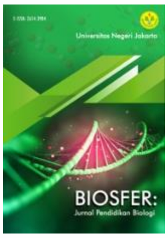
Biosfer: Jurnal Pendidikan Biologi

Journal homepage: http://journal.unj.ac.id/unj/index.php/biosfer



Analysis items of the four-tier immune system multiple choice test instrument using rasch analysis

Feni Andriani*, **Meti Indrowati, Bowo Sugiharto**

Biology Education, Faculty of Teacher Training and Education, Universitas Sebelas Maret, Indonesia

*Corresponding author: feniandrianibio1@student.uns.ac.id

ARTICLE INFO

Article history
Received: 16 November 2020
Revised: 24 April 2021
Accepted: 30 April 2021

Keywords:
Four-tier
Instruments
Immune system
Misconceptions
Rasch analysis

ABSTRACT

The purpose of this study was to analyze the feasibility of the items of the four-tier multiple-choice test immune system instrument that had been developed. The development of the instrument using the Treagust (1988) model, namely defining content, collecting student misconceptions information, and developing a diagnostic test. A total of 25 items have been developed. The results of the instrument development were tested on 142 students of grade XI from several high schools in Surakarta who were selected by simple random sampling. The data analysis technique was performed using Rasch analysis in the Winstep application. The results of the construct validity test showed items number 5, 7, and 9 did not fit the validity standards. The reliability test shows that the value of Cronbach Alpha reliability is bad (n = 0.51), the value of the reliability item is special (no = 0.97), the value of person reliability is sufficient (n = 0.68), the value of person separation is weak (n = 1.44), and the item separation value is special (n = 5.38). The person discrimination test showed student 056P31 has the highest ability and student 098P51 has the lowest ability. The item discrimination test shows item number 1 is the best item and the bad item is number 14. The item difficulty analysis showed less proportionality because there were too many items in the easy and difficult categories. An expansion of the sample is needed to see a more comprehensive and diverse range of responses to instruments.

Andriani, F., Indrowati, M., & Sugiharto, B. (2021). Analysis items of the four-tier immune system multiple choice test instrument using rasch model. *Biosfer: Jurnal Pendidikan Biologi, 14*(1), 99-119. https://doi.org/10.21009/biosferjpb.18020

INTRODUCTION

Data on the results of the National Examination of Senior High School level in Indonesia from 2015 to 2019 shows that the average Biology score is less than the minimum completeness criteria of the national exam, which is 55.00 (The Center for Educational Assessment, Ministry of Education and Culture, 2019). According to The Center for Educational Assessment, Ministry of Education the average score for the Biology National Exam score for the last five years (2015-2019) was 64.48; 55.89; 46.56; 45.30; and 47.36. The proportion of immune system question items in the Biology National Exam in 2015 and 2019 was 2.5% of the total questions. The results of the indicator analysis showed that the immune system material had always been the material with the lowest percentage of students who answered correctly and never even reached 26%. In the 2019 national exam, the immune system had a percentage of students who answered correctly only 25.19%, while in 2015 the percentage of students who answered correctly was only 24.90%. From 2016 to 2018 the immune system was not included in the National Exam material (The Center for Educational Assessment, Ministry of Education and Culture, 2019).

The low percentage of students' ability to answer correctly on the immune system material is caused by many things, including difficulty understanding the concept of the immune system material (Al-zoubi, 2015). The cause of students' difficulty understanding the concept of the immune system according to Faggioni et al., (2019) can be caused by the difficulty in visualizing molecular phenomena, complex immune system material, and inappropriate teaching methods (Lazarowitz & Penso, 1992). Siqueira-batista et al., (2009) and Su, Cheng, & Lin, (2014) stated that other causes of students' difficulty understanding immune system material are due to the use of foreign and specific terms, development of immune system material, lack of basic knowledge about the immune system at previous educational levels, and learning duration that is too short. One of the effective teaching methods to teach the immune system in a short time is the just-in-time teaching (JITT) method, which is a form of flip classroom. This method is effective because it can shorten the time by giving students homework to learn material concepts at home and discuss the results and solve the problems in the classroom (Stranford, Owen, Mercer, & Pollock, 2020).

Difficulty understanding of the immune system material does not only occur in Indonesia, but also in other countries such as Malaysia, Taiwan, and America (Lukin, 2013; Su et al., 2014; Subari, 2017). Novice students in Malaysia find it difficult to understand the material on the immune system due to limited knowledge and misinterpretation of immune system terms and phenomena (Subari, 2017). In Taiwan, incomplete and unfinished concepts, abstract and interconnected phenomena, have been triggered a lack of understanding of the immune system so that is still below average (Su et al., 2014). Students need to learn and understand the basic concept of the immune system not only to improve learning outcomes but also useful in daily life applications such as maintaining a healthy body.

Difficulty understanding the concept of material can cause students to experience misconceptions (Assaraf, Dodick, & Tripto, 2013; Hasyim, Suwono, & Susilo, 2018; Stylos, Evangelaksis, & Kotsis, 2008; Tekkaya, 2002). According to Hammer (1996), a misconception is a change in students' understanding of a scientific concept that is different from the actual scientific concept. Tekkaya (2002) uses the term misconception to explain each student's thoughts on a concept that is contrary to the concept that has been described by the experts. The relationship between difficulties in understanding a concept and misconceptions according to Senocak, Taskesenligil, & Sozbilir, (2007); and Stylos et al., (2008) is when students have difficulty understanding a concept, students tend to build concepts based on their perceptions. The concept that students build comes from the results of student interactions with the environment. Condition suitability causes students to continue maintaining a concept even though the truth is not true.

According to the National Research Council (1997) based on the source, misconceptions are divided into several types, namely preconceived notions, nonscientific beliefs, conceptual misunderstanding, vernacular misconceptions, and factual misunderstandings (Patil, Chavan, & Khandagale, 2019; Sarimanah, Dewi, & Sabri, 2019). Based on other literature, misconceptions are divided into factual misconceptions and oncology misconceptions (Verkade et al., 2017). Factual misconceptions are misconceptions that occur due to misinformation received as a result of interactions with everyday environments, such as false information on social media. Oncology misconceptions are misconceptions that occur as a result of personal experiences in a phenomenon.

Oncology misconceptions are more difficult to change (Verkade et al., 2017). Other literature distinguishes misconceptions into inaccurate misconceptions and incommensurate misconceptions (Chi, 2013). Inaccurate misconceptions are divided into false beliefs and flawed mental models. Incommensurate misconception divided into category mistakes and missing schema.

Misconceptions found in students about the immune system material include diseases caused by microbes, contact with people who are sick means getting the disease directly, diseases such as coughs and fever only occur due to exposure of hot or cold conditions, and the use of antibiotics can overcome all types of the diseases that are caused by microbes (Allen, 2014; Kurt, 2013; Subari, 2017). Students need to avoid misconceptions and form concepts correctly according to scientific explanations (Akamca, Ellez, & Hamurcu, 2009; Suliyanah, Putri, & Rohmawati, 2018). Toka & Askar (2002) stated that misconceptions affect the low achievement of student learning outcomes and become a barrier for students to learn the next material (Suliyanah et al., 2018). According to Singh (2016), misconceptions can also interfere with students' ability to solve problems and develop scientific reasons. Misconceptions are difficult to change because they are profound, stable, and believed to be true (Adeniyi, 1985; Fisher, 1985). The potential for misconceptions that can cause negative impacts needs to be known early so that a test is needed to immediately diagnose misconceptions in students (National Research Council, 1997).

A four-tier multiple-choice test (4TMCT) has been developed to address the deficiencies of previously developed misconception tests. The four-tier multiple-choice test according to Caleon & Subramaniam (2010) is divided into four stages. The first stage contains multiple choices, the second stage contains the belief in the choice of the first answer, the third stage shows the reasons for the answers chosen in the first stage, and the fourth stage shows the confidence in the answers in the third stage. The advantages of the four-tier multiple-choice test are that it can show differences in the concept and understanding of each student, and can distinguish incorrect student answers due to misconceptions or lack of understanding of the material (Caleon & Subramaniam, 2010a; Gurel, Eryılmaz, & Mcdermott, 2015; Milenkovic, Hrin, Segedinac, & Horvat, 2016; Pujyanto et al., 2018; Sarimanah et al., 2019). The drawback of a four-tier multiple-choice test is that it only takes a long time to make test materials (Gurel et al., 2015; Sarimanah et al., 2019).

Rasch is one of the techniques used to develop instruments such as surveys and tests, monitor the quality of the instruments, and calculate the performance of respondents (Boone, 2016). Rasch analysis can provide outcome measures that offer a guide for researchers in interpreting the quality of the instruments and research subjects (Linacre, 2017). The advantages of Rasch are being able to compare everything consistently, and knowing the right number of samples to be used for research so that the data becomes more valid (Boone, 2016; Linacre, 2017; Sumintono & Widiharso, 2015).

Based on the ability of the four-tier multiple-choice test to measure misconceptions, a Four-Tier Immune System Multiple Choice Test (FTISTMCT) instrument was developed. The test was developed specifically to identify high school student's misconceptions and understanding of the immune system material. The whole stage is designed to see students' understanding, reasoning, and belief in immune system material. The language and level of the material are adjusted to the high school material. The structure and number of questions are adjusted so that students do not get bored with doing the test which can affect the results. The FTISTMCT development was analyzed using Rasch to obtain consistent results. The formulation of the problem in this study is: how are the results of the analysis of the four-tier immune system multiple choice test using Rasch analysis?

METHODS

Research design

The development model used in this study follows the development model conducted by Treagust (1988). The Treagust model was chosen because the stages were specifically developed to identify students' understanding of alternative conceptions. This model has also been widely used in developing misconception tests and other similar tests (Chandrasegeran, Treagust, & Mocerino, 2007). The development procedure is divided into three main stages, namely (1) defining content, (2) gathering information about student misconceptions, and (3) developing diagnostic instruments. The defining content stage consists of four sub-stages, namely (a) Identifying content proportion upon scientific statements, (b) developing a concept map, (c) linking content proportions to the concept map, and (d) content validation. The stage of gathering information about student misconception consists of three sub-stages, namely, (a) examining literature related to misconceptions, (b) conducting student interviews, and (c) developing multiple-choice items. The stage of developing diagnostic instruments consists of three sub-stages, namely, (a) developing a four-tier immune system multiple choice test, (b) ensuring the feasibility of the instrument, and (c) continuing the improvement. These details of the stages will explain in the procedure.

Population and sample

The population consists of grade XI students who come from both public senior high schools and private senior high schools in Surakarta. The total population is 220 people. The sample was selected by a simple random sampling technique. The samples were determined using the Slovin formula with a significance level of 5%. The calculation results show that the number of research samples used is 142 samples. The reasons for using simple random sampling are because the large population and the entire population are considered to have the same basic knowledge in the immune system material. After all, they were studying immune system material in the classroom before the test is carried out.

Procedure

1. Defining content

a. Identifying content proportion

This sub-stage was carried out based on the curriculum and syllabus (Chandrasegeran et al., 2007; Hasyim et al., 2018; Treagust, 1988). The curriculum that is being used in the learning process is the 2013 revision of the 2017 curriculum. Analysis of the principles of the immune system and cognitive outcomes results in the development of indicators. All material on indicators is taught during three weeks of learning meeting with a total of nine hours of lesson time. The indicators developed are shown in Table 1.

Table 1
Indicators of Immune System Material

No.	Indicator
3.14.1.	Distinguish antibodies and antigens.
3.14.2.	Determine the types of antibodies.
3.14.3.	Criticize how antibodies work.
3.14.4.	Determine the mechanism of action of antibodies.
3.14.5.	Describe the non-specific and specific mechanism of action of the body's immune system.
3.14.6.	Categorizes the mechanism of action of the body's specific and non-specific immune systems.
3.14.7.	Analyze statements regarding the mechanism of the body's immune systems.
3.14.8.	Describes the humoral and cellular immune systems.
3.14.9.	Categorizes the immunity pathways.
3.14.10.	Ordering inflammatory events
3.14.11.	Explain how immunization works.
3.14.12.	Distinguish the concept of vaccination and immunization.
3.14.13.	Determine infectious and non-infectious diseases
3.14.14.	Determine the type of disorder/disease in the body's immune system.
3.14.15.	Determine the factors that affect immunity.
3.14.16.	Describe the cells involved in the components of the body's immune system.

b. Developing concept map

Done to consider or determine the content of the material to be developed. The process of developing a concept map was based on the textbooks used by students. The use of student textbooks was used as the basis for drafting concept maps to see the boundaries of the learning material students are learning. Books that are generally used are the Biology Book Exploring the World for 11th Grade High School Students (*Buku Biologi Menjelajah Dunia*), published by Tiga Serangkai, and the Biology

Book for 11th Grade High School Students (*Buku Biologi*), published by Erlangga. The concept map is in Appendix 1.

c. Linking indicators with concept map

done to see whether the selected material is consistently used in learning material for the immune system.

d. Conducting content validation

this was a validation stage carried out by experts and practitioners to see the extent to which the selected material indicators are valid and suitable for misconceptions in the immune system. Validation of the content of expert material is carried out by doctors or masters who have a field of science in immunology, evaluation of biology learning. A practitioner is a biology teacher who has at least 10 years of teaching experience.

2. Gathering information of students' misconception

a. Checking literature

Which deals with misconceptions, including literature related to student learning outcomes. The process of examining literature related to misconceptions begins by looking at the results of the students' National Biology Exam scores for the past 5 years. The literature on the results of students' National Biology Exam scores can be accessed through the website page of The Center for Educational Assessment, Ministry of Education and Culture (*Puspendik Kemendikbud*). The process of collecting misconception literature is also carried out by looking at related research in journals.

b. Gathering students misconception information

This was done to see the students' initial understanding. Information gathering was conducted using unstructured interviews. The information-gathering process was carried out when the COVID-19 pandemic broke out so that schools were closed and classroom learning was carried out online. The information-gathering followed government recommendations so that it was done through the WhatsApp voice note application with the help of the Google Form. Students were given 10 questions about the immune system adapted from (Astutik, 2018). Interviews were conducted to explore the answers to each question answered by students.

c. Developing multiple-choice items

This was carried out based on the results of gathering information on student misconceptions by compiling multiple-choice questions equipped with reasons for choosing the answer choices, then after developing multiple-choice, the material is tested on students.

3. Developing diagnostic instrument

a. Developing four-tier multiple-choice immune system test

This step was done by adapting the format to the development of the Four-tier multiple-choice test that has been developed by the expert (Anggrayni & Ermawati, 2019; Caleon & Subramaniam, 2010; Maharani et al., 2019; Pujyanto et al., 2018).

b. Ensuring instrument eligibility

Conducted by analyzing the feasibility of the validity test items, reliability test, difference power test, and items difficulty level test.

c. Continuing improvement

The results of the feasibility test for instruments that unsuitable were then corrected and the final results are tested again on the students. Examples of test development results are shown in Appendix 2. The research procedure is shown in Figure 1.

Data Analysis Techniques

The research data consisted of item analysis data and student understanding data. Item analysis data were in the form of content and construct validation data, reliability data, difference power data, and item difficulty level data. The item analysis data were analyzed using Rasch analysis in the Winstep application.

1. Data analysis of content validation by experts

Data from the expert validity test were analyzed using the formula:

$$\mathbf{N = \frac{JP}{JPT} \times 100}$$

(Anggrayni & Ermawati, 2019)

Note :

N	= Score
JP	= Number of points
JPT	= Total Number of Points

The score categories can be seen in Table 2.

Table 2

Expert Validation Data Result Category

No.	Score	Category
1.	100 ≥ 76	The instrument is feasible to be applied without revision
2.	56-75	The instrument is feasible to be applied with a few revisions
3.	41-55	The instrument still needs a lot of improvement with notes and revisions.
4.	≤ 40	The instrument cannot be applied

(Source : Astutik, 2018; Rahman et al., 2018; Zaini & Rusmini, 2016)

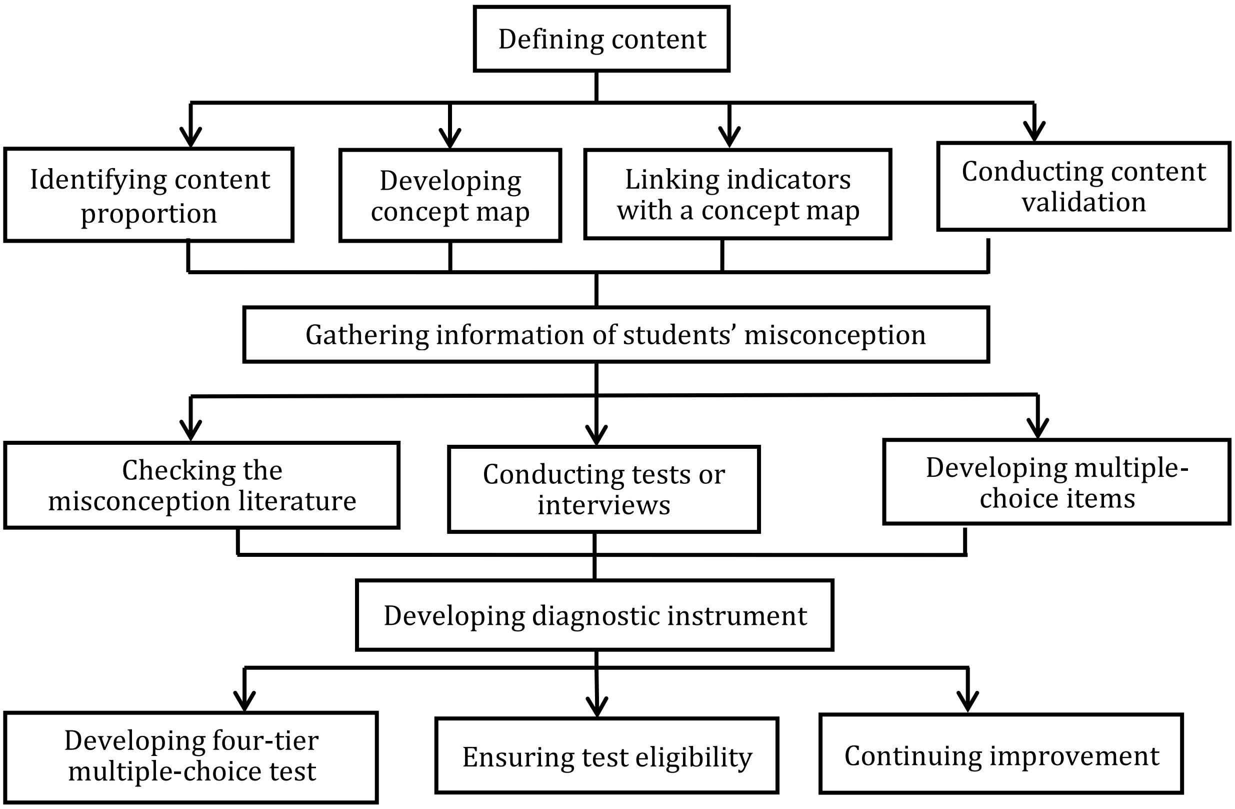

Defining content
Identifying content proportion
Developing concept map
Linking indicators with a concept map
Conducting content validation
Gathering information of students' misconception
Checking the misconception literature
Conducting tests or interviews
Developing multiple-choice items
Developing diagnostic instrument
Developing four-tier multiple-choice test
Ensuring test eligibility
Continuing improvement

Figure 1. Development Procedure (Source: Treagust, 1988)

2. Data analysis of interview about student misconception

Interview data were analyzed descriptively to explain students' understanding of choosing answers and the reasons for choosing answers. Interview data were also used to see how students responded to the items and about students' beliefs in determining answer choices. The results of the interview analysis were used to develop multiple-choice as part of the four-tier multiple-choice test instrument component.

3. Data analysis of instrument

a. Grouping and Coding of Student Answers

The data analysis process began with classifying students' answers based on their level of understanding. Mapping data on students' answers are divided into Understand the Concept, Less Understand the Concept, Misconception, Misunderstanding the Concept, and Incomplete Answers (Anggrayni & Ermawati, 2019; Kaltacki, 2012; Suliyanah et al., 2018). The results of grouping the level of understanding of students were then grouped in the form of coding (Fratiwi, Ramalis, & Samsudin, 2019; Kaltacki, 2012; Septiantini et al., 2020). The coding was used for data analysis using the winstep application. Guidelines for grouping and coding the level of understanding of students are shown in Table 3.

Table 3
Guidelines for Grouping Student Understanding Levels

Concept Level	First Level (tier 1)	Second Level (tier 2)	Third Level (tier 3)	Fourth Level (tier 4)	Code
Understand the concept	Correct	Sure	Correct	Sure	4
Less understand the concept	Correct	Sure	Correct	Not Sure	3
	Correct	Not Sure	Correct	Sure	
	Correct	Not Sure	Correct	Not Sure	
	Correct	Sure	Incorrect	Sure	
	Correct	Sure	Incorrect	Not Sure	
	Correct	Not Sure	Incorrect	Sure	
	Correct	Not Sure	Incorrect	Not Sure	
	Incorrect	Sure	Correct	Sure	
	Incorrect	Sure	Correct	Not Sure	
	Incorrect	Not Sure	Correct	Sure	
	Incorrect	Not Sure	Correct	Not Sure	
Misconception	Incorrect	Sure	Incorrect	Sure	2
Misunderstand the concept	Incorrect	Sure	Incorrect	Not Sure	1
	Incorrect	Not Sure	Incorrect	Sure	
	Incorrect	Not Sure	Incorrect	Not Sure	
Incomplete answer					0

b. Data Processing of Student Answer

The data that had been coded were then analyzed using the Winstep application to obtain data on the validity, reliability, discrimination, and item difficulty level. Guidelines for data analysis of validity test results adjusted based on the value of Outfit Mean Square (MNSQ), Outfit Z Standard (ZSTD), and Point Measure Correlation (Pt Mean Corr). Sumintonoet al. (2014) had grouped item analysis criteria to show valid items. Classification of the item analysis criteria is shown in Table 4.

Table 4
Items Validity Criteria

Criteria	Score
Outfit Mean Square (MNSQ)	0,5 < Nilai MNSQ < 1,5
Outfit Z Standard (ZSTD)	-2,0 <Nilai ZSTD <2,0
Point Measure Correlation (Pt Mean Corr).	0,4 < Nilai Pt mean corr < 0,85

(Source: Sumintono et al., 2014)

The reliability score was seen from the Cronbach Alpha score. The reliability score was also seen from the person reliability (sample reliability and item reliability). Data reliability test results according to Sumintono et al., (2014) were grouped into five levels; weak, sufficient, good, excellent, and special. The table for grouping the reliability values is in Table 5 and Table 6.

Table 5
Reliability Criteria Based on Alpha Cronbach Score

Score	Category
Alpha Cronbach score > 0,80	Excellent
0,70 < Alpha Cronbach score ≤ 0,80	Good
0,60 < Alpha Cronbach score ≤ 0,70	Sufficient
0,50 < Alpha Cronbach score ≤ 0,60	Bad
Alpha Cronbach score ≤ 0,50	Worse

(Source: Sumintono et al., 2014)

Table 6
Item Reliability and Person Reliability Criteria

Score	Category
> 0,94	Special
0,91- 0,94	Excellent
0,81 - 0,90	Good
0,67 - 0,80	Sufficient
≤ 0,67	Weak

(Source: Sumintono et al., 2014)

Reliability according to Sumintono et al., (2014) and Sumintono & Widiharso, (2015) could also be seen from the value of item separation and person separation. Apart from that, item separation and person separation can also be used to see the discrimination. The criteria for item separation and person separation are grouped into weak, sufficient, good, excellent, and special (Sumintono et al., 2014). Interval grouping of item separation and person separation criteria is shown in Table 7.

Table 7
Item Separation and Person Separation Criteria

Score Interval	Criteria
< 2	Weak
2 - 3	Sufficient
3 - 4	Good
4 - 5	Excellent
> 5	Special

(Source: Sumintono et al., 2014)

The item difficulty level could be seen based on the item measure output table. The item difficulty level was divided into four categories based on the logit score. Grouping of difficulty score is shown in Table 8.

Table 8.
Difficulty Level Criteria

Measure Score (Logit)	Category
measure score < –1	Very Easy
–1 ≤ measure score < 0	Easy
0 ≤ measure score ≤1	Difficult
measure score > 1	Very Difficult

(Source: Sumintono et al., 2014)

RESULTS AND DISCUSSION

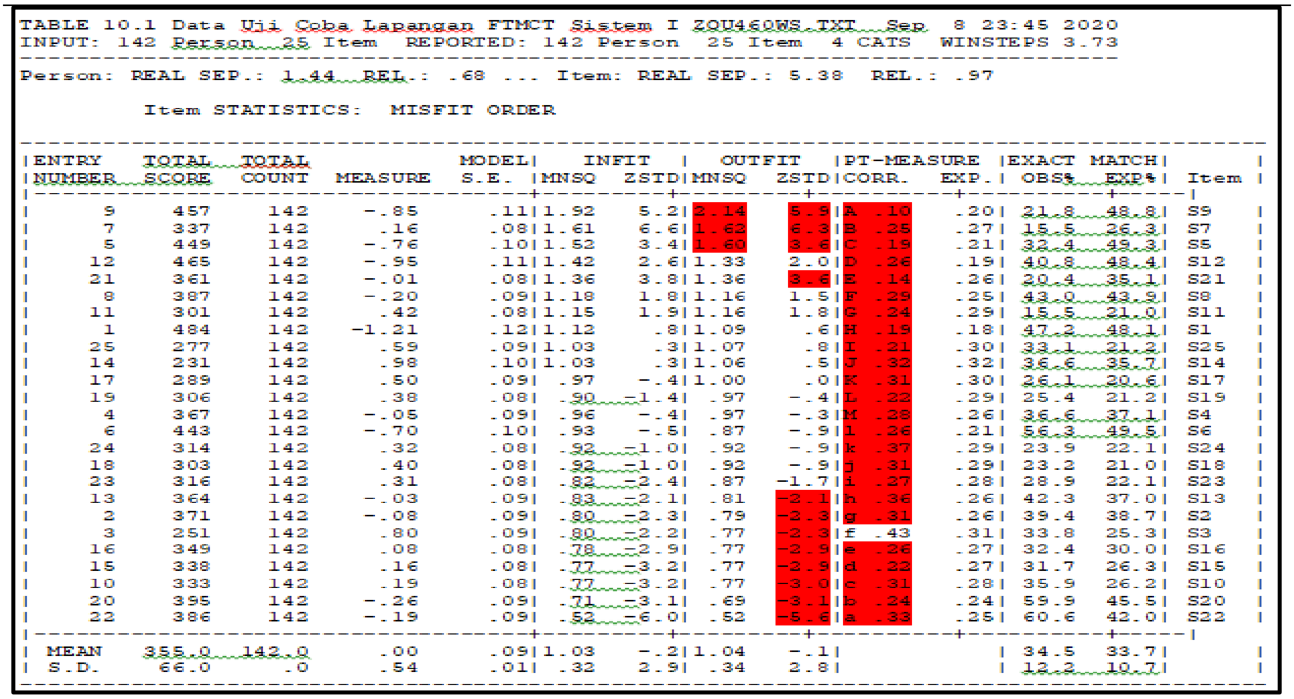
a) Validation

The results of the validity test according to Riyantono & Hatmawan (2020) are strongly influenced by the test sample so that they are not universal. The instrument can be valid on the object or research subject at a certain place and at a certain time, but if tested at different times it is possible to obtain different validity results. The purpose of the validity test is to test whether a test can properly test the sample to be tested. The instrument in the form of a test according to Mamik (2015) must meet content validity and construct validity.

The results of the validation from physiology and immunology experts provide an overview of the accuracy of the immune system concepts compiled in the narrative statement and answer choices at level 1 and level 3 with the concept of immunity according to experts. Educational evaluation experts provide reviews at the cognitive level and question design, including suggestions to reduce the choice of answers to the confidence level of answers (tier 2) and the reason confidence level (tier 4). The reduction of answer choices considers the effectiveness of each answer choice by scoring guidelines. The answer choices at level 2 and level 4 which initially consisted of six answer choices were then reduced to four choices, namely, (a) very unsure, (b) unsure, (c) sure, and (d) very sure. The results of content validation by material experts are shown in Table 9.

Table 9
Validation score from experts

Evaluator	Score	Category
Human Physiology Expert	95.4	The instrument is feasible to be applied without revision
Virology and Microbiology Expert	68.1	The instrument is feasible to be applied with a few revisions
Biology Learning Evaluation Expert 1	95	The instrument is feasible to be applied without revision
Biology Learning Evaluation Expert 2	92.5	The instrument is feasible to be applied without revision
Biology Teacher 1	96	The instrument is feasible to be applied without revision
Biology Teacher 2	71.6	The instrument is feasible to be applied with a few revisions



```
TABLE 10.1 Data Uji Coba Lapangan FTMCT Sistem I ZOU460WS.TXT  Sep  8 23:45 2020
INPUT: 142 Person  25 Item  REPORTED: 142 Person  25 Item  4 CATS  WINSTEPS 3.73
------------------------------------------------------------------------------------
Person: REAL SEP.: 1.44  REL.: .68 ... Item: REAL SEP.: 5.38  REL.: .97

        Item STATISTICS:  MISFIT ORDER

-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PT-MEASURE |EXACT  MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----------+-----------+------|
|     9    457    142    -.85    .11|1.92   5.2|2.14   5.9|A .10   .20| 21.8  48.8| S9   |
|     7    337    142     .16    .08|1.61   6.6|1.62   6.3|B .25   .27| 15.5  26.3| S7   |
|     5    449    142    -.76    .10|1.52   3.4|1.60   3.6|C .19   .21| 32.4  49.3| S5   |
|    12    465    142    -.95    .11|1.42   2.6|1.33   2.0|D .26   .19| 40.8  48.4| S12  |
|    21    361    142    -.01    .08|1.36   3.8|1.36   3.6|E .14   .26| 20.4  35.1| S21  |
|     8    387    142    -.20    .09|1.18   1.8|1.16   1.5|F .29   .25| 43.0  43.9| S8   |
|    11    301    142     .42    .08|1.15   1.9|1.16   1.8|G .24   .29| 15.5  21.0| S11  |
|     1    484    142   -1.21    .12|1.12    .8|1.09    .6|H .19   .18| 47.2  48.1| S1   |
|    25    277    142     .59    .09|1.03    .3|1.07    .8|I .21   .30| 33.1  21.2| S25  |
|    14    231    142     .98    .10|1.03    .3|1.06    .5|J .32   .32| 36.6  35.7| S14  |
|    17    289    142     .50    .09| .97   -.4|1.00    .0|K .31   .30| 26.1  20.6| S17  |
|    19    306    142     .38    .08| .90  -1.4| .97   -.4|L .22   .29| 25.4  21.2| S19  |
|     4    367    142    -.05    .09| .96   -.4| .97   -.3|M .28   .26| 36.6  37.1| S4   |
|     6    443    142    -.70    .10| .93   -.5| .87   -.9|l .26   .21| 56.3  49.5| S6   |
|    24    314    142     .32    .08| .92  -1.0| .92   -.9|k .37   .29| 23.9  22.1| S24  |
|    18    303    142     .40    .08| .92  -1.0| .92   -.9|j .31   .29| 23.2  21.0| S18  |
|    23    316    142     .31    .08| .82  -2.4| .87  -1.7|i .27   .28| 28.9  22.1| S23  |
|    13    364    142    -.03    .09| .83  -2.1| .81  -2.1|h .36   .26| 42.3  37.0| S13  |
|     2    371    142    -.08    .09| .80  -2.3| .79  -2.3|g .31   .26| 39.4  38.7| S2   |
|     3    251    142     .80    .09| .80  -2.2| .77  -2.3|f .43   .31| 33.8  25.3| S3   |
|    16    349    142     .08    .08| .78  -2.9| .77  -2.9|e .26   .27| 32.4  30.0| S16  |
|    15    338    142     .16    .08| .77  -3.2| .77  -2.9|d .22   .27| 31.7  26.3| S15  |
|    10    333    142     .19    .08| .77  -3.2| .77  -3.0|c .31   .28| 35.9  26.2| S10  |
|    20    395    142    -.26    .09| .71  -3.1| .69  -3.1|b .24   .24| 59.9  45.5| S20  |
|    22    386    142    -.19    .09| .52  -6.0| .52  -5.6|a .33   .25| 60.6  42.0| S22  |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN   355.0  142.0     .00    .09|1.03   -.2|1.04   -.1|           | 34.5  33.7|      |
| S.D.    66.0     .0     .54    .01| .32   2.9| .34   2.8|           | 12.2  10.7|      |
-------------------------------------------------------------------------------------------
```

Figure 2. Results Data of FTISMCT Construct Validity Test

The data from the validity analysis shows that from the 25 questions developed, three questions do not meet the three validity criteria, namely questions number 5, 7, and 9. Questions number 5, 7, and 9 have scores of the MNSQ outfit and the ZSTD outfit that are greater than the standard criteria and have a lower score of Pt. Correct Measure. The scores of the MNSQ outfit for the three questions are 1.60; 1.62; and 2.14 while the ZSTD outfit scores for the three questions are 3.6; 6.3; and 5,9. The three questions each have a Pt. Measure Corr score of 0.19; 0.25; and 0.10.

Another question number that also has a ZSTD outfit score greater than the standard criteria is question number 21 with a score of 3.6. Eight questions have ZSTD outfit scores lower than the criteria, namely questions number 2, 3, 10, 13, 15, 16, 20, and 22. The eight-question numbers have a ZSTD outfit scores of -2.3; -2.3; -2.6; -3.0; -2.1; -2.9; -2.9; -3.1; and -5.6. According to Khine (2020), the score of the ZSTD outfit can be neglected as long as the MNSQ outfit scores are within the range of scores that match the criteria. The negligence of the ZSTD outfit score according to Sumintono et al. (2014) is due to the strong influence of the sample size upon ZSTD outfit score. The larger the sample size, the greater the result of the ZSTD outfit score.

The analysis results of Pt. Measure Corr score from 25 questions shows that only one question which has Pt. Measure Corr scores match the criteria, which is question number 3 with a score of 0.43. The other 24 questions do not meet the standard criteria of Pt. Measure Corr and have a Pt. Measure

Corr score range of 0.10 to 0.37.

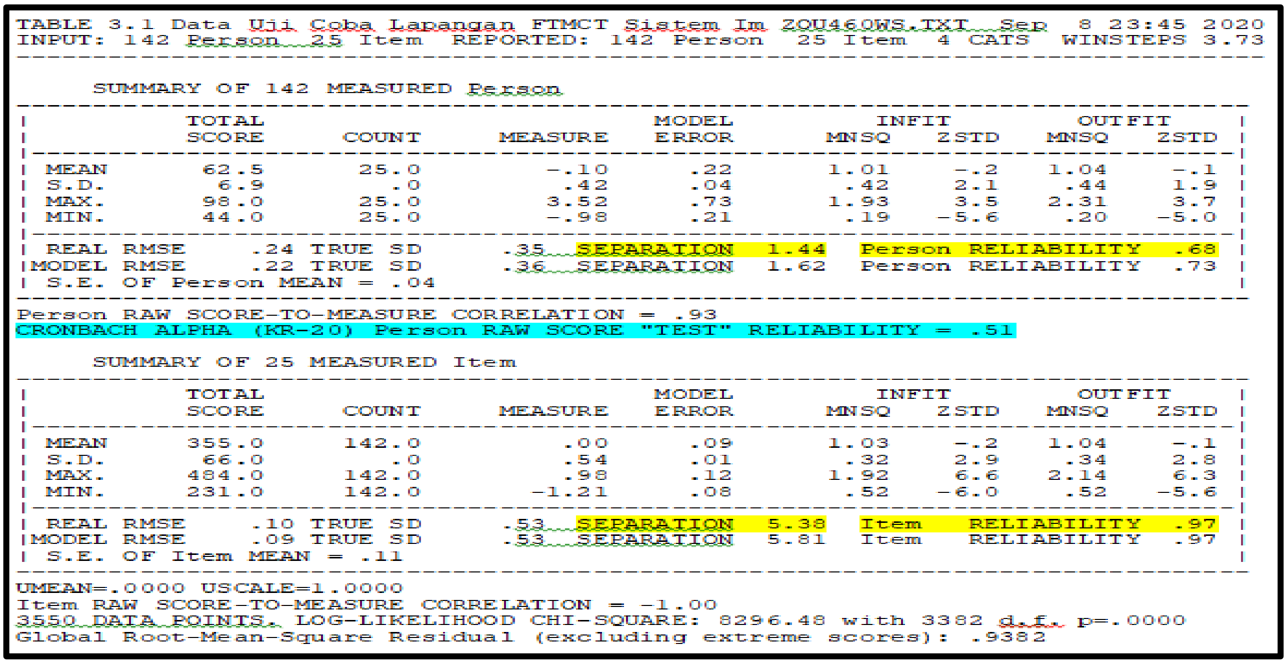
The score of Pt. Measure Corr refers to the relationship between the difficulty of each individual with the overall difficulty of the test (Khine, 2020b). Good criteria of Pt. Measure Corr according to Smiley (2016) can differentiate the ability of each student to answer questions. If the score of Pt. Measure Corr is more than 1, which indicates that the item is better at differentiating the abilities of each student. Zero scores on Pt. Measure Corr shows that there is no clear relationship between the response of a particular item and the overall test, in other words, whether students choose the right or wrong answer is random. Negative scores indicate defective test items because students with lower ability can get high scores on difficult items. A negative score of Pt. Measure Corr indicates that the component of the question must be checked again to keep it used or removed from the test component (Smiley, 2016).

Validity test data in the Pt. Measure Corr shows that no question has a negative score of Pt. Measure Corr, all items have positive scores. No negative items are indicating that there are no defective items. Lower Pt Measure Corr score indicates that each item on the FTISMCT is less sensitive in differentiating the abilities of each student.

The conclusion from the validity test results is that three items must be reduced because they do not meet all the criteria so they are invalid for measuring the sample. The three questions are questions number 5, 7, and 9. The cause of the three questions that are not fit to measure can be due to various factors such as the special knowledge of each individual in the sample, the sample guessing answers, errors in scoring or data entry, or various random factors that affect the data (Boone, Noltemeyer, Boone, & Noltemeyer, 2017; Linacre, 1999, 2017).

b) Reliability

Reliability analysis is seen from the output summary statistical table in the Cronbach Alpha value section. The results of the reliability test are shown in Figure 3.



```
TABLE 3.1 Data Uji Coba Lapangan FTMCT Sistem Im ZOU460WS.TXT  Sep  8 23:45 2020
INPUT: 142 Person  25 Item  REPORTED: 142 Person  25 Item  4 CATS  WINSTEPS 3.73
--------------------------------------------------------------------------------

     SUMMARY OF 142 MEASURED Person
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT          OUTFIT    |
|          SCORE      COUNT    MEASURE   ERROR     MNSQ   ZSTD    MNSQ   ZSTD  |
|-----------------------------------------------------------------------------|
| MEAN      62.5       25.0      -.10     .22      1.01    -.2    1.04    -.1  |
| S.D.       6.9         .0       .42     .04       .42    2.1     .44    1.9  |
| MAX.      98.0       25.0      3.52     .73      1.93    3.5    2.31    3.7  |
| MIN.      44.0       25.0      -.98     .21       .19   -5.6     .20   -5.0  |
|-----------------------------------------------------------------------------|
| REAL RMSE    .24 TRUE SD     .35  SEPARATION  1.44  Person RELIABILITY  .68  |
|MODEL RMSE    .22 TRUE SD     .36  SEPARATION  1.62  Person RELIABILITY  .73  |
| S.E. OF Person MEAN = .04                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .93
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .51

     SUMMARY OF 25 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT          OUTFIT    |
|          SCORE      COUNT    MEASURE   ERROR     MNSQ   ZSTD    MNSQ   ZSTD  |
|-----------------------------------------------------------------------------|
| MEAN     355.0      142.0       .00     .09      1.03    -.2    1.04    -.1  |
| S.D.      66.0         .0       .54     .01       .32    2.9     .34    2.8  |
| MAX.     484.0      142.0       .98     .12      1.92    6.6    2.14    6.3  |
| MIN.     231.0      142.0     -1.21     .08       .52   -6.0     .52   -5.6  |
|-----------------------------------------------------------------------------|
| REAL RMSE    .10 TRUE SD     .53  SEPARATION  5.38  Item   RELIABILITY  .97  |
|MODEL RMSE    .09 TRUE SD     .53  SEPARATION  5.81  Item   RELIABILITY  .97  |
| S.E. OF Item MEAN = .11                                                     |
-------------------------------------------------------------------------------
UMEAN=.0000 USCALE=1.0000
Item RAW SCORE-TO-MEASURE CORRELATION = -1.00
3550 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 8296.48 with 3382 d.f. p=.0000
Global Root-Mean-Square Residual (excluding extreme scores): .9382
```

Figure 3. Results Data of FTISMCT Reliability Test

The reliability test data shows the Cronbach Alpha score of the instrument is 0.51 which means bad. The item reliability score is 0.97 which is classified as special and the person reliability score is 0.68 which is classified as sufficient. According to Sumintono et al. (2014), reliability can be seen from the score of item separation and person separation. The test results show that the person separation score is 1.44 which is classified as weak, while the item separation score is 5.38 which is special. The data of person separation and item separation can be seen in Figure 3. High item reliability indicates the adequacy of the sample size with excellent quality, meaning that the sample size is very representative, while the low reliability of people indicates the insufficiency of items targeting the range of ability levels assessed (Khine, 2020a; Sumintono & Widiharso, 2015). Inadequacy of targeting the level of understanding because of the four expected levels of understanding, most samples only dominate at one level of understanding, for example, misconceptions in number 9, or understanding concept number 1.

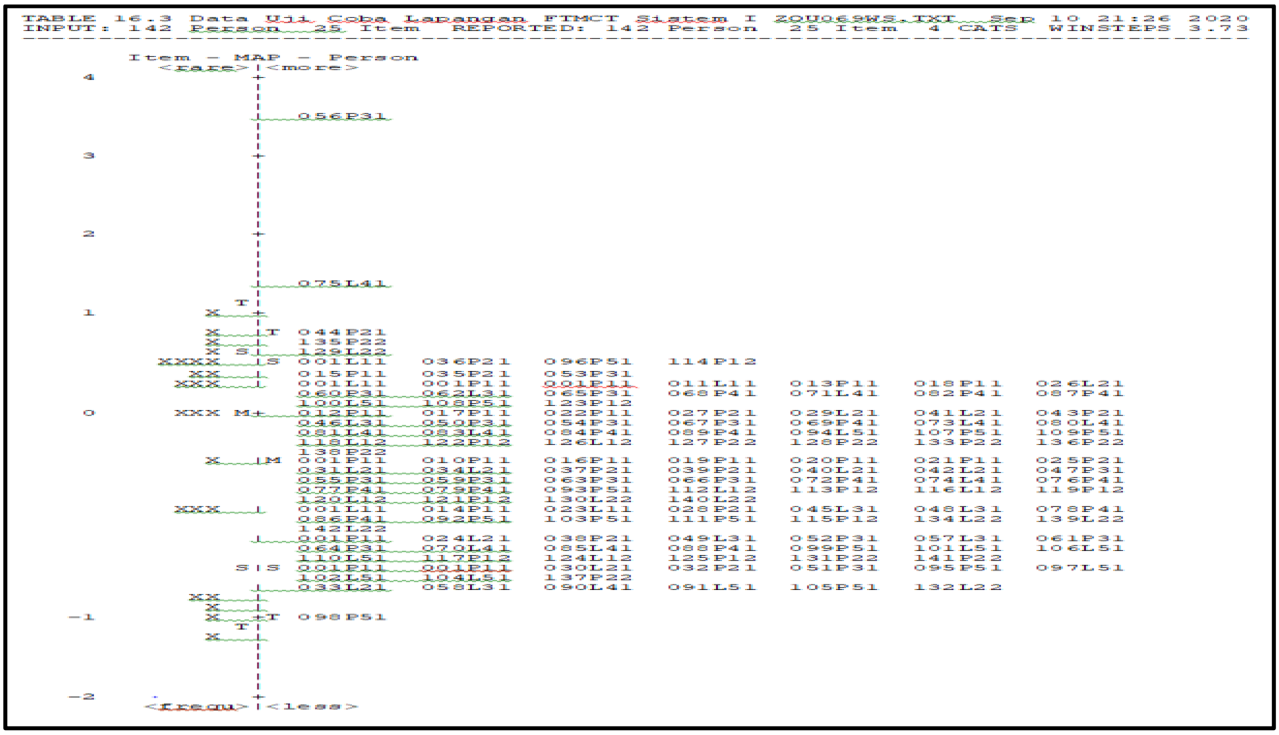
The higher the item separation score, the better the test is arranged because the test items can distinguish the high to low distribution of individual abilities, while the higher the score of item separation, the better the sample proportion on the measurements taken or the more scattered the sample is at each level of understanding (Khine, 2020a). If it is correlated to the person separation test score which is classified as weak (n = 1.44), it shows that the proportion sample distribution is not too good. On the other hand, the item separation score is classified as special (n = 5.38), which indicates that the test items are very good in measuring the ability of the sample. The difference in individual abilities from the results of the person separation is important to know because the teacher can immediately help students who do not understand the material or misconceptions to find the correct concept. The results of this test are also a benchmark for improving learning in class, using certain methods, increasing the improvement of learning outcomes.

Khine (2020a) said the factors related to reliability results include the length of the test and the score of the sample test results. Gronlund (1985) in Arifin (2012)stated that the longer the test and the size of the measured sample score, the greater the reliability score because it is possible to have a greater proportion of answers. Gronlund also explained that the distribution of test results can affect the reliability score. The greater the chance of the difference in score between each sample, the better the reliability score. Apart from the length of the test and the score results, Gronlund (1985) also mentioned two other factors that influence the reliability value; objectivity and the level of difficulty of the questions. Objectivity depends on several criteria including the personal ability of each sample and the test procedure. The FTISMCT research implementation procedure was carried out using an online research system using Google form. Online research was carried out because schools were closed due to the COVID-19 pandemic, so the research data was taken online. The data collection process began with a command to do the test and an explanation of the test mechanism to all samples.

Online research provides several advantages such as more environmentally friendly, cheaper, obtain more specific and programmatic data accuracy, and able to reach more samples without being limited by demographics (Padayachee, 2016; Sinclair, Toole, Malawaraarachchi, & Leder, 2012; Wu, Sun, & Tan, 2013). The drawback of online research is the low response rate of respondents which affects the validity and reliability scores (Manzo & Burke, 2012; Sinclair et al., 2012). The low-reliability score also indicates that the responses from students are too random so that there is too much data variance and the possibility of error data.

c) Discrimination

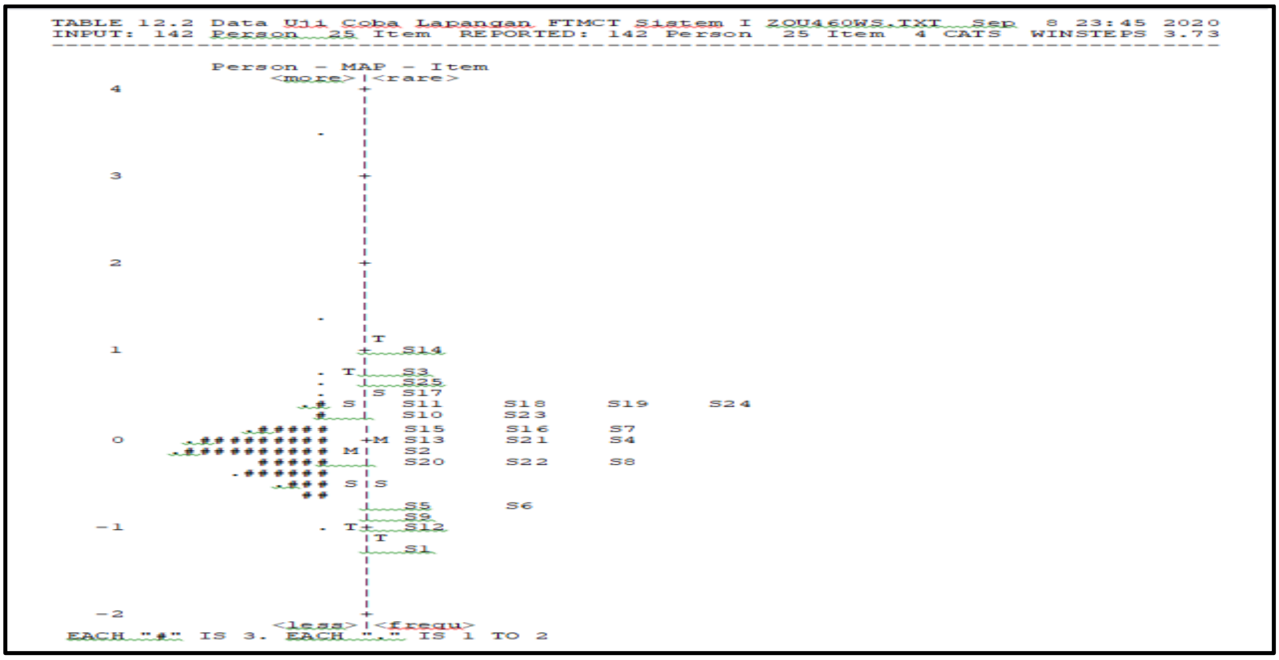
The discrimination test shows the ability of the instrument to measure the difference in the ability of each sample. The better the instrument, the better it can differentiate between competent samples and those who are incompetent in doing the test (Arifin, 2012). The test results analyzed by a summary statistical table show the score of the person separation instrument is 1.44 which is classified as weak. To see the distribution of differences in the ability of each student in answering questions, it can be seen using the Person Map and Item Map tables. The item map table is useful for clarifying the question number given the code "#" in the person map test results. The results of testing the person discrimination using a person map can be seen in Figure 4.



```
TABLE 16.3 Data Uji Coba Lapangan FTMCT Sistem I ZOU069WS.TXT  Sep 10 21:26 2020
INPUT: 142 Person  25 Item  REPORTED: 142 Person  25 Item  4 CATS  WINSTEPS 3.73
--------------------------------------------------------------------------------

          Item - MAP - Person
            <rare>|<more>
    4             +
                  |
                  |  056P31
                  |
    3             +
                  |
                  |
                  |
    2             +
                  |
                  |
                  |  075L41
                 T|
    1          X  +
               X  |T 044P21
               X  |  135P22
               X S|  129L22
            XXXX  |S 001L11  036P21  096P51  114P12
              XX  |  015P11  035P21  053P31
             XXX  |  001L11  001P11  001P11  011L11  013P11  018P11  026L21
                     060P31  062L31  065P31  068P41  071L41  082P41  087P41
                     100L51  108P51  123P12
    0        XXX M+  012P11  017P11  022P11  027P21  029L21  041L21  043P21
                     046L31  050P31  054P31  067P31  069P41  073L41  080L41
                     081L41  083L41  084P41  089P41  094L51  107P51  109P51
                     118L12  122P12  126L12  127P22  128P22  133P22  136P22
                     138P22
               X  |M 001P11  010P11  016P11  019P11  020P11  021P11  025P21
                     031L21  034L21  037P21  039P21  040L21  042L21  047P31
                     055P31  059P31  063P31  066P31  072P41  074L41  076P41
                     077P41  079P41  093P51  112L12  113P12  116L12  119P12
                     120L12  121P12  130L22  140L22
             XXX  |  001L11  014P11  023L11  028P21  045L31  048L31  078P41
                     086P41  092P51  103P51  111P51  115P12  134L22  139L22
                     142L22
                  |  001P11  024L21  038P21  049L31  052P31  057L31  061P31
                     064P31  070L41  085L41  088P41  099P51  101L51  106L51
                     110L51  117P12  124L12  125P12  131P22  141P22
                 S|S 001P11  001P11  030L21  032P21  051P31  095P51  097L51
                     102L51  104L51  137P22
                  |  033L21  058L31  090L41  091L51  105P51  132L22
              XX  |
               X  |
   -1          X  +T 098P51
                 T|
               X  |
                  |
                  |
   -2             +
            <frequ>|<less>
```

Figure 4. Results Data of FTISMCT Person Discrimination Using Person Map

Item test using a Person Map shows that students with the codes 056P31 and 075L41 have a person logit score that is higher than the logit item score. The person measure data shows that the logit person scores of the two students are 3.52 and 1.33, while the highest logit item value is 0.98. Student 056P31 is the student with the highest ability and student 098P51 is the student with the lowest ability to answer questions.



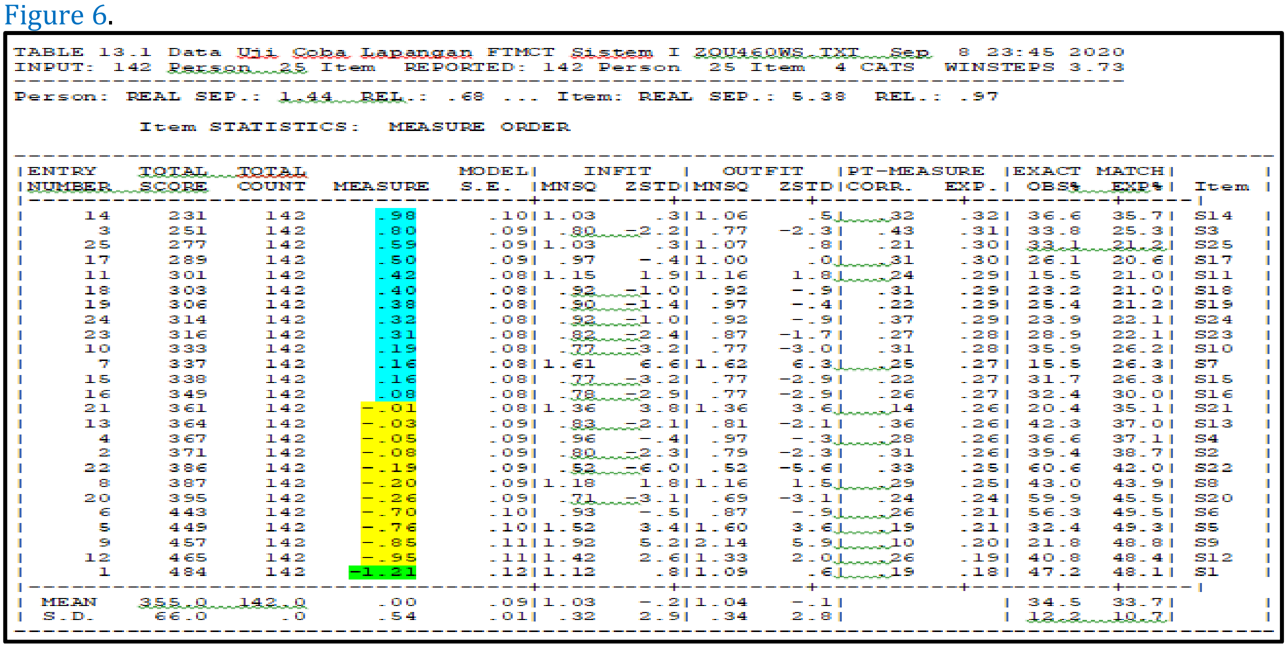
```
TABLE 12.2 Data Uji Coba Lapangan FTMCT Sistem I ZOU460WS.TXT  Sep  8 23:45 2020
INPUT: 142 Person  25 Item  REPORTED: 142 Person  25 Item  4 CATS  WINSTEPS 3.73
--------------------------------------------------------------------------------

           Person - MAP - Item
                 <more>|<rare>
    4                  +
                       |
                   .   |
                       |
    3                  +
                       |
                       |
                       |
    2                  +
                       |
                   .   |
                       |T
    1                  +  S14
                   .  T|  S3
                   .   |  S25
                   .   |S S17
                  .#  S|  S11      S18      S19      S24
                   #   |  S10      S23
              .#####   |  S15      S16      S7
    0     .#########   +M S13      S21      S4
         .##########  M|  S2
               #####   |  S20      S22      S8
             .######   |
                .###  S|S
                  ##   |
                       |  S5       S6
                       |  S9
   -1              .  T+  S12
                       |T
                       |  S1
                       |
                       |
   -2                  +
                 <less>|<frequ>
 EACH "#" IS 3. EACH "." IS 1 TO 2
```

Figure 5. Results Data of FTISMCT Item Discrimination Using Item Map

When viewed from the distribution of person map data and item maps, overall students have a logit person value that does not scatter and accumulated. The position of the logit person scores of students who do not scatter and accumulated is related to the ability of students who are almost the same in answering the questions. The results data of the discrimination test using item maps can be seen in Figure 4. The logit item score which is lower than the logit person score on a question shows that in general the student can answer the question well or the question has a difficulty level that is easy for almost all students to work on.

d) Item Difficulty Level

The item difficulty level shows a number that indicates whether an item is classified as easy or difficult (Ismail, 2020). A good item difficulty level must have a balanced proportion of the difficulty of the items tested so that the difficult items don't dominate or vice versa (Arifin, 2012). The difficulty level can be seen from the output measure item table. Testing the difficulty level of 25 items was carried out on 142 samples. The measure score column is used to view the logit score for classifying the difficulty level of each question. The test data for the difficulty level of the items can be seen in Figure 6.



```
TABLE 13.1 Data Uji Coba Lapangan FTMCT Sistem I ZOU460WS.TXT  Sep  8 23:45 2020
INPUT: 142 Person  25 Item  REPORTED: 142 Person  25 Item  4 CATS  WINSTEPS 3.73
------------------------------------------------------------------------------------
Person: REAL SEP.: 1.44  REL.: .68 ... Item: REAL SEP.: 5.38  REL.: .97

         Item STATISTICS:  MEASURE ORDER

------------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PT-MEASURE |EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----------+-----------+------|
|    14    231    142      .98    .10|1.03    .3|1.06    .5|  .32   .32| 36.6  35.7| S14  |
|     3    251    142      .80    .09| .80  -2.2| .77  -2.3|  .43   .31| 33.8  25.3| S3   |
|    25    277    142      .59    .09|1.03    .3|1.07    .8|  .21   .30| 33.1  21.2| S25  |
|    17    289    142      .50    .09| .97   -.4|1.00    .0|  .31   .30| 26.1  20.6| S17  |
|    11    301    142      .42    .08|1.15   1.9|1.16   1.8|  .24   .29| 15.5  21.0| S11  |
|    18    303    142      .40    .08| .92  -1.0| .92   -.9|  .31   .29| 23.2  21.0| S18  |
|    19    306    142      .38    .08| .90  -1.4| .97   -.4|  .22   .29| 25.4  21.2| S19  |
|    24    314    142      .32    .08| .92  -1.0| .92   -.9|  .37   .29| 23.9  22.1| S24  |
|    23    316    142      .31    .08| .82  -2.4| .87  -1.7|  .27   .28| 28.9  22.1| S23  |
|    10    333    142      .19    .08| .77  -3.2| .77  -3.0|  .31   .28| 35.9  26.2| S10  |
|     7    337    142      .16    .08|1.61   6.6|1.62   6.3|  .25   .27| 15.5  26.3| S7   |
|    15    338    142      .16    .08| .77  -3.2| .77  -2.9|  .22   .27| 31.7  26.3| S15  |
|    16    349    142      .08    .08| .78  -2.9| .77  -2.9|  .26   .27| 32.4  30.0| S16  |
|    21    361    142     -.01    .08|1.36   3.8|1.36   3.6|  .14   .26| 20.4  35.1| S21  |
|    13    364    142     -.03    .09| .83  -2.1| .81  -2.1|  .36   .26| 42.3  37.0| S13  |
|     4    367    142     -.05    .09| .96   -.4| .97   -.3|  .28   .26| 36.6  37.1| S4   |
|     2    371    142     -.08    .09| .80  -2.3| .79  -2.3|  .31   .26| 39.4  38.7| S2   |
|    22    386    142     -.19    .09| .52  -6.0| .52  -5.6|  .33   .25| 60.6  42.0| S22  |
|     8    387    142     -.20    .09|1.18   1.8|1.16   1.5|  .29   .25| 43.0  43.9| S8   |
|    20    395    142     -.26    .09| .71  -3.1| .69  -3.1|  .24   .24| 59.9  45.5| S20  |
|     6    443    142     -.70    .10| .93   -.5| .87   -.9|  .26   .21| 56.3  49.5| S6   |
|     5    449    142     -.76    .10|1.52   3.4|1.60   3.6|  .19   .21| 32.4  49.3| S5   |
|     9    457    142     -.85    .11|1.92   5.2|2.14   5.9|  .10   .20| 21.8  48.8| S9   |
|    12    465    142     -.95    .11|1.42   2.6|1.33   2.0|  .26   .19| 40.8  48.4| S12  |
|     1    484    142    -1.21    .12|1.12    .8|1.09    .6|  .19   .18| 47.2  48.1| S1   |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN   355.0  142.0      .00    .09|1.03   -.2|1.04   -.1|           | 34.5  33.7|      |
| S.D.    66.0     .0      .54    .01| .32   2.9| .34   2.8|           | 12.2  10.7|      |
------------------------------------------------------------------------------------------------
```

Figure 6. Results Data of FTISMCT Item Difficulty Level Test

The data from the difficulty level test shows that there is one question that is classified as very easy, namely question number 1 with a logit measure score of -1.21. There are eleven questions in the easy category, namely questions number 2, 4, 5, 6, 8, 9, 12, 13, 20, 21, and 22. There are 13 questions in the difficult category, namely questions number 3, 7, 10, 11 , 14, 15, 16, 17, 18, 19, 23, 24, and 25. There are no questions in the very difficult category.

Based on the results of the difficulty level test, the overall item difficulty level is not good because it tends to be disproportionate. Rasch's analysis classifies the items into four categories, namely very easy, easy, difficult, and very difficult. If the item difficulty level is good and proportional, then each category should contain 5 to 6 items. The data shows that the questions with the easy and difficult categories dominate the test compared to the questions with the very easy and very difficult categories.

The results of the easy test item difficulty level according to Matondang, Djulia, Sriadhi, & Simarmata (2019) probably mean that students have understood the material being asked or the answer choices that are not functioning properly. If the question items are difficult, then the possibility of interpretation is that there is an error in the answer key, there are two correct answer options in the answer choices, students are not competent enough because the learning material has not been thoroughly taught, the question form is not suitable for testing the sample, or the statement and question narrative are too long (Matondang et al., 2019).

A good test must have a balanced category between very easy, easy, difficult, and very difficult (Hartati & Yogi, 2019). A very easy or very difficult item cannot reflect the differences in abilities between students so that they are less informative, but items that are too easy do not mean that they have to be deleted from the questions. These very easy items also can increase confidence for the student doing the test (Musa, Shaheen, & Elmardi, 2018; Quaigrain & Arhin, 2017).

CONCLUSION

The results of the construct validity test using Rasch analysis showed that three items had to be reduced, namely items number 5, 7, and number 9 because the three items did not meet all the validity criteria so they were not valid to measure the sample. The reliability test data showed that the Cronbach Alpha score was 0.51 (bad). The score of item reliability was 0.97 (special) while the score of person reliability was 0.68 (sufficient). The result of the person separation score was 1.44 (weak), while the score of the item separation was 5.38 (special). Students with the highest ability were the student with code 056P31 while the student with the lowest ability is students with code 098P51. The results of the item difficulty level test showed that the most difficult question was question number 14 while the easiest question was question number 1. There was 1 question with the very easy category, namely question number 1, questions with easy category totaled 11, namely questions number 2, 4, 5, 6, 8, 9, 12, 13, 20, 21, and 22, and there were 13 questions with difficult categories, namely questions number 3, 7, 10, 11, 14, 15, 16, 17, 18, 19, 23, 24, and 25. Based on the results of the difficulty level test, the overall difficulty level of the items is not good because it tends to be disproportionate.

Content validation by material experts showed that several items needed to deepen the material and improve the choice of words so that they were more in line with more general language and concepts intended by the experts. The evaluation expert revised the section on the cognitive level which was contained in the indicators of the grid, the structure of the questions, and errors in writing procedures. The biology teacher provided an evaluation of how students might respond to a narrative question that is too long and difficult to understand. The overall results of suggestions and validation by the experts had been used to correct the questions.

REFERENCES

Adeniyi, E. O. (1985). Misconceptions of selected ecological concepts held by some Nigerian students Misconceptions of selected ecological concepts held by some Nigerian students. *Jurnal of Biology Education*, *19*(4), 311–316.

Akamca, G. Ö., Ellez, A. M., & Hamurcu, H. (2009). Effects of computer-aided concept cartoons on learning outcomes. *Procedia-Social and Behavioral Sciences*, *1*(1), 296–301. https://doi.org/10.1016/j.sbspro.2009.01.054

Al-zoubi, S. M. (2015). Low Academic Achievement : Causes and Results. *Theory and Practice Language Studies*, *5*(11), 2262–2268.

Allen, M. (2014). *Misconceptions in Primary Science*. *Mc Graw Hill Education Open University Press*. https://doi.org/10.1152/advances.2000.24.1.62

Anggrayni, S., & Ermawati, F. U. (2019). The validity of Four-Tier's s misconception diagnostic test for Work and Energy concepts The validity of Four- Tier's misconception diagnostic test for Work and Energy concepts. *Journal of Physics: Conference Series*, *1171*(1), 0–13. https://doi.org/10.1088/1742-6596/1171/1/012037

Arifin, Z. (2012). Menganalisis Kualitas Tes. In *Evaluasi Pembelajaran* (pp. 311–367). Jakarta: Direktorat Jenderal Pendidikan Islam Kementerian Agama RI.

Assaraf, O. B., Dodick, J., & Tripto, J. (2013). High School Students ' Understanding of the Human Body System. *Research in Science Education*, *43*(1), 33–56. https://doi.org/10.1007/s11165-011-9245-2

Astutik, W. (2018). *Pengembangan instrumen three-tier multiple choice diagnostic test untuk mengidentifikasi miskonsepsi siswa sma materi gerak melingkar beraturan*. Universitas Islam Negeri Walisongo.

Boone, W. J. (2016). Rasch Analysis for Instrument Development : Why, When, and How ? *CBL-Life Sciences Education*, *15*(4), 1–7. https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., Noltemeyer, A., Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis : A primer for school psychology researchers and practitioners Rasch analysis : A primer for school psychology researchers and practitioners. *Cogent Education*, *1*(4), 1–13. https://doi.org/10.1080/2331186X.2017.1416898

Caleon, I. S., & Subramaniam, R. (2010a). Do Students Know What They Know and What They Don't Know ? Using a Four-Tier Diagnostic Test to Assess the Nature of Students ' Alternative Conceptions. *Research in Science Education*, *40*(3), 313–337. https://doi.org/10.1007/s11165-009-9122-4

Caleon, I. S., & Subramaniam, R. (2010b). Do Students Know What They Know and What They Don't Know ? Using a Four-Tier Diagnostic Test to Assess the Nature of Students ' Alternative Conceptions Do Students Know What They Know and What They Don't Know ? Using a Four-Tier Diagnostic Test to Ass. *Research in Science Education*, *40*, 313–337. https://doi.org/10.1007/s11165-009-9122-4

Chandrasegeran, A. ., Treagust, D. F., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students ' ability to describe and explain chemical reactions using multiple levels using multiple levels of representation. *Chemistry Education Research and Practice*, *8*(3), 293–307. https://doi.org/10.1039/B7RP90006F

Chi, M. T. . (2013). Two Kinds and Fur Sub-Types of Misconceived Knowledge, Ways to Change it, and Learning Outcomes. In *International Handbook of Research on Conceptual Change Routledge* (pp. 49–70). https://doi.org/10.4324/9780203154472.ch3

Faggioni, T., Ferreira, N. C. da S., Lopes, R. M., Fidalgo-neto, A. A., Cotta-de-almeida, V., & Alves, L. A. (2019). Open educational resources in immunology education. *Advances in Physiology Education*, *43*(2), 103–109. https://doi.org/10.1152/advan.00116.2018

Fisher, K. M. (1985). A Misconception In Biology : Amino Acids AND TRANSLATION. *Journal Od Research In Science Teaching*, *22*(1), 53–62.

Fratiwi, N. J., Ramalis, T. R., & Samsudin, A. (2019). The Three-tier Diagnostic Instrument : Using Rasch Analysis to Develop and Assess K-10 Students ’ Alternative Conceptions on Force Concept. In *RSU International Research Conference* (pp. 654–663).

Gronlund, N. E. (1985). *Stating objectives for classroom instruction*. Macmillan Publishing Company.

Gurel, D. K., Eryılmaz, A., & Mcdermott, L. C. (2015). A Review and Comparison of Diagnostic Instruments to Identify Students ’ Misconceptions in Science. *Eurasia Journal of Mathematics, Science & Technology Education*, *11*(5), 989–1008. https://doi.org/10.12973/eurasia.2015.1369a

Hammer, D. (1996). More than misconceptions : Multiple perspectives on student knowledge and reasoning, and an appropriate role for education research. *American Journal Od Physics*, *64*, 1316–1325. https://doi.org/10.1119/1.18376

Hartati, N., & Yogi, H. P. S. (2019). Item Analysis for a Better Quality Test. *English Language in Focus (ELIF)*, *2*(1), 59. https://doi.org/10.24853/elif.2.1.59-70

Hasyim, W., Suwono, H., & Susilo, H. (2018). Three-tier Test to Identify Students’ Misconception of Human Reproduction System. *Jurnal Pendidikan Sains*, *6*(2), 48–54.

Ismail, I. (2020). Analisis Tingkat Kesukaran Hasil Tes. In *Asesmen dan Evaluasi Pembelajaran* (pp. 144–145). Makasar: Cendekia.

Kaltacki, D. (2012). *Development and application of a four-tier test to assess pre-service physics teachers" misconceptions about geometrical optics*. Middle East Technical University.

Khine, M. S. (2020a). *Rasch Measurement : Applications in Quantitive Educational Research*. (M. S. Khine, Ed.). Jerman: Springer.

Khine, M. S. (2020b). *Rasch Measurement Applications in Quantitative Educational*. Singapore: Springer. https://doi.org/10.1007/978-981-15-1800-3

Kurt, H. (2013). Turkish student biology teachers' conceptual structures and semantic attitudes towards microbes. *Journal of Baltic Science Education*, *12*(5), 608–639.

Lazarowitz, R., & Penso, S. (1992). High school students' difficulties in learning biology concepts. *Journal of Biological Education*, *26*(3), 37–41.

Linacre, J. M. (1999). Understanding Rasch Measurement: Estimation Methods for Rasch Measures. *Journal of Outcome Measurement*, *3*(4), 382–405.

Linacre, J. M. (2017). Rasch Measurement Transactions. *Autmn*, *31*(2), 1629–1642.

Lukin, K. (2013). Exciting middle and high school students about immunology: An easy, inquiry-based lesson. *Immunologic Research*, *55*(1–3), 201–209. https://doi.org/10.1007/s12026-012-8363-x

Maharani, L., Rahayu, D. I., Amaliah, E., Rahayu, R., & A, S. (2019). Diagnostic Test with Four-Tier in Physics Learning : Case of Misconception in Newton's Law Material Diagnostic Test with Four-Tier in Physics Learning : Case of Misconception in Newton's s Law Material. *Journal of Physics: Conference Series*, *115*. https://doi.org/10.1088/1742-6596/1155/1/012022

Mamik. (2015). *Metode Kualitatif*. Sidoarjo: Zifatama.

Manzo, A. N., & Burke, J. M. (2012). *Increasing Response Rate in Web-Based/Internet Surveys*. New York: Springer. https://doi.org/10.1007/978-1-4614-3876-2

Matondang, Z., Djulia, E., Sriadhi, & Simarmata, J. (2019). Evaluasi Hasil Belajar. In *Evaluasi Hasil Belajar* (p. 59).

Milenkovic, D. D., Hrin, T. N., Segedinac, M. D., & Horvat, S. (2016). Development of a Three-Tier Test as a Valid Diagnostic Tool for Identification of Misconceptions Related to Carbohydrates. *Journal of Chemical Education*, *93*(9), 1514–1520. https://doi.org/10.1021/acs.jchemed.6b00261

Musa, A., Shaheen, S., & Elmardi, A. (2018). Item difficulty & item discrimination as quality indicators of physiology MCQ examinations at the Faculty of Medicine Khartoum University Abstract : *Khartoum Medical Journal*, *11*(2), 1477–1468.

National Research Council. (1997). Misconceptions as Barriers to Understanding Science. In *Science Teaching Reconsidered : A Handbook* (pp. 27–32). Washington DC: The National Academies Press. https://doi.org/10.17226/5287

Padayachee, K. (2016). Internet-mediated research : Challenges and issues. *SACJ*, *28*(December), 25–45.

Patil, S. J., Chavan, R. L., & Khandagale, V. S. (2019). Identification of Misconceptions in Science: Tools, Techniques & Skills for Teachers. *Aarhat Multidisciplinary International Education Research Journal (AMIERJ)*, *8*(2), 466–472.

Pujyanto, Budiharti, R., Radiyono, Y., Rizky Amalia Nurani, N., Putri, H. V., Saputro, D. E., & Adhitama, E. (2018). Pengembangan Tes Diagnostik Miskonsepsi Empat Tahap Tentang Kinematika. *Cakrawala Pendiidkan*, *2*, 237–249.

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). https://doi.org/10.1080/2331186X.2017.1301013

Rahman, A., Ahmar, A. S., Arifin, A. N. M., Upu, H., Mulbar, U., Alimuddin, … Ihsan, H. (2018). The Implementation of APIQ Creative Mathematics Game Method in the Subject Matter of Greatest Common Factor and Least Common Multiple in Elementary School. *Journal of Physics: Conference Series*, *954*(1). https://doi.org/10.1088/1742-6596/954/1/012011

Riyantono, S., & Hatmawan, A. A. (2020). Metode Riset Penelitian Kuantitatif. In *Metode Riset Penelitian Kuantitatif Penelitian Di Bidang Manajemen, Teknik, Pendidikan Dan Eksperimen* (p. 63). Yogyakarta: Deepublish.

Sarimanah, E., Dewi, F. I., & Sabri, T. (2019). A Review Of Students ' Common Misconceptions In Science And Their Diagnostic Assessment Tools. *Jurnal Pendidikan IPA Indonesia*, *8*(2), 247–266. https://doi.org/10.15294/jpii.v8i2.18649

Senocak, E., Taskesenligil, Y., & Sozbilir, M. (2007). A Study on Teaching Gases to Prospective Primary Science Teachers Through Problem-Based Learning A Study on Teaching Gases to Prospective Primary Science Teachers Through Problem-Based Learning. *Research in Science Education*, *37*(3), 279–290. https://doi.org/10.1007/s11165-006-9026-5

Septiantini, T., Samsudin, A., Aminudin, A. H., Rasmitadila, Rachmatullah, R., Costu, B., & Nurtanto, M. (2020). Static Fluid Four-Tier Instrument (SFFTI): Develop and Identify K-11 Brebes- Scholars ' Alternative Conception with Rasch Analysis Static Fluid Four-Tier Instrument (SFFTI): Develop and Identify K-11 Brebes-Scholars ' Alternative Conception with Rasch. *International Journal of Advanced Science and Technology*, *9*(7), 3190–3199.

Sinclair, M., Toole, J. O., Malawaraarachchi, M., & Leder, K. (2012). Comparison of response rates and cost-effectiveness for a community-based survey : postal, internet and telephone modes with generic or personalised recruitment approaches. *BMC Medical Research Methodology*, *12*(132), 1–8.

Singh, C. (2016). Effect of Misconception on Transfer in Problem Solving. *In AIP Conference Proceedings*, *951*(1), 196–199.

Siqueira-batista, R., Gomes, A. P., Albuquerque, V. S., Madalon-fraga, R., Aleksandrowicz, A. M. C., & Geller, M. (2009). Lições de Akira Kurosawa. *Revista Brasileira de Educação Médica*, *33*(2), 186–190.

Smiley, J. (2016). Classical test theory or Rasch : A personal account from a novice user. *Shiken*, *9*(1), 16–29.

Stranford, S. A., Owen, J. A., Mercer, F., & Pollock, R. R. (2020). Active Learning and Technology Approaches for Teaching Immunology to Undergraduate Students. *Frontiers in Public Health*, *8*(May), 1–17. https://doi.org/10.3389/fpubh.2020.00114

Stylos, G., Evangelaksis, G. A., & Kotsis, K. T. (2008). Misconceptions on classical mechanics by freshman university students : A case study in a Physics Department in Greece Misconceptions on classical mechanics by freshman university students : A case study in a Physics Department in

Greece. *Themes In Science And Technology Education*, *1*(2), 157–177.

Su, T., Cheng, M., & Lin, S. (2014). Investigating the Effectiveness of an Educational Card Game for Learning How Human Immunology Is Regulated. *CBE—Life Sciences Education*, *13*(3), 504–515. https://doi.org/10.1187/cbe.13-10-0197

Subari, K. (2017). Improving Understanding and Reducing Matriculation Students' Misconceptions in Immunity Using the Flipped Classroom Approach. In *Overcoming Students' Misconceptions in Science: Strategies and Perspectives from Malaysia* (pp. 1–344). Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-10-3437-4

Suliyanah, Putri, H., & Rohmawati, L. (2018). Identification student's misconception of heat and temperature using three-tier diagnostic test. *Journal of Physics: Conference Series*, *997*(1), 1–10.

Sumintono, B., Widhiarso, W., & Mada, U. G. (2014). *Aplikas Model Rasch untuk Penelitian Ilmu-Ilmu Sosial (edisi revisi)*. Cimahi: Trim Komunikata Publishing House.

Sumintono, B., & Widiharso, W. (2015). *Aplikasi Permodelan Rasch Pada Assessment Pendidikan*. Cimahi: Trim Komunikata Publishing House.

Tekkaya, C. (2002). Misconceptions As Barrier To Understanding Biology. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *23*, 259–266.

Toka, Y., & Askar, P. (2002). The Effect Of Cognitive Conflict And Conceptual Change Text On Students' Achievement Related To First Degree Equations With One Unknown. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *23*(23), 211–217.

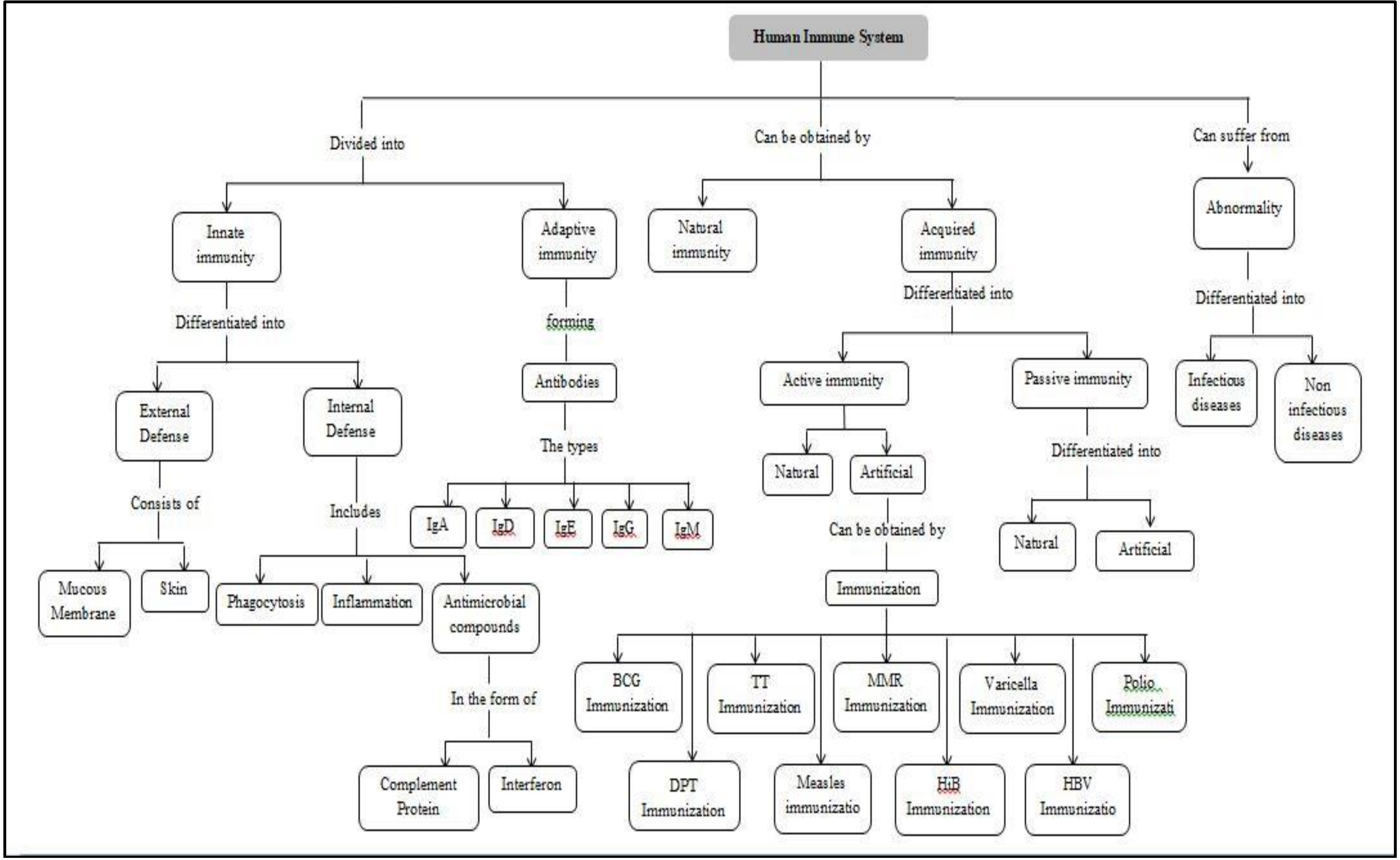
Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students ' misconceptions in science. *International Journal of Science Education*, *10*(2), 159–169. https://doi.org/10.1080/0950069880100204

Verkade, H., Mulhern, T. D., Lodge, J. M., Elliott, K., Cropper, S., Rubinstein, B. I. P., … Mulder, R. (2017). *Misconceptions as a trigger for enhancing student learning in higher education*. (T. U. of Melbourne, Ed.), *The University of Melbourne*. Melbourne.

Wu, J., Sun, H., & Tan, Y. (2013). Social Media Research : A review. *Journal of Systems Science and Systems Engineering*, *22*(Sep), 257–282. https://doi.org/10.1007/s11518-013-5225-6

Zaini, M., & Rusmini. (2016). Pengembangan Perangkat Pembelajaran Konsep Klasifikasi Benda Terhadap Keterampilan Berpikir Kritis Siswa SMP Developing Learning Instrument Concept Of Classification Of Objects Of Critical Thinking Skills Smp Students. *Biologi, Sains, Lingkungan Dan Pembelajarannya.*, *13*(1), 102–111. Retrieved from https://jurnal.uns.ac.id/prosbi/article/view/5668

Appendix 1. Development of the Immune System Concept Map



Appendix 2. Example of Four-Tier Immune System Multiple Choice Test

Four-Tier Immune System Multiple Choice Test Questions

Number 1

Statement:

Ani's family cleans the house on Sundays. They replaced worn window curtains, cleaned dusty floors and tables, and painted moldy walls. Ani felt that her nose was itchy, sneezing, her eyes were red and watery and had red bumps on her skin after she finished cleaning the house, whereas previously Ani was healthy and had no bumps on her skin.

Answer Options (*Tier* 1)

The correct statement regarding the reaction on Ani's body is....

a. Ani's body reaction is caused by exposure to antigens such as dust and microorganisms.
b. Ani's body reaction is caused by exposure to antibodies such as dust and microorganisms.
c. The reactions in Ani's body are not related to antibodies or antigens.
d. There is no correct answer regarding the reactions that occurred in Ani's body.

Reason Confidence Level (*Tier* 2)

How sure are you with the reasons you have chosen?

a. 1 (Very Unsure)
b. 2 (Unsure)
c. 3 (Sure)
a. 4 (Very Sure)

Answer Options (*Tier* 3)

a. Dust and microorganisms contain substances that are recognized by the body so that the body reacts.
b. Dust and microorganisms contain substances that are unrecognized by the body so that the body reacts.
c. The reaction occurred because Ani was too tired to clean the house.
d. There is no appropriate reason why Ani's body reaction can occur.

Reason Confidence Level (*Tier* 4)

How sure are you with the reasons you have chosen?

a. 1 (Very Unsure)
b. 2 (Unsure)
c. 3 (Sure)
d. 4 (Very Sure)

Number 2

Statement:

Study Case 1

Ani conducts an experiment to check blood type. She drops Anti A serum onto the blood sample. The blood changed to clot after being given the Anti A serum...

Study Case 2

Andi fell off the bike, injuring his leg and bleeding. A week later, the wound appeared pus and scabs.

Answer Options (*Tier* 1)

The exact statement regarding the two cases above is....

a. Anti A serum is the antigen for blood in the Ani sample test.
b. Clotting is a form of antigen for blood in the Ani sample test.
c. Clotting is a form of antibody response.
d. Wound and serum Anti A are antibodies.

Reason Confidence Level (*Tier* 2)

How sure are you with the reasons you have chosen?

a. 1 (Very Unsure)
b. 2 (Unsure)
c. 3 (Sure)
d. 4 (Very Sure)

Answer Reasons (*Tier* 3)

a. Anti A serum and pus are natural substances produced by the body so that the body recognizes these substances and responds in the form of clots and pus.
b. Anti A serum and pus are natural substances that not produced by the body so that the body recognizes these substances and responds in the form of clots and pus.
c. The clots and pus are different reactions because the anti-A serum does not come from the body while the pus is excreted by the body.
d. There is no appropriate reason for the clotting reaction to occur.

Reason Confidence Level (*Tier* 4)

How sure are you with the reasons you have chosen?

a. 1 (Very Unsure)
b. 2 (Unsure)
c. 3 (Sure)
d. 4 (Very Sure)