Al-Ishlah: Jurnal Pendidikan
Vol.16, 3 (September, 2024), pp. 3924-3937
ISSN: 2087-9490 EISSN: 2597-940X, DOI: 10.35445/alishlah.v16i3.5621

Assessing Undergraduate Cognitive System Thinking Instruments in Genetics Lectures: A Rasch Model Analysis

Iffa Ichwani Putri [1,2], Adi Rahmat[3], Riandi[4], and Lala Septem Riza[5]

[1] Universitas Pendidikan Indonesia, Bandung, Indonesia; iffa.ichwani@upi.edu
[2] Universitas Islam Riau, Pekanbaru, Indonesia; iffa.ichwani@edu.uir.ac.id
[3] Universitas Pendidikan Indonesia, Bandung, Indonesia; adirahmat@upi.edu
[4] Universitas Pendidikan Indonesia, Bandung, Indonesia; rian@upi.edu
[5] Universitas Pendidikan Indonesia, Bandung, Indonesia; lala.s.riza@upi.edu

ARTICLE INFO

Keywords:

Biology;
Cognitive;
Genetics;
Rasch;
Thinking.

Article history:

Received 2024-07-04
Revised 2024-09-13
Accepted 2024-09-27

ABSTRACT

This study aims to test the reliability, validity and wright person-items of the test instrument. It used to measure students' cognitive system thinking in genetics lectures. This cognitive system dimension (New Marzano Taxonomy) consists of 4 levels: retrieval, comprehension, analysis, and knowledge utilization. The research method used was descriptive quantitative, with respondents as many as 116 undergraduate students who had attended genetics lectures. The instrument used consisted of 20 multiple-choice questions developed from 4 levels of the cognitive dimension of the system. Data analysis using the Rasch Model-Dichotom to generate more accurate estimates based on individual abilities and difficulty levels of question items. The results showed that the Alpha Cronbach and Item Reliability value on the instrument was 0.81 (Very Good) to 0.92 (Best). Furthermore, person reliability is 0.63 (Moderate). It can be concluded the interaction between the question items and the quality of the question items on the test instrument indicates that the instrument effectively assesses constructs and is relevant for measuring the level of cognitive system thinking. Revisions are made to items that are too easy or too difficult, as well as a review of items with similar difficulties. So it can be assessed that this test has the potential to be widely used in measuring cognitive ability, but some improvements are needed to further improve its reliability for other population.

Corresponding Author:
Adi Rahmat
Universitas Pendidikan Indonesia, Bandung, Indonesia; adirahmat@upi.edu

1. INTRODUCTION

Higher education has an important role in developing students' cognitive thinking skills. Cognitive thinking skills allow students to not only memorize information, but also to understand biological concepts at a deeper level. In biology, students are often faced with data and must be able to analyze and synthesize the information to make logical conclusions. This ability is very important in biological research and practice. According to (Aramendiz & Córdova, 2019; Betu, 2023; Sari & Nada, 2022) analytical and synthesis skills help students solve problems and produce innovative solutions.

In biology lectures, many require decision-making with strong scientific evidence. This requires students to have good cognitive and critical thinking skills. Students can assess the reliability and validity of information sources and arguments. Good cognitive thinking skills prepare students for the world of work. This improves the ability to apply knowledge and skills in a variety of practical situations and collaborate with peers from different disciplines. Active learning methods that improve cognitive thinking skills can improve students' academic performance and job readiness (Freeman et al., 2014).

One of the challenging courses in this regard is Genetics. This course requires a deep understanding and the ability to apply the concept in a variety of contexts. Genetics requires a deep conceptual understanding of complex biological structures and functions. The ability to think about cognitive systems helps students deconstruct and understand this complexity. Students must be able to identify the components of the system and understand the causal relationships between components to understand complex systems.

Based on (Aivelo & Uitto, 2021) there is a potential relationship between the emphasis or teaching material used and the understanding of students' conceptions of genetics. In addition, other studies have also shown that genetics remains one of the main and important topics in learning, but continues to be a difficult concept for learners (Aivelo & Uitto, 2018; Etobro & Banjoko, 2017; Prochazkova et al., 2019). In their teaching, educators modify the way they teach certain topics that are considered to cause negative reactions or misunderstandings. The presentation of the teaching of genetics concepts can be adjusted to reduce controversy and improve understanding.

Misconceptions about the concept of genetics are common among prospective teachers. In the context of teacher education, a correct understanding of scientific concepts is critical, especially since teachers are the main influence in shaping students' scientific understanding. Research by (Etobro & Banjoko, 2017) was conducted to identify and analyze specific misconceptions that future teachers have in the field of genetics, covering topics such as heredity, DNA, and mutations. Genetics often involves analyzing complex data, such as DNA sequence data or experimental results.

Skills in cognitive systems thinking allow students to evaluate data, identify patterns, and make evidence-backed conclusions. (Tsui, 2002) found that the ability to analyze and evaluate information is one of the important skills that biology students must master to succeed in research and professional practice. Research by (Pane, Steiner, Baird, Hamilton, & Pane, 2017) shows that adaptive approaches in education can improve learning outcomes by organizing content that suits students' skill level. In genetics, educational technology can be used to adapt learning materials or use systems that monitor student understanding and provide feedback in real time.

The learning process always involves assessment as an important thing to do. Without going through an assessment, it is difficult to know for sure whether progress and learning goals have been achieved. The findings (Prochazkova et al., 2019) suggest that the assessment will deepen knowledge about topics from previous lectures and practical classes such as inheritance modes, gene structures, mutations, methods in molecular biology and many others, which are shown in practice from different perspectives. An evaluation procedure that is usually carried out by a teacher about the knowledge and skills of students to find out their performance using certain instruments.

Failure of instruments that are not accurate in measuring what is intended results in incorrect assessment of students' abilities. These failures can influence important decisions of educational interventions, learning and educational objectives. (DeLuca, Chapman-Chin, LaPointe-McEwan, & Klinger, 2018) found that teachers who lack assessment literacy are unable to design or choose the right evaluation instruments. This can reflect the effectiveness of learning objectives. The research of (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Taylor, Rapp, & Brunye, 2007) added that improperly designed questions will interfere with learners' ability to demonstrate true understanding.

However, the assessment instruments used in lectures are not fully capable of measuring students' thinking skills. The instrument measures students' factual knowledge more than their ability to analyze and solve problems. Teaching systems that involve students such as in doing short assignments and

explanations are significantly able to improve learning and create positive attitudes towards students' understanding and cognition (Machluf, Gelbart, Ben-Dor, & Yarden, 2017)

Based on the theory of Marzano and Kendall, there are four levels of thinking ability in the realm of the cognitive system. The thinking ability of prospective biology education teachers is very relevant in the context of genetics lectures. High thinking skills can solve problems that occur in the context of genetic material. Skills are very useful in educating biology students to think systematically and deeply about scientific concepts. By applying this model, educators can help students develop their cognitive abilities more effectively and thoroughly, which will be very useful in both academic and professional contexts. Each level of cognitive system thinking ability has a different goal. Each level plays an important role in the formation and application of cognitive abilities, providing a framework for educators to design learning experiences that facilitate holistic intellectual development.

Retrieval (Level 1), includes the process of remembering and executing knowledge that is often encountered during the learning process. This process involves recovering information that has been stored in memory, allowing for the quick and efficient use of knowledge in different situations. Understanding Information (Level 2), is the integration and symbolization of information into two interrelated processes. At this level, a deep understanding of information is emphasized, allowing individuals to interpret and restructure the knowledge gained in a way that makes it more relevant or applicable in new contexts.

Analysis (Level 3) requires further ability to generalize new information that has not been previously processed in learning activities. This level involves more complex analytical activities such as matching, classification, error analysis, and generalization. This process is important for developing a deeper understanding and for applying knowledge in a variety of situations. Knowledge Utilization (Level 4) refers to the use of learning outcomes as a tool to complete special tasks. This process is not only about using existing knowledge but also involves decision-making, problem-solving, experimentation, and investigation. This level demonstrates the application of knowledge in real-life situations, demanding proficiency in integrating and applying knowledge and skills from various domains to achieve concrete results.

The Rasch model is a highly effective model for analyzing test and survey data, and has been used extensively in a variety of fields, including education. The Rasch model takes an in-depth look at the characteristics of items and how they contribute to overall measurements. Furthermore, the efficiency of these items in measuring at various levels of respondents' abilities was also examined (Septi Purnama, Farozin, & Astuti, 2022; Zhang, Liu, & Feng, 2023).

In this study, the Rasch Model dichotomous method was used to analyze the validity of the test instruments used in Genetics lectures. The Rasch model of the dichotomous method is used to measure the extent to which the test instrument is capable of measuring what it is supposed to measure. In other words, this model can be used to evaluate the validity of test instruments. Validity is a measure of the extent to which the test instrument is able to measure the construct in question. It was added by (Raof, Mustaâ, Zamzuri, & Salleh, 2021) that the Rasch model is used to measure the construct of the instrument used in order to achieve the right measurements and decisions according to the purpose.

This study aims to evaluate the feasibility or effectiveness (reliability, validity and wright person-items) of test instruments in the form of multiple choice in properly measuring students' cognitive system thinking skills in genetics lectures. Proper and effective evaluation is expected to contribute to the development and use of strategies, methods, and the implementation of meaningful learning.

2. METHODS

This study uses a survey design. This design is in accordance with the purpose of the research, which is to measure the level of thinking ability of students' cognitive systems based on *Marzano's taxonomy*. This design allows for the collection of data widely from a large number of participants in a relatively short period of time, so that it can provide a representative picture of the student population

taking Genetics courses. Furthermore, survey design can help in understanding the perception and capabilities of the cognitive system quantitatively, which is consistent with the use of multiple-choice tests as research instruments.

The sample in this study consisted of 106 students who attended Genetics lectures. The selection of samples was carried out using *a purposive sampling technique*, with the criteria for students who were actively involved in lectures to be selected specifically to ensure that respondents had a knowledge background that had been fulfilled beforehand. Another criterion is the determination of samples of activeness in lectures, attendance, and participation in learning activities. The number of 106 students was chosen because it was considered sufficient to achieve an acceptable level of confidence and margin of error in statistical analysis, as well as representing a proportionately larger population.

The study's instrument was a multiple-choice test based on Marzano's taxonomy. The test development process consists of multiple stages, including a review of the literature on Marzano's taxonomy and conversations with education professionals to ensure the questions are appropriate for the level of cognitive thinking tested. Instrument validation was carried out through trials on a small group of students who were not included in the research sample to test the reliability and validity of the preliminaries. The results of the trial showed that this instrument was able to measure cognitive abilities as expected. The preparation of the multiple-choice test consists of cognitive thinking levels, ranging from *retrieval*, *understanding information*, *analysis*, to *knowledge utilization* (Table 1).

Table 1. Cognitive System Thinking Level

Level	Cognitive System Abilities	Multiple Choice Question Order
Level 1	Retrieval	1,2,3,4,5
Level 2	Understanding Information	6,7,8,9,10,11,12
Level 3	Analysis	13,14,15, 16, 17
Level 4	Knowledge Utilization	18,19,20

Data was collected by distributing multiple-choice tests to samples. This test is held online through the Google Forms platform. The distribution to samples is also carried out by direct supervision from the lecturer in charge of the course to ensure uniform conditions during the implementation of the test. Students are required to complete the test within the specified time to minimize bias. Additional measures are taken to maintain data integrity, such as the use of a randomization system for question sequences on the test platform to prevent cheating.

The data was analyzed using *Winstep-Rasch Model software* with a dichotomous method. The Rasch model was chosen because it has the ability to evaluate the reliability and validity of the test instrument in a probabilistic manner, which allows for a deeper analysis of the interaction between the individual (test subject) and the test item. The model also provides a more accurate estimate of the participant's abilities, regardless of the difficulty level of the item. In this analysis, reliability will be evaluated through *item fit statistics* and *person reliability*, while validity will be seen from *item difficulty* and *person ability estimates*. The results of the analysis will be used to determine the extent to which this instrument is able to measure the thinking ability of the student's cognitive system according to the cognitive level identified by *Marzano's taxonomy*.

3. FINDINGS AND DISCUSSION

3.1. Reliability of Multiple Choice Test Instruments

Reliability refers to the consistency of measurements. Reliability can be analyzed by looking at the estimated reliability for people and items.

```
-------------------------------------------------------------------------------
|          TOTAL                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       8.9      20.0        -.25     .51      1.00    .01   1.00    .00 |
|  SEM        .3        .0         .08     .00       .02    .09    .03    .09 |
| P.SD       3.3        .0         .84     .05       .17    .95    .28    .97 |
| S.SD       3.3        .0         .85     .05       .17    .95    .28    .97 |
| MAX.      18.0      20.0        2.40     .76      1.47   2.21   2.30   2.05 |
| MIN.       3.0      20.0       -1.91     .47       .71  -2.04    .56  -1.98 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .53 TRUE SD     .66  SEPARATION  1.24  Person RELIABILITY  .61 |
|MODEL RMSE    .51 TRUE SD     .67  SEPARATION  1.31  Person RELIABILITY  .63 |
| S.E. OF Person MEAN = .08                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = 1.00 (approximate due to missing data)
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .62  SEM = 2.04 (approximate due to missing data)
STANDARDIZED (50 ITEM) RELIABILITY = .81

     SUMMARY OF 20 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      51.9     116.0         .00     .21      1.00    .00   1.00    .04 |
|  SEM        3.9       .0         .17     .00       .02    .21    .03    .21 |
| P.SD      17.2        .0         .73     .01       .09    .93    .12    .91 |
| S.SD      17.6        .0         .75     .01       .09    .95    .12    .93 |
| MAX.      83.0     116.0        1.40     .25      1.30   3.14   1.39   2.99 |
| MIN.      22.0     116.0       -1.31     .20       .89  -1.36    .82  -1.29 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .21 TRUE SD     .70  SEPARATION  3.28  Item   RELIABILITY  .92 |
|MODEL RMSE    .21 TRUE SD     .70  SEPARATION  3.33  Item   RELIABILITY  .92 |
| S.E. OF Item MEAN = .17                                                     |
-------------------------------------------------------------------------------
Item RAW SCORE-TO-MEASURE CORRELATION = -1.00 (approximate due to missing data)
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000
```

Figure 1. Results of Reliability Measurement of Multiple Choice Questions

Based on the data obtained in Figure 1, it can be seen that individual reliability has a score of 0.63 with a sufficient category. This shows that the developed questions consistently measure the characteristics or abilities of individuals are at a moderate but acceptable level. Reliability scores in this range may indicate inconsistencies in answers or that test items have not been fully able to distinguish the level of ability in the population.

The reliability of the question items is in the very good category with a value of 0.92, which indicates that the items are very consistent in measuring the construct in question. This score acquisition means that the questions effectively capture the desired dimension of the respondents. This high reliability of the items indicates that the items are tested well overall to provide consistent and accurate measurements between individuals. An item separation index of 3.33 indicates that the test is quite good at distinguishing difficulty levels between items, with an average item size of 116.0 and a standard deviation of 3.9. The average score of the respondents was 8.9 with a standard deviation of 3.3, which indicates that there is a significant variation among the respondents. However, there are some items that show a higher MNSQ Outfit value 3.16 and 3.95. The value indicates that the question items require further review to ensure their suitability to measure cognitive system thinking ability.

So it is known that with individual reliability scores showing moderate consistency in assessing individual differences, high item reliability scores ensure that the test items themselves are highly reliable. The implication of these findings is that further refinements are needed in the test design to improve individual reliability, namely by improving existing items to better distinguish individual abilities.

3.2. Validity of Multiple Choice Test Instruments

Validity analysis can be used to evaluate questions for the purpose of measuring the differentiation of each respondent in each respondent's ability and the validity of the question item.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
14	22	116	1.40	.25	.92	-.50	1.00	.06	.40	.33	84.5	82.1	S14
13	26	116	1.16	.24	1.11	.84	1.02	.17	.25	.34	77.6	79.1	S13
9	32	116	.84	.22	.92	-.70	.93	-.40	.43	.35	79.3	75.1	S9
7	39	116	.52	.21	1.30	3.14	1.39	2.99	.02	.36	60.3	70.8	S7
20	39	116	.52	.21	.95	-.55	.95	-.42	.41	.36	70.7	70.8	S20
1	41	116	.43	.21	.89	-1.36	.86	-1.29	.48	.36	78.4	69.7	S1
5	41	116	.43	.21	.99	-.10	.98	-.17	.37	.36	75.0	69.7	S5
4	45	116	.26	.20	1.06	.86	1.10	.96	.28	.36	66.4	67.7	S4
12	45	116	.26	.20	1.02	.30	1.01	.13	.34	.36	66.4	67.7	S12
16	48	116	.14	.20	1.01	.14	1.01	.16	.35	.36	65.5	66.5	S16
17	49	116	.09	.20	1.00	.08	1.03	.34	.35	.36	66.4	66.2	S17
15	54	116	-.11	.20	.94	-.97	.94	-.63	.42	.36	72.4	64.9	S15
2	57	116	-.23	.20	1.00	.07	.99	-.13	.36	.36	60.3	64.4	S2
10	57	116	-.23	.20	1.04	.61	1.03	.32	.32	.36	63.8	64.4	S10
6	62	116	-.42	.20	1.01	.12	1.00	.02	.35	.35	62.9	64.7	S6
8	70	116	-.74	.20	.96	-.51	.94	-.53	.39	.34	65.5	66.4	S8
19	70	116	-.74	.20	1.01	.14	.98	-.11	.34	.34	63.8	66.4	S19
3	77	116	-1.04	.21	.97	-.37	.90	-.73	.38	.33	69.8	69.7	S3
18	81	116	-1.22	.21	.99	-.11	1.19	1.26	.31	.32	69.8	71.8	S18
11	83	116	-1.31	.22	.90	-1.07	.82	-1.20	.44	.31	74.1	73.1	S11
MEAN	51.9	116.0	.00	.21	1.00	.00	1.00	.04			69.7	69.6	
P.SD	17.2	.0	.73	.01	.09	.93	.12	.91			6.6	4.7	

Figure 2. Results of Measurement of Validity of Multiple Choice Question Items

Figure 2 shows the results of the analysis of the validity of multiple-choice test instruments to measure the cognitive system thinking ability of biology education students. Instrument validity analysis can be described as follows:

a) Measure dan Standar Error

The *Measure* column shows the difficulty level of each item on the logit scale (the unit of measurement in the Rasch model). Positive values indicate that question items are more difficult than average, while negative values indicate that they are easier. Like item 14 had a measure value of 1.40, making it the most difficult item. In contrast, item 11 had a measure value of -1.31, so it was the easiest item. The scale value range from 1.40 to -1.31 shows a fairly good variation in difficulty and can assess abilities at various levels.

Furthermore, the Standard Value error for each item is relatively small. Based on table 2, the score is around 0.21, which shows that the estimated difficulty level of this item is quite stable. A low error standard reinforces confidence in the item's difficulty estimation, which means these results are reliable.

b) INFIT and OUTFIT

The Infit Mean Square (MNSQ) measures how feasible the response on the item matches the predictions of the Rasch model. Ideally, the MNSQ value should be around 1.0. In this table, most items show MNSQ Infit values that are close to 1.0, such as item 14 (0.92), indicating that the data on these items can measure thinking ability. From the table, it is known that the minimum infit value of 0.89 and a minimum outfit of 0.86 in question 1. Furthermore, the maximum value on infit 1.30 and outfit 1.39 is found in question 7.

The Outfit Mean Square (MNSQ) measures the mismatch in the extremities (unexpected responses, either too high or too low). The MNSQ outfit value for question 7 is 3.14, which is well above the acceptable range (usually between 0.5 to 1.5 for a good fit). A high MNSQ outfit indicates that the response to the item has greater variability. These results show that this question cannot be understood well for respondents at very high or very low levels of ability.

ZSTD gives an indication of the statistical significance of the MNSQ Infit and Outfit. The ZSTD outfit value for question 7 is 3.99. This suggests that the unexpected variation observed is statistically

significant. Obtaining a ZSTD score above 2.0 is considered problematic, which means this question may not align with the objectives being measured.

c) Measuring Point Correlation

PTMEASURE-CORR is a correlation between the score on each item and the total score on all instruments. This reflects how well the item contributes to the overall measurement. Based on Figure 2, it can be seen that the overall positive correlation value (+) is obtained, so it is known that each question item is rated well in measuring each respondent's ability. Correlation values ranged from 0.25 to 0.48, indicating that most items contributed positively to the intended capability measurement. The item with a lower correlation in item 13 (0.25), indicates that this item should be improved to measure students' cognitive thinking ability.

d) Exact Match

Observed match shows the percentage of match between the observed and predicted responses. The value obtained ranged from 60.3% (item 7) to 84.5% (item 14), with an average of 69.7%. Item 14 had the highest percentage of matches, indicating that respondents answered these items as expected by the Rasch model.

Expectation Correlation shows the percentage of matches expected by the Rasch model. The result table shows that there is a problem 7 with a point-to-point correlation value of 0.02, which is much smaller than the expectation correlation value of 0.36. The same thing can also be seen in question items 14, 9, 20, 1, 15, and 11, where the correlation value of the measuring point is greater and the range is far from the expected correlation value.

3.3. Wright Person-Item Map Analysis

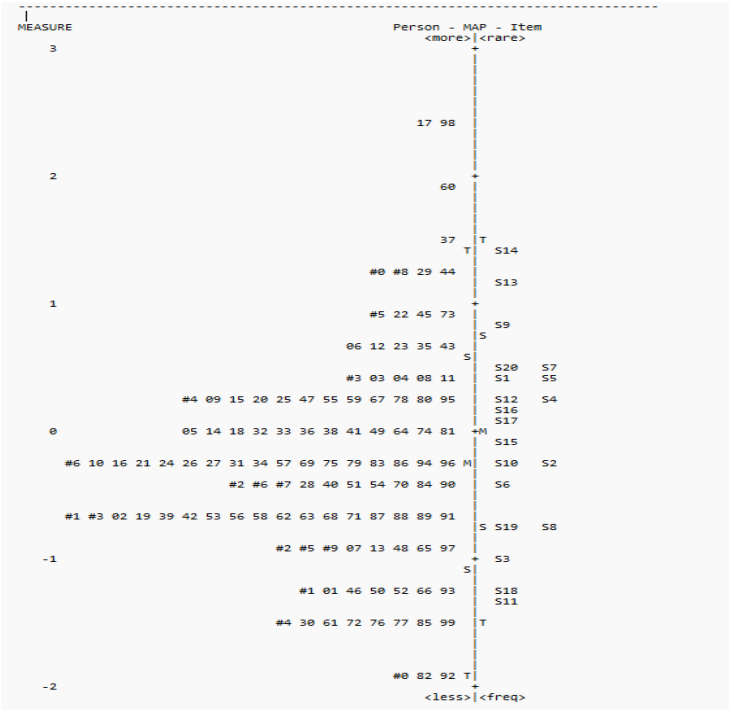

Figure 3. Results of Wright Map Analysis Person-Item Map

The Wright map visually represents the distribution of respondents' abilities and the difficulty of items in the test on the same scale. This analysis aims to understand how well these test instruments cover the different levels of respondents' abilities and how the difficulty of items is spread across those range of abilities.

The Wright Person-Item Map shows two sides of the analysis, namely the distribution of respondents (*logit person*) and the right side is the distribution of *logit* of question items. This analysis can provide an overview of the results of the distribution of students' abilities and the distribution of problem difficulties.

Based on Figure 3, The Wright Map shows how well the test matches student abilities with question difficulty. it is known that the left wright map (*logit* person) shows that there are several students as respondents with high ability, namely exceeding two standard deviations (T) between +2 to +3 *logits.* Students at the top of the map, like number 17 and 98, did very well and answered many difficult questions correctly. On the other hand, students at the bottom, such as number 82 and 92, struggled with the harder questions. It is rated low because it has a small *logit value* of -2 and is limited by two standard deviations.

On the right side of the map, the questions are listed by difficulty. This map shows that the test has a good mix of easy and hard questions, allowing it to challenge students with different ability levels. However, some of the hardest questions may have been too difficult for many students, while the easiest questions might not have been challenging enough for the top-performing students. The harder questions, like S14 and S13, are at the top and were only answered correctly by the stronger students. Easier questions, like S18 and S11, are at the bottom and were answered correctly by most students, even those who had trouble with the harder ones. Overall, the test works well, but it could be improved by adjusting some of the very easy or very hard questions to better fit the abilities of the students taking it.

The item map provides information on the difficulty level of the item (on a *logit scale*) for each item of the question. The item logit map shows that the S14 question item has the highest *logit* value, which is between 1 to 2 *logits*, but it is still in the range of two standard deviations. This data shows that this problem is quite difficult and not all respondents can do it. The level of difficulty of the questions plays an important role in determining the accuracy of the measurement of various levels of respondents' abilities. With this, it is clear that it can accurately measure the respondents who are in the top position on the map.

In contrast, the question with *the lowest log it* is S11, which shows that this question item has a low level of difficulty or is considered easy and almost all respondents can answer correctly. The lowest logit score, is needed to assess basic ability, so the test is effective for identifying respondents with lower ability.

Some questions have the same *logit* score, such as S20-S7, S1-S5, S12-S4, S10-S2; S19-S8, which can be interpreted that the two questions have the same level of difficulty. These findings help maintain a balance of difficulty in the test, too many items of the same difficulty level can reduce the test's ability to effectively distinguish respondents' abilities.

Discussion

The results of the overall test and its validity are highly dependent on the difficulty level of the question items according to the ability to resonate. If only a few questions are difficult, high-ability students may get almost the same score, making it difficult to distinguish their abilities. On the other hand, if there are only a few easy questions, low-ability students do not get a fair opportunity to demonstrate their abilities. Thus, the inclusion of questions with varying levels of difficulty is very important to accurately measure students' abilities.

Item reliability index refers to similarity in terms of item difficulty compared to other samples of equivalent abilities. So it can be seen that with the acquisition of the item reliability index value in the category is very good.The high-reliability index of the items obtained in this data, indicates that the items in the test have a stable level of difficulty. Furthermore, it is reliable when applied to different populations with comparable capabilities.

So it can be seen that reliability items that are included in the Very Good category, it indicates that the questions developed to measure the ability to think consistently in various situations. High item

reliability results ensure that test results do not depend on irrelevant factors, such as samples or test conditions, but rather truly reflect the capabilities you want to measure. Tests with high item reliability provide consistent and accurate results. It shows that the questions developed to measure thinking ability are considered reliable.

Supported by (Raof et al., 2021) items with sufficient to good reliability are essential to measure that the developed test consistently measures the cognitive construct to be measured. Item reliability is an important indicator to show the feasibility of items in the test with the aim of measuring the intended ability. The high reliability value of the item explains that the variation in responses from participants is not the result of inconsistencies in the item design. This gives meaning to the difference in objectively measured abilities.

The results of another study emphasized that the reliability of acceptable items, as well as the reliability of separation, demonstrated the ability of the items in the test to distinguish characteristics appropriately. The items developed are able to group participants based on different skill levels in a structured and reliable way. Furthermore, good reliability also ensures that the test results can be used for equivalent samples. These findings provide high consistency of results over time and in different contexts. This is very important to maintain the external validity of the test, as the instrument must still be able to measure the same across different groups of participants (Schmiemann, Nehm, & Tornabene, 2017).

The high reliability value of the items in this study can be seen that the items developed to measure thinking ability have been well designed. These items work effectively to differentiate participants based on the student's ability level, while maintaining high reliability. These gains show that the test is not only consistent in measuring the desired construct, but also has the potential to be widely used in a variety of populations with similar results. As has been proven in various previous studies, the high reliability of the item serves as a solid basis for ensuring that the test can serve as a valid and reliable measuring tool (Raof et al., 2021; Schmiemann et al., 2017).

INFIT and OUTFIT are statistics used to evaluate the extent to which the data matches the model. Fit items for measuring the developed construction can see fit items in measuring construction as observations on study data (Raof et al., 2021; B Sumintono, 2017). INFIT focuses on the suitability of responses from respondents who have a level of ability close to the difficulty level of the item. Data acquisition from INFIT is more sensitive to the expected response. Furthermore, OUTFIT will provide information against unexpected extreme responses. Responses came from respondents with very high or very low abilities compared to the difficulty level of the item (Laliyo, Tangio, Sumintono, Jahja, & Panigoro, 2020; B Sumintono, 2017; Bambang Sumintono, 2015, 2016; Wicaksono, Roebianto, & Sumintono, 2021). These statistics are used to determine whether the item under test is functioning according to the intended purpose in the measurement.

INFIT and OUTFIT values close to 1 indicate that the item matches the model. A value much greater than 1 indicates that the item may be too difficult or too easy, while a value much smaller than 1 indicates that the item may be redundant. INFIT and OUTFIT values close to 1 indicate that the item fits the model, meaning that the item performs well in measuring the desired construct (Raof et al., 2021; B Sumintono, 2017; Bambang Sumintono, 2018). If the INFIT or OUTFIT value is much greater than 1, it could indicate that the item is too difficult or too easy, which causes the item to not be able to distinguish respondents well. For example, if an item is too difficult, many participants may answer it randomly, which results in unexpected data variability and makes the item not fit the model. Conversely, if the INFIT or OUTFIT value is much less than 1, it indicates that the item may be too easy or too similar to other items, so it does not provide new information and instead introduces redundancy in the measurement.

Evaluation of INFIT and OUTFIT is essential to ensure that each item is able to accurately measure the participant's thinking ability and does not generate bias. It is an important tool for assessing the quality of items and ensuring that they are suitable for valid and reliable measurements (Shiau Wei Chan, Ismail, & Sumintono, 2014; Raof et al., 2021; Bambang Sumintono, 2015, 2016).

Criteria for good questions based on fit items:

1) The value of the outfit mean square (MNSQ) is acceptable in the value range of 0.5< MNSQ <1.5
2) Outfit Z-standard (ZSTD) grades are acceptable in the value range of -2.0< ZSTD <2.0
3) The value of point measure correlation (Pt Mean Corr) is accepted in the range of 0.4< Pt Measure Corr <0.85.

Questions are accepted if they meet at least 2 of the provisions of the good question criteria. If not, the questions must be discarded or replaced (S W Chan, Looi, & Sumintono, 2021; B Sumintono, 2017).

Measure-Point Correlation (PTMEASURE-CORR.) is the correlation between an item's score and a total score. A positive value (+) indicates that the item is able to distinguish between respondents with different abilities. If the value obtained is otherwise negative (-), it means that the item being developed has failed to measure the construction of interest. Therefore, it is necessary to correct or drop because the item does not lead to a question or is difficult for respondents to answer.

Point Measure Correlation (PTMEA CORR) value analysis is used to detect polarity items in testing the extent to which the construction construction achieves its objectives. Correlation of Expectations (EXP. CORR.) is the expected correlation between the item score and the total score based on the model. If the Measure Point Correlation is close to the Expectation Correlation, it indicates that the item is valid (B Sumintono, 2017).

A positive PTMEA CORR value (+) indicates that the item is effective in distinguishing respondents with different abilities. In contrast, a negative PTMEA CORR value (-) indicates that the item fails to measure the construct in question (Shiau Wei Chan et al., 2014; Bambang Sumintono, 2016; Wicaksono et al., 2021). The acquisition of PTMEA CORR is negative, it must be revised or replaced as needed, because this item is irrelevant to the measurement or too difficult to answer by most respondents (Raof et al., 2021).

PTMEA CORR analysis is also used to detect the polarity of items, i.e. to ensure that they perform in accordance with the purpose for which they were designed. Correlation of Hope (EXP. CORR.) is the expected correlation between the item score and the total score, based on the measurement model. If the PTMEA CORR value is close to the EXP value. CORR., indicating that the item has good validity and functions as expected and in accordance with the intended purpose (B Sumintono, 2017; Bambang Sumintono, 2018).

Wright Map presents participants' abilities and item difficulty on the same scale. This analysis allows researchers to evaluate whether the items in the test are well spread out to measure a wide range of participants' abilities. According to (Bond, Yan, & Heene, 2020; Kristyasari, Asmaningrum, & Hanif, 2022), the good distribution of items in the Wright Map shows that the test can capture variations in participants' abilities more effectively. However, if many items have similar logit values, the test can lose its discriminatory power, meaning that the test may not be effective enough in distinguishing participants based on abilities within the same difficulty range (Boone, 2019; Boone, Staver, & Yale, 2014a, 2014b). Instead, items that are spread evenly across the logit scale help accurately measure participants' abilities at different levels. This is important to ensure that participants with low, medium, or high abilities face items that match their ability level, which will ultimately improve the overall reliability and validity of the test.

The findings of the study showed that items with positive PTMEA CORR were close to the correlation of expectations (EXP. CORR.) In general, it shows that most of the items work well in distinguishing the abilities of participants. However, the presence of several items with similar logit values indicates that these items may be less effective at distinguishing participants at certain skill levels, which need to be improved. In accordance with the findings of (Boone, 2019), a less varied distribution of items can reduce the discriminatory strength of the test, so it is necessary to review the difficulty distribution of items to be more accurate and useful.

The results of the item map in figure 3 are considered ideal, because each interval in the map is represented by a test item, starting from low to high logit. This question can represent a measurement

of the respondent's ability from low to high. Several questions with the same logem need to be considered or revised in order to measure the goal well. The average item difficulty is set at 0 logits; A higher logit score indicates a more difficult item and a lower logit score (negative) indicates an easier item (Schmiemann et al., 2017).

Based on the analysis that has been carried out, namely reliability, validity, and the Wright Map, this study shows that the multiple-choice test instrument developed has a good level of reliability. This category is found in reliability for both individuals and items. The reliability gains in the findings of this study show that the developed test consistently measures the cognitive thinking ability of participants with a high level of reliability. Some items need to be refined to improve individual reliability, especially in distinguishing abilities in participants with more varied abilities.

The validity of the instrument shows that the difficulty level of the item varies well. INFIT and OUTFIT analysis shows that most items have a good match with Rasch's model. PTMEA CORR shows a positive correlation between item scores and total scores. While most items do a good job of distinguishing participants' abilities, there are items like item 13 that have a lower correlation (0.25). This needs to be improved to improve its ability to effectively measure participants' cognitive thinking skills.

The Wright Map analysis shows that the test development has a good enough distribution of items to measure the different levels of respondents' abilities. The more difficult items, such as items 14 and 13, are located at the top of the map and are only answered correctly by participants with higher abilities. In contrast, the easier items, such as items 11 and 18, are located at the bottom of the map and are answered correctly by most participants, including those with lower abilities. Although the distribution of items varies well, multiple items with the same logit score can reduce the discriminatory power of the test. So these items need to be reviewed to improve the effectiveness of the test in distinguishing participants' abilities.

The results of this study are in line with previous research which emphasizes the importance of reliability and validity in measurement instruments, especially in the context of measuring cognitive abilities (Pratiwi, Kuntjoro, Sunarti, & Budiyanto, 2023; Raof et al., 2021; Schmiemann et al., 2017; B Sumintono, 2017). These findings support the research of (Raof et al., 2021) which states that items with good reliability are essential to ensure consistency in the measurement of cognitive constructs. These findings also reinforce the research of (Schmiemann et al., 2017) which emphasized the importance of item distribution in the test to effectively capture variations in participants' abilities.

However, there are different results from previous studies. The finding that some items have a high OUTFIT value and need improvement. This shows the importance of more in-depth review and revision of items to increase the discriminatory power of the test. It is affirmed by (Boone, 2019) that less varied distribution of items can reduce the discriminatory power of the test.

From these findings, it can be concluded that the multiple choice test instrument used has good reliability and validity, but there is still room for improvement in some items, especially those with high OUTFIT values and low PTMEA correlation. To improve the test, revisions are made to items that are too easy or too difficult, as well as a review of items with similar difficulties. So it can be assessed that this test has the potential to be widely used in measuring cognitive ability, but some improvements are needed to further improve its reliability. Future research may focus on further development of underfunctional items and conducting additional trials on different populations to ensure consistency of results.

4. CONCLUSION

Based on the results, it can be concluded that the multiple-choice test instrument developed in measuring students' thinking ability in genetics lectures has good quality and can be used in accordance with these objectives. The use of the Rasch Model in analyzing this multiple-choice test instrument can determine many components in evaluating the quality of the instruments used, including: 1) Reliability of person-items, which ensures consistency in individual ability measurements, 2) Overall validity such

as Item Fit, Measuring Point Correlation and Expectation Correlation, and 3) Wright Person-Item Map. Overall, the purpose of this study has been achieved to evaluate test instruments to measure the level of students' thinking ability in the context of genetics. Using the Rasch Model, tests can be analyzed in more detail, so that improvements to the instrument can be made more precisely.

There are limitations in this research, there are items with high *Outfit* values. This suggests that it may not be suitable for all students, especially those with very high or low levels of ability. In addition, the study was conducted on only one population group. Further research is needed on more diverse populations so that the test can be consistent across different groups of students. In addition, improvements to items that do not fit are needed to improve the test's ability to distinguish students abilities. Future research may also expand the use of the Rasch Model to develop assessment instruments.

REFERENCES

Aivelo, T., & Uitto, A. (2018). *Teachers ' approaches to genetics teaching mirror their perceptions of teaching controversial , societal and sensitive issues Teachers ' approaches to genetics teaching mirror their perceptions of teaching controversial , societal and sensitive issues*. (July). https://doi.org/10.1101/350710

Aivelo, T., & Uitto, A. (2021). Factors explaining students' attitudes towards learning genetics and belief in genetic determinism. *International Journal of Science Education*, *43*(9), 1408–1425. https://doi.org/10.1080/09500693.2021.1917789

Aramendiz, V. R. R., & Córdova, K. E. G. (2019). Decisions in evaluation: Virtual postgraduate environments, ex post-facto study. *Revista de Educación a Distancia*, *1*(59). https://doi.org/10.6018/red/59/06

Betu, F. S. (2023). KOMPONEN TUJUAN DALAM NEW TAXONOMY MARZANO & KENDALL DAN RELEVANSINYA BAGI PENDIDIK. *Atma Reksa : Jurnal Pastoral Dan Kateketik*, *7*(1). https://doi.org/10.53949/ar.v7i1.142

Bond, T. G., Yan, Z., & Heene, M. (2020). Applying the rasch model: Fundamental measurement in the human sciences. In *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. https://doi.org/10.4324/9780429030499

Boone, W. J. (2019). Rasch methods for beginners. *Tidsskriftet "Pædagogisk Psykologisk Tidsskrift"S*, *1*.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014a). Item Measures. In *Rasch Analysis in the Human Sciences*. https://doi.org/10.1007/978-94-007-6857-4_5

Boone, W. J., Staver, J. R., & Yale, M. S. (2014b). Understanding Person Measures. In *Rasch Analysis in the Human Sciences*. https://doi.org/10.1007/978-94-007-6857-4_4

Chan, S W, Looi, C. K., & Sumintono, B. (2021). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. *Journal of Computers in Education*. https://doi.org/10.1007/s40692-020-00177-2

Chan, Shiau Wei, Ismail, Z., & Sumintono, B. (2014). A Rasch Model Analysis on Secondary Students' Statistical Reasoning Ability in Descriptive Statistics. *Procedia - Social and Behavioral Sciences*, *129*. https://doi.org/10.1016/j.sbspro.2014.03.658

DeLuca, C., Chapman-Chin, A. E. A., LaPointe-McEwan, D., & Klinger, D. A. (2018). Student perspectives on assessment for learning. *Curriculum Journal*, *29*(1), 77–94. https://doi.org/10.1080/09585176.2017.1401550

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, Supplement*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Etobro, A. B., & Banjoko, S. O. (2017). Misconceptions of genetics concepts among pre-service teachers. *Global Journal of Educational Research*, *16*(2), 121. https://doi.org/10.4314/gjedr.v16i2.6

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Kristyasari, M. L., Asmaningrum, H. P., & Hanif, R. F. (2022). Wright-Map to Analyze Students' Abilities on Chemical Bond Test. *SHS Web of Conferences*, *149*. https://doi.org/10.1051/shsconf/202214901050

Laliyo, L. A. R., Tangio, J. S., Sumintono, B., Jahja, M., & Panigoro, C. (2020). Analytic approach of response pattern of diagnostic test items in evaluating students' conceptual understanding of characteristics of particle of matter. *Journal of Baltic Science Education*, *19*(5). https://doi.org/10.33225/jbse/20.19.824

Machluf, Y., Gelbart, H., Ben-Dor, S., & Yarden, A. (2017). Making authentic science accessible-the benefits and challenges of integrating bioinformatics into a high-school science curriculum. *Briefings in Bioinformatics*, *18*(1), 145–159. https://doi.org/10.1093/bib/bbv113

Pane, J. F., Steiner, E. D., Baird, M. D., Hamilton, L. S., & Pane, J. D. (2017). *Informing Progress*. (July), Ahmed, M. & Change, O. (2017) Academics' e-learnin.

Pratiwi, M. K., Kuntjoro, S., Sunarti, T., & Budiyanto, M. (2023). TOSLS Cognitive Instrument to Measure Students' Scientific Literacy Abilities. *IJORER : International Journal of Recent Educational Research*, *4*(6). https://doi.org/10.46245/ijorer.v4i6.432

Prochazkova, K., Novotny, P., Hancarova, M., Prchalova, D., & Sedlacek, Z. (2019). Teaching a difficult topic using a problem-based concept resembling a computer game: Development and evaluation of an e-learning application for medical molecular genetics. *BMC Medical Education*, *19*(1), 1–8. https://doi.org/10.1186/s12909-019-1817-2

Raof, S. A., Mustaâ, A. H., Zamzuri, F. K., & Salleh, M. H. (2021). Validity and reliability of students perceptions on OBE approach in Malaysian VC using Rasch model. *Journal of Innovation in …*.

Sari, W. K., & Nada, E. I. (2022). Marzano Taxonomy-Based Assessment Instrument to Measure Analytical and Creative Thinking Skills. *Jurnal Pendidikan Kimia Indonesia*, *6*(1). https://doi.org/10.23887/jpk.v6i1.40117

Schmiemann, P., Nehm, R. H., & Tornabene, R. E. (2017). Assessment of Genetics Understanding: Under What Conditions Do Situational Features Have an Impact on Measures? *Science and Education*, *26*(10), 1161–1191. https://doi.org/10.1007/s11191-017-9925-z

Septi Purnama, D., Farozin, M., & Astuti, B. (2022). The ryff's psychological well-being scale for indonesian higher education students: A RASCH model analysis how to cite. *IRJE |Indonesian Research Journal in Education*, *6*(2).

Sumintono, B. (2017). *Rasch model measurement as tools in assessment for learning*. eprints.um.edu.my.

Sumintono, Bambang. (2015). Pemodelan Rasch pada Asesmen Pendidikan: Suatu Pengantar. *Konferensi Guru Dan Dosen Nasional (KGDN) 2015*, (November 2015).

Sumintono, Bambang. (2016). Aplikasi Pemodelan Rasch pada asesmen pendidikan: Implementasi penilaian formatif (assessment for learning). *Makalah Dipresentasikan Dalam Kuliah Umum Pada Jurusan Statistika, Institut Teknologi Sepuluh November, Surabaya, 17 Maret 2016.*, (March).

Sumintono, Bambang. (2018). *Rasch Model Measurements as Tools in Assesment for Learning*. https://doi.org/10.2991/icei-17.2018.11

Taylor, H. A., Rapp, D. N., & Brunye, T. A. D. T. (2007). Repetition and Dual Coding in Procedural Multimedia Presentations. *Applied Cognitive Psychology*, *22*(September 2007), 877–895. https://doi.org/10.1002/acp

Tsui, L. (2002). Fostering critical thinking through effective pedagogy: Evidence from four institutional case studies. *Journal of Higher Education*, *73*(6). https://doi.org/10.1080/00221546.2002.11777179

Wicaksono, D. A., Roebianto, A., & Sumintono, B. (2021). Internal Validation of the Warwick-Edinburgh Mental Wellbeing Scale:Rasch Analysis in the Indonesian Context. *Journal of*

Educational, Health and Community Psychology, *10*(2). https://doi.org/10.12928/jehcp.v10i2.20260

Zhang, L., Liu, X., & Feng, H. (2023). Development and validation of an instrument for assessing scientific literacy from junior to senior high school. *Disciplinary and Interdisciplinary Science Education Research*, *5*(1). https://doi.org/10.1186/s43031-023-00093-2