Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 9, Special Issue, 2025

" Emerging Trends, Challenges, and Innovative Practices in Education"

Investigating High School Student Misconceptions: A Rasch-Based Three-Level Diagnostic Evaluation of Osmoregulation and Excretion Systems

Ridwan [1*], Sunarto A. Sura [1], Diki Koerniadi [1], Lia Zaradiva [1], Soeharto Soeharto [2, 3, 4], Najira [1], Ghurrotul Bariroh [1], Fadil Muhammad [5]

[1] Universitas Pendidikan Indonesia, Bandung, Indonesia.

[2] Research Center for Education, National Research and Innovation Agency, Indonesia.

[3] Research Center of Educational Technologies, Azerbaijan State University of Economics, Azerbaijan.

[4] Linz School of Education, Johannes Kepler University, Austria.

[5] Biology Education, Faculty of Mathematics and Science, Universitas Negeri Makassar, Makassar, Indonesia.

Abstract

This study aimed to develop and validate a three-tier multiple-choice diagnostic assessment instrument to identify misconceptions related to osmoregulation and the excretion system among Indonesian high school students. A total of 281 students from West Sumatra and Jambi Province participated in the research. Employing a quantitative approach, the psychometric properties of a 20-item test were analyzed using Rasch modeling. The analysis revealed that the instrument had strong item reliability (0.84), though person reliability was relatively low (0.55), indicating variability in students' response consistency. Despite this, the test demonstrated high internal consistency, as shown by a Cronbach's Alpha of 0.90. The mean student ability level (-2.37) was significantly lower than the item difficulty level (0.00), suggesting widespread conceptual gaps among participants. All items met the model's expectations, with average Outfit Mean Square (MNSQ) at 1.02 and Z-Standard at 0.1. The findings highlight the diagnostic tool's effectiveness in detecting prevalent misconceptions in biology education. This study contributes to the field by offering a structured and psychometrically sound instrument, supporting more targeted instructional strategies to enhance conceptual understanding in science education.

Keywords:

Misconceptions;
Osmoregulation & Excretory Systems;
Diagnostic Assessment;
Rasch Analysis.

Article History:

Received:	05	April	2025
Revised:	04	July	2025
Accepted:	17	July	2025
Published:	12	August	2025

1- Introduction

Learning the excretory system and osmoregulation are fundamental components of the biology curriculum at the high school level, but this topic remains an area that often leads to conceptual misunderstandings among students. This misunderstanding of complex biological concepts not only hinders further learning but also has implications for the overall construction of scientific knowledge [1, 2]. The identification of such misunderstandings through valid and reliable diagnostic instruments, thus, becomes an important prerequisite for designing effective pedagogical interventions.

Previous empirical explorations have documented various misconceptions about the excretory system that are widespread among students. Enochson et al. [3] identified that many students experience fundamental confusion

[*] **CONTACT**: ridwantanjung@upi.edu

DOI: http://dx.doi.org/10.28991/ESJ-2025-SIED1-05

regarding the function and location of organs in the human body system. This is reinforced by Aydın's findings [4], which suggested that students often mistakenly identify organs such as the liver and pancreas as part of the excretory system, rather than understanding their true role in the digestive and metabolic processes. This kind of misunderstanding, as identified by Ameyaw & Okyer [5], can come from teaching methods that do not clarify the fundamental differences between the digestive and excretory systems. The factors that influence misconceptions are not only limited to students' cognitive inaccuracies, but are also influenced by their attitudes towards subjects, teaching approaches, and teachers' qualifications and competencies. Assem et al. [6] emphasized that even if school facilities are complete, learning outcomes can remain low if teaching methods are not appropriate and teachers are less competent in dismantling existing misconceptions.

Previous research on the learning environment has also revealed that certain learning models can reinforce or overcome misunderstandings. Kibirige & Mamashela [7] proved that approaches like Flipped Classroom can significantly stimulate students' cognitive engagement, encouraging them to confront conceptual misconceptions. On the contrary, research by Reichert et al. [8] has shown that traditional lecture-based methods often fail to address misconceptions effectively, leading to the reinforcement of improper concepts in students' cognitive structures.

To comprehensively address this problem, the development of reliable diagnostic instruments has become crucial. The Three-Level Test has emerged as a potential assessment methodology, consisting of three levels of questions: (1) conventional answer choices, (2) justifications for such choices, and (3) respondents' confidence level assessments [9, 10]. This methodology allows for a more precise differentiation between actual misconceptions and errors derived from random guessing or lack of knowledge [11]. Kirbulut & Geban [11] have demonstrated the advantages of multi-level diagnostic tests in revealing the complexity of students' conceptual understanding compared to traditional methods.

Current literature also indicates that misconceptions in biology often show resistance to change and can come from a variety of sources, including previous learning experiences and suboptimal instructional strategies [12, 13]. Integration of formative assessment into the teaching of biological concepts, as evidenced by Çakmak & Bulunuz [14], has been shown to improve understanding and retention of concepts. The three-level diagnostic approach allows for the collection of empirical data on student-specific misconceptions regarding osmoregulation, such as misconceptions about the role of the kidneys and the mechanisms of water balance in the body [15, 16].

In order to improve the construct validity and reliability of diagnostic instruments, this study adopts Rasch measurements as a framework for psychometric analysis. Rasch measurements, introduced by Georg Rasch, provide a robust statistical methodology for converting ordinal data into interval data, allowing for more robust analysis compared to classical test theory approaches [17, 18]. One of the fundamental advantages of Rasch measurements is their ability to assess the unidimensionality of constructs, ensuring that the instrument measures a single underlying attribute, thereby increasing the validity of the measurement [19, 20]. Recent research by Ma et al. [21] suggests that multi-level-based tests such as five-level tests combined with Rasch's model have high capabilities in identifying complex conceptual misconceptions. This study emphasizes the importance of layered and contextual testing to explore the sources of misconceptions and interactions between causes.

Rasch measurements also facilitate the identification of Differential Item Functioning (DIF), a phenomenon in which individuals from different demographic groups (e.g. by gender) respond differently to certain items despite having the same level of ability [22, 23]. This aspect is critical in educational research to ensure equity of measurement across diverse student populations [24]. Its application in the context of diagnostic instruments helps researchers identify and address potential biases in measurements, improving the fairness and accuracy of assessments [25-27].

Although there are various studies that address students' misconceptions about the excretory system and osmoregulation, there are significant gaps in the literature that need to be filled. First, the majority of research is still limited to a descriptive approach without a standardized quantitative methodology. Second, the application of Rasch measurements in developing diagnostic instruments for this topic is still very limited. Third, studies integrating a three-level approach with Rasch's analysis to evaluate misconceptions in osmoregulation and excretory systems have not been empirically proven.

The study offers an innovative contribution by integrating a Three-Level Test approach that not only identifies students' misunderstandings, but also considers their confidence level in answering, thus providing a more comprehensive insight into the misunderstandings that occur [28]. In addition, the study adopted Rasch measurements to quantitatively assess the quality of items, allowing for more objective instrument validation than traditional methods [28, 29]. The results of this study are expected to provide evidence-based recommendations for more effective teaching strategies in overcoming students' misconceptions about the osmoregulation and excretion system. The main objectives of this study are to:

- Analyze the reliability and validity of the developed instruments using the Rasch measurement approach to ensure consistency and accuracy in evaluating students' misconceptions about osmoregulation and excretion systems.
- Describe the interaction between items and respondents in the developed instrument, to identify the suitability of the item's difficulty level with the student's abilities as well as the response patterns that emerge.
- Identify the development pattern of students' misunderstandings in understanding the concept of osmoregulation and the excretory system, as well as contributory factors.

- Analyzed the potential bias of gender-based instruments using the Differential Item Functioning (DIF) approach in the Rasch model.

The structure of this article is organized as follows: the second part describes the research methodology including the design of the three-level instrument and the Rasch analysis procedure; the third part presents empirical results and data analysis; the fourth section discusses the implications of the findings in the context of the existing literature; and the fifth section concludes the study with recommendations for further research and pedagogical practice.

Recent bibliometric studies by Amiruddin et al. [30] it shows that research trends regarding misconceptions and conceptual changes have developed significantly in the last three decades, but there has been a post-pandemic decline. This opens up opportunities for further research with stronger evaluative and diagnostic approaches, including the use of the Rasch model in three-level tests.

2- Methods

2-1- Research Design

A quantitative approach is used, in which a three-level multiple-choice test is administered to understand students' misconceptions in osmoregulation and excretory systems and Rasch modeling is used to analyze psychometric properties.

2-2- Participants

The participants of this preliminary study were 281 students at public high schools in Mentawai, Payakumbuh, South Coast, South Solok, Padang, parts of West Sumatra province and Jambi Province, Indonesia. The sample was recruited using random sampling stratification based on student scores. Data was collected from 281 students. Data collection was carried out from September to October 2024. Students spend 60 minutes completing the test under the supervision of researchers and teachers.

2-3- Instruments

Three-level multiple-choice diagnostic test:

To identify students' misunderstandings in the material of the osmoregulation system and the excretory system, 20 items were developed. A three-level multiple-choice diagnostic test to identify common misconceptions in this material. Then the chosen concept has been adjusted to the Indonesian education curriculum, namely the Independent Curriculum, especially at the high school level. All items in the test were translated using a back-and-forth translation from English to Indonesian and then from Indonesian to English by the researchers. Examples of questions in Indonesian and English can be seen in Table 1.

Table 1. Sample Assignments in English and English Versions

Indonesian Version	Pak Budi memiliki dua akuarium di rumahnya. Akuarium pertama diisi dengan ikan hias air tawar, sedangkan akuarium kedua diisi dengan ikan hias air asin. Setelah beberapa waktu, Pak Budi memperhatikan bahwa ikan di akuarium air asin tampaknya tidak menghadapi banyak masalah ketika ia sedikit terlambat mengganti airnya, tetapi ikan di akuarium air tawar menjadi tidak aktif ketika waktu penggantian airnya terlewat. Apa yang paling mungkin menjelaskan perbedaan reaksi ikan-ikan tersebut terhadap kondisi akuariumnya? • Ikan air asin lebih tahan terhadap perubahan suhu dibandingkan ikan air tawar. • Ikan air tawar telah terbiasa dengan jadwal pemberian pakan yang rutin, sementara ikan air asin lebih mandiri. • **Ikan air asin bisa menyesuaikan dengan kondisi air yang berubah lebih baik daripada ikan air tawar.** • Ikan air tawar membutuhkan level oksigen yang lebih tinggi dibanding ikan air asin untuk tetap aktif. • Ikan air tawar secara alami menghasilkan lebih banyak limbah daripada ikan air asin, sehingga memerlukan penggantian air yang lebih sering. Alasan: • Karena ikan air asin tidak perlu melakukan pergantian air karena bisa ber adaptasi dengan mudah di bandingkan ikan air tawar • Ikan air tawar cenderung menghasilkan limbah yang sangat banyak di bandingkan dengan ikan air asin • Karena ikan air tawar hidup dalam lingkungan dengan kadar garam yg rendah sehingga mereka membutuhkan lebih banyak oksigen terlarut dalam air untuk proses resprirasi • Disebabkan oleh adanya perbedaan tingkah laku, yang dimana ikan di akuarium air tawar memiliki tingkah laku yang lebih rendah ketika waktu penggantian airnya terlewat, sementara ikan di akuarium air asin tidak memperlihatkan masalah ketika airnya sedikit terlambat mengganti • **Ikan air asin menyesuaikan diri dengan perubahan kadar garam dalam air dengan lebih baik, karena lingkungan asin cenderung lebih stabil daripada lingkungan air tawar yang lebih rentan terhadap perubahan.** Tingkat keyakinan: i. Yakin ii. Tidak yakin

English Version	Pak Budi has two aquariums in his house. The first aquarium is filled with freshwater ornamental fish, while the second aquarium is filled with saltwater ornamental fish. After some time, Mr. Budi noticed that the fish in the saltwater aquarium did not seem to have much difficulty when he was a little late changing the water, but the fish in the freshwater aquarium became dormant when the water change time was missed. What is most likely to explain the different reactions of fish to aquarium conditions? • Saltwater fish are more resistant to temperature changes than freshwater fish. • Freshwater fish are accustomed to a regular feeding schedule, while saltwater fish are more independent. • **Saltwater fish can adapt to changing water conditions better than freshwater fish.** • Freshwater fish require higher oxygen levels than saltwater fish to stay active. • Freshwater fish naturally produce more waste than saltwater fish, so they require more frequent water changes. Reason: • Because saltwater fish do not need to make water changes because they can adapt easily compared to freshwater fish. • Freshwater fish tend to produce a lot of waste compared to saltwater fish • Because freshwater fish live in a low-saline environment, they need more dissolved oxygen in the water for respiration. • Due to the differences in behavior, fish in freshwater aquariums have lower behaviors when water changes are missed, whereas fish in saltwater aquariums show no problems when water changes are slightly delayed. • **Saltwater fish adapt to changes in salt content in the water better, as salty environments tend to be more stable than freshwater environments that are more susceptible to change.** Confidence level: i. Confident ii. Not sure

2-4- Rasch Procedures, Data Analysis, and Measurement

Before conducting data collection in schools, the researchers sought permission to conduct tests in schools and obtained ethical approval for the study. Online-based tests with google forms are conducted in students' classrooms with the guidance and supervision of researchers and teachers. Winsteps software version 3.0.0 was used in this study. Winsteps was used to perform data analysis using Rasch modeling. Winsteps performs Rasch analysis from a simple rectangular dataset. Winsteps can be used to analyze multiple-choice, dichotomous, and double-scoring questions as well as partial credit questions. In summary, the research procedure can be seen in Figure 1.

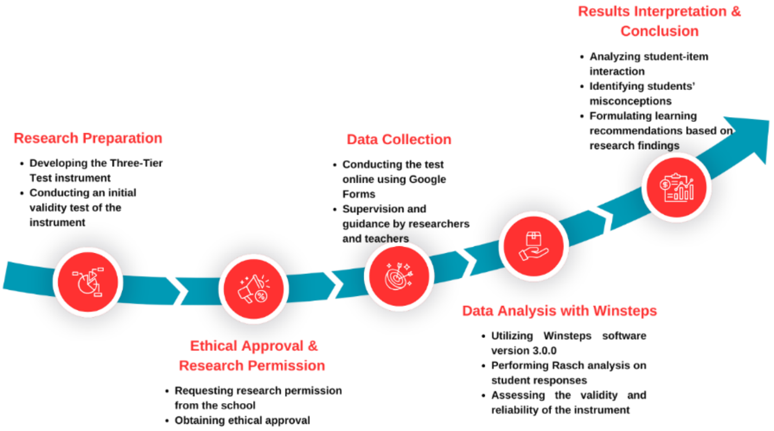

Results Interpretation & Conclusion
• Analyzing student-item interaction
• Identifying students' misconceptions
• Formulating learning recommendations based on research findings
Research Preparation
• Developing the Three-Tier Test instrument
• Conducting an initial validity test of the instrument
Data Collection
• Conducting the test online using Google Forms
• Supervision and guidance by researchers and teachers
Data Analysis with Winsteps
• Utilizing Winsteps software version 3.0.0
• Performing Rasch analysis on student responses
• Assessing the validity and reliability of the instrument
Ethical Approval & Research Permission
• Requesting research permission from the school
• Obtaining ethical approval

Figure 1. Research Procedure

To enrich the quantitative data, a limited qualitative follow-up was conducted in the form of semi-structured interviews with eight students who showed high confidence in the wrong answers in items Q6 and Q13. The results of the interviews show that these beliefs are based on previous learning experiences that prioritize memorization, not comprehension. For example, some students associate excretion only with "the process of removing substances from the body", without understanding the role of the kidneys in filtration and reabsorption in detail.

In addition, some students admitted to answering confidently because they were used to choosing answers that they thought "sounded scientific", even though their understanding was superficial. This phenomenon is in line with the Dunning-Kruger effect, in which individuals with low competence tend to overestimate their abilities.

3- Results and Discussion

3-1- Result

3-1-1- Descriptive Analysis

The description of the data from the research results can be used to enrich the discussion, through the description of the respondent response data it can be seen how the respondents respond to each variable studied. To find out the description of students' conceptual abilities related to the material, you can see the Tier 1 and 2 test results for each item. To facilitate research, scores were calculated from students' answers. The principle of score calculation is based on correct and incorrect answers in Tier 1 and 2. Both Tiers must be correct to be able to get a score of 1, and if they don't, they will be given a score of 0.

Table 2 presents the principle of scoring based on students' answers in Tier 1 and Tier 2. If the student answers correctly in both tiers, then a score of 1 is given. If it is wrong in one or both of them, then the score is 0. This principle allows for the analysis of pure conceptual understanding because it eliminates the influence of guesswork.

Table 2. Score Calculation Principle

Not.	Thing		Score
	Level 1	Level 2	
1	B	B	1
2	B	S	0
3	S	B	0
4	S	S	0

3-2-Descriptive Analysis of Tier 1 and Tier 2 Variables

In the test results with 20 items in 281 students, the following are the results of the recapitulation based on the number of students who successfully answered true/wrong on each item (Tier 1 and 2).

Table 3 shows the proportion of students who answered correctly and incorrectly for each question item. This data shows that Q9 questions are answered correctly the most, while Q6 is the most difficult. Figure 2 then represents a graph of the number of students who answered with high (Y) and low (TY) confidence levels. Item Q2 shows the highest level of confidence, while Q16 shows the opposite.

Table 3. Recapitulation of the Number of Students Based on Item Value

Item No.	Number of Students		
	1	0	Entire
Question 1	39	242	281
Question 2	63	218	281
Question 3	52	229	281
Question 4	37	244	281
Question 5	61	220	281
Question 6	22	259	281
Question 7	30	251	281
Question 8	43	238	281
Question 9	68	213	281
Question 10	38	243	281
Question 11	33	248	281
Question 12	43	238	281
Question 13	23	258	281
Question 14	39	242	281
Question 15	60	221	281
Question 16	32	249	281
Question 17	38	243	281
Question 18	40	241	281
Question 19	30	251	281
Question 20	42	239	281

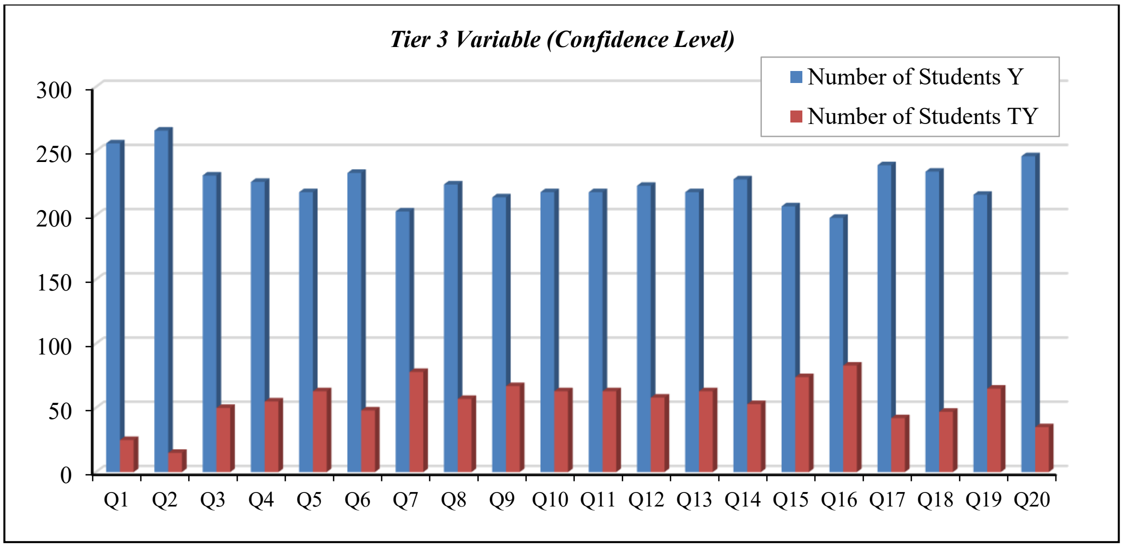


Figure 2. Recapitulation of the number of students based on the level of confidence of the answer

3-3- *Descriptive Analysis of Tier 3 Variables*

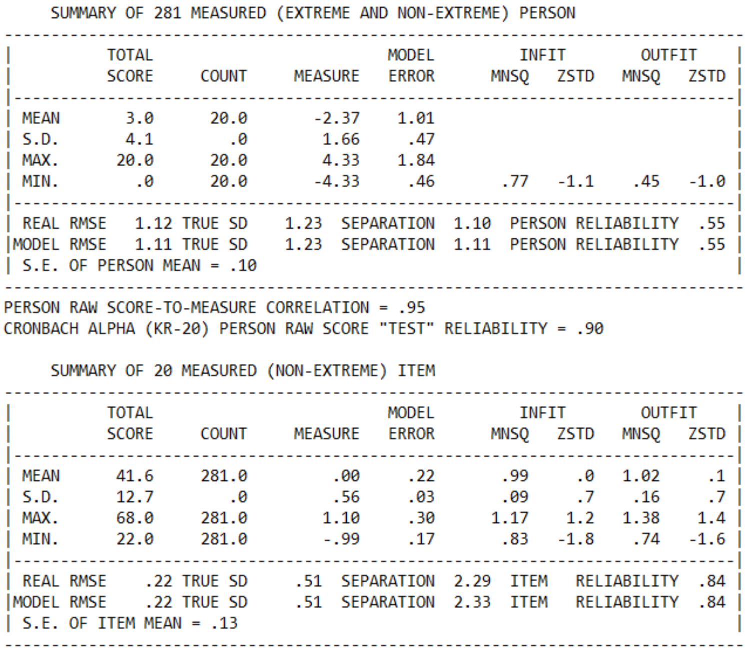
In the test results with 20 items in 281 students, the following are the recapitulation results based on the number of students who chose Y (definitely) and TY (not sure) for the answers given. Based on Table 4, it can be seen that students tend to be confident in their answers to item Q2 and unsure for Q16 answers.

Table 4. Recapitulation of the number of students based on the level of confidence of the answer

Item No.	Number of Students		
	Y	TY	Entire
Question 1	256	25	281
Question 2	266	15	281
Question 3	231	50	281
Question 4	226	55	281
Question 5	218	63	281
Question 6	233	48	281
Question 7	203	78	281
Question 8	224	57	281
Question 9	214	67	281
Question 10	218	63	281
Question 11	218	63	281
Question 12	223	58	281
Question 13	218	63	281
Question 14	228	53	281
Question 15	207	74	281
Question 16	198	83	281
Question 17	239	42	281
Question 18	234	47	281
Question 19	216	65	281
Question 20	246	35	281

3-4- *Summary of Measurements on 281 Students and 20 Items*

This summary is just a preliminary overview of the results of Rasch's analysis for 281 students and the 20 items used. Based on Figure 3, it can be seen that the average score of the student size is -2.37 and the average item size is 0.0, which means that the student's ability level is lower than the difficulty level of the question so that it can be said that the question or item is too difficult for the average student. The reliability value of the person is 0.55 and the reliability value of the item is 0.84. This shows that the consistency of the students' answers is still relatively weak, but the quality of the items is quite good. The Cronbach Alpha value is 0.90. This shows that the reliability of the test is generally quite good. In general, the reliability value of >0.6 is considered good.



```
     SUMMARY OF 281 MEASURED (EXTREME AND NON-EXTREME) PERSON
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       3.0      20.0       -2.37    1.01                                 |
| S.D.       4.1        .0        1.66     .47                                 |
| MAX.      20.0      20.0        4.33    1.84                                 |
| MIN.        .0      20.0       -4.33     .46       .77   -1.1    .45   -1.0 |
|-----------------------------------------------------------------------------|
| REAL RMSE    1.12 TRUE SD    1.23  SEPARATION  1.10  PERSON RELIABILITY  .55 |
|MODEL RMSE    1.11 TRUE SD    1.23  SEPARATION  1.11  PERSON RELIABILITY  .55 |
| S.E. OF PERSON MEAN = .10                                                   |
-------------------------------------------------------------------------------
PERSON RAW SCORE-TO-MEASURE CORRELATION = .95
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .90

     SUMMARY OF 20 MEASURED (NON-EXTREME) ITEM
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      41.6     281.0         .00     .22       .99     .0   1.02     .1 |
| S.D.      12.7        .0         .56     .03       .09     .7    .16     .7 |
| MAX.      68.0     281.0        1.10     .30      1.17    1.2   1.38    1.4 |
| MIN.      22.0     281.0        -.99     .17       .83   -1.8    .74   -1.6 |
|-----------------------------------------------------------------------------|
| REAL RMSE     .22 TRUE SD     .51  SEPARATION  2.29  ITEM   RELIABILITY  .84 |
|MODEL RMSE     .22 TRUE SD     .51  SEPARATION  2.33  ITEM   RELIABILITY  .84 |
| S.E. OF ITEM MEAN = .13                                                     |
-------------------------------------------------------------------------------
```

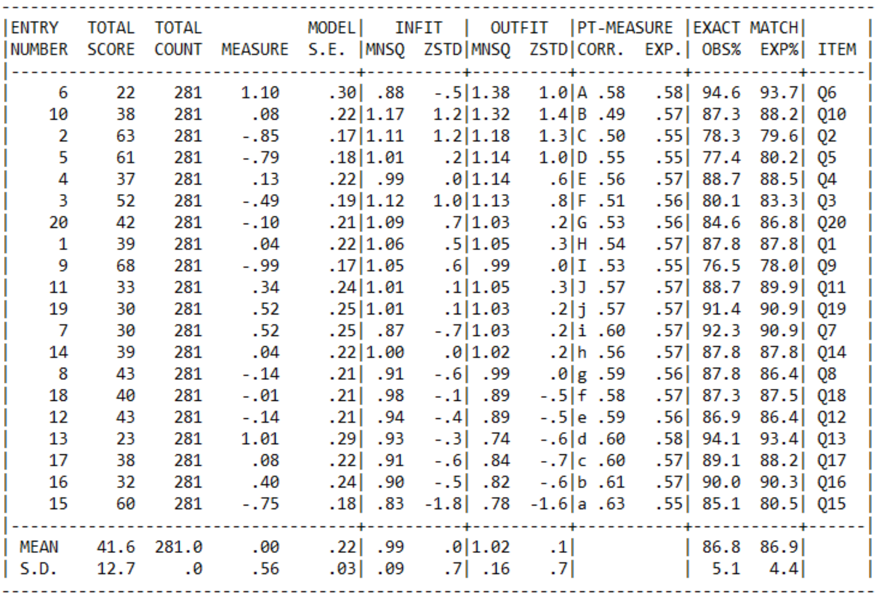
Figure 3. **Summary of People and Goods Measured**

3-5-Item Fit Rate

A well-fitting item means that it behaves consistently with what the model expects. If it is found that the item does not match, this is an indication that there is a misunderstanding on the part of the student about the item. According to the criteria used to check the appropriate items are [31]:

- Average Square Clothing Value (MNSQ) Accepted: 0.5 < MNSQ < 1.5
- Z standard Clothing Value (ZSTD) accepted: -2.0 < ZSTD < +2.0

Based on Figure 4, all items have a Mean Square outfit (MNSQ) value of more than 0.5 and less than 1.5 with an average of 1.02. The value of Z-Standard (ZSTD) clothing obtained by each item is more than -2.0 and less than +2.0 with an average of 0.1. From these results, it can be seen that all items show conformity with the requirements of the criteria. From Figure 5, it can be seen that the items (circles) tend to be around a vertical line which shows that most of the items fit the IRT (Item Response Theory) model and it can be said that the quality of the test items is quite good. From the two images, it can be concluded that all instrument items function normally in taking measurements. Figure 5 shows a pie chart of item suitability based on the Rasch Model. All items are within the tolerance range (0.5 < MNSQ < 1.5 and -2 < ZSTD < +2), indicating that the quality of the items is quite good. The graph shows the item is close to the vertical line of the model, which signifies the item is measuring consistently.



```
-------------------------------------------------------------------------------------------
|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 |
|------------------------------------+----------+----------+-----------+-----------+------|
|     6     22    281    1.10    .30| .88   -.5|1.38   1.0|A .58   .58| 94.6  93.7| Q6   |
|    10     38    281     .08    .22|1.17   1.2|1.32   1.4|B .49   .57| 87.3  88.2| Q10  |
|     2     63    281    -.85    .17|1.11   1.2|1.18   1.3|C .50   .55| 78.3  79.6| Q2   |
|     5     61    281    -.79    .18|1.01    .2|1.14   1.0|D .55   .55| 77.4  80.2| Q5   |
|     4     37    281     .13    .22| .99    .0|1.14    .6|E .56   .57| 88.7  88.5| Q4   |
|     3     52    281    -.49    .19|1.12   1.0|1.13    .8|F .51   .56| 80.1  83.3| Q3   |
|    20     42    281    -.10    .21|1.09    .7|1.03    .2|G .53   .56| 84.6  86.8| Q20  |
|     1     39    281     .04    .22|1.06    .5|1.05    .3|H .54   .57| 87.8  87.8| Q1   |
|     9     68    281    -.99    .17|1.05    .6| .99    .0|I .53   .55| 76.5  78.0| Q9   |
|    11     33    281     .34    .24|1.01    .1|1.05    .3|J .57   .57| 88.7  89.9| Q11  |
|    19     30    281     .52    .25|1.01    .1|1.03    .2|j .57   .57| 91.4  90.9| Q19  |
|     7     30    281     .52    .25| .87   -.7|1.03    .2|i .60   .57| 92.3  90.9| Q7   |
|    14     39    281     .04    .22|1.00    .0|1.02    .2|h .56   .57| 87.8  87.8| Q14  |
|     8     43    281    -.14    .21| .91   -.6| .99    .0|g .59   .56| 87.8  86.4| Q8   |
|    18     40    281    -.01    .21| .98   -.1| .89   -.5|f .58   .57| 87.3  87.5| Q18  |
|    12     43    281    -.14    .21| .94   -.4| .89   -.5|e .59   .56| 86.9  86.4| Q12  |
|    13     23    281    1.01    .29| .93   -.3| .74   -.6|d .60   .58| 94.1  93.4| Q13  |
|    17     38    281     .08    .22| .91   -.6| .84   -.7|c .60   .57| 89.1  88.2| Q17  |
|    16     32    281     .40    .24| .90   -.5| .82   -.6|b .61   .57| 90.0  90.3| Q16  |
|    15     60    281    -.75    .18| .83  -1.8| .78  -1.6|a .63   .55| 85.1  80.5| Q15  |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN    41.6  281.0     .00    .22| .99    .0|1.02    .1|           | 86.8  86.9|      |
| S.D.    12.7     .0     .56    .03| .09    .7| .16    .7|           |  5.1   4.4|      |
-------------------------------------------------------------------------------------------
```

Figure 4. **MNSQ and ZSTD values**

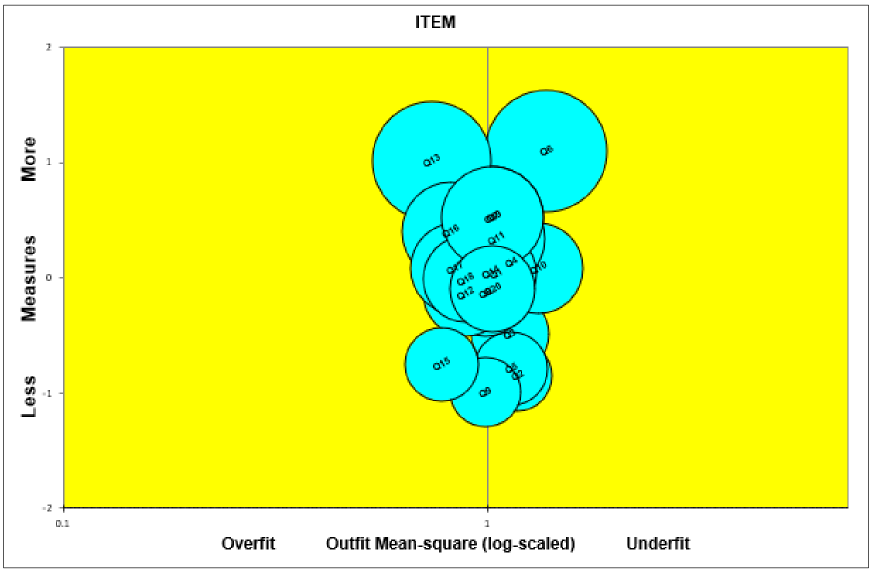


Figure 5. Item fit order pie chart

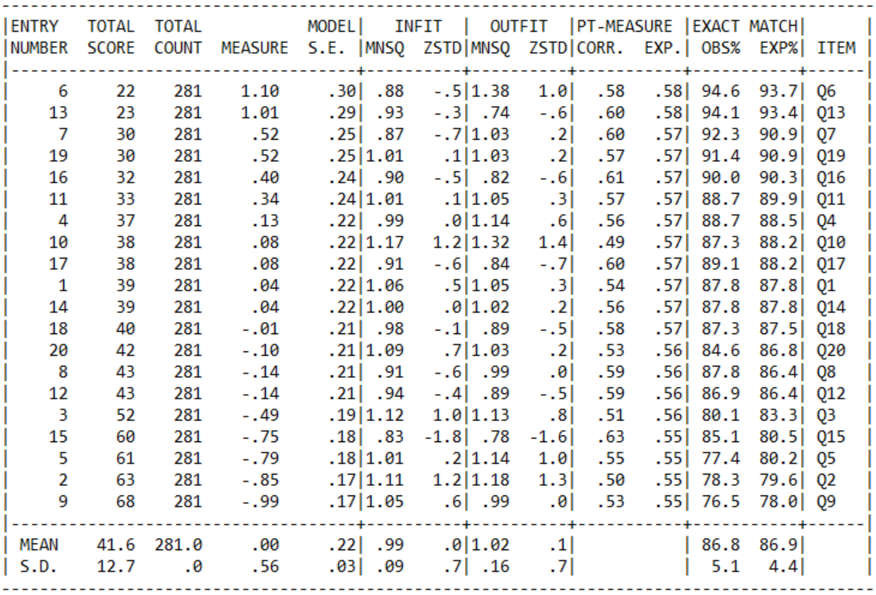
3-6- Rasch Discriminating Power (Point Size Correlations)

Rasch's discriminating power is used to measure whether the items used in the study can differentiate between participants who have high and low abilities. The acceptable Point Size Correlation value is 0.4 < pt the corr size < 0.85.

Based on Figure 4, all items have a point Measurement Correlation value of more than 0.4 and less than 0.85, which means that all items meet the criteria so that it can be concluded that all instrument items have good discriminating power.

3-7- Item Difficulty (Item Size)

Item difficulty is used to measure which items have the highest and lowest difficulty levels. The higher the size value, the more difficult the item will be. Based on Figure 6, it can be seen that the item with the highest difficulty level is Q6 and the item with the lowest difficulty level is Q9.



```
--------------------------------------------------------------------------------------------
|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 |
|------------------------------------+----------+----------+-----------+-----------+------|
|     6     22    281    1.10     .30| .88   -.5|1.38   1.0|  .58   .58| 94.6  93.7| Q6   |
|    13     23    281    1.01     .29| .93   -.3| .74   -.6|  .60   .58| 94.1  93.4| Q13  |
|     7     30    281     .52     .25| .87   -.7|1.03    .2|  .60   .57| 92.3  90.9| Q7   |
|    19     30    281     .52     .25|1.01    .1|1.03    .2|  .57   .57| 91.4  90.9| Q19  |
|    16     32    281     .40     .24| .90   -.5| .82   -.6|  .61   .57| 90.0  90.3| Q16  |
|    11     33    281     .34     .24|1.01    .1|1.05    .3|  .57   .57| 88.7  89.9| Q11  |
|     4     37    281     .13     .22| .99    .0|1.14    .6|  .56   .57| 88.7  88.5| Q4   |
|    10     38    281     .08     .22|1.17   1.2|1.32   1.4|  .49   .57| 87.3  88.2| Q10  |
|    17     38    281     .08     .22| .91   -.6| .84   -.7|  .60   .57| 89.1  88.2| Q17  |
|     1     39    281     .04     .22|1.06    .5|1.05    .3|  .54   .57| 87.8  87.8| Q1   |
|    14     39    281     .04     .22|1.00    .0|1.02    .2|  .56   .57| 87.8  87.8| Q14  |
|    18     40    281    -.01     .21| .98   -.1| .89   -.5|  .58   .57| 87.3  87.5| Q18  |
|    20     42    281    -.10     .21|1.09    .7|1.03    .2|  .53   .56| 84.6  86.8| Q20  |
|     8     43    281    -.14     .21| .91   -.6| .99    .0|  .59   .56| 87.8  86.4| Q8   |
|    12     43    281    -.14     .21| .94   -.4| .89   -.5|  .59   .56| 86.9  86.4| Q12  |
|     3     52    281    -.49     .19|1.12   1.0|1.13    .8|  .51   .56| 80.1  83.3| Q3   |
|    15     60    281    -.75     .18| .83  -1.8| .78  -1.6|  .63   .55| 85.1  80.5| Q15  |
|     5     61    281    -.79     .18|1.01    .2|1.14   1.0|  .55   .55| 77.4  80.2| Q5   |
|     2     63    281    -.85     .17|1.11   1.2|1.18   1.3|  .50   .55| 78.3  79.6| Q2   |
|     9     68    281    -.99     .17|1.05    .6| .99    .0|  .53   .55| 76.5  78.0| Q9   |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN    41.6  281.0     .00     .22| .99    .0|1.02    .1|           | 86.8  86.9|      |
| S.D.    12.7     .0     .56     .03| .09    .7| .16    .7|           |  5.1   4.4|      |
--------------------------------------------------------------------------------------------
```

Figure 6. Measure the value of an item

3-8- Item Information Function

The information function indicates the reliability of the measurement. The Rasch model emphasizes the separation coefficient (the separation of items). The higher the peak of information that can be achieved, the higher the reliability of the measurement. Based on Figure 7, the peak of the curve is at a size value of 0, which means that the item provides the most information about students with ability levels around the size value of 0.

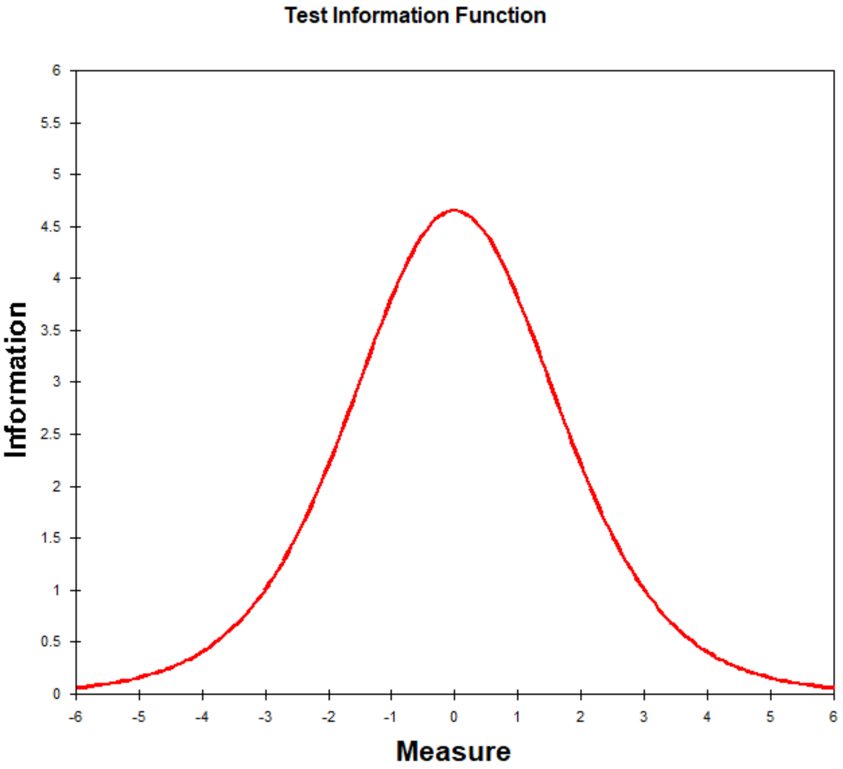


Figure 7. Item Information Function

3-9- Item Bias

DIF (*Differential Item Functioning)* analysis is performed to check if there is an item bias based on gender.

From Figure 8, it is known that there are two curves based on gender, namely L (male) and P (female). From the chart, there are indications that there are some items that are biased, especially for men. As can be seen in the Q4 item, the blue (male) plot is well below average, which means that the item is much easier for men to answer than women. The Q5 item looks very difficult for men than women, the Q11 item looks very easy for men than women. From this it can be concluded that there is an indication of item bias for men.

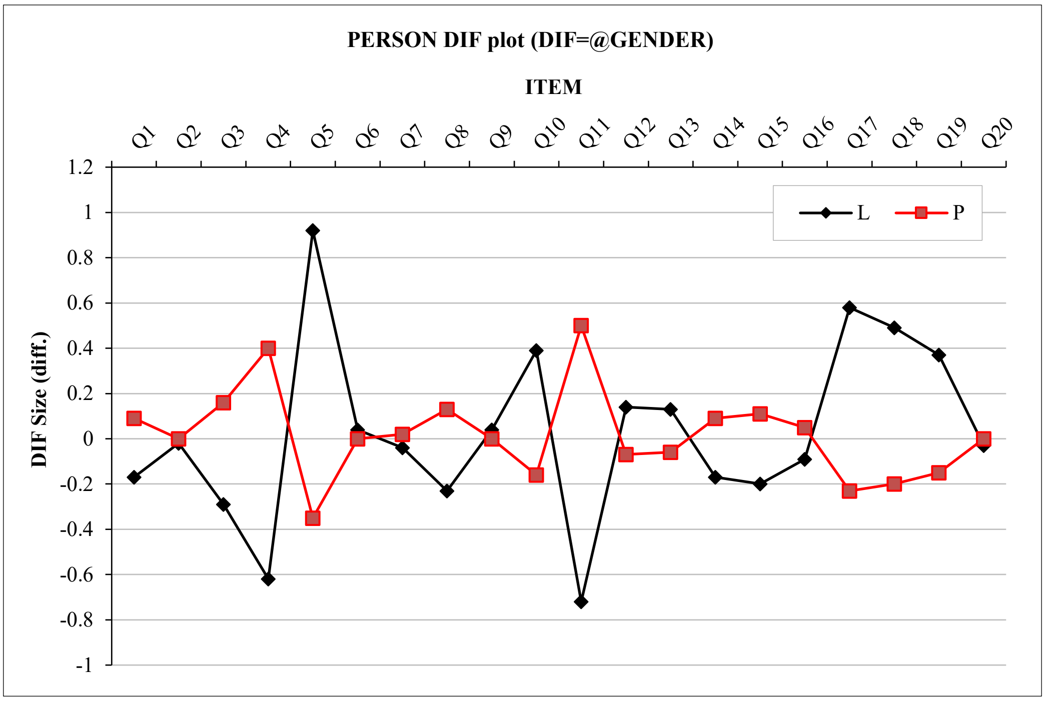


Figure 8. DIF (Differential Item Functioning) *graph* by gender

In the analysis of Differential Item Functioning (DIF) that has been applied in Figure 8, there is an indication of gender bias based on the results obtained from several specific items. Recent research shows that DIF analysis can effectively identify items that provide an advantage or disadvantage to one gender over another, an important issue considered in the development of assessment instruments [32-34].

Based on the analysis, items Q4 and Q11 exhibited highly negative DIF values for men, indicating that these items were considerably easier for men than for women. This aligns with the findings of Fahmi et al., who reported significant gender bias in certain items of the work-life balance scale, illustrating how test items can be influenced by the socio-cultural context of each gender. In contrast, item Q5 showed a positive DIF value, suggesting that it was more difficult for men than for women; however, no relevant references were found to support this observation. Consequently, the corresponding references have been removed [35, 36].

Most other items that have a DIF value close to zero indicate that these questions tend to be fair for both sexes. explains that the practice of eliminating items with large DIFs can reduce differences in gender-specific mean values within a given domain, emphasizing the importance of ensuring that test items are unbiased. This illustrates the importance of DIF analysis in maintaining balance and fairness in testing [37, 38].

These findings suggest that gender bias in questions can arise due to differences in cognitive and learning styles between men and women. Research has found that these factors can affect how participants interpret and answer test items, thus influencing outcomes [39, 40]. In addition, differences in confidence levels also contributed to these results, as revealed by Twiss et al. [40], who linked psychological aspects to performance in different test formats.

Thus, the results of this DIF analysis show that there are indications of bias against men in several items, and this needs to be considered in the preparation of the instrument to be more fair and inclusive. The researchers emphasized the importance of bias analysis in the context of measurement to ensure appropriate and non-discriminatory instrumentation. This raises awareness of the need to actively address and correct biases that may arise in the context of testing [41, 42].

3-10- Student Ability Level

The following student ability levels are measured using an item map to determine the student's ability level at the difficulty level of the question.

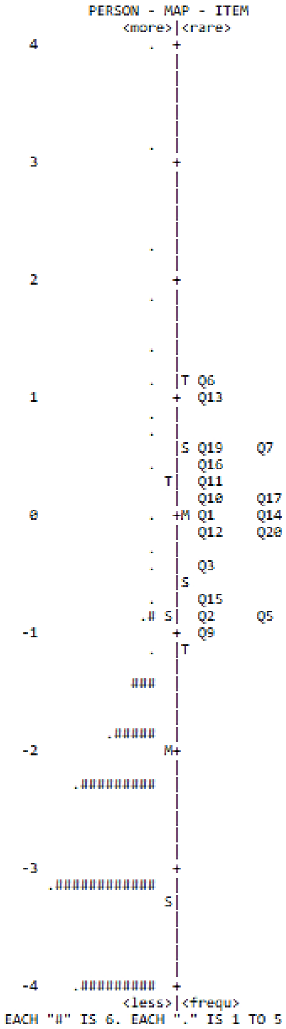
Based on the analysis conducted using the Person-Item Map, it can be seen that the mapping between the student's ability and the difficulty level of the item provides important insights into the suitability of the measuring tool for the target population. Students are shown on the left side of the graph with the symbols "#" and ".", while the right side shows the location of the item based on its difficulty level on the logit scale. With the majority of students being in the range of -3 to -1 logits, this suggests that many students have low to moderate levels of ability. This is in line with findings that show that the distribution of students' abilities is much lower than the difficulty level of the existing items, making it a challenge for students in achieving the expected results [32, 33, 43, 44].

Certain items, such as Q6 and Q13 which are at the logite level +1, are included as the most difficult questions and can only be answered by students with high abilities. In contrast, items like Q5, Q2, and Q15 are located below -1 logit, suggesting that these questions are relatively straightforward. This reality shows the need to review the difficulty level of the question in the context of the student's existing abilities. The fit between the student's location and the item is an important indicator in the DIF analysis, and ideally, the distribution of the item should be balanced with the distribution of the student's ability. However, in this map it seems that many of the questions are around logit 0, while the majority of students are below it. This suggests that the difficulty level of the questions tends to be higher than the general ability of the students [34, 45].

As stated by Lee et al., mapping out students' abilities and the difficulty of the right items is essential to ensure effective and proportionate assessment instruments. Therefore, the conclusions of this analysis confirm that most students have below-average abilities, as well as some questions that are too difficult for the existing student population to answer. This is an important input in the evaluation and improvement of test instruments to better suit students' abilities [43, 46].

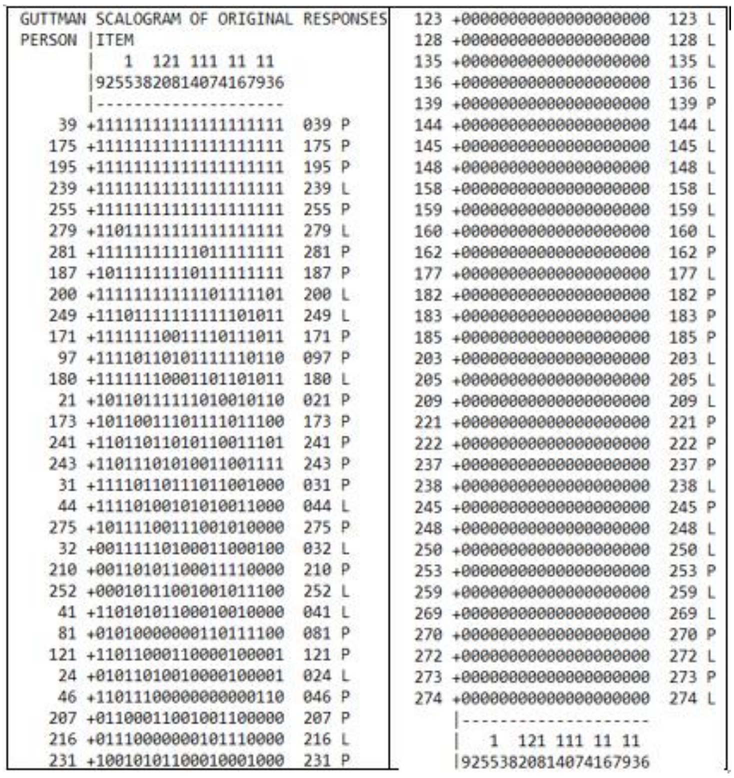
Based on Figure 9, it can be seen that overall the items are distributed higher than the average distribution of students, so it can be said that these students tend to have difficulty working on the questions or items given.

Figure 10 is the output of the Gutman Scale used to measure the number of questions that have been answered correctly and incorrectly by students. The left image is the top row of the Gutman Scalagram analysis, and the right is the bottom row of the analysis results. From the picture, it can be seen that the majority of students answered incorrectly, even for relatively easy questions (Q9).



```
           PERSON - MAP - ITEM
                <more>|<rare>
    4              .  +
                      |
                      |
                      |
                      |
                      |
                   .  |
    3                 +
                      |
                      |
                      |
                      |
                   .  |
                      |
    2                 +
                   .  |
                      |
                      |
                   .  |
                      |
                   .  |T Q6
    1                 +  Q13
                   .  |
                   .  |
                      |S Q19    Q7
                   .  |  Q16
                     T|  Q11
                      |  Q10    Q17    Q4
    0              .  +M Q1     Q14    Q18
                      |  Q12    Q20    Q8
                   .  |
                   .  |  Q3
                      |S
                   .  |  Q15
                  .# S|  Q2     Q5
   -1                 +  Q9
                   .  |T
                      |
                 ###  |
                      |
                      |
              .#####  |
   -2                M+
                      |
          .#########  |
                      |
                      |
                      |
                      |
   -3                 +
       .############  |
                     S|
                      |
                      |
                      |
                      |
   -4     .#########  +
                <less>|<frequ>
EACH "#" IS 6. EACH "." IS 1 TO 5
```

Figure 9. Map of People-Items



```
GUTTMAN SCALOGRAM OF ORIGINAL RESPONSES
PERSON |ITEM
       |   1  121 111 11 11
       |92553820814074167936
       |--------------------
    39 +11111111111111111111  039 P
   175 +11111111111111111111  175 P
   195 +11111111111111111111  195 P
   239 +11111111111111111111  239 L
   255 +11111111111111111111  255 P
   279 +11011111111111111111  279 L
   281 +11111111111011111111  281 P
   187 +10111111110111111111  187 P
   200 +11111111111101111101  200 L
   249 +11101111111111101011  249 L
   171 +11111110011110111011  171 P
    97 +11110110101111110110  097 P
   180 +11111110001101101011  180 L
    21 +10110111111010010110  021 P
   173 +10110011101111011100  173 P
   241 +11011011010110011101  241 P
   243 +11011101010011001111  243 P
    31 +11110110111011001000  031 P
    44 +11110100101010011000  044 L
   275 +10111100111001010000  275 P
    32 +00111110100011000100  032 L
   210 +00110101100011110000  210 P
   252 +00010111001001011100  252 L
    41 +11010101100010010000  041 L
    81 +01010000000110111100  081 P
   121 +11011000110000100001  121 P
    24 +01011010010000100001  024 L
    46 +11011100000000000110  046 P
   207 +01100011001001100000  207 P
   216 +01110000000101110000  216 L
   231 +10010101100010001000  231 P

   123 +00000000000000000000  123 L
   128 +00000000000000000000  128 L
   135 +00000000000000000000  135 L
   136 +00000000000000000000  136 L
   139 +00000000000000000000  139 P
   144 +00000000000000000000  144 L
   145 +00000000000000000000  145 L
   148 +00000000000000000000  148 L
   158 +00000000000000000000  158 L
   159 +00000000000000000000  159 L
   160 +00000000000000000000  160 L
   162 +00000000000000000000  162 P
   177 +00000000000000000000  177 L
   182 +00000000000000000000  182 P
   183 +00000000000000000000  183 P
   185 +00000000000000000000  185 P
   203 +00000000000000000000  203 L
   205 +00000000000000000000  205 L
   209 +00000000000000000000  209 L
   221 +00000000000000000000  221 P
   222 +00000000000000000000  222 P
   237 +00000000000000000000  237 P
   238 +00000000000000000000  238 L
   245 +00000000000000000000  245 P
   248 +00000000000000000000  248 L
   250 +00000000000000000000  250 L
   253 +00000000000000000000  253 P
   259 +00000000000000000000  259 L
   269 +00000000000000000000  269 L
   270 +00000000000000000000  270 P
   272 +00000000000000000000  272 L
   273 +00000000000000000000  273 P
   274 +00000000000000000000  274 L
       |--------------------
       |   1  121 111 11 11
       |92553820814074167936
```

Figure 10. Skalagram Gutman

3-11- Student Confidence Level

In this analysis, students are divided into two groups based on the test scores obtained. The average score is 0.67 on a scale of 200. The division of groups is based on average scores, Grade A is the group with good scores (above average) and grade B is the group with bad scores (below average). Tier 3 data containing definite/non-existent questions with the answers given is sought because of its relationship to the values obtained by the chi square method. From Table 5, it can be seen that 13 of the 20 items have a p-value of <0.05 so it can be concluded that in general there is a significant relationship between the level of student confidence and the score obtained. It can be seen that from all the details, the students who answered "sure" were more in class B, which means that many students were confident with the wrong answers.

Table 5. The Relationship of Level 3 Variables to Student Scores

Thing		Degree		P-Value
		A	B	
Question 1	Y	91	165	0.246
	TY	6	19	
Question 2	Y	92	174	0.921
	TY	5	10	
Question 3	Y	82	149	0.458
	TY	15	35	
Question 4	Y	80	146	0.53
	TY	17	38	
Question 5	Y	89	129	<.001
	TY	8	55	
Question 6	Y	87	146	0.029
	TY	10	38	
Question 7	Y	79	124	0.12
	TY	18	60	
Question 8	Y	86	138	0.007
	TY	11	46	
Question 9	Y	80	134	0.071
	TY	17	50	
Question 10	Y	85	133	0.003
	TY	12	51	
Question 11	Y	84	134	0.008
	TY	13	50	
Question 12	Y	85	138	0.013
	TY	12	46	
Question 13	Y	82	136	0.042
	TY	15	48	
Question 14	Y	84	144	0.089
	TY	13	40	
Question 15	Y	77	130	0.114
	TY	20	54	
Question 16	Y	77	121	0.017
	TY	20	63	
Question 17	Y	89	150	0.022
	TY	8	34	
Question 18	Y	87	147	0.036
	TY	10	37	
Question 19	Y	82	134	0.027
	TY	15	50	
Question 20	Y	88	158	0.242
	TY	9	26	

This is in line with the purpose of the article, which is to evaluate misunderstandings, not only based on correct or incorrect answers, but also through the metacognitive dimension of belief in answers. Without Tier 3 data, misunderstandings like these are difficult to detect because students may answer incorrectly but feel doubtful, or conversely, feel confident when they are wrong. Therefore, the Tier 3 data not only complement, but reinforce the diagnostic value of the three-level tests used in this study. Theoretically, these findings are also supported by the Dunning-Kruger effect, where students with low comprehension often have an exaggerated perception of their abilities. High confidence in incorrect answers reflects a lack of strong metacognitive skills, which in the long run can hinder the learning process . So, it is important for teachers to not only focus on the correct answer, but also on how the student arrived at the answer and how confident they are in it. Thus, the Tier 3 analysis contributes directly to the validity of the diagnostic instruments used in this article, and supports efforts to understand and address students' misconceptions in the concepts of osmoregulation and excretion systems more comprehensively [47-49].

The findings showed that the average ability of students was -2.37 logits, far below the difficulty level of the question (0.0 logit). Most students show misconceptions, especially in items Q6 and Q13 which are classified as difficult. It was also found that high confidence in incorrect answers (the Dunning-Kruger effect) showed a weak metacognition of students. These results are in line with the findings by Treagust [50] and Kirbulut & Geban [11] which suggests that students often develop an unscientific understanding of physiological processes. In addition, the three-level test method supports previous studies by Lim & Poo [2], which states that the confidence dimension enriches diagnostic assessments.

Although the main focus of this study is the development and validation of the instrument, the data collected from several regions and grade levels allows for spatial exploration. Preliminary analysis showed that students from urban areas with adequate biology laboratory facilities (e.g. Padang and Payakumbuh) tended to have higher Tier 1–2 scores than students from remote areas such as Mentawai and South Solok.

In addition, there is a tendency that grade XI students show a slightly better understanding of concepts than grade X students, especially in understanding the mechanism of osmoregulation. This suggests that factors of cognitive maturity and previous learning experience play a role in the formation or correction of misconceptions.

3-12-Discussion

Based on the research conducted, the main objective is to identify and analyze students' misconceptions about osmoregulatory material and excretory systems using the Three-Level Diagnostic Test instrument analyzed using the Rasch Model approach. The results showed that the understanding of the basic concept of the student excretory system was still low, with the average ability of students recorded at -2.37 logits, while the average difficulty of the questions was at the level of 0.00 logits. These findings show a significant gap between students' abilities and item difficulty levels, which is consistent with the results of previous research that emphasized that misconceptions are a common problem in learning biology, especially in complex concepts such as the physiological processes of the human body [41, 43].

Low respondent reliability at a score of 0.55 indicates inconsistency in comprehension between individual students, although item reliability is classified as high at 0.84, and a good internal consistency level is indicated with an Alpha Cronbach of 0.90. This shows that the instruments used have validity in measuring concept understanding, even though students' knowledge backgrounds are diverse. As Treagust revealed, misunderstandings are not only caused by misunderstandings but also by the way students construct meaning from previous learning experiences [40].

The Three-Level Diagnostic Test used in this study was effective in classifying the types of misunderstandings based on students' answers and their confidence levels, supporting the finding that the addition of confidence levels in diagnostic assessments helped distinguish between strong misunderstandings, weak misunderstandings, and honest ignorance. Thus, this method not only provides information about correct or incorrect answers, but also provides a deeper understanding of students' conceptual errors [11, 41].

The misconceptions identified, such as the assumption that the liver is the main organ of excretion, confusion between osmoregulation and excretion, and inaccuracies in understanding kidney function, are very similar to previous findings that suggest that students often mistakenly associate organs as a result of learning that emphasizes memorization rather than understanding concepts [43, 44].

In a pedagogical perspective, these findings point to the need for a more diagnostic and constructivistic approach to learning. One of the practical recommendations from the results of this study is the application of formative assessment based on periodic misunderstandings. Teachers can use the results of this instrument to conduct targeted remedial learning. According to Duit and Treagust, a strategy that works in overcoming misunderstandings is one that is able to shake students' beliefs about wrong concepts and replace them with more precise scientific knowledge through cognitive conflict [45, 51-53].

The Rasch model used also provides additional advantages in the form of diagnostic information regarding the quality of question items and the suitability between questions and student responses. All items in this study showed the clothing values of MNSQ and ZSTD within the tolerance limits of the Rasch model, suggesting that the instrument measured one major construct consistently. In the context of instrument development, this strengthens the validity of the construction and the reliability of diagnostics according to Bond and Fox's research [46].

Furthermore, this study emphasizes the need to increase teacher capacity in diagnostic assessment and literacy. Adequate training is required so that teachers can use diagnostic instruments based on the Rasch model and understand the follow-up to misunderstanding findings. As stated by Sadler, the effectiveness of formative assessment is highly dependent on the ability of teachers to utilize assessment outcomes to improve teaching and support students' cognitive development [54].

In the local context, these findings also make an important contribution to mapping the quality of biology learning in West Sumatra and Jambi Provinces. Given the diverse socio-cultural backgrounds and educational facilities, it is important to ensure that the diagnostic approach used is not only theoretically correct, but also adaptive to the realities of the field. The study as a whole confirms that students' conceptual understanding of osmoregulation and the excretory system is still weak, and that the Three-Level Diagnostic Test approach supported by Rasch's analysis is a promising method for detecting and addressing misconceptions effectively.

Indications of gender bias in items Q4, Q5, and Q11 can be traced into the dimensions of students' cognitive style and conceptual representation. For example, the Q4 item, which demands the ability to understand the regulation of homeostasis under extreme conditions, is likely to be more in line with the systematic approach more often associated with male students. In contrast, Q5, which contains an applicative context in household or microbiological situations, may be closer to the daily experiences of female students. Whereas Q11, which emphasizes the relationship between structure and kidney function, may be better understood by male students because they are familiar with structural visual models.

This hypothesis is in line with previous research by Arnup et al. [34] which suggests that differences in contextual experience and cognitive approaches can lead to different responses to item content, even with similar levels of ability. Therefore, the study of context, language style, and depiction in items becomes important in the cross-gender validation process. Cronbach's Alpha value of 0.90 indicates that overall the instrument has excellent internal consistency in measuring one conceptual construct. However, the respondent reliability (person reliability) which only reached 0.55 indicates that the variability of responses between students is quite high. This means that even though the question items work well in general, the students' responses to the item do not show a stable pattern.

This difference can be interpreted as a reflection of the heterogeneity of student understanding. Some students have strong and consistent misconceptions, while others show ignorance or guessing, which results in inconsistencies in answers. This reinforces the diagnostic value of the three-level test because it is able to reveal the diversity of levels of conceptual understanding and irregularities in students' metacognition.

The misconceptions found in this study reinforce the systematic mapping presented by Guerra-Reyes et al. [55], which categorizes misconceptions based on disciplines and themes such as thermodynamics, sound, and mechanics. The main causes were identified in memorization-based learning patterns, lack of teacher training, and the use of the transmission-reception model in teaching.

Research by Batlolona & Jamaludin [56] Suggests that student misconceptions can also be rooted in a lack of integration between local natural phenomena and science learning. In the Indonesian context, misconceptions about the basic concepts of excretion and osmoregulation can be exacerbated by non-contextual teaching and minimal exploration of the surrounding environment.

4- Conclusion

Based on the results of the research and discussions that have been conducted, the conclusions that can be drawn are as follows: The student's ability level is lower than the difficulty level of the item with the average value of the student size being -2.37 and the average value of the item size is 0.0. All instrument items function normally according to the criteria in making measurements with the average value of square clothing is 1.02 and the average value of Z-Standard clothing (ZSTD) is 0.1. In addition, the Correlation value of the item points is more than 0.4 and less than 0.85. The item with the highest difficulty level is Q6 with a size value of 1.10 and the item with the lowest difficulty is Q9 with a size value of -0.99. There is item bias for male students on items Q4, Q5, and Q11. The majority of students answered the wrong answer even for the easiest item, Q9. The majority of students who answered "sure" were in grade B, which means they were sure of the wrong answer.

5- Declarations

5-1- Author Contributions

Conceptualization, R. and D.K.; methodology, R. and S.S; software, N., G., and L.Z.; validation, F.M., D.K., and S.A.S.; formal analysis, S.A.S. and L.Z.; investigation, R. and D.K.; resources, N., D.K., and S.A.S.; data curation, R., D.K., and L.Z.; writing—original draft preparation, F.M.; writing—review and editing, N., G., and S.S.; visualization, D.K., S.S., and S.A.S.; supervision, G.; project administration, R., D.K., and L.Z.; funding acquisition, S.A.S., L.Z., N., G., and F.M. All authors have read and approved the published version of the manuscript.

5-2- Data Availability Statement

The data presented in this study are available on request from the corresponding author.

5-3- Funding and Acknowledgements

We would like to thank all participants who have actively and voluntarily contributed to the completion of this research, both as assessors and students. We are also grateful to the Indonesian Education Scholarship (BPI), the Indonesian Endowment Fund for Education (LPDP), the Center for Higher Education Funding and Assessment (PPAPT) of the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia for sponsoring the research publication.

5-4- Institutional Review Board Statement

This research was conducted in accordance with the Helsinki Declaration and approved by the Institutional Review Board (or Ethics Committee) of Universitas Pendidikan Indonesia (26/UN40.K/PT.01.01/2025).

5-5- Informed Consent Statement

Informed consent was obtained from all subjects involved in the study.

5-6- Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

6- References

[1] Mapulanga, T., & Bwalya, A. (2025). Gender Differences in Secondary School Students' Perceptions of Teaching Practices Used in Biology Classrooms. African Journal of Research in Mathematics, Science and Technology Education, 29(1), 42–53. doi:10.1080/18117295.2024.2444793.

[2] Lim, H. L., & Poo, Y. P. (2021). Diagnostic test to assess misconceptions on photosynthesis and plant respiration: Is it valid and reliable? Jurnal Pendidikan IPA Indonesia, 10(2), 241–252. doi:10.15294/jpii.v10i2.26944.

[3] Enochson, P. G., Redfors, A., Dempster, E. R., & Tibell, L. A. E. (2015). Ideas about the human body among secondary students in South Africa. African Journal of Research in Mathematics, Science and Technology Education, 19(2), 199–211. doi:10.1080/10288457.2015.1050804.

[4] Aydın, S. (2016). To what extent do Turkish high school students know about their body organs and organ systems? International Journal of Human Sciences, 13(1), 1094. doi:10.14687/ijhs.v13i1.3498.

[5] Ameyaw, Y., & Okyer, M. (2018). Concept mapping instruction as an activator of students' performance in the teaching and learning of excretion. Annals of Reviews and Research, 1(4), 93-102. doi:10.19080/arr.2018.01.555568.

[6] Assem, H. D., Nartey, L., Appiah, E., & Aidoo, J. K. (2023). A Review of Students' Academic Performance in Physics: Attitude, Instructional Methods, Misconceptions and Teachers Qualification. European Journal of Education and Pedagogy, 4(1), 84–92. doi:10.24018/ejedu.2023.4.1.551.

[7] Kibirige, I., & Mamashela, D. (2022). Learners' prevalent misconceptions about force and experiences of flipped classes. Journal for the Education of Gifted Young Scientists, 10(1), 109–120. doi:10.17478/jegys.1058677.

[8] Reichert, C., Cervato, C., Niederhauser, D., & Larsen, M. D. (2015). Understanding atmospheric carbon budgets: Teaching students conservation of mass. Journal of Geoscience Education, 63(3), 222–232. doi:10.5408/14-055.1.

[9] Uzunhasanoğlu, Ö., Çakır, M., & Avcı, S. (2020). Developing and Implementing Two-Stage Diagnostic Test to Measure the Concept Concepts of Biology Teacher Candidates. Turkish Studies-Educational Sciences, 15(4), 2407–2423. doi:10.47423/turkishstudies.44131. (In Turkish).

[10] usilaningsih, E., Aprilia, N., & Kristiansi, F. (2023). Misconception analysis of sub-microscopic level in chemistry learning of reaction rate using three-tier multiple-choice test (TTMCT) for class XI students. International Conference on Applied Computational Intelligence and Analytics (ACIA-2022), 2705, 030008. doi:10.1063/5.0125980.

[11] Kirbulut, Z. D., & Geban, O. (2014). Using three-tier diagnostic test to assess students' misconceptions of states of matter. Eurasia Journal of Mathematics, Science and Technology Education, 10(5), 509–521. doi:10.12973/eurasia.2014.1128a.

[12] Karataş, A. (2020). Preservice Science Teachers' Misconceptions About Evolution. Journal of Education and Training Studies, 8(2), 38. doi:10.11114/jets.v8i2.4690.

[13] Machová, M., & Ehler, E. (2023). Secondary school students' misconceptions in genetics: origins and solutions. Journal of Biological Education, 57(3), 633–646. doi:10.1080/00219266.2021.1933136.

[14] Çakmak, T., & Bulunuz, N. (2022). Teaching Seventh Graders About the Digestive System Using Formative Assessment to Evaluate Comprehension Levels. Academy Journal of Educational Sciences, 6(1), 59–67. doi:10.31805/acjes.1116921.

[15] Çuçin, A., Özgür, S., & Güngör Cabbar, B. (2020). Comparison of Misconceptions about Human Digestive System of Turkish, Albanian and Bosnian 12th Grade High School Students. World Journal of Education, 10(3), 148. doi:10.5430/wje.v10n3p148.

[16] Kummer, T. A., Whipple, C. J., & Jensen, J. L. (2016). Prevalence and Persistence of Misconceptions in Tree Thinking. Journal of Microbiology & Biology Education, 17(3), 389–398. doi:10.1128/jmbe.v17i3.1156.

[17] Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch Analysis in the Development and Application of Quality of Life Instruments. Value in Health, 7, S22–S26. doi:10.1111/j.1524-4733.2004.7s106.x.

[18] Andrich, D. (2011). Rasch Models for Measurement. Rasch Models for Measurement. SAGE Publications, Thousand Oaks, United States. doi:10.4135/9781412985598.

[19] Lu, Y. M., Wu, Y. Y., Hsieh, C. L., Lin, C. L., Hwang, S. L., Cheng, K. I., & Lue, Y. J. (2013). Measurement precision of the disability for back pain scale-by applying Rasch analysis. Health and Quality of Life Outcomes, 11(1), 119. doi:10.1186/1477-7525-11-119.

[20] Aryadoust, V., Tan, H. A. H., & Ng, L. Y. (2019). A scientometric review of rasch measurement: The rise and progress of a specialty. Frontiers in Psychology, 10(Oct), 02197. doi:10.3389/fpsyg.2019.02197.

[21] Ma, H., Liu, W., & Li, G. (2025). Development and Application of a Five-Tier Diagnostic Test to Assess Misconceptions on Respiration and Photosynthesis among Senior High School Students in Mainland China. Research in Science Education, 4. doi:10.1007/s11165-025-10232-6.

[22] Zhong, J., Ma, H. Y., Wang, X. M., Huang, X. J., & Xu, M. Z. (2023). Rasch analysis of the Chinese version of the clinically useful depression outcome scale in patients with major depressive disorder. BMC Psychology, 11(1), 1–9,. doi:10.1186/s40359-023-01255-7.

[23] Salzberger, T., & Sinkovics, R. R. (2006). Reconsidering the problem of data equivalence in international marketing research: Contrasting approaches based on CFA and the Rasch model for measurement. International Marketing Review, 23(4), 390–417. doi:10.1108/02651330610678976.

[24] Farlie, M., Johnson, C., Wilkinson, T., & Keating, J. (2021). Refining assessment: Rasch analysis in health professional education and research. Focus on Health Professional Education: A Multi-Professional Journal, 22(2), 88–104. doi:10.11157/fohpe.v22i2.569.

[25] Raja, P., Setiadi, B., & Abdurrahman, A. (2019). Developing and Validating an Instrument of In-service Teachers Responses to Knowledge-Based Teacher, Engagement, and Expectation in Teacher Profession Education Program in Indonesia: Integrating factor analysis with Rasch modeling. Proceedings of the International Conference on Educational Sciences and Teacher Profession (ICETeP 2018), 74. doi:10.2991/icetep-18.2019.74.

[26] Long, C., Bansilal, S., & Debba, R. (2014). An investigation of mathematical literacy assessment supported by an application of rasch measurement. Pythagoras, 35(1), 1–17. doi:10.4102/pythagoras.v35i1.235.

[27] Rowe, V. T., Winstein, C. J., Wolf, S. L., & Woodbury, M. L. (2017). Functional Test of the Hemiparetic Upper Extremity: A Rasch Analysis With Theoretical Implications. Archives of Physical Medicine and Rehabilitation, 98(10), 1977–1983. doi:10.1016/j.apmr.2017.03.021.

[28] Taslidere, E. (2016). Development and use of a three-tier diagnostic test to assess high school students' misconceptions about the photoelectric effect. Research in Science & Technological Education, 34(2), 164–186. doi:10.1080/02635143.2015.1124409.

[29] Laliyo, L. A. R., La Kilo, A., Paputungan, M., Kunusa, W. R., Dama, L., & Panigoro, C. (2022). Rasch Modelling To Evaluate Reasoning Difficulties, Changes of Responses, and Item Misconception Pattern of Hydrolysis. Journal of Baltic Science Education, 21(5), 817–835. doi:10.33225/jbse/22.21.817.

[30] Amiruddin, M. Z. B., Samsudin, A., Suhandi, A., & Costu, B. (2024). Bibliometric investigation in misconceptions and conceptual change over three decades of science education. International Journal of Educational Methodology, 10(3), 367-385. doi:10.12973/ijem.10.3.367.

[31] Boone, W. J. (2016). Rasch analysis for instrument development: Why,when,and how? CBE Life Sciences Education, 15(4), 1-7. doi:10.1187/cbe.16-04-0148.

[32] Ginther, D. K., Kahn, S., & Schaffer, W. T. (2016). Gender, race/ethnicity, and national institutes of health R01 research awards: Is there evidence of a double bind for women of color'. Academic Medicine, 91(8), 1098–1107. doi:10.1097/ACM.0000000000001278.

[33] Brown, A. J., & Goh, J. X. (2016). Some evidence for a gender gap in personality and social psychology. Social Psychological and Personality Science, 7(5), 437–443. doi:10.1177/1948550616644297.

[34] Arnup, J. L., Murrihy, C., Roodenburg, J., & McLean, L. A. (2013). Cognitive style and gender differences in children's mathematics achievement. Educational Studies, 39(3), 355–368. doi:10.1080/03055698.2013.767184.

[35] Fahmi, E. F. F. El, Astutik, F., & Ibrahim, F. F. (2023). Analysis of Differential Item Functioning (DIF) on the Work-Life Balance Scale. Atlantis Press SARL, 728. doi:10.2991/978-2-38476-032-9_31.

[36] Moradi, E., Ghabanchi, Z., & Pishghadam, R. (2022). Reading comprehension test fairness across gender and mode of learning: insights from IRT-based differential item functioning analysis. Language Testing in Asia, 12(1), 39. doi:10.1186/s40468-022-00192-3.

[37] Pässler, K., Beinicke, A., & Hell, B. (2014). Gender-Related Differential Validity and Differential Prediction in Interest Inventories. Journal of Career Assessment, 22(1), 138–152. doi:10.1177/1069072713492934.

[38] Maričić, M., Đoković, A., & Jeremić, V. (2019). The validity of student evaluation of teaching: Is there a gender bias? Croatian Journal of Education, 21(3), 743–775. doi:10.15516/cje.v21i3.3177.

[39] Kuzu, Y., & Gelbal, S. (2023). Investigation of Differential Item and Step Functioning Procedures in Polytomus Items*. Journal of Measurement and Evaluation in Education and Psychology, 14(3), 200–221. doi:10.21031/epod.1221823.

[40] Twiss, J., McKenna, S. P., Graham, J., Swetz, K., Sloan, J., & Gomberg-Maitland, M. (2016). Applying Rasch analysis to evaluate measurement equivalence of different administration formats of the Activity Limitation scale of the Cambridge Pulmonary Hypertension Outcome Review (CAMPHOR). Health and Quality of Life Outcomes, 14(1), 1–9,. doi:10.1186/s12955-016-0462-2.

[41] Hagquist, C., & Andrich, D. (2017). Recent advances in analysis of differential item functioning in health research using the Rasch model. Health and Quality of Life Outcomes, 15(1), 1–8,. doi:10.1186/s12955-017-0755-0.

[42] Astegiano, J., Sebastián-González, E., & Castanho, C. D. T. (2019). Unravelling the gender productivity gap in science: A meta-analytical review. Royal Society Open Science, 6(6), 181566. doi:10.1098/rsos.181566.

[43] Malonisio, M. O., & Malonisio, C. C. (2023). Validation of the teacher education institution's entrance test using the Rasch model. International Journal of Innovative Research and Scientific Studies, 6(3), 644–655. doi:10.53894/ijirss.v6i3.1726.

[44] Barcelo, J. (2024). Development and Rasch Analysis of the Prior Knowledge of Chemistry Concepts Test for Pre-medical Students in the Philippines. KIMIKA, 34(2), 14–33. doi:10.26534/kimika.v34i2.14-33.

[45] Lee, S. C., Lee, Y. C., & Chiu, E. C. (2023). Psychometric validation of the Cognitive Abilities Screening Instrument using Rasch analysis in people with dementia. Medicine (United States), 102(32), E34093. doi:10.1097/MD.0000000000034093.

[46] Jimam, N. S., Ismail, N. E., Dangiwa, D. A., Dapar, M. L. P., Sariem, C. N., Paul, L. A., Mohammed, S. G., & Dayom, D. W. (2021). Use of Rasch Wright map to understand the quality of Healthcare Workers' Knowledge, Attitudes, and Practices for Uncomplicated Malaria (HKAPIUM). Journal of Pharmacy & Bioresources, 18(3), 237–244. doi:10.4314/jpb.v18i3.8.

[47] Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77(6), 1121–1134. doi:10.1037/0022-3514.77.6.1121.

[48] Kuczmann, I. (2017). The structure of knowledge and students' misconceptions in physics. AIP Conference Proceedings, 1916, 050001. doi:10.1063/1.5017454.

[49] Swanson, H., Anton, G., Bain, C., Horn, M., Wilensky, U. (2019). Introducing and Assessing Computational Thinking in the Secondary Science Classroom. Computational Thinking Education. Springer, Singapore. doi:10.1007/978-981-13-6528-7_7.

[50] Treagust, D. F. (1993). The evolution of an approach for using analogies in teaching and learning science. Research in Science Education, 23(1), 293–301. doi:10.1007/BF02357073.

[51] Luz, M. R. M. P., Oliveira, G. A., & Poian, A. T. D. (2013). Glucose as the sole metabolic fuel: Overcoming a misconception using conceptual change to teach the energy-yielding metabolism to Brazilian high school students. Biochemistry and Molecular Biology Education, 41(4), 224–231. doi:10.1002/bmb.20702.

[52] Fulmer, G. W., Liang, L. L., & Liu, X. (2014). Applying a force and motion learning progression over an extended time span using the force concept inventory. International Journal of Science Education, 36(17), 2918–2936. doi:10.1080/09500693.2014.939120.

[53] Rost, J. (1990). Rasch Models in Latent Classes: An Integration of Two Approaches to Item Analysis. Applied Psychological Measurement, 14(3), 271–282. doi:10.1177/014662169001400305.

[54] Aydeniz, M., & Kotowski, E. L. (2012). What Do Middle and High School Students Know About the Particulate Nature of Matter After Instruction? Implications for Practice. School Science and Mathematics, 112(2), 59–65. doi:10.1111/j.1949-8594.2011.00120.x.

[55] Guerra-Reyes, F., Guerra-Dávila, E., Naranjo-Toro, M., Basantes-Andrade, A., & Guevara-Betancourt, S. (2024). Misconceptions in the Learning of Natural Sciences: A Systematic Review. Education Sciences, 14(5), 497. doi:10.3390/educsci14050497.

[56] Batlolona, J. R., & Jamaludin, J. (2024). Students' misconceptions on the concept of sound: a case study about Marinyo, Tanimbar Islands. Journal of Education and Learning, 18(3), 681–689. doi:10.11591/edulearn.v18i3.21135.