagnostik sebagai alat evaluasi kesulitan belajar fisika. *Prosiding Seminar Nasional Fisika dan Pendidikan Fisika (SNFPF) ke-6 2015* (pp. 1-10). Program Studi Pendidikan Fisika, Universitas Negeri Semarang.

Sari, E. D. K. & Mahmudi, I. (2024). *Analisis pemodelan Rasch pada assessment pendidikan (Analisis dengan menggunakan aplikasi Winstep*). PT. Pena Persada Kerta Utama.

Silaban, O. R. (2021). *Analisis miskonsepsi siswa pada materi sel sebagai unit terkecil kehidupan di kelas XI IPA SMA Negeri 1 Pagaran T.P 2020/2021.* Undergraduate thesis, Universitas Negeri Medan, Medan.

Streiner, D. L. (2010). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99-103. https://doi.org/10.1207/S15327752JPA8001_18

Sumintono. B, & Widhiarso, W. (2015). *Aplikasi pemodelan RASCH pada assessment pendidikan.* Trim Komunikata.

Suwono, H., Prasetyo, T. I., Lestari, U., Lukiati, B., Fachrunnisa, R., Kusairi, S., Saefi, M., Fauzi, A., & Atho'Illah, M. F. (2021). Cell biology diagnostic test (CBD-Test) portrays pre-service teacher misconceptions about biology cell. *Journal of Biological Education, 55*(1), 1-24. https://doi.org/10.1080/00219266.2019.1643765

Tambo, E. M. Z., Mukaro, J. P., & Mahaso, J. (2003). Some misconceptions on cell structure and function held by a-level biology students: Implications for curriculum development. *Zimbabwe Journal of Educational Research, 15*(2), 122–131. https://opendocs.ids.ac.uk/ndownloader/files/48257272

Tavakol, M., & Dennick, R. (2012). Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE Guide No. 72. *Medical Teacher, 35*(1), e838–e848. https://doi.org/10.3109/0142159X.2012.737488

Tekkaya, C. (2002). Misconceptions as barrier to understanding biology. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 23*(23), 259-266. http://www.efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/971-published.pdf

Wandersee, J. H., Mintzes, J. J., & Novak, J. D. (1994). Research on alternative conceptions in science. In D. L. Gabel (Eds.), *Handbook of research on science teaching and learning* (pp. 177-210). McMillan.

Zimmerman, D. W. & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of *multiple*-choice tests. *Applied Psychological Measurement*, *27*(5), 357-371. https://doi.org/10.1177/0146621603254799

Zulfianto, M. R. & Abduh, M. (2023). How do science content misconceptions occur in primary school teachers with teaching certificates?. *Mimbar Sekolah Dasar*, *10*(3), 595-613. https://doi.org/10.53400/mimbar-sd.v10i3.62887