Examinee Characteristics and their Impact on the Psychometric Properties of a Multiple Choice Test According to the Item Response Theory (IRT)

Deyab Almaleki
Department of Evaluation, Measurement, and Research
Umm Al-Qura University
Makkah, Saudi Arabia
damaleki@uqu.edu.sa

Abstract-The aim of the current study is to provide improvement evaluation practices in the educational process. A multiple choice test was developed, which was based on content analysis and the test specification table covered some of the vocabulary of the applied statistics course. The test in its final form consisted of 18 items that were reviewed by specialists in the field of statistics to determine their validity. The results determine the relationship between individual responses and the student ability. Most thresholds span the negative section of the ability. Item information curves show that the items provide a good amount of information about a student with lower or moderate ability compared to a student with high ability. In terms of precision, most items were more convenient with lower ability students. The test characteristic curve was plotted according to the change in the characteristics of the examinees. The information obtained by female students appeared to be more than the information obtained by male students and the test provided more information about students who were not studying statistics in an earlier stage compared with students who did. This test clearly indicated that, based on the level of the statistics course, there should be a periodic review of the tests in line with the nature and level of the course materials in order to have a logical judgment about the level of the students' progress at the level of their ability.

Keywords-item response theory; item characteristics; multiple-choice; psychometric properties

I. INTRODUCTION

A test is an educational tool that is frequently used to evaluate students' academic achievement and progress. Tests also provide an opportunity to verify students' skills in many educational situations when it is not possible to use other assessment methods. Despite the known problems of indirect measurement, a lot of traits such as mathematical abilities, verbal skills, resistance to stress, intelligence, dissatisfaction, different opinions about a particular topic, etc. cannot be directly observed and measured [1-3]. These are known as the latent traits, and they can be measured only indirectly, often using specially prepared questionnaires where the responses are closely related to the specific traits being studied. Tests are frequently used to assess students' cognitive progress and to build question banks. As a result, the so-called latent trait models have been developed and are used to estimate the parameter values associated with the human personality [4-5]. These models provide a different type of information that in turn helps to develop and improve tests accordingly. Many researchers rely on the information from data that are analyzed as a result of the subjects' responses. But it is important to ask whether the formulation of stimuli (questions) may provide another type of data or information that serves the research process [1, 6-8].

Many researchers and specialists in the field of measurement and evaluation, are interested in the basic concepts and organizational theoretical frameworks of measurement and evaluation and ways to apply them [9-11], due to the great role this science plays in various fields of scientific research in general and educational and psychological research in particular. Research activity according to the needs of educational institutions will positively affect development and improvement in accordance with Saudi Arabia's Vision 2030. Furthermore, the means of developing tests and measurement methods are extremely important because the data issued from the measurement processes have to be valid and accurate, as some crucial decisions such as admission or promotion may be based upon them [12-16]. In addition, it is the responsibility of specialists in the field of measurement and evaluation to enrich the literature, develop tests used in the educational field, and reduce the possibility of potential measurement errors during the evaluation process [17, 18]. Postgraduate tests in Saudi universities have not yet been subjected to much scrutiny by local and international evaluation institutions, because most of the quality assurance agencies, such as the National Commission for Evaluation and Accreditation (NCAAA), has only recently included postgraduate programs in its plans, providing the opportunity for universities to apply for accreditation for these programs. Midterm and final exams and the way they are administered are some of the indicators used by the NCAAA or other agencies to accurately judge the progress of an academic program. Therefore, improving these tests has become a necessary requirement.

Corresponding author: Deyab Almaleki

II. SIGNIFICANCE OF THE CURRENT STUDY

The significance of the current study stems from the importance of the evaluation processes in the educational process. The process of improving tests and identifying their psychometric characteristics is the task of those working in the field of measurement and evaluation in order to provide a comprehensive understanding and a deep descriptive analysis of the advantages and disadvantages observed in those tests [19-21]. Furthermore, the scarcity of this type of scientific studies has widened the gap between the tools currently used to measure the level of achievement of students and what is hoped that these tools should be like. The quality of these tools has not been determined or reviewed, and therefore they have not been assessed or evaluated [22-25]. The practical significance of this type of research lies in the use of Item Response Theory (IRT) models in analyzing students' responses to the achievement test in a more objective way, showing whether there is an effect of the multiplicity of characteristics of the participants on the test items (multiple choice) in terms of the accuracy of the estimates of the items' parameters and the individuals' ability parameters. It can also guide the composers of the test questions to take into account some points that may affect the psychometric properties of the items and the test and the accuracy of its results [26-29].

III. ITEM RESPONSE THEORY

There are many ways (i.e. models) to determine the relationship between individual responses and student ability. Within the framework of modern measurement theory, many models and applications have been formulated and applied to real test data. This includes the measurement assumptions about the characteristics of the test item, the performance of the subject, and how this performance is related to knowledge [27-37-40]. Tests and evaluation processes in general form the basis of the education system, and their importance lies in improving educational planning, developing a mechanism for enhancing curricular content, measuring learners' competence, and comparing student performance or achievement data. Evaluations also have a role for schools and teachers [30-33]. Tests are a tool of assessment, and their quality depends on a large extent on the nature and quality of the information collected during the preparation of the assessment. Over the decades, the test building system has undergone a lot of development through the emergence of many test building theories focused in many types of tests, such as oral tests, standardized tests, and realistic evaluation. Until today, theories have continued to develop in order to keep up with the changes in policies and new educational practices [2]. The modern theoretical methods were largely developed in the sixties to the late eighties. The IRT is a general statistical theory considering the characteristics of the test item, the subject's performance on the item, and how the performance is related to the abilities that are measured by the test items [12, 34-36]. The IRT provides a rich statistical tool for analyzing educational tests and psychometric measures. The IRT assumes the following:

- The test performance of the subjects can be predicted (or explained) by a set of factors called traits or latent traits and abilities.
- The relationship between the subject's performance and the properties of the test item can be described through a monotonic increasing function called the item information function.
- The response on the test item can be either isolated or continuous and it can be binary or bimodular. Item score categories can be ordered or unordered, and there can be one or many abilities behind the test performance.

IV. CHARACTERISTICS OF THE IRT MODELS

- The IRT model should be defined as the relationship between the observed response and the unobserved infrastructure (latent trait).
- The model should provide a method for estimating the degrees of the latent trait.
- The subjects' scores will be the basis for the assessment of the basic construction of the model.
- The IRT model assumes that the subject's performance can be predicted or explained by one latent trait or more.

In IRT is often assumed that the examinee has some unobservable latent trait (also called latent ability), which cannot be studied directly. The purpose of the IRT is to propose models that allow linking these underlying traits to some of the characteristics that can be observed on the subject [41]. There are many models in the IRT and they have been classified into two types: models that use the cumulative natural curve and logistic models. Logistic models are currently more widespread, they are suitable for two-stage items, and differ according to the number of the estimated item parameters [6, 31, 42-45]. However, there are three commonly used models for binary data, which use (1) for the correct response and (0) for the wrong response, and these models are the one parameter and the two-parameter logistic models, which will be examined below.

A. One-Parameter Logistic Model

The concept of information availability plays an important role in IRT as it can be used to evaluate how the item included in the test accurately measures the level of the latent trait with (parameter value θi). This latent trait could include, for example, the level of the student's knowledge, intelligence, ability, satisfaction, stress, etc. For example, in educational tests, the item parameter represents the difficulty of the item while the subject parameter represents the ability level of the people being evaluated. The greater the subject's ability in relation to the difficulty of the item (the parameter αj describes the degree of difficulty of the item and the level of influence of the item on the subject), the greater the probability of a correct response to that item. Whereas, when the subject's position on the latent trait is equal to the difficulty of the item, according to Rasch's model, there is a 0.5 probability that a subject's response is correct. Accurate information about the value of θi depends on a number of factors, the most important of which is the properties of the questions (items) used to evaluate the parameter (the latent trait) [2, 30, 33, 46].

B. Two-Parameter Logistic Model

In the Two-Parameter Logistic (2PL) model, the situation is different from the one in the one-parameter model. The one-parameter model assumes that questions differ only with respect to item difficulty, whereas, in the two-parameter logistic model, two parameters are assumed to be connected to the test item: the parameter αj which describes the difficulty of the item (question) and the additional parameter βj, which describes the discrimination of the item. The parameter β (the slope of the curve) describes the degree to which the question helps to distinguish between the subjects with the highest level of a trait compared to those with a lower level of the same trait. This parameter also shows the extent of the relevance of the item to the overall score of the test. The higher the value of that parameter, the greater the discrimination of the items (and the easier it is to select subjects with a high level and those with a low concentration of the same trait). It should also be noted that the most difficult test item is not necessarily the test item with the highest potential to discriminate between the subjects [2, 19, 36, 47, 48].

C. Three-Parameter Logistic Model

The Three-Parameter Logistic (3PL) model is used in IRT, and it determines the probability of a correct response for a dichotomously scored multiple-choice item as a logistic distribution. The 3PL model is an extension of the 2PL logistic model as it introduces the guessing parameter. Items now differ in terms of discrimination, difficulty, and probability of guessing the correct response [47]. After adding the guessing parameter, denoted Ci, in the 3PL model, this parameter is the lower asymptote of the item characteristic curve and represents the probability of subjects who have a low ability to answer the item correctly. The parameter is included in the model to account for item response data from low-ability subjects, where guessing is a factor in test performance [48-50]. The basic equation for the 3PL model is the probability that a randomly selected examinee with a certain proficiency level on scale k will respond correctly to item j, which is characterized by discrimination (αj), difficulty (βj), and guessing probability (Ci) [27, 35, 37, 38, 51].

V. MULTIPLE CHOICE TEST ANALYSIS

Understanding how to interpret and use the information based on students' test scores is just as important as knowing how to create a well-designed test. An essential part of building tests is using the feedback from a good test analysis. Among the most important statistical information provided by a good analysis of a multiple-choice test are the following:

A. Item Difficulty

The test item difficulty factor βj represents the percentage of the respondents who answered the item correctly. The difficulty factor ranges from 0.0 to 1.00. The higher the value of the difficulty factor, the easier the test item is. For example, when the value of the difficulty factor βj is higher than 0.90, the test item is described as very easy and should not be used again in subsequent tests since almost all students are able to properly respond to it. Whereas when the value of the βj is less than 0.20, the test item is described as extremely difficult and should be reviewed in subsequent tests. The optimal test item difficulty factor is 0.50, and it insures maximum discrimination between high and low ability [52-54]. To maximize item discrimination, the desired difficulty levels are slightly higher than halfway between the probability of answering correctly by chance (1.00 divided by the number of alternatives for the item) and the ideal score for the item (1.00) [55-58]. For example, if the test item contains four alternatives to the answer, the probability of answering it correctly by chance would be 0.25 (1.00/4=0.25), and the ideal degree of difficulty for the item can be calculated by substituting in the following rule:

((Ideal score for item - probability of a correct answer by chance) / 2) + probability of a correct answer by chance

TABLE I. THE IDEAL DEGREE OF DIFFICULTY GUIDELINE

Design of the test item	Ideal degree of difficulty for the test item
Multiple choice (5 alternatives)	0.60
Multiple choice (4 alternatives)	0.62
Multiple choice (3 alternatives)	0.66
Multiple choice (2 alternatives)	0.75

B. Item Discrimination

The test item discrimination factor is referred to using the symbol αj, as it represents the point relationship between the respondent's performance on the item and the respondents' total scores. The discrimination factor value ranges from -1.00 to 1.00. When the value of the test item discrimination factor is high, it indicates that the test item is able to distinguish between respondents. It distinguishes between those who scored high in the tests and were able to answer the test item correctly and those who obtained low test scores and were not able to respond to the item correctly [54, 59]. Test items that have point values close to or less than zero should be removed. Moreover, further consideration should be given to the item which was responded to better by those who generally performed poorly on the test than those who performed better on the test as a whole. The test item may be confusing in some way to top-performing respondents [52, 53, 58, 59].

TABLE II. THE IDEAL DEGREE OF DISCRIMINATION GUIDELINE

Discrimination factor value	Description of the test item
≥ 0.4	A very good test item
0.3 - 0.39	A good test item. Possible improvements may be considered
0.2 - 0.29	A fairly good test item. It is recommended to improve it
≤ 0.2	A weak test item, with the recommendation of deleting it
≤ 0	It is recommended to directly delete the item

VI. PSYCHOMETRIC PROPERTIES OF THE TEST

Psychometric properties are a statistical mechanism to verify the fairness, objectivity, and relevance of the test for the phenomenon to be measured. The most important psychometric properties are the following:

- The individual data series for all test components.
- The characteristics of the data collected for all test components. The fairness of the test lies in its freedom from

any bias and its suitability for the target group, regardless of gender, race, and religion. The psychometric properties of the test are tested to verify that it is objectively constructed and free of any bias.

Studying the rules for formulating multiple-choice test items is important because it has an impact on the level of performance on the items or the test as a whole. This means that the good construction of the test and the verification of all its psychometric properties ensures that the test avoids any violations in the structure of the items, which in turn affects the individual's performance on the test items [36, 60-62].

VII. METHODS

The descriptive survey method was used to obtain data from a real-life scenario of giving postgraduate-level midterm and final exams to assess master's students' achievement level in the subject of applied statistics. This analytical study aimed to determine the quality of the test, its efficiency, and the reliability of its results despite the varying circumstances in which it is given.

A. Measurements

A Criterion Referenced Test (CRT) was used to evaluate students' achievement in the course of applied statistics in a master's degree level in order to verify the quality of the test as a tool to evaluate the level of students' achievement. The test was developed and based on content analysis. The test specification table covered some of the course vocabulary for the applied statistics course. The test in its first form consisted of 25 test items that were reviewed by specialists in the field of statistics to determine their face validity, and 7 items were omitted as a result. So, in its final form, the test had 18 items, and they were applied to the study sample to verify their quality (Table III). The results of the thorough analysis of the test items were handed over to the central question bank in order to compare the performance of the test items and verify their lifespan when re-performing statistical operations on them later. To verify the test reliability, the Kuder Richardson (KR-20) method was used because the binary data are coded using 0 and 1 after correcting the items, and because the test items differ in their difficulty parameter. The results indicated that the test has a high reliability coefficient of KR-20 = 0.842.

B. Sample

The current study population consisted of all students of the applied statistics course at the master's level at Umm Al-Qura University on the main campus in Makkah and all branches of the University. The size of the population was rather large, estimated at about 400 male and female students registered during the second semester of the academic year 2020. It was difficult to reach all the sample members because of the financial cost, time, and effort required. Moreover, the educational and population environment conditions for all students are very similar and the previous studies using samples from students in the Saudi universities' population did not show any clear bias. Therefore, the current study used a random sample consisting of 338 students, equivalent to the 84.5% of the study population, studying different disciplines in the College of Education.

TABLE III. TEST ITEMS

Test Items	Alternatives (answers)
1. Which of the following is true for the interval scale level?	- Classification of individuals - Ranking of individuals and identifying differences - Both
2. If the value of the correlation coefficient is equal to (-0.8) this is an indication that the relationship is	- Weak - Nonexistent - Strong
3. Estimates that are calculated by studying the sample members are called	- Variables - Parameters - Statistics
4. When the population is homogeneous and its number is very large, we use the following type of sampling	- Simple Random Sampling - Stratified Random Sampling - Cluster Random Sampling
5. When studying the relationship between job performance and job satisfaction, job satisfaction is called	- Dependent variable - Independent variable - Intruder valuable
6. Which of the following statements is true for the relationship between the sample and the population?	- Population parameters are a good estimate of sample statistics - Population parameters are a good estimate of raw scores - Sample statistics are a good estimate of population parameters
7. If the relationship between chronological age and academic achievement is (r = 0.91) then this is evidence of	- The greater the age, the greater the achievement - The younger the age, the lesser the achievement - The one or the other
8. If the sum of the squares of the deviations from the mean = 80 and the number of students = 21 then the standard deviation equals	- 2 - 4 - 6
9. The range of the relationship that exists between two quantitative variables is called	- Slope - Connection - Range
10. The mode for the values (1, 4, 9, 12) is	- Zero - 12 - No mode
11. The median of the values (9, 15, 7, 10, 12) is	- 7 - 9 - 10
12. In one of the regions in KSA, a study was conducted on the pros and cons of the e-learning system at the undergraduate level. In this study the academic level is	- Variable - Constant - Other
13. The measure of central tendency not affected by outliers is	- The median - The mode - The one or the other
14. The number of training courses during a whole semester is a	- Continuous quantitative variable - Discrete quantitative variable - Descriptive variable
15. Achievement tests depend on the following level of measurement scale	- Ordinal - Interval - Ratio
17. When the sample size is increased	- It increases the probability of a normal distribution - It makes the sample to not comply with the normal distribution - There is no relation between them
18. Setting the confidence level at 95% is a prerequisite for educational sciences.	- Yes - It may take different values according to the nature of the study - Other

The sample distribution when responding to the test is represented in Tables IV and V. The members of the study

sample were contacted by e-mail, and a test link was created and made available on the student's electronic page. The link was made available for one hour, representing the testing period, and prior coordination with the study sample was made to select a time appropriate for everyone. Cases which had any type of technological problems were not recorded. Electronic reminders via the university's electronic system were used before the test to alert the study sample about the test time.

TABLE IV. DISTRIBUTION OF THE STUDY SAMPLE ACCORDING TO GENDER AND CURRENT MAJOR

Gender	Current major in the master's study					
	Psychology	Islamic Education	Curriculum and Instruction	Educational Administration	Special Education	Total
Male	45	38	37	30	30	180
Female	29	31	29	32	37	158
Total	74	69	66	62	67	338

TABLE V. DISTRIBUTION OF THE STUDY SAMPLE ACCORDING TO STUDYING STATISTICS IN EARLIER STAGES AND CURRENT MAJOR

Studied statistics in earlier stage	Current major in the master's study					
	Psychology	Islamic Education	Curriculum and Instruction	Educational Administration	Special Education	Total
Yes	56	51	52	42	34	235
No	18	18	14	20	33	103
Total	74	69	66	62	67	338

VIII. RESULTS

Figure 1 show the eigenvalue scree plot. It is clear that the first eigenvalue is much greater than the others, suggesting that a unidimensional model is reasonable for this data.

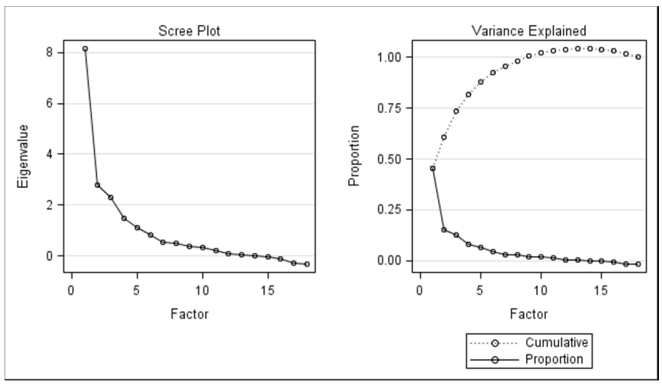

Fig. 1. Eigenvalue scree plot.

This test has items with one correct alternative answer that worths a single point, the item difficulty was simply the percentage of students who answer an item correctly, which is equal to the item mean. In this case, the item difficulty index in Tables VI to X shows the ranges, based on the examinees' characteristics. The item difficulty ranged from 0.647 to 0.928. For students who studied statistics in earlier stages, the items' difficulties ranged from 0.878 to 0.970, however for students who did not, the items' difficulties ranged from 0.310 to 0.893, while the items' difficulties based on gender ranged from 0.650 to 0.966 for male and 0.645 to 0.924 for the female students. Regarding the students' current major, the item difficulty varied over subjects: for special education, it ranged from 0.522 to 0.985, for educational administration it ranged from 0.50 to 0.870, for curriculum and instruction it ranged from 0.651 to 0.999, for Islamic education it ranged from 0.521 to 0.927, and for psychology it ranged from 0.837 to 0.999. For higher GPA the items' difficulties ranged from 0.857 to 0.994, whereas for moderate GPA they ranged from 0.400 to 0.936, and for low GPA they ranged from 0.021 to 0.869.

TABLE VI. OVERALL ITEM DIFFICULTY

Items	Overall
Q1	0.837
Q2	0.668
Q3	0.843
Q4	0.751
Q5	0.647
Q6	0.928
Q7	0.798
Q8	0.763
Q9	0.917
Q10	0.857
Q11	0.757
Q12	0.792
Q13	0.784
Q14	0.739
Q15	0.695
Q16	0.825
Q17	0.786
Q18	0.828

TABLE VII. ITEM DIFFICULTY ACCORING TO STUDYING STATISTICS IN AN EARLIER STAGE

Items	Studying statistics in earlier stages	
	Yes	No
Q1	0.927	0.631
Q2	0.787	0.398
Q3	0.970	0.553
Q4	0.842	0.543
Q5	0.795	0.310
Q6	0.944	0.893
Q7	0.897	0.572
Q8	0.765	0.757
Q9	0.974	0.786
Q10	0.944	0.660
Q11	0.834	0.580
Q12	0.868	0.621
Q13	0.880	0.563
Q14	0.872	0.436
Q15	0.825	0.398
Q16	0.910	0.631
Q17	0.842	0.660
Q18	0.893	0.679

Figures 2 - 19 present the combined curve for the 18 items based on the overall data. Each item has one threshold. Most thresholds span the negative section of the ability. Item information curves show that the items provide a good amount of information for students who had a lower or moderate ability compared to high ability students. In terms of precision, most of the items were more convenient to lower ability students (e.g. items 9, 10, 17), while items 2, 5, 14 gathered more information for students who had moderate ability.

TABLE VIII. ITEMS DIFFICULTY ACCORING TO GENDER

Items	Gender	
	Male	Female
Q1	0.844	0.829
Q2	0.661	0.677
Q3	0.872	0.810
Q4	0.761	0.740
Q5	0.650	0.645
Q6	0.933	0.924
Q7	0.827	0.765
Q8	0.688	0.848
Q9	0.966	0.860
Q10	0.861	0.854
Q11	0.772	0.740
Q12	0.816	0.765
Q13	0.850	0.708
Q14	0.777	0.696
Q15	0.755	0.626
Q16	0.888	0.753
Q17	0.850	0.715
Q18	0.838	0.816

TABLE IX. ITEMS DIFFICULTY ACCORING TO THE CURRENT MAJOR

Items	Current master's study major				
	Special Education	Educational Administration	Curriculum and Instruction	Islamic Education	Psychology
Q1	0.761	0.822	0.939	0.768	0.891
Q2	0.522	0.645	0.712	0.521	0.918
Q3	0.567	0.838	0.999	0.840	0.959
Q4	0.731	0.693	0.848	0.623	0.851
Q5	0.507	0.500	0.651	0.623	0.918
Q6	0.985	0.725	0.954	0.956	0.999
Q7	0.776	0.741	0.772	0.753	0.932
Q8	0.820	0.822	0.893	0.608	0.689
Q9	0.865	0.838	0.969	0.927	0.972
Q10	0.656	0.870	0.984	0.782	0.986
Q11	0.656	0.661	0.939	0.521	0.986
Q12	0.761	0.709	0.954	0.637	0.891
Q13	0.776	0.741	0.787	0.652	0.945
Q14	0.567	0.596	0.863	0.710	0.932
Q15	0.582	0.516	0.742	0.739	0.864
Q16	0.791	0.629	0.878	0.869	0.932
Q17	0.776	0.677	0.772	0.855	0.837
Q18	0.791	0.725	0.893	0.840	0.873

Figure 20 represents the test characteristic curve that was the functional relation between the true score and the ability scale. As we can see the probability of the correct response was near to 1 at the lowest levels of ability and it increased until it came to the highest levels of ability. The probability of correct response in this test is about 18 for high ability students. Figure 21 presents the total amount of the information that has been obtained from the test. It appears clearly that the test gives good indicators to assess lower levels of agreement.

TABLE X. ITEM DIFFICULTY ACCORING TO THE GPA

Items	GPA		
	≥ 3.75	3.30-3.75	2.5 – 3.30
Q1	0.950	0.809	0.456
Q2	0.950	0.427	0.130
Q3	0.989	0.872	0.195
Q4	0.928	0.581	0.456
Q5	0.956	0.400	0.021
Q6	0.994	0.845	0.869
Q7	0.939	0.700	0.478
Q8	0.857	0.645	0.673
Q9	0.983	0.936	0.608
Q10	0.939	0.854	0.304
Q11	0.972	0.554	0.282
Q12	0.956	0.627	0.543
Q13	0.956	0.627	0.478
Q14	0.972	0.627	0.086
Q15	0.939	0.527	0.130
Q16	0.950	0.745	0.521
Q17	0.923	0.700	0.456
Q18	0.934	0.781	0.521

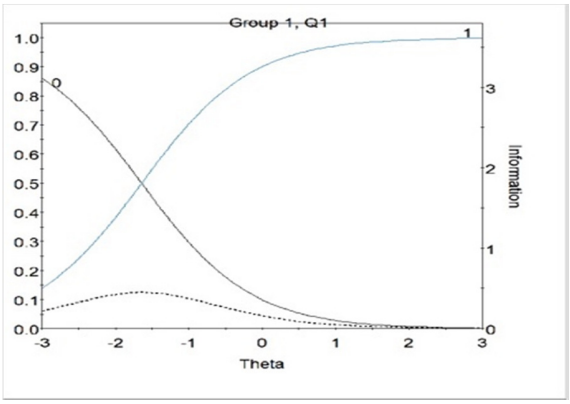

Fig. 2. Characteristic curve of item 1.

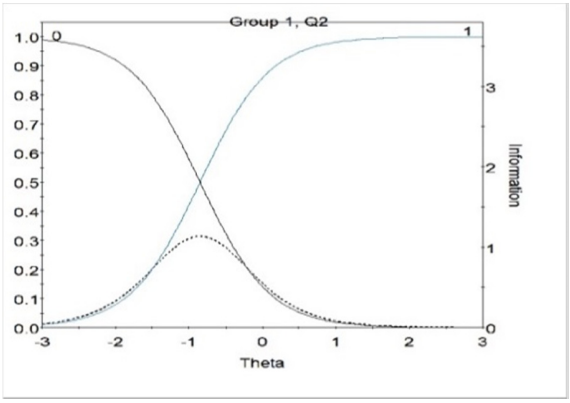

Fig. 3. Characteristic curve of item 2.

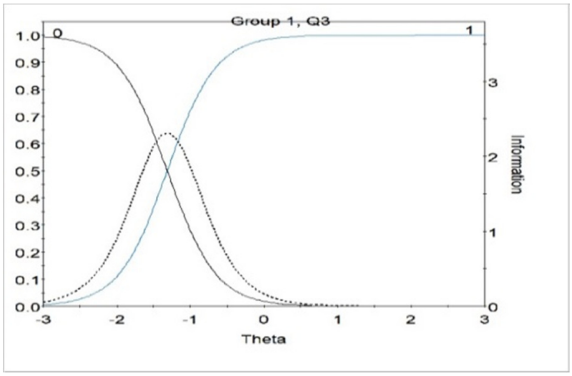

Fig. 4. Characteristic curve of item 3.

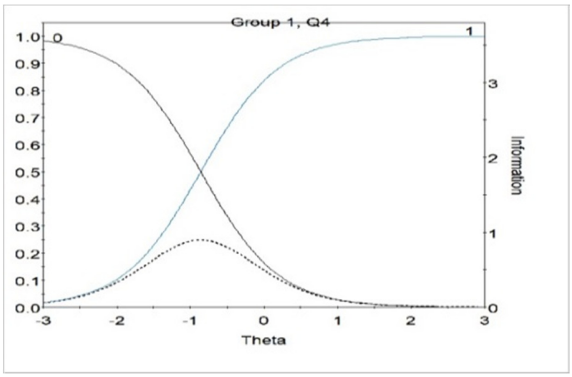

Fig. 5. Characteristic curve of item 4.

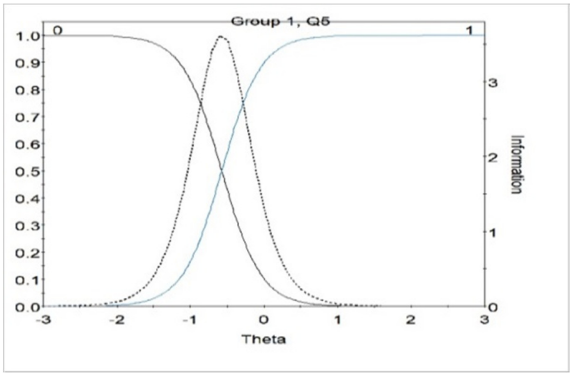

Fig. 6. Characteristic curve of item 5.

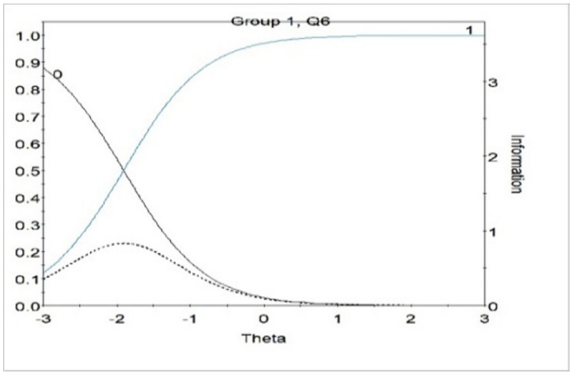

Fig. 7. Characteristic curve of item 6.

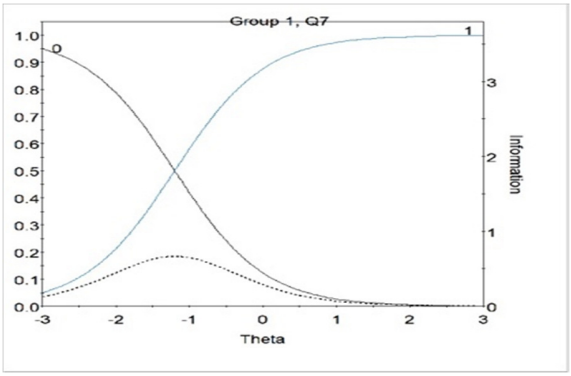

Fig. 8. Characteristic curve of item 7.

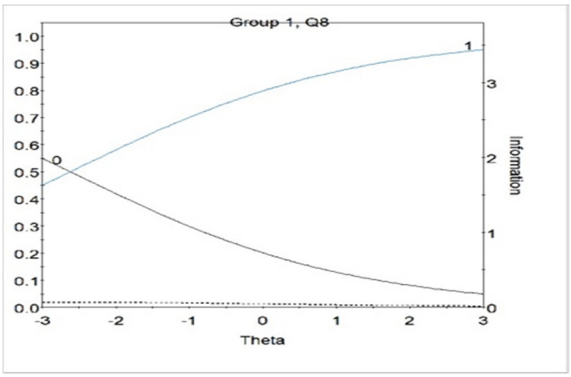

Fig. 9. Characteristic curve of item 8.

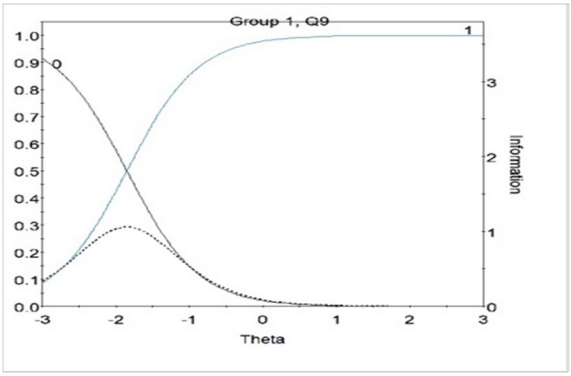

Fig. 10. Characteristic curve of item 9.

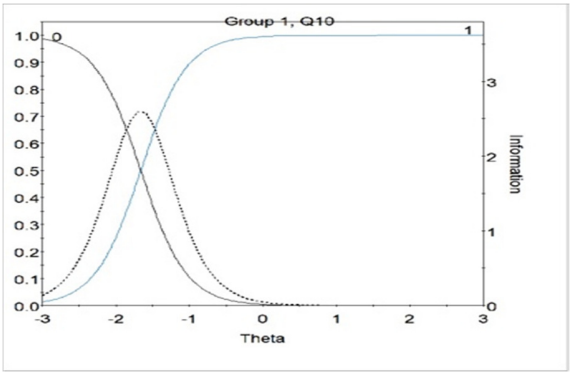

Fig. 11. Characteristic curve of item 10.

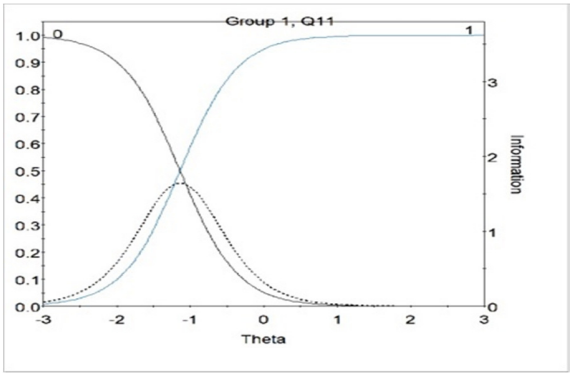

Fig. 12. Characteristic curve of item 11.

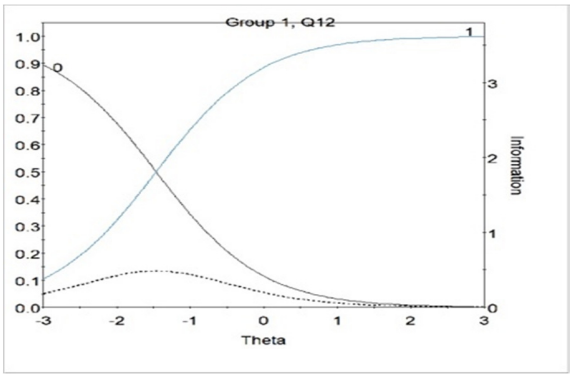

Fig. 13. Characteristic curve of item 12.

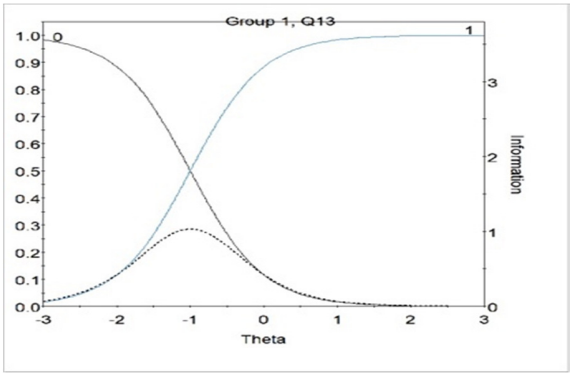

Fig. 14. Characteristic curve of item 13.

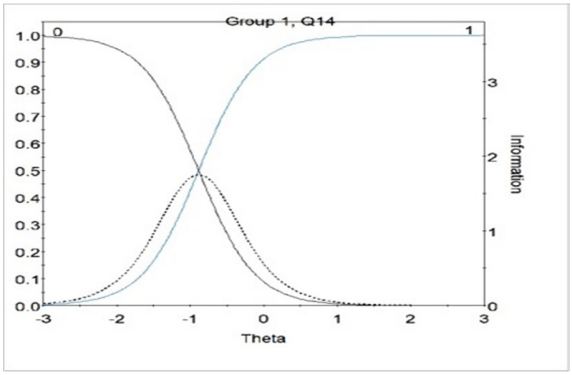

Fig. 15. Characteristic curve of item 14.

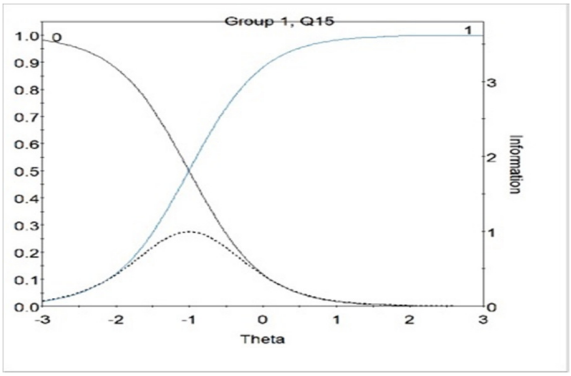

Fig. 16. Characteristic curve of item 15.

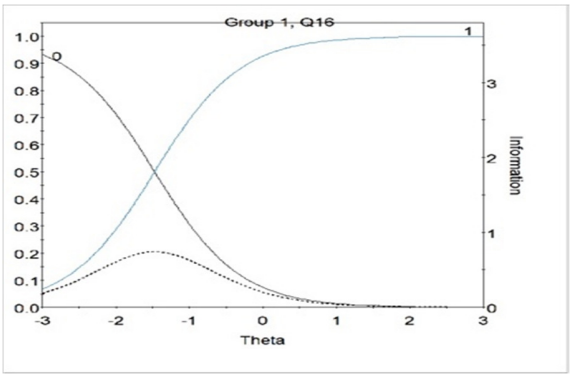

Fig. 17. Characteristic curve of item 16.

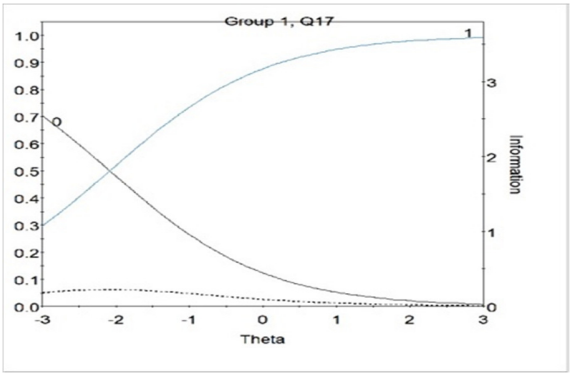

Fig. 18. Characteristic curve of item 17.

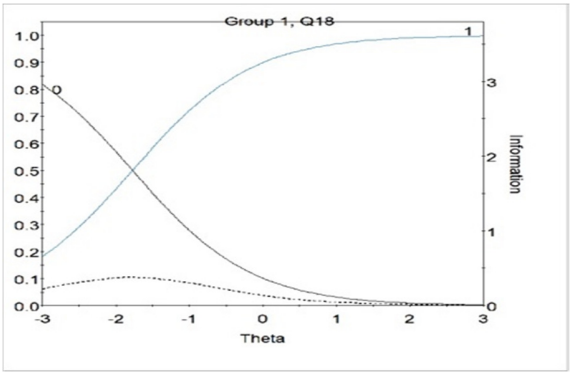

Fig. 19. Characteristic curve of item 18.

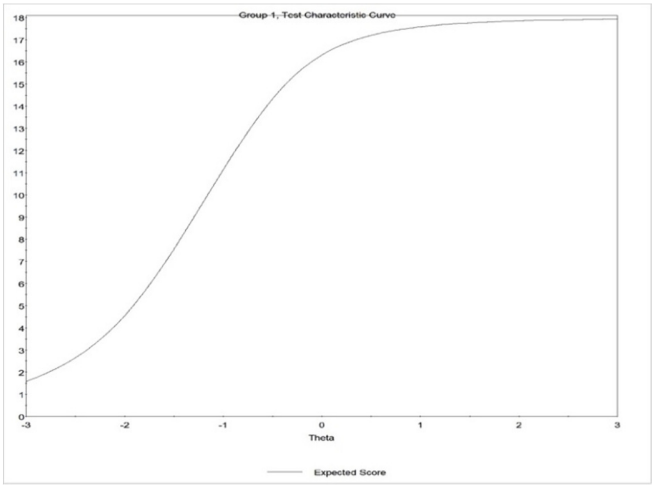

Fig. 20. Test characteristic curve between the true score and the ability scale.

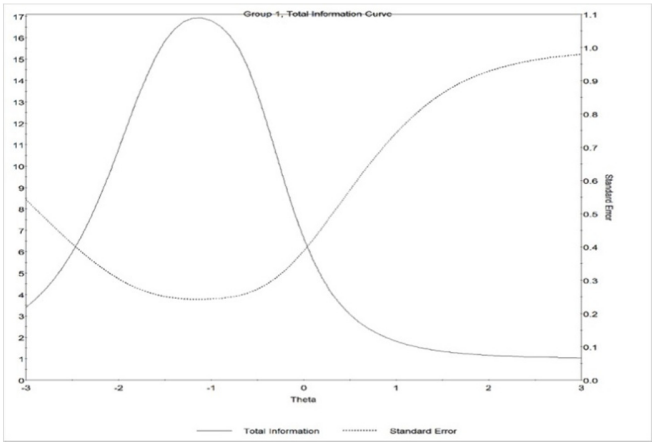

Fig. 21. Test characteristic curve.

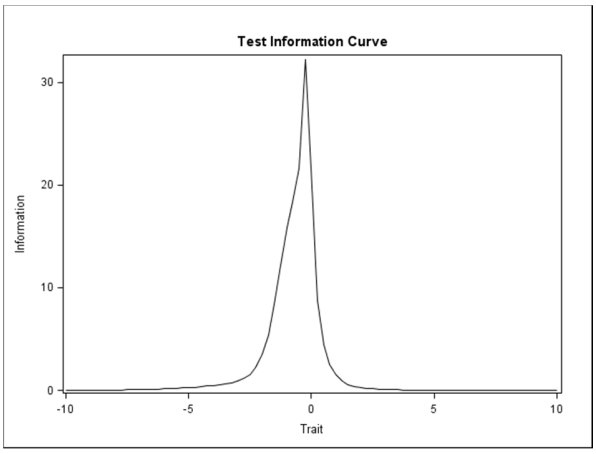

Fig. 22. Test information curve according to male examinees.

Figures 22 to 33 represent the test characteristic curve according to the change in the examinees' characteristics. The amount of information that has been obtained by female students appeared to be more than the information extracted by male and the test provided more information about students who were not studying statistics in earlier levels compared to students who were. Figures 27 and 28 also confirm that the test provides a good amount of information for students who had a lower or moderate ability compared to students who had high ability.

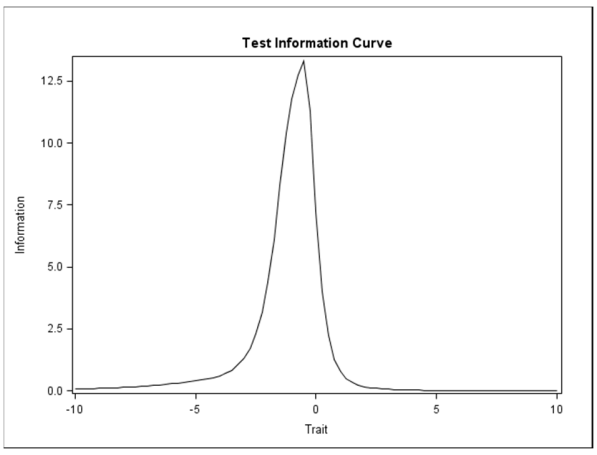

Fig. 23. Test information curve according to female examinees.

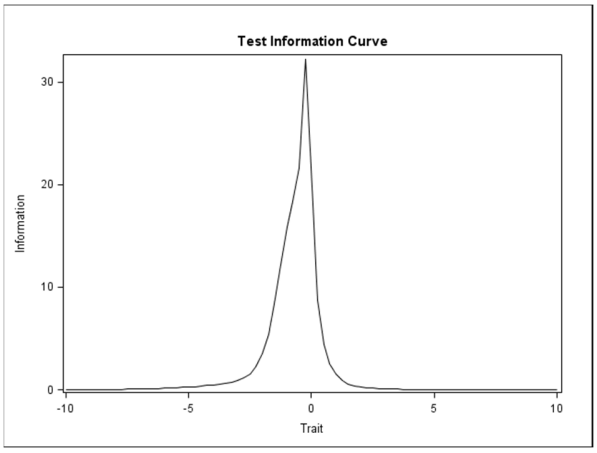

Fig. 24. Test information curve according to examinees who were studying statistics in earlier stage.

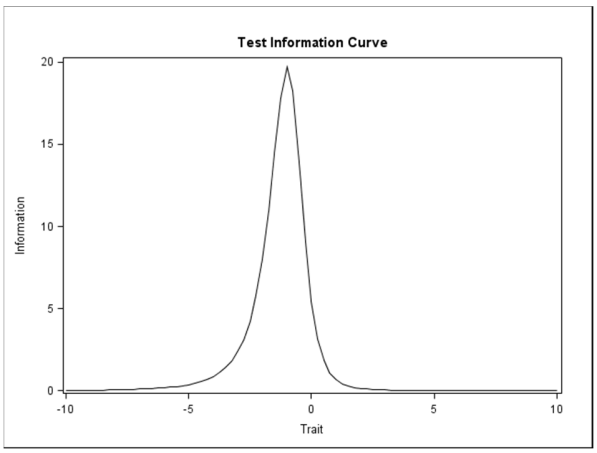

Fig. 25. Test information curve according to examinees who were not studying statistics in an earlier stage.

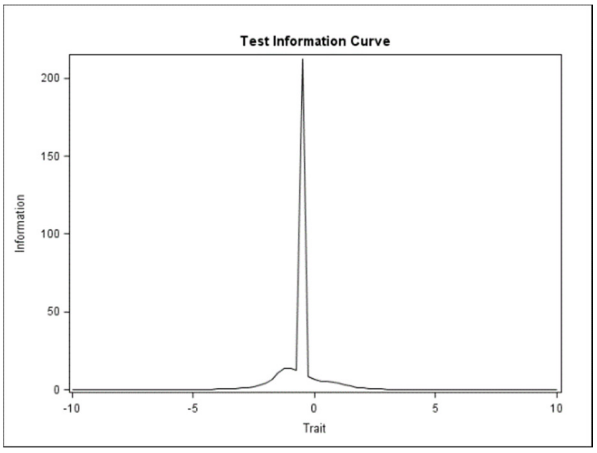

Fig. 26. Test information curve according to examinees who had a GPA of 3.75 and above.

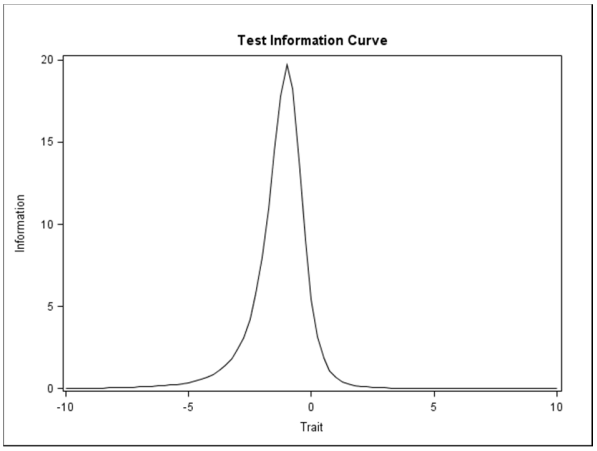

Fig. 27. Test information curve according to examinees who had a GPA in the range from 3.30 to less than 3.75.

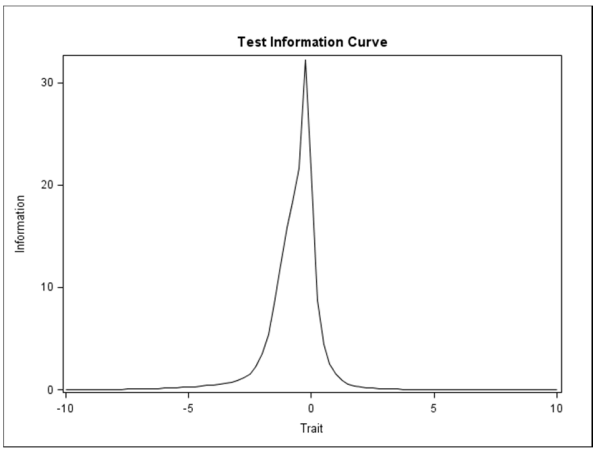

Fig. 28. Test information curve according to examinees who had a GPA in the range from 2.50 to less than 3.30.

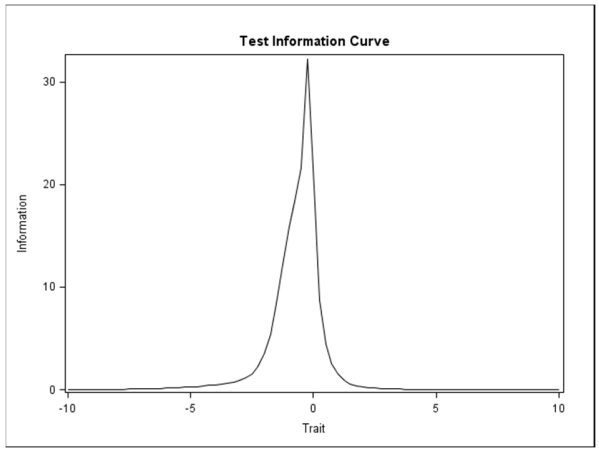

Fig. 29. Test information curve according to current major (Special Education).

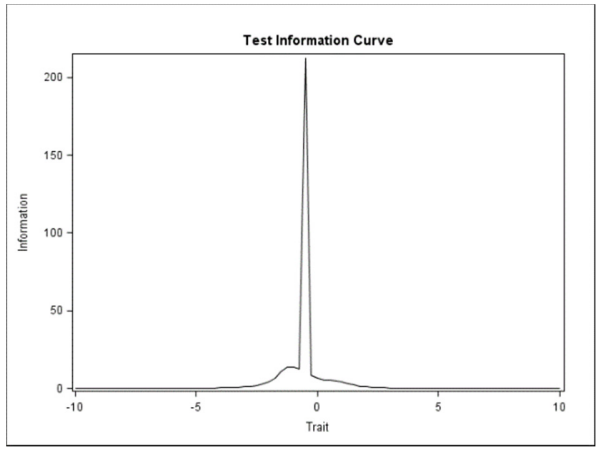

Fig. 30. Test information curve according to current major (Educational Administration).

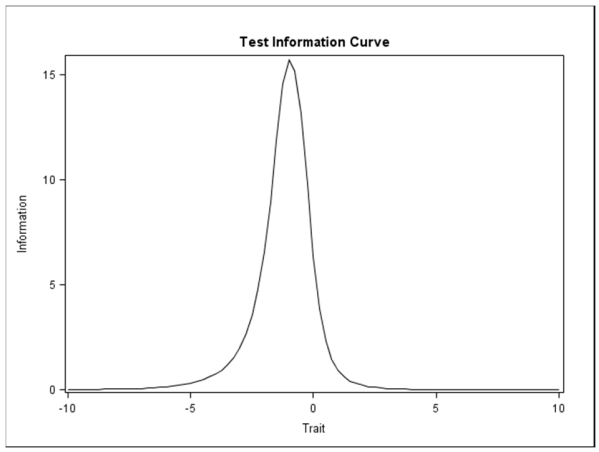

Fig. 31. Test information curve according to current major (Curriculum and Instruction).

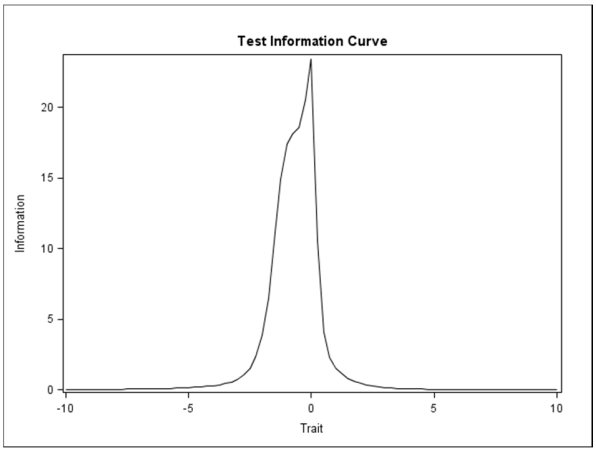

Fig. 32. Test information curve according to current major (Islamic Education).

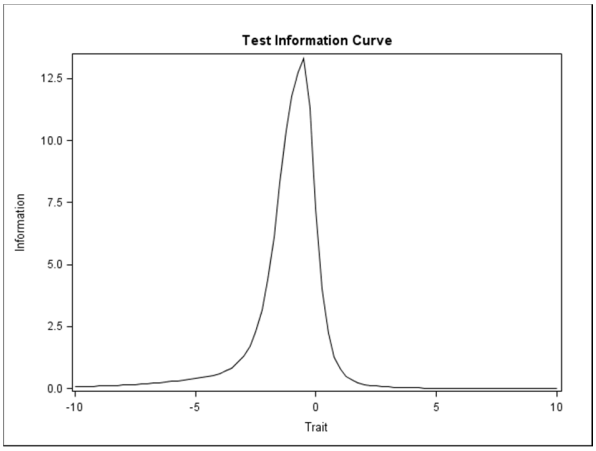

Fig. 33. Test information curve according to current major (Psychology).

IX. DISCUSSION

The study findings support previous researches [5-7, 63] for the use of IRT models in analyzing students' responses to an achievement test in a more objective way, showing whether there is an effect of the multiplicity of characteristics of the participants on the test items (multiple choice) in terms of the accuracy of the estimates of the items' parameters and the individuals' ability parameters. The results also help determine the relationship between individual responses and their basic ability. Within the framework of modern measurement theory, many models and applications have been formulated and applied to real test data. This includes the measurement assumptions about the characteristics of the test items, the performance of the subjects, and how performance is related to knowledge. This test clearly indicated that based on the level of the statistics course, there should be a periodic review of the tests in line with the nature and level of the course materials in order to have a logical judgment about the level of students' progress and the level of their ability. This conclusion is consistent with the findings of [50, 64].

X. CONCLUSION

In general, this study seeks to further improve the evaluation practices in the educational process. The tests that describe the students' progress in the educational process should be subject to review by evaluation and measurement specialists in order to ensure that we have valid and reliable evaluation tools. The administrators of the educational system need to find a mechanism to review question banks and align them with the requirements of the scientific material of the courses that are subject to continuous development.

REFERENCES

[1] B. Zhuang, S. Wang, S. Zhao, and M. Lu, "Computed tomography angiography-derived fractional flow reserve (CT-FFR) for the detection of myocardial ischemia with invasive fractional flow reserve as reference: systematic review and meta-analysis," *European Radiology*, vol. 30, no. 2, pp. 712–725, Feb. 2020, https://doi.org/10.1007/s00330-019-06470-8.

[2] Y. A. Wang and M. Rhemtulla, "Power Analysis for Parameter Estimation in Structural Equation Modeling:A Discussion and Tutorial," in *Advances in Methods and Practices in Psychological Science*, California, USA: University of California, 2020.

[3] H. Zhu, W. Gao, and X. Zhang, "Bayesian Analysis of a Quantile Multilevel Item Response Theory Model," *Frontiers in Psychology*, vol. 11, Jan. 2021, Art. no. 607731, https://doi.org/10.3389/fpsyg.2020.607731.

[4] M. R. Szeles, "Examining the foreign policy attitudes in Moldova," *PLOS ONE*, vol. 16, no. 1, 2021, Art. no. e0245322, https://doi.org/10.1371/journal.pone.0245322.

[5] D. Almaleki, "The Precision of the Overall Data-Model Fit for Different Design Features in Confirmatory Factor Analysis," *Engineering, Technology & Applied Science Research*, vol. 11, no. 1, pp. 6766–6774, Feb. 2021, https://doi.org/10.48084/etasr.4025.

[6] D. Almaleki, "Empirical Evaluation of Different Features of Design in Confirmatory Factor Analysis," Ph.D. dissertation, Western Michigan University, MC, USA, 2016.

[7] C. S. Wardley, E. B. Applegate, A. D. Almaleki, and J. A. Van Rhee, "A Comparison of Students' Perceptions of Stress in Parallel Problem-Based and Lecture-Based Curricula," *The Journal of Physician Assistant Education*, vol. 27, no. 1, pp. 7–16, Mar. 2016, https://doi.org/10.1097/JPA.0000000000000060.

[8] C. Wardley, E. Applegate, A. Almaleki, and J. V. Rhee, "Is Student Stress Related to Personality or Learning Environment in a Physician Assistant Program?," *The Journal of Physician Assistant Education*, vol. 30, no. 1, pp. 9–19, Mar. 2019, https://doi.org/10.1097/JPA.0000000000000241.

[9] A. C. Villa Montoya *et al.*, "Optimization of key factors affecting hydrogen production from coffee waste using factorial design and metagenomic analysis of the microbial community," *International Journal of Hydrogen Energy*, vol. 45, no. 7, pp. 4205–4222, Feb. 2020, https://doi.org/10.1016/j.ijhydene.2019.12.062.

[10] N. M. Moo-Tun, G. Iniguez-Covarrubias, and A. Valadez-Gonzalez, "Assessing the effect of PLA, cellulose microfibers and CaCO3 on the properties of starch-based foams using a factorial design," *Polymer Testing*, vol. 86, Jun. 2020, Art. no. 106482, https://doi.org/10.1016/j.polymertesting.2020.106482.

[11] K. M. Marcoulides, N. Foldnes, and S. Grønneberg, "Assessing Model Fit in Structural Equation Modeling Using Appropriate Test Statistics," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 27, no. 3, pp. 369–379, May 2020, https://doi.org/10.1080/10705511.2019.1647785.

[12] M. D. H. Naveiras, "Using Auxiliary Item Information in the Item Parameter Estimation of a Graded Response Model for a Small to Medium Sample Size: Empirical versus Hierarchical Bayes Estimation," Ph.D. dissertation, Vanderbilt University, Nashville, TN, USA, 2020.

[13] M. N. Morshed, M. N. Pervez, N. Behary, N. Bouazizi, J. Guan, and V. A. Nierstrasz, "Statistical modeling and optimization of heterogeneous Fenton-like removal of organic pollutant using fibrous catalysts: a full factorial design," *Scientific Reports*, vol. 10, no. 1, Sep. 2020, Art. no. 16133, https://doi.org/10.1038/s41598-020-72401-z.

[14] W. van Lankveld, R. J. Pat-El, N. van Melick, R. van Cingel, and J. B. Staal, "Is Fear of Harm (FoH) in Sports-Related Activities a Latent Trait? The Item Response Model Applied to the Photographic Series of Sports Activities for Anterior Cruciate Ligament Rupture (PHOSA-ACLR)," *International Journal of Environmental Research and Public Health*, vol. 17, no. 18, Sep. 2020, Art. no. 6764, https://doi.org/10.3390/ijerph17186764.

[15] C. Shin, S.-H. Lee, K.-M. Han, H.-K. Yoon, and C. Han, "Comparison of the Usefulness of the PHQ-8 and PHQ-9 for Screening for Major Depressive Disorder: Analysis of Psychiatric Outpatient Data," *Psychiatry Investigation*, vol. 16, no. 4, pp. 300–305, Apr. 2019, https://doi.org/10.30773/pi.2019.02.01.

[16] C. W. Ong, B. G. Pierce, D. W. Woods, M. P. Twohig, and M. E. Levin, "The Acceptance and Action Questionnaire – II: an Item Response Theory Analysis," *Journal of Psychopathology and Behavioral Assessment*, vol. 41, no. 1, pp. 123–134, Mar. 2019, https://doi.org/10.1007/s10862-018-9694-2.

[17] A. Acevedo-Mesa, J. N. Tendeiro, A. Roest, J. G. M. Rosmalen, and R. Monden, "Improving the Measurement of Functional Somatic Symptoms With Item Response Theory," *Assessment*, Aug. 2020, Art. no. 1073191120947153, https://doi.org/10.1177/1073191120947153.

[18] J. Xia, Z. Tang, P. Wu, J. Wang, and J. Yu, "Use of item response theory to develop a shortened version of the EORTC QLQ-BR23 scales," *Scientific Reports*, vol. 9, no. 1, Feb. 2019, Art. no. 1764, https://doi.org/10.1038/s41598-018-37965-x.

[19] Y. Liu and J. S. Yang, "Interval Estimation of Latent Variable Scores in Item Response Theory," *Journal of Educational and Behavioral Statistics*, vol. 43, no. 3, pp. 259–285, Jun. 2018, https://doi.org/10.3102/1076998617732764.

[20] U. Gromping, "Coding invariance in factorial linear models and a new tool for assessing combinatorial equivalence of factorial designs," *Journal of Statistical Planning and Inference*, vol. 193, pp. 1–14, Feb. 2018, https://doi.org/10.1016/j.jspi.2017.07.004.

[21] P. J. Ferrando and U. Lorenzo-Seva, "Assessing the Quality and Appropriateness of Factor Solutions and Factor Score Estimates in Exploratory Item Factor Analysis," *Educational and Psychological Measurement*, vol. 78, no. 5, pp. 762–780, Oct. 2018, https://doi.org/10.1177/0013164417719308.

[22] X. An and Y.-F. Yung, "Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It," SAS Institute Inc., Paper SAS364-2014.

[23] K. Coughlin, "An Analysis of Factor Extraction Strategies: A Comparison of the Relative Strengths of Principal Axis, Ordinary Least Squares, and Maximum Likelihood in Research Contexts that Include both Categorical and Continuous Variables," Ph.D. dissertation, University of South Florida, Tampa, FL, USA, 2013.

[24] D. L. Bandalos and P. Gagne, "Simulation methods in structural equation modeling," in *Handbook of structural equation modeling*, New York, NY, USA: The Guilford Press, 2012, pp. 92–108.

[25] J. C. F. de Winter, D. Dodou, and P. A. Wieringa, "Exploratory Factor Analysis With Small Sample Sizes," *Multivariate Behavioral Research*, vol. 44, no. 2, pp. 147–181, Apr. 2009, https://doi.org/10.1080/00273170902794206.

[26] J. D. Kechagias, K.-E. Aslani, N. A. Fountas, N. M. Vaxevanidis, and D. E. Manolakos, "A comparative investigation of Taguchi and full factorial design for machinability prediction in turning of a titanium alloy," *Measurement*, vol. 151, Feb. 2020, Art. no. 107213, https://doi.org/10.1016/j.measurement.2019.107213.

[27] G. Kuan, A. Sabo, S. Sawang, and Y. C. Kueh, "Factorial validity, measurement and structure invariance of the Malay language decisional balance scale in exercise across gender," *PLOS ONE*, vol. 15, no. 3, 2020, Art. no. e0230644, https://doi.org/10.1371/journal.pone.0230644.

[28] M. J. Allen and W. M. Yen, *Introduction to measurement theory*. Monterey, CA, USA: Cole Publishing, 1979.

[29] O. P. John and S. Srivastava, "The Big Five trait taxonomy: History, measurement, and theoretical perspectives," in *Handbook of personality: Theory and research*, New York, NY, USA: Guilford Press, 1999, pp. 102–138.

[30] S.-H. Joo, L. Khorramdel, K. Yamamoto, H. J. Shin, and F. Robin, "Evaluating Item Fit Statistic Thresholds in PISA: Analysis of Cross-Country Comparability of Cognitive Items," *Educational Measurement: Issues and Practice*, Nov. 2020, https://doi.org/10.1111/emip.12404.

[31] H. Bourdeaud'hui, "Investigating the effects of presenting listening test items in a singular versus dual mode on students' critical listening performance," in *Upper-primary school students' listening skills: Assessment and the relationship with student and class-level characteristics*, Ghent, Belgium: Ghent University, 2019.

[32] D. M. Dimitrov and Y. Luo, "A Note on the D-Scoring Method Adapted for Polytomous Test Items," *Educational and Psychological Measurement*, vol. 79, no. 3, pp. 545–557, Jun. 2019, https://doi.org/10.1177/0013164418786014.

[33] J. Suarez-Alvarez, I. Pedrosa, L. Lozano, E. Garcia-Cueto, M. Cuesta, and J. Muniz, "Using reversed items in Likert scales: A questionable practice," *Psicothema*, vol. 30, no. 2, pp. 149–158, 2018, https://doi.org/10.7334/psicothema2018.33.

[34] J. P. Lalor, H. Wu, and H. Yu, "Learning Latent Parameters without Human Response Patterns: Item Response Theory with Artificial Crowds," in *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, Nov. 2019, pp. 4240–4250, https://doi.org/10.18653/v1/D19-1434.

[35] B. Couvy-Duchesne, T. A. Davenport, N. G. Martin, M. J. Wright, and I. B. Hickie, "Validation and psychometric properties of the Somatic and Psychological HEalth REport (SPHERE) in a young Australian-based population sample using non-parametric item response theory," *BMC Psychiatry*, vol. 17, no. 1, Aug. 2017, Art. no. 279, https://doi.org/10.1186/s12888-017-1420-1.

[36] P. M. Bentler and D. G. Bonett, "Significance tests and goodness of fit in the analysis of covariance structures," *Psychological Bulletin*, vol. 88, no. 3, pp. 588–606, 1980, https://doi.org/10.1037/0033-2909.88.3.588.

[37] A. Schimmenti, L. Sideli, L. L. Marca, A. Gori, and G. Terrone, "Reliability, Validity, and Factor Structure of the Maladaptive Daydreaming Scale (MDS–16) in an Italian Sample," *Journal of Personality Assessment*, vol. 102, no. 5, pp. 689–701, Sep. 2020, https://doi.org/10.1080/00223891.2019.1594240.

[38] C.-Y. Lin, V. Imani, M. D. Griffiths, and A. H. Pakpour, "Validity of the Yale Food Addiction Scale for Children (YFAS-C): Classical test theory and item response theory of the Persian YFAS-C," *Eating and Weight Disorders - Studies on Anorexia, Bulimia and Obesity*, Jul. 2020, https://doi.org/10.1007/s40519-020-00956-x.

[39] L. Jiang *et al.*, "The Reliability and Validity of the Center for Epidemiologic Studies Depression Scale (CES-D) for Chinese University Students," *Frontiers in Psychiatry*, vol. 10, 2019, Art. no. 315, https://doi.org/10.3389/fpsyt.2019.00315.

[40] S. Doi, M. Ito, Y. Takebayashi, K. Muramatsu, and M. Horikoshi, "Factorial validity and invariance of the Patient Health Questionnaire (PHQ)-9 among clinical and non-clinical populations," *PLOS ONE*, vol. 13, no. 7, 2018, Art. no. e0199235.

[41] T. Tsubakita, K. Shimazaki, H. Ito, and N. Kawazoe, "Item response theory analysis of the Utrecht Work Engagement Scale for Students (UWES-S) using a sample of Japanese university and college students majoring medical science, nursing, and natural science," *BMC Research Notes*, vol. 10, no. 1, Oct. 2017, Art. no. 528, https://doi.org/10.1186/s13104-017-2839-7.

[42] S. C. Smid, D. McNeish, M. Miocevic, and R. van de Schoot, "Bayesian Versus Frequentist Estimation for Structural Equation Models in Small Sample Contexts: A Systematic Review," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 27, no. 1, pp. 131–161, Jan. 2020, https://doi.org/10.1080/10705511.2019.1577140.

[43] M. K. Cain and Z. Zhang, "Fit for a Bayesian: An Evaluation of PPP and DIC for Structural Equation Modeling," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 26, no. 1, pp. 39–50, Jan. 2019, https://doi.org/10.1080/10705511.2018.1490648.

[44] D. Garson, "StatNotes: Topics in Multivariate Analysis," *North Carolina State University*. https://faculty.chass.ncsu.edu/garson/PA765/statnote.htm (accessed Feb. 10, 2021).

[45] H. W. Marsh, K.-T. Hau, and D. Grayson, "Goodness of Fit in Structural Equation Models," in *Contemporary psychometrics: A festschrift for Roderick P. McDonald*, Mahwah, NJ, USA: Lawrence Erlbaum Associates Publishers, 2005, pp. 275–340.

[46] I. Williams, "A speededness item response model for associating ability and speededness parameters," Ph.D. dissertation, Rutgers University, New Brunswick, NJ, USA, 2017.

[47] B. Shamshad and J. S. Siddiqui, "Testing Procedure for Item Response Probabilities of 2Class Latent Model," *Mehran University Research Journal of Engineering and Technology*, vol. 39, no. 3, pp. 657–667, Jul. 2020, https://doi.org/10.22581/muet1982.2003.20.

[48] K. M. Williams and B. D. Zumbo, "Item Characteristic Curve Estimation of Signal Detection Theory-Based Personality Data: A Two-Stage Approach to Item Response Modeling," *International Journal of Testing*, vol. 3, no. 2, pp. 189–213, Jun. 2003, https://doi.org/10.1207/S15327574IJT0302_7.

[49] D. Tafiadis *et al.*, "Using Receiver Operating Characteristic Curve to Define the Cutoff Points of Voice Handicap Index Applied to Young Adult Male Smokers," *Journal of Voice*, vol. 32, no. 4, pp. 443–448, Jul. 2018, https://doi.org/10.1016/j.jvoice.2017.06.007.

[50] L. Lina, D. Mardapi, and H. Haryanto, "Item Characteristics on Pro-TEFL Listening Section," presented at the First International Conference on Advances in Education, Humanities, and Language, ICEL 2019, Malang, Indonesia, 23-24 March 2019, Jul. 2019, https://dx.doi.org/10.4108/eai.11-7-2019.159630.

[51] D. L. Moody, "The method evaluation model: a theoretical model for validating information systems design methods," in *European Conference on Information Systems*, Naples, Italy, Jun. 2003, pp. 1–17.

[52] H. Davis, T. M. Rosner, M. C. D'Angelo, E. MacLellan, and B. Milliken, "Selective attention effects on recognition: the roles of list context and perceptual difficulty," *Psychological Research*, vol. 84, no. 5, pp. 1249–1268, Jul. 2020, https://doi.org/10.1007/s00426-019-01153-x.

[53] L. Sun, Y. Liu, and F. Luo, "Automatic Generation of Number Series Reasoning Items of High Difficulty," *Frontiers in Psychology*, vol. 10, 2019, Art. no. 884, https://doi.org/10.3389/fpsyg.2019.00884.

[54] T. O. Abe and E. O. Omole, "Difficulty and Discriminating Indices of Junior Secondary School Mathematics Examination; A Case Study of Oriade Local Government, Osun State," *American Journal of Education and Information Technology*, vol. 3, no. 2, pp. 37–46, Oct. 2019, https://doi.org/10.11648/j.ajeit.20190302.12.

[55] G. Nelson and S. R. Powell, "Computation Error Analysis: Students With Mathematics Difficulty Compared To Typically Achieving Students," *Assessment for Effective Intervention*, vol. 43, no. 3, pp. 144–156, Jun. 2018, https://doi.org/10.1177/1534508417745627.

[56] H. Retnawati, B. Kartowagiran, J. Arlinwibowo, and E. Sulistyaningsih, "Why Are the Mathematics National Examination Items Difficult and What Is Teachers' Strategy to Overcome It?," *International Journal of Instruction*, vol. 10, no. 3, pp. 257–276, Jul. 2017.

[57] T. A. Holster, J. W. Lake, and W. R. Pellowe, "Measuring and predicting graded reader difficulty," vol. 29, no. 2, pp. 218–244, Oct. 2017.

[58] S. Gaitas and M. A. Martins, "Teacher perceived difficulty in implementing differentiated instructional strategies in primary school," *International Journal of Inclusive Education*, vol. 21, no. 5, pp. 544–556, May 2017, https://doi.org/10.1080/13603116.2016.1223180.

[59] J. L. D'Sa and M. L. Visbal-Dionaldo, "Analysis of Multiple Choice Questions: Item Difficulty, Discrimination Index and Distractor Efficiency," *International Journal of Nursing Education*, vol. 9, no. 3, pp. 109–114, 2017.

[60] A. H. Blasi and M. Alsuwaiket, "Analysis of Students' Misconducts in Higher Education using Decision Tree and ANN Algorithms," *Engineering, Technology & Applied Science Research*, vol. 10, no. 6, pp. 6510–6514, Dec. 2020, https://doi.org/10.48084/etasr.3927.

[61] N. Sharifi, M. Falsafi, N. Farokhi, and E. Jamali, "Assessing the optimal method of detecting Differential Item Functioning in Computerized Adaptive Testing," *Quarterly of Educational Measurement*, vol. 9, no. 33, pp. 23–51, Oct. 2018, https://doi.org/10.22054/jem.2019.11109.1323.

[62] J. J. Hox, C. J. M. Maas, and M. J. S. Brinkhuis, "The effect of estimation method and sample size in multilevel structural equation modeling," *Statistica Neerlandica*, vol. 64, no. 2, pp. 157–170, 2010, https://doi.org/10.1111/j.1467-9574.2009.00445.x.

[63] G. Makransky, L. Lilleholt, and A. Aaby, "Development and validation of the Multimodal Presence Scale for virtual reality environments: A confirmatory factor analysis and item response theory approach," *Computers in Human Behavior*, vol. 72, pp. 276–285, Jul. 2017, https://doi.org/10.1016/j.chb.2017.02.066.

[64] J. A. Costa, J. Maroco, and J. Pinto-Gouveia, "Validation of the psychometric properties of cognitive fusion questionnaire. A study of the factorial validity and factorial invariance of the measure among osteoarticular disease, diabetes mellitus, obesity, depressive disorder, and general populations," *Clinical Psychology & Psychotherapy*, vol. 24, no. 5, pp. 1121–1129, 2017, https://doi.org/10.1002/cpp.2077.