International Journal of Instruction *July 2022 ● Vol.15, No.3*
e-ISSN: 1308-1470 ● www.e-iji.net *p-ISSN: 1694-609X*
pp. 1111-1130

Article submission code: 20210918131303 | Received: 18/09/2021 | Revision: 30/03/2022 | Accepted: 23/04/2022 | OnlineFirst: 24/06/2022

The Effect of Item Form on Estimating Person's Ability, Item Parameters, and Information Function According to Item Response Theory (IRT)

Taha Okleh ALKursheh
Dr., Department of Education and Psychology, University of Tabuk, Saudi Arabia, *talkursheh@ut.edu.sa*

Habis Saad Al-zboon
Dr., Faculty of Educational Sciences, Al-Hussein Bin Talal University, Jordan, *habes_zbn77@yahoo.com*

Mo'en Salman AlNasraween
Dr., Faculty of Educational and Psychological Sciences, Amman Arab University, Jordan, *mueen@aau.edu.jo*

This study aimed at comparing the effect of two test item formats (multiple-choice and complete) on estimating person's ability, item parameters and the test information function (TIF).To achieve the aim of the study, two format of mathematics(1) test have been created: multiple-choice and complete, In its final format consisted of (31) items. The test has been applied on (300) students in Tabuk University. The responses of the examinees were analyzed by BILOG– MG 3 programs for each item form according to the two parameter logistic model. The study findings revealed the following: there were statistically significant differences at the level of significance (α=0.05) among the standard error means for estimating item parameters (difficulty and discrimination) due to the format of test item in favor of the complete test. And the results showed there were statistically significant differences at the level of significance (α=0.05) among the standard error means for estimating person's ability due to the format of test item in favor of the complete test. On the other hand, the results showed there were statistically significant differences (α=0.05) among the standard error means of test information function (TIF) due to the format of test item in favor of the complete test. The study recommends educating teachers to diversify the forms of items of the tests they use so that they contain the two forms (multiple-choice and complete), and using complete items on a larger scale. than it is now being more discrimination.

Keywords: format of test item, item parameter, person's ability, test information function (TIF), item response theory (IRT)

Citation: ALKursheh, T. O., Al-zboon, H. S., & ALNasraween, M. S. (2022). The effect of item form on estimating person's ability, item parameters, and information function according to item response theory (IRT). *International Journal of Instruction*, *15*(3), 1111-1130. https://doi.org/10.29333/iji.2022.15359a

INTRODUCTION

Tests are among the important evaluation tools, as they are used in various forms to measure students' achievement in various academic subjects. Tests' results help in identifying the extent to which educational goals are achieved, comparing students to each other, diagnosing the strengths and weaknesses of students, or selecting persons for different jobs.(Hambelton & Cook, 1977; Moghadamzadeh et al,2011).

Test is considered an educational situation, as it is expected to include questions that measure the basic goals. It also covers the most important parts of the academic content, and the student is expected to think deeply about the test situation compared to other situations. Therefore, the ideas and information contained in the test have a distinct share of thinking, thus, leading to the possibility of retaining them (Thawabieh,2016; AlKhatib et al,2020)

Items are of many forms *that* can be used in evaluating students' achievement. Some of them require that the examinee chooses the answer from multiple choices, such as multiple-choice tests; true and false test; multiple true and false test; and matching test, while some others require that the examinee gives the answer in his own words such as essay tests, short answer test, and complete test (Haladyna and Downing,2002; Falani et al,2020).

Several factors control the selection of a specific type of items, including: the scope of the subject material, the level of the objective statement, the purpose of the evaluation, the ages and mental levels of students, and the extent to which the psychometric characteristics of the test are affected by the method of correction.(Haladyna et al,2002; Thawabieh,2016)

Audeh (1999) pointed out that a type of items does not pertain to a specific mental level or to certain psychometric characteristics such as validity and reliability, as it is possible to obtain a test with satisfactory characteristics regardless of the type of items if they are well prepared. However, this does not mean that any type of items can replace another type. While Aiken (1987) (referred to in Allen & Yen, 1979) indicated that multiple-choice tests became the most prevalent among all forms of tests in education, and what increased in the popularity and the spread of this type of items is its superiority over all other objective forms of items due to their efficiency and versatility. Therefore, that it is possible to measure simple and complex goals in various subjects.(Currie et al,2010).

It has been common for teachers to use two types of paragraphs, which are multiple choice paragraphs and supplementary paragraphs with a short and specific answer, because of their efficiency, as they can measure academic achievement in a more consistent and honest way, as well as they cover the largest amount of content vocabulary in which the student is to be tested, as well as the ease of correction and objectivity, And also the possibility of collecting data for his results and correcting them automatically. (Thawabieh,2016 ; Primi et al,2020)

On the other hand, according to Al-Kahlout (2002), despite the widespread use of the items of multiple-choice in measuring many educational purposes, it still faces increasing criticism. The guessing factor is still one of the factors that threaten its

validity and reliability, in addition to the insufficient empirical evidence to indicate that this type of items is more valid and reliable than all other objective tests. In contrast, the complete item is not affected by the random guessing because it requires that the examinee recalls the response and not the chosen answer, as it is easy to be prepared and does not require lengthy answers. However, it encourages the memorization of information and that it measures simple mental levels in addition to its being the least reliable types of objective questions (Murad & Suleiman, 2002; Shin & Gierl,2020).

It should be noted that most of studies relied on the traditional theory of measurement in analyzing the results, The study of Hambleton and Jones (1993) compared the two Classical Test Theory (CTT) and Item Response Theory (IRT). The study showed the shortcomings of the CCT, and how the IRT addressed these aspects. Moreover, it indicated the superiority shown by IRT in solving problems that the CCT cannot solve. Therefore, based on what the decision-making process requires of applying the measures through which a person's abilities can be accurately estimated(Subal et al,2021). This requires building measuring tools that are accurate and objective. Hence, the current study relied on the modern theory of measurement, or what is known as the Item Response Theory as it constitutes a framework for the current and prospective future attitude in the choice of items (Anastasi, 1982 ; O'Neill et al,2020; .Hassan & Miller,2019).

Purpose of the Study

There is a great debate about how the test-building process has evolved, and is it better to use items of the selected response type or the formulated response? There is a rationale for the assumption that the cognitive requirements of the selected response items differ from the formulated response items. Given the widespread use of both the multiple-choice and the complete items in measuring students' achievement and the contrast of studies' results in providing qualitative evidence that the two types of items present different mental processes. Therefore, the current study is a modest attempt to contribute to the investigation of the relationship of items' form to the accuracy of the assessment of the ability of persons, the parameters of the items, and the information function, specifically, this study will attempt to answer the following questions:

1. Does the accuracy of the estimates for the items' parameters differ (difficulty, discrimination) according to the form of the item (multiple-choice, complete)?
2. Does the accuracy of the ability estimates of persons differ according to the form of the item (multiple-choice, complete)?
3. Does the information function of the test and the standard error of the measurement differ according to the form of the item (multiple-choice, complete)?

The significance of the study:

The significance of the study lies in its extent of contribution to knowledge, both theoretical and practical sides. Consequently, the significance of the current study could be determined in that it will provide a new benchmark for increasing the accuracy of measuring tools (testing), as achievement tests are one of the most widespread and common measurement tools in universities and other educational institutions.

As the research findings of preference for one of the forms of the test items will guide researchers to the form of the test items that will render reliability and validity to the test results on which decisions are based on the students' progress from one academic level to another, and the level of students' mastery of the skills and knowledge they acquired during their study at university. This renders an accurate and valid picture of the level of university graduates and reflects positively on the ability of students to compete in the labor market. The importance of accurately estimating the performance levels in universities and their reflection on job opportunities, and the educational decisions taken on them concerning students. The present study could help those in charge of developing exams in educational institutions to take care of the test structure and preparation.

Terminology of the study:

Item parameters: These are the parameters of difficulty, discrimination, derived from the two-parameter logistic model. (Wang & Weiss,2018)

Item difficulty parameter: A point representing the position of the item on a power continuum corresponds to the probability $\left(1+\frac{c_i}{2}\right)$ for answering the correct answer item, where $(c_i$). represents the guessing parameter, and the item difficulty parameter is denoted by the symbol P_i. (Hambleton, Swaminthan, 1985).

Item discrimination parameter: The slope ratio of the item characteristics curve (ICC), which corresponds to the point at which the parameter of the ability on characteristics' continuum equal to the difficulty of the item. It is also defined as the ability of the item to distinguish between the different levels of the examinees on the power continuum and denoted by the symbol a_i. . (Wang & Weiss,2018).

Test information function: It is a mathematical correlation that expresses a set of information functions for the test items. (Moghadamzadeh et al,2011)

Person's ability: A value assessed that maximizes the likelihood of the person's responses to the test items. (Hambleton, Swaminthan, 1985).

Theoretical framework and previous studies:

The utilization of Item Response Theory in investigating educational and psychological test information is an answer toavoid the majority of the inadequacies of the Classical Test Theory (CTT) that overwhelmed the development of tests in the 20th century.(Al-zboon et al,2021).CTT has been pillar of educational and psychological test development for must of 20th century. However, since lord and Novick's (1968) classic book introduced model-based measurement, revolution has occurred in test theory. (Borgata et al,2015, Yalçın,2018)

Item Response Theory(IRT) shows the relationships between the ability or trait (ɵ) estimated by the tool and an item response. The item may be dichotomous when we manage two classifications, or it may be polytomous for more than two classification.

Since traits are not directly measurable, they are alluded to as latent abilities An item response model indicates a relation between the observable person performance and the unobservable abilities expected to underlie performance person. nside the wide framework of item response theory, numerous models can be operationalized due to the large number of choices available for the mathematical form of the item characteristic curves(ICC). But, while item response theory can't be show to be correct or incorrect, the appropriateness of specific models with any group of test data can be established by conducting a suitable goodness of fit checking (Al-zboon et al,2021). the relationship between unobservable and the observable quantities is described by a mathematical function. For this reason, Item Response Theory models are mathematical models, which depend on specific assumptions that must be achieved in the test data in order to obtain correct results.(Brzezińska,2020 ; Borgata et al,2015)

There are three assumptions of Item Response Theory (IRT),The first assumption is the unidimensionality, which means that there is one trait that explains the person's performance on the test, (i.e. that the person's score on the test reflects the trait that the test measures only). The second assumption the local independence means the responses of the respondents to the test should be statistically independent at a certain ability(ɵ) level. In other words, that the person's response to one item does not negatively or positively affect his response to the other items. (Crocker & Algina, 1986). The third assumption of IRT is called the Item Characteristics Curve (ICC). The concept of the curve of Item Response Theory is a mathematical association that relates the probability of the success of the person on the item with the ability measured by a set of items composing the test.(Brzezińska,2020; Al-zboon et al,2021)

The item parameters may include four distinct types of parameters: a difficulty parameter(b_i), discrimination parameter (ai), guessing parameter(ci), and carelessness (di). Depending on the number of parameters included in the model equation, there are several of IRT mathematical models distinguished.(Brzezińska,2020; Ul Hassan& Miller,2020)

One Parameter (1PLM), also known as the Rasch Model, is the simplest IRT model and one of the broadest models used in the IRT, which assumes that all items do not differ from each other except with the difficulty parameter of the item (b_i) It also assumes that the discrimination parameter(ai) is equal for all items, while the guessing parameter(ci) for items approaches zero, and the model takes the following mathematical formula to express the probability of the correct answer to the item: .(Brzezińska,2020 ; Borgata et al,2015)

$$P_{i\,(\theta)} = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}} \quad , (i = 1 \dots\dots\dots\dots\dots\dots\dots n)$$

Another IRT model is two-parameter (2PL) also known as the Birnbaum model, assumes that the items differ in the discrimination and difficulty parameters, while the

guessing parameter approaches zero, and the probability of a correct answer to the item is given by the following mathematical formula:

$$p_i(\theta) = \frac{e^{Dai(\theta - b_i)}}{1 + e^{Dai(\theta - b_i)}}$$

Another 3-parameter (3PL) Birnbaum IRT model , The 3-Parameter Model can be obtained from the 2- parameter model by adding a guessing parameter(ci), the mathematical formula of the 3- Parameter model is written: (Borgata et al,2015)

$$P_i(\theta) = C_i + (-C_i)\frac{e^{D_{ai}(\theta - b_i)}}{1 + e^{D_{ai}(\theta - b_i)}}, i = 1, 2, \ldots\ldots, n$$

Test Information Function:

The Test information function (TIF) is viewed one of the fundamental ideas that play an important role in the IRT. The TIF determines the amount of information given by the test or the item as a whole when estimating the ability of persons or respondents, and through which the standard error in the estimation could be determined. (Samajima, 1994 ; Moghadamzadeh et al, 2011)

Since it is all the test items in a test that are utilized to estimate a testee's ability, the information value obtained through the test can be determined for each ability level. A test is a group of items; hence, test information function can, for given ability level, be obtained from the sum of the information function of all the items. Can be expressed the value of the test information function is calculated at each ability level (Ɵ) through the following (Moghadamzadeh et al, 2011):

$$I(\theta) = \sum_{i=1}^{N} I_i(\theta)$$

$I(\theta)$: Value information test

$I_i(\theta)$: Value information item

The Information function of the entire test will a lot more prominent than that of the individual items. Consequently, a total test measures person's ability more accurately and more comprehensively than each individual item. The important point is that the very presence of more items in a test increases the information function of the test. Therefore, we can generally contend that longer tests measure a testee's ability more accurately than shorter ones do.(Brzezińska,2020).

The utilization of tests for other explicit purposes requires the utilization of different types of information function. Although the fact that it is conceivable to get the information function of each item, this is rarely done. The information value gained form each item is insignificant, so the information value gained from the entire test is of prime significance. That is the reason the information value at one ability level and its information function are going. Obviously, since a test's information acquired through the amount of all he items' information value at a given level, the item information value

should initially be characterized. It is to be noticed that the numerical meaning of the item information value relies upon the particular curve utilized.(Samajima,1994; Moghadamzadeh et al,2011,).

Previous Studies

Many studies were conducted to explore the effect of item formats on the characteristics of the psychometric test. the study of Abdel-Aal (2019) indicated that there were statistically significant differences in the estimation of the discrimination of items due to the form of the items and in favor of the items of true and false.

Thawabieh (2016) conducted a study aimed at investigating the effect of item formats on examinees' performance and item psychometric characteristics, The students of evaluation and measurement 68 students were chosen to collect the data, 2 test format were applied; Multiple Choice items test and Completion items test. The results indicated that students performance was effected by the type of the test items, female students' performed higher than male students' on the 2 forms, and the psychometric characteristics of the test were found to be affected by test format.

The study of Culligan (2015) aimed to compare three common vocabulary test formats, the Yes/ No test, the vocabulary knowledge scale, and the vocabulary level test, as measures of vocabulary difficulty. The three tests were given to 165 Japanese students, the results indicated that the three tests measured one major latent trait (unidimensional) and they were significantly correlated in estimating their item difficulty.

The study of Simbak, Aung, Ismail, Joush, Ali, Yaseein, Haque, and Rebuan (2014) aimed to compare between examinees' performance on two evaluation techniques: multiple true-false and single best answer test formats, and correlated them with other assessment outcomes. The study analyzed the data for 20 item formats for each type of the questions, the participants were 3rd year medicine students at Sultan Zainal Abdin University in Malaysia. The results indicated that students got higher marks in single best answer than multiple true false quiz. Single best answer test results were found to be well correlated with clinical marks.

Hudson (2012), investigated the effectiveness of two forms of questions: Multiple Choice and short answer questions used in the State University Entrance Examination for Chemistry and the effect of gender on performance, data was collected from 192 year 11 students from 4 secondary colleges. The researcher constructed short test that asked similar questions but in both types: multiple choice and short answer form. The results indicated that male students achieved higher scores than female students with respect to mean scores on both tests and subtests.

The results of ALmasri (2009) showed that there were no statistically significant differences in estimating the abilities of persons due to the form of the test items, as well as there were statistically significant differences in estimating the difficulty of the items due to the form of the items, and in favor of the complete items. Similarly.

In another study by Al-Kahlout (2002), the results showed that the reliability factor of the complete test calculated by the Cronbach Alpha and the repetition methods is greater than that of the multiple-choice test. The results also indicated that the average difficulty

factors for a multiple-choice test are greater than those of the complete test, and that the average of the discrimination factors for the complete test is greater than those of the multiple-choice test.

The results of the study of Bock (1993) showed that multiple-choice test items had a high reliability factor, nevertheless, they indicated that one item of the free response type is equivalent to four items of the multiple-choice type in terms of the information function provided by the test.

METHOD

The sample of the study

The sample of the study consisted of all students of the preparatory year who study the mathematics course (1) at the University of Tabuk, Ummluj Branch, and the number of the study sample was (350) students distributed over seven sections. A section was chosen randomly from among the seven sections, and the number of students enrolled in it was (50) who formed the pilot study sample. The remaining (300) students formed the experimental study sample.

The tool of the study

To achieve the purpose of the study, two tests were prepared in the mathematics course (1), one of which was a multiple-choice test, for each item (4) options and the other of the complete type, each of which consists of (34) items. The complete test was obtained by deleting the options for the multiple-choice test items. Then the two tests (multiple-choice and complete) had the same items. The two tests were presented to a group of teachers specialized in teaching the course and they were asked to express their opinion on the items in terms of their linguistic integrity and accuracy in measuring the goals for which they were set and the effectiveness of the distracters. There were no substantive suggestions except in some formatting which were addressed.

The study tools were applied according to repeated measures design, and this design is stronger than the random group design that was adopted by most of the previous studies (Ferguson & Takane, 1989). After the study sample sat for the test, question papers were distributed so that one student would start with the multiple-choice test questions and the other next to him starts with the complete questions and after completing the answers, students who started with multiple-choice questions were given complete questions, and students who started with the complete questions were given multiple-choice questions.

Statistical Processing:

1- The (BILOG-MG3) program was used to find the items parameters and information function of the items of the three test forms according to the three-parameter model in the item response theory.

2- Arithmetic means, and standard deviations, Sample paired t-test, factor analysis

Statistical analysis

Before starting to process the data statistically, it was necessary to examine the basic assumptions that must be present in the data when applying the item response theory, the most important of which were:

1- **Unidimensionality**

Unidimensionality was verified for its effect on the accuracy of the estimates by using factor analysis for each of the two test forms using the Principle Component Analysis method. the unidimensionality depends on that the ratio of the eigen value of the first factor to the eigen value of the second factor is a great percentage and is not less than (2), and the explained variance by the first factor is more than 20% approximately (Hambleton & Swaminathan, 1985). Table (1) shows the eigen values and Explained Variance ratios for the first and second factors and the result of dividing the first root by the second factor in the two test forms.

Table 1
Eigen values and explained variance ratios for the first and second factors

Test		Factor		λ1 / λ2
		First	second	
Multiple-choice	Eigen Values	11.713	2.341	5.003
	Explained Variance	20.412%	4.029%	
Complete	Eigen Values	10.750	2.360	4.55
	Explained Variance	20.186%	3.826%	

Note. λ1 / λ2 = ratio of the first to the second Eigen value

It is noticed from table (1) that the eigen value of the first factor in the multiple-choice test is (11.713), with a percentage of (20.412%) of the total variance of the test. While the eigen value of the second factor was (2.341), with a percentage of (4.029 %) of the total variance of the test, and it is also noticed that the eigen value of the first factor in the complete test is (10.750) with a percentage of (20.186%) of the total variance of the test. Further, the eigen value of the second factor was (2.360) with a percentage of (3.826). %) of the test's total variance.

It is also noticed that the variance ratio explained by the first factor of the two test forms is high and this is an indication of unidimensionality. Reckase (1997) suggested that if more than (20%) of the variance was explained by the first factor and the ratio of the eigen root of the first factor to that of the second factor was large, that was an indication of unidimensionality. The result of dividing the eigen value of the first factor by the that of the second factor in the two test forms was greater than (2) which means that the test measures one dimension.

2- Goodness of fit

In order to verify the goodness of fit of the two-parameter logistic model, the (BILOG-MG3) program was used and the goodness of fit process was checked by using the statistic chi-square. The results showed that three items of the multiple-choice test were poorly matched to the model at the level of significance ($\alpha = 0.05$) where the correlation coefficient was weak between the item and the test (less than 0.3), and the non-

conforming items in the multiple-choice test were (4, 16, 27). In the complete test, there were two items that did not match the model, which were (16, 22), the same items on the multiple-choice test. These items were deleted, and the items matching to the model and common between the two models were retained and re-analysis after deleting the items was carried out. Thus, (31) items for each of the two forms of the test have undergone appropriate statistical analysis.

FINDINGS AND DISCUSSION

In order to answer the first question related to the accuracy of the estimates of the parameters of the items (difficulty(b), discrimination(a)) according to the shape of the items (multiple-choice, complete) and to study the statistical differences between them, the BILOG-MG3 program was used to find the values of the items' parameters and the standard error in their estimation, as shown in the table (2).

Table 2
The values of the items' parameters (b, a) and the standard error in their estimation for each form of the test

item	multiple choice test				complete test			
	b	S.E	A	S.E	b	S.E	A	S.E
1	-0.202	0.15	0.502	0.101	-0.334	0.118	0.689	0.089
2	-1.16	0.371	0.26	0.114	-0.318	0.109	0.798	0.061
3	0.089	0.374	0.181	0.131	1.856	0.255	0.747	0.046
4	-1.878	0.804	0.111	0.081	1.227	0.267	0.419	0.03
5	2.15	0.345	0.755	0.293	3.065	0.688	0.841	0.177
6	0.997	0.66	0.112	0.044	-0.044	0.374	0.18	0.03
7	0.247	0.103	0.852	0.122	0.186	0.098	0.84	0.139
8	0.051	0.214	0.329	0.085	0.02	0.138	0.538	0.069
9	-0.896	0.277	0.325	0.083	-0.526	0.168	0.485	0.071
10	0.418	0.15	0.538	0.152	0.372	0.088	1.103	0.09
11	-0.246	0.115	0.714	0.108	-0.319	0.106	0.797	0.121
12	1.05	0.359	0.254	0.09	1.188	0.219	0.517	0.058
13	-0.278	0.156	0.494	0.109	-0.253	0.111	0.718	0.086
14	-0.397	0.294	0.248	0.091	-0.507	0.151	0.559	0.059
15	-1.226	0.494	0.179	0.056	0.067	0.277	0.247	0.046
16	-1.845	0.313	0.585	0.187	-1.399	0.153	1.15	0.116
17	0.66	0.644	0.108	0.045	-0.043	0.365	0.185	0.029
18	0.775	0.116	1.02	0.248	0.858	0.093	1.463	0.177
19	0.265	0.188	0.393	0.09	0.608	0.161	0.548	0.076
20	-0.73	0.156	0.595	0.11	-0.671	0.13	0.742	0.101
21	2.437	0.471	0.561	0.229	3.405	0.808	0.717	0.133
22	-1.374	0.597	0.146	0.06	0.49	0.272	0.271	0.038
23	-0.301	0.641	0.105	0.037	0.164	0.46	0.146	0.028
24	2.16	0.313	1.003	0.433	2.052	0.248	1.376	0.243
25	-0.25	0.091	1.046	0.379	-0.147	0.058	2.318	0.176
26	-0.448	0.642	0.106	0.05	0.488	0.334	0.217	0.028
27	-3.59	1.169	0.117	0.052	0.431	0.317	0.227	0.032
28	-1.101	0.402	0.224	0.063	-1.693	0.421	0.294	0.054
29	-0.519	0.134	0.664	0.12	-0.655	0.118	0.866	0.114
30	-2.422	0.497	0.458	0.137	-1.992	0.265	0.771	0.106
31	0.289	0.353	0.196	0.072	0.476	0.189	0.407	0.049
Mean	-0.235	0.374	0.425	0.128	0.260	0.244	0.683	0.086

It is evident from table (2) that the complete test rendered high difficulty and discrimination coefficients compared to the multiple-choice test, where the means of the items' parameters (difficulty and discrimination) of the complete test were (0.260, 0.683), respectively and those of the multiple-choice test were (- 0.235, 0.425). Table (2) also shows a marked discrepancy in the items' parameters (difficulty, discrimination) between the two forms of the test, where the range in the difficulty of the items was the highest in the multiple-choice test and their values ranged from (-3.59 to 2.437), and the range in the difficulty of the items in the complete test ranged from (1.992 to 3.405). The range in discrimination items was the highest on the test of its kind.

Further, the values of the complete test ranged from (0.146 to 2.318) and those of the multiple-choice test ranged from (0.105 to 1.046). Figure (1) illustrates this variation in the difficulty of discrimination items.

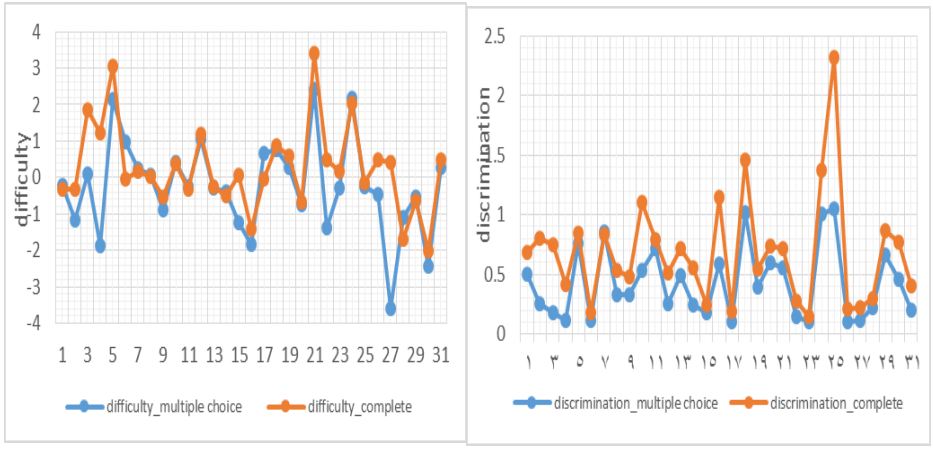


Figure 1
An illustrative diagram for evaluating the parameters of the items (difficulty and discrimination) in the two test forms

It also shown in Table (2) that the complete test produced the lowest standard errors for the estimates of items' parameters compared to the multiple-choice test, where the means of the standard errors of the estimates of items' parameters (difficulty and discrimination) of the complete test were (0.244, 0.086), and those of the multiple-choice test were (0.374, 0.128). Figure (2) shows an illustrative diagram of the standard error values for estimates of the items' parameters in the two test forms..

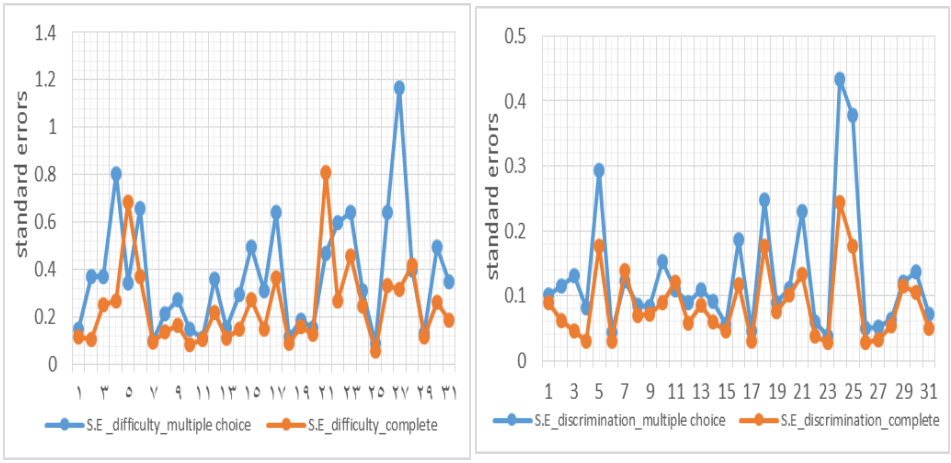


Figure 2
An illustrative diagram of the standard error values for estimates of the items' parameters in the two test forms

To study the differences in the accuracy of the estimation of the parameters of the items according to the form of the test items, the Sample paired t-test was used to reveal the differences between the means of the standard errors of the estimates of the parameters of the items (difficulty and discrimination). Table (3) shows the results of this test.

Table 3
Sample paired t-test results for the effect of the form of the item on estimating the parameters of the items

Parameters	Test	Mean	Std. Deviation	df	T	Sig
Difficulty	Multiple-choice	.374	.247	30	3.328	.002
	Complete	.244	.172			
Discrimination	Multiple-choice	.128	.096	30	3.536	.000
	Complete	.086	.054			

The results presented in table (3) indicate that there are statistically significant differences at the level of significance (α=0.05) between the means of the standard errors of the estimates of the parameters of the items (difficulty and discrimination) in favor of the complete test, where the mean of the standard errors of estimates of the difficulty parameters of the items in the multiple-choice test was (.374) and that of the complete test was (.244). This means that the complete test gives a better and more accurate estimate of the item difficulty factor compared to the multiple-choice test. This could be attributed to the fact that the complete test items, despite their similarity in the text with the multiple-choice test items, require higher abilities to be answered correctly, as they require the student to use his thinking and formulate the answer in his own way and thus leave no room for the student to guess the correct answer as in the multiple-choice test items in which the student chooses to answer from a group of options.

Moreover, the mean of the standard errors of estimates of the discrimination parameters of the items in the multiple-choice test was (.128) and that of the complete test was (.086). This means that the complete test gives a better and more accurate estimate of the factor of discrimination compared to the multiple-choice test. This could be due to the fact that the multiple-choice tests include the random guessing factor as well as they allow intelligent guessing, and this could lead to parity in performance between high-performing and low-performing students besides reducing the variance in students' performance on the item, and thus reduce its discrimination factor. While the complete test reflects the actual realistic variance of students' performance on the item and this leads to high discrimination factors for the items compared to those of the multiple-choice test. This result is consistent with the results of (Anderson & Hyers, 1991; ALmasri, 2009; and Al-Kahlout, 2002).

In order to answer the second question about the accuracy of the ability estimates of persons according to the form of the item (multiple-choice, complete), the abilities of persons were estimated(ɵ) and the standard error in estimating them in each form of the test, and figure (3) shows the distribution of the abilities of persons according to the form of the item (multiple-choice, complete).

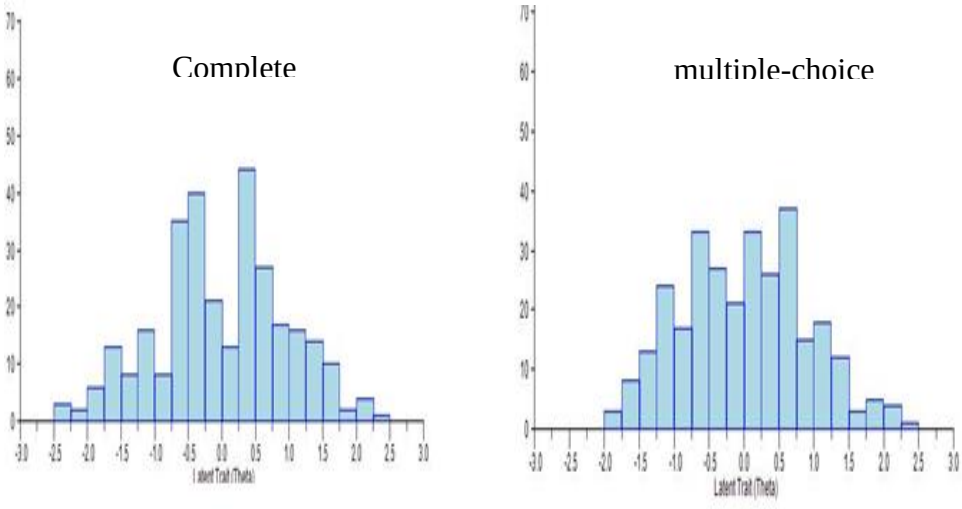


Figure 3
The distribution of the abilities of persons according to the form of the item (multiple-choice, complete)

To find the significance of the differences between the means of the standard error in estimating the ability parameter of persons according to the test form, the statistic (t) test was used to compare between the means, and table (4) shows the results of the statistic (t) test comparing between the values of the standard error of the estimation attributed to the form of the test.

Table 4
The results of the statistic (t)test comparing between the values of the standard error of the estimation attributed to the form of the test

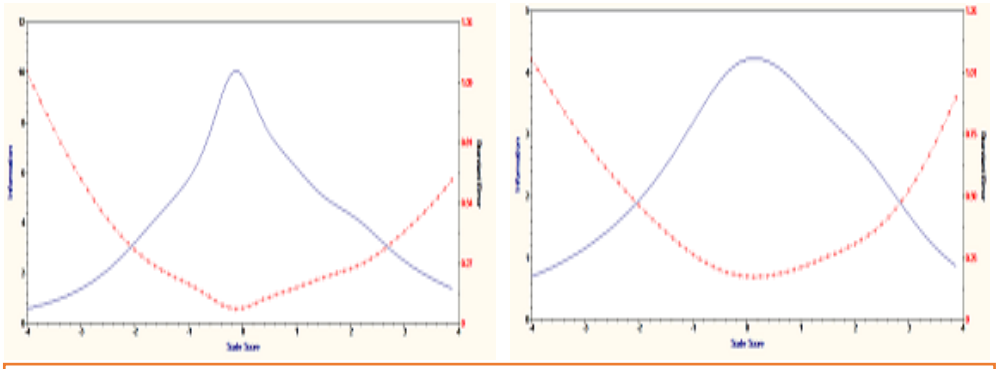
Test	Mean	Std. Deviation	Df	T	Sig
Multiple-choice	.5212	.0356	299	27.979	.000
Complete	.3743	.0910			

It is evident from table (4) that the value of the difference between the mean value of the standard error of estimation for the complete test and the mean value of the standard error of estimation for the multiple-choice test was statistically significant at the level of significance (α=0.05). This means that the complete test gives a better and more accurate estimate of the abilities of persons compared to the multiple-choice test.

As for the third question concerning the effect of the items form on the information function and the standard error of the items and the scale as a whole, where the Test Information Function (TIF) is considered one of the indicators from which the reliability parameter of the scale is inferred in the item response theory (Moghadamzadeh et al,2011). The test information function curve works against the standard error curve for the measurement, therefore, increasing the amount of information leads to a decrease in the standard error of the measurement. Reeve (2004) showed that the information function of the test correlates with the reliability of the scale through the following relationship $r = 1 - \frac{1}{\sum_{i=1}^{I} I(\theta)}$

It is expected that the more the information function of the test at a certain level of ability increases the reliability in the sense that it reduces the standard error of the measurement, which provides an opportunity to estimate the standard error at each level of ability and to identify the extent of the contribution of each item to determine the accuracy of the measurement (Hambleton, 1994).

To study the effect of the item shape on the test information function, the information functions curves of the test and the standard error of measurement were used for the two test forms (multiple-choice, complete) using the BILOGMG-MG3 program as shown in figure (4).



complete Test multiple-choice

Figure 4
Information functions curves of the test and the standard error of measurement

To facilitate the comparison between the two curves of the information function, these curves were placed in one figure, as shown in figure (5).

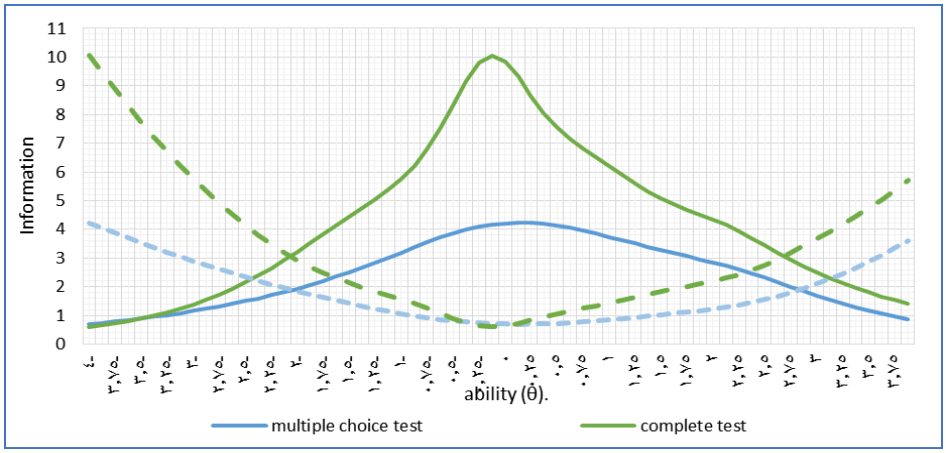


Figure 5
The comparison between the two curves of the information function

It is evident from figure (5) that the largest amount of the information function is for the complete test, where the maximum value of the amount of information provided by the multiple-choice test was approximately (4), and the maximum value of the amount of information provided by the complete test was approximately (10). This could be due to the discrimination parameter for the items, where the discrimination values for the items in the complete test were the highest. The higher the discrimination parameter for the item is, the more information the item contributes about the difficulty parameter. To test

the significance of the differences between the means of the information function, the statistic (t) test was used, and table (5) illustrates that the results of the statistic (t) test to compare the values of the means of the information function that is attributed to the form of the item.

Table 5
The statistic (t)test value comparing between the means of the information function according to the form of the item

Test	Mean	Std. Deviation	df	t	Sig
Multiple choice	2.450	1.162	63	-9.150	.000
Complete	4.243	2.665			

The results presented in table (5) indicate that there are statistically significant differences at the level of significance ($\alpha = 0.05$). Among the information function means in favor of the complete test. This result is due to the measurement error resulting from guessing or correct random answers in the multiple-choice tests, and this error affects the standard error of the ability estimation and this would reduce the value of the information function in the multiple-choice tests. While in the complete tests that do not suffer from the problem of guessing this makes it less susceptible to measurement error, so its mean standard errors are less, and this in turn leads to increasing the information function of the test, which is considered an indicator of the accuracy of estimating the abilities of persons, and this is what the result of the second question shows.

For the correlation of the test information function as a whole with the item information functions, and to determine the effect of the difference in the form of the item on the information function of the item, the maximum values of the item information function for each form of the test were summarized as shown in table (6).

Table 6
The maximum values of the item information function for each form of the test

Item	multiple choice	complete	Item	multiple choice	Complete
1	0.1819	0.3431	17	0.0084	0.0247
2	0.0489	0.4606	18	0.7514	1.546
3	0.0237	0.4034	19	0.1119	0.2174
4	0.0088	0.127	20	0.2558	0.3981
5	0.4116	0.5105	21	0.2276	0.3711
6	0.009	0.0235	22	0.0155	0.0529
7	0.5247	0.5097	23	0.0079	0.0155
8	0.0781	0.2093	24	0.7267	1.3674
9	0.0765	0.1699	25	0.7901	3.8818
10	0.2092	0.8789	26	0.0081	0.034
11	0.3679	0.4594	27	0.0099	0.0373
12	0.0466	0.193	28	0.0364	0.0625
13	0.1764	0.3723	29	0.3187	0.5422
14	0.0444	0.2257	30	0.1515	0.429
15	0.0231	0.0441	31	0.0278	0.1198
16	0.2476	0.9547			

It is evident from the results of table (6) that the maximum values of the information function for most of the complete test items were higher than the maximum values of the item information function for the multiple-choice test, and figure (6) illustrates this.

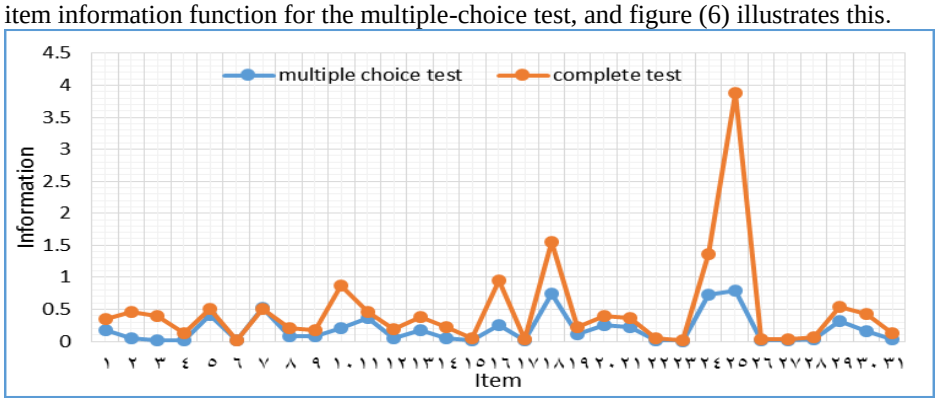


Figure 6
Comparison between the maximum vales of the information function of the two test forms

By tracking the maximum values of the information function for the items of the multiple-choice test, it is noticed that the values are low compared to the complete test, as the values of the information function for the item ranged between (.0079 to .7901). The reason for this is that the multiple-choice test produced a wide range of the difficulty parameter as compared to the complete test.

CONCLUSION

It is evident from the findings of the present study that there is an effect of the item form on the parameters of persons, items, and the information function of the test. As the parameters of the items varied between the two test forms, this signifies the variability of the parameters of the items according to the form of the item. Furthermore, because the results of the study are affected by a group of factors such as: the length of the test, the sample size, and the method of estimation, there is an urgent need to conduct similar studies dealing with the effect of the difference in the form of the item on the parameters of the items and the function of information in relation to these factors. Therefore;

RECOMMENDATIONS

The study recommends educating teachers to diversify the forms of items of the tests they use so that they contain the two forms (multiple-choice and complete), and using complete items on a larger scale than it is now since they have the stronger discrimination effect factor.

LIMITATIONS

One of the limitations of this study is the Two different test formats which are Completion items, and multiple choice tests measured the same behaviour. In addition, the sample size can be limitations too. Especially IRT can be affected by the sample size

when the item statistics are the topic, On the other hand this research is limited to difficulty and discrimination items parameters.

REFERENCES

Abdel- Aal, Sabri. (2019). The Effect of Item Format The Multiple choice And True-false On Psychometrics Properties According To Item Response Theory (IRT) of The Computer Test for The First Secondary Grade in The City of Tabuk. International *Journal of Educational Psychological Studies, 6*(1),1-17, https://doi.org/DOI:10.31559/EPS2019.6.1.1

ALkahlout, Ahmed (2002). Comparison of psychometric properties for both multiple-choice and complementary tests. *Journal of the Educational Research Center, 22*(2),127-153.

AlKhatib, H. S., Brazeau, G., Akour, A., & Almuhaissen, S. A. (2020). Evaluation of the effect of items' format and type on psychometric properties of sixth year pharmacy students clinical clerkship assessment items. *BMC Medical Education, 20*(1), 1-8. https://doi.org /10.21203/rs.2.17768/v1

Allen. M. & Yen. W. (1979). *Introduction To measurement theory*. California. Education. Brooks Cole Publishing Company Monterey.

ALmasri, Ahmed. (2009). *The effect of test item format in estimating examinee's and Item Parameters According to Item Response Theory*. Ph.D. dissertation, yarmouk university.

Al-zboon, H. S., & Alkursheh, T. O. (2021). The Effect of the Percentage of Missing Data on Estimating the Standard Error of the Items' Parameters and the Test Information Function According to the Three-Parameter Logistic Model in the Item Response Theory. *Elementary Education Online, 20*(1), 887-898. https://doi.org/10.17051/ilkonline.2021.01.82

Anastasi. A (1982). *Psychological testing* (6th ed.). New York: Macmillan.

Anderson, P. S., & Hyers, A. D. (1991). Quantitative Comparisons of Difficulty, Discrimination and Reliability of Machine-Scored Completion Items and Tests (in the MDT Un-Cued Answer-Bank Format) in Contrast with Statistics from Comparable *Multiple Choice Questions: The First Round of Results.* ERIC (ED 349319).

Audeh, Ahmed. (2010). *Measurement and evaluation in the teaching process*, Irbid, Jordan: Dar Al-Amal.

Bock, R. D. (1993). *Report on Models for Educational Assessment Involving Multiple-Choice and Free Response Exercises*. Open-Ended Exercises in Secondary School Science Assessment. Project 2.4: Quantitative Models To Monitor the Status and Progress of Learning and Performance and Their Antecedents.

Borgata, A. F., Azevedo, C., Pinheiro, A., & Andrade, D. (2015). Comparison of ability estimation methods using IRT for tests with different degrees of difficulty.

Communications in Statistics-Simulation and Computation, 44(2), 474-488. https://doi.org/10.1080/03610918.2013.781630.

Brown, Frederick G.(1983). *Principles of educational and psychological testing* (3ed) ed) , New York: Holt, Rinehart and Winston.

Brzezińska, J. (2020). Item response theory models in the measurement theory. *Communications in Statistics-Simulation and Computation, 49*(12), 3299-3313. https://doi.org/10.1080/03610918.2018.1546399.

Culligan, B. (2015). A comparison of three test formats to assess word difficulty. *Language Testing, 32*(4), 503-520.

Currie, M., & Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing, 27*(4), 471-491.

Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of Students' AbilityEstimation on Combinations of Item Response Theory Models. *International Journal of Instruction, 13*(4), 545-558. https://doi.org/10.29333/iji.2020.13434a.

Ferguson. G. & Takane. Y. (1989). *Statistical analysis in psychology and education* (6th ed.). New York: McGraw-Hill.

Frisbie, D. A. (1974). The effect of item format on reliability and validity: A study of multiple choice and true-false achievement tests. *Educational and Psychological Measurement, 34*(4), 885-892

Gronlund, N., and Linn, R. (1990). *Measurement and Evaluation in Teaching*. New York: Macmillan publishing Co., Inc.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education, 15*(3), 309-333. https://doi.org /10.1207/S15324818AME1503_5

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff Publishing.

Hambleton, R., K.(1994). Item Response Theory: Abroad psychometric frame work for measurement advances. *Psicothema, 6*, 535-556.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Moghadamzadeh, A., Salehi, K., & Khodaie, E. (2011). A comparison the information functions of the item and test on one, two and three parametric model of the item response theory (IRT). *Procedia-Social and Behavioral Sciences, 29*, 1359-1367

Murad, Ahmed& Suleiman,A. (2002). *Testing and measurements in psychological and educational sciences, steps to prepare and characteristics.* Cairo, Egypt: Modern Book

O'Neill, T. R., Gregg, J. L., & Peabody, M. R. (2020). Effect of sample size on common item equating using the dichotomous Rasch model. *Applied Measurement in Education, 33(*1), 10-23. https://doi.org/10.1080/08957347.2019.1674309.

Primi, R., De Fruyt, F., Santos, D., Antonoplis, S., & John, O. P. (2020). True or false? Keying direction and acquiescence influence the validity of socio-emotional skills items in predicting high school achievement. *International Journal of Testing, 20*(2), 97-121. https://doi.org/10.1080/15305058.2019.1673398 .

Reckase, M.D.(1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 1*, 25-36.

Samajima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18*(3), 229-244.

Shin, J., Bulut, O., & Gierl, M. J. (2020). The Effect of the Most-Attractive-Distractor Location on Multiple-Choice Item Difficulty. *The Journal of Experimental Education, 88*(4), 643-659. https://doi.org/10.1080/00220973.2019.1629577.

Simbak, N. B., Aung, M. M. T., Ismail, S. B., Jusoh, N. B. M., Ali, T. I., Yassin, W. A. K., ... & Rebuan, H. M. A. (2014). Comparative Study of Different Formats of MCQs: Multiple True-False and Single Best Answer Test Formats, in a New Medical School of Malaysia. *International Medical Journal, 21*(6), 562-566.

Subali, B., Kumaidi. & Aminah, N. S. (2021). The Comparison of Item Test Characteristics Viewed from Classic and Modern Test Theory. *International Journal of Instruction, 14*(1), 647-660. https://doi.org/10.29333/iji.2021.14139a

Thawabieh, A. M. (2016). A comparison between two test item formats: Multiple-choice items and completion items. *British Journal of Education, 4*(8), 63-74.

Ul Hassan, M., & Miller, F. (2020). Discrimination with unidimensional and multidimensional item response theory models for educational data. *Communications in Statistics-Simulation and Computation*, 1-21. DOI: 10.1080/03610918.2019.1705344.

Wang, C., Su, S., & Weiss, D. J. (2018). Robustness of parameter estimation to assumptions of normality in the multidimensional graded response model. *Multivariate behavioral research, 53*(3), 403-418. https://doi.org/10.1080/00273171.2018.1455572.

Yalcin, S. (2018). Data fit comparison of mixture item response theory models and traditional models. *International Journal of Assessment Tools in Education, 5*(2), 301-313. DOI: 10.21449/ijate.402806