Relationship between Test Item Arrangements and Testee's Performance and Test Usefulness Criteria

Gholam-Reza Abbasian[1,*] & Hanieh Zadsar[2]

[1]Imam Ali University, Tehran, Iran

[2]Islamic Azad University, South Tehran Branch, Iran

*Corresponding author: Imam Ali University, Tehran, Iran. E-mail: gabbasian@gmail.com

Received: June 24, 2019 Accepted: July 20, 2019 Online Published: July 25, 2019

doi:10.5430/ijelt.v6n2p52 URL: https://doi.org/10.5430/ijelt.v6n2p52

Abstract

Test method, test content and test type are supposed to affect test taker's performance and have close connections with test characteristics. However, these issues have not been subject to much in-depth investigations. To shed some lights on the issues like test item arrangements in relation to test taker's performance, test usefulness criteria including *test Validity, Reliability, Impact, Interactiveness, Authenticity*, and *Practicality,* and also test characteristics such as Item Facility (IF), Item Discrimination (ID), and Choice Distribution (CD), this ex post facto design study was conducted with a group of Iranian EFL learners. For this purpose, some university randomly selected students majoring in English Language Teaching, English Translation Studies, and English Literature and learners from different language institutes received a version of the Nelson proficiency test. Then, two different versions of a same researcher-made test in terms of item arrangements were administered to 116 students obtaining an acceptable score in the Nelson proficiency test. Simultaneously, a Test Usefulness Criteria Questionnaire developed and validated by Abbasian and Nassirian (2015) based on Bachman and Palmer's (1996) framework, was also attempted by the participants. Respective statistical analyses revealed contradictorily that test item arrangement did not have any significant effect on the performance of the test takers. However, test usefulness criteria, though with varying extent, proved to be subject to test item arrangement. In the same vein, IF and ID were also affected in the light of test method facet. The first contradictory finding leaves room open for further research, while the reaming findings offer insights to test developers, classroom teachers and all practitioners to pay due attention to test method facets in their educational assessment decisions as the learners' performance, nature of the measurement devices as well as the nature of the construct itself are all affected test method factors.

Keywords: test item arrangements, test methods, testee's performance

1. Introduction

The choice and sequencing of test items is supposed to have an influence on the level of group performance on a solitary test scores both for the group as a whole and for specific individuals inside the group in particular. However, few studies have demonstrated that diverse item arrangement influences the level of group performance on a solitary test. It has been widely observed that after a test, both teachers and students have stated that the test failed to measure the true level of their ability, and it is sometimes said that a student with a high level of knowledge, skill and ability, has obtained a score rather lower than his perceived expectation, and it has mutually been seen that some students with a level of knowledge and skills that are not at an acceptable level have been able to score better than others. Status such as these can be attributed to several factors, including the learners' mental and physical condition at the time of the test, to test usefulness characteristics and functioning of each individual item amongst many other parameters.

Among many test method factors contributing to or affecting test taker's performance one may refer to test item arrangement which has been, but scarcely, investigated. MacNichol (1956) found that, under almost immaculate power conditions, a hard-to-easy arrangement was significantly more troublesome than an easy-to-hard arrangement (cited in Weiss and Betz, 1973). It might also be assumed or hypothesized that test item arrangements may play a role in test

usefulness criteria, meaning that notions like reliability, validity, authenticity, interactiveness, impact and practicality (Bachman and palmer, 1996) maybe a function of the manner a test is structured and arranged, which altogether maybe related to the test taker's performance and input processing. Additionally, testee's performance can be subject to item characteristics including item facility, item discrimination and choice distribution (Farhady, Jafarpur, Birjandi, 1994). Obviously, a network of multiple variables can be visualized are supposed to interact. However, these three issues have not been much studied interactively and empirically.

2. Review of the Literature

Testing is an important part of each language teaching and language learning experience and well-made tests can help students in at least two ways. First, testing will encourage the students and will motivate them in learning the issue. Second, testing will enable the students get ready themselves and therefore learn the materials (Madsen, 1983). Amongst many others, test method factor can play a crucial role in this process. One of the determinants of test method is the manner through which (i.e., form of a test) test items are presented. According to Farhady, Jafarpur, Birjandi (1994, p.26), "the form of a test refers to its physical appearance".

Thawabieh (2016) suggests that test constructers select the best items to test students' achievement. Many other scholars (Allam, 2007), (Alzude and Alaan, 2005), (Kufahi, 2003), and (Thorndike and Hagen, 1986) have focused on certain guidelines for choosing the item format including: item format must be reasonable to the learning results, students' age, item difficulty, proper to the evaluation goals, substance and teachers' experience. How to measure is closely connected to what to measure. For example, Crocker and Algina (1986) described the test developer's task as requiring two main kinds of decisions: what to measure and also how to measure it. This is justified on the ground that test format affects performance on a test. As a proof, Thawabieh (2016) noted that Phipps and Brackbil (2009) assessed the relationship between assessment item formats (case– based and non-case– based) and item performance. They gathered 1575 items from examinations managed in a few therapeutics courses more than four scholastic years. The outcomes showed that non-case-based items exhibited a higher discrimination index than case-based items, while case-based items were lengthier, included increasingly definite data and not progressively difficult.

Similarly, McNeill (1956) concluded in his research that if items are arranged from easy to hard, the performance of testees would be better than the time to arrange items from hard to easy. Nonetheless, according to Plake (1980), no significant effects were found for knowledge of the orderings or for the interaction of knowledge of arrangement and order. Vander schee (2009) noted that Chidomere (1989) utilized a Principles of Marketing course to examine test item arrangement and student performance. He concluded from his examination, which included four multiple choice tests with forward and random-sequential forms, that there was no significant contrast in student performance dependent on test item arrangement. This supports past examinations by Sax and Cromack (1966) and Schmitt and Scheirer (1977).

Abbasian and Farhady (2000) conducted a study to explore the basic structure of language ability in connection to the levels of language proficiency and test method. The findings support effect test method factor on testees performance. In the same vein, according to Margaret and Victor (2017), item arrangement assumes an essential job in deciding the performance of students in examinations. It may be the case that while a few students may think that it's difficult in speculating answers in multi-choice Item for example, others then again may think that its less demanding while some may absolutely lean toward essay questions. In all these, the various item arrangement designs in which they are presented could also assume a fundamental job in deciding students' responses.

As Tei-Firstman (2008) noted, giving students straightforward question first will help continue their advantage and enthusiasm towards moving toward more questions. Then again, it could likewise be that test item arrangement dependent on sliding order of difficulty (i.e from complex to simple) may likewise have impacts on the performance of the students. Moreover, Margaret and Victor (2017) argued that students likewise may appear to recall things dependent on the order in which they are educated in the class. Consequently, it may be the case that test item arrangement dependent on order of point introduction in the class (The way and order in which the topics were acquainted with them in the class) may influences their performance.

As Ollennu (2011) argued, in the literature overview, it was found that specialists are not consistent in their discoveries with respect to regardless of whether changing item arrangement in a multiple-choice test would influence performance adversely. MacNicol (1956) researched the impacts of changing a "simple to-hard" arrangement to either hard-to-simple or a random arrangement. He discovered that the hard-to-simple arrangement was altogether more difficult than the first simple to-hard order while the random arrangement was not

fundamentally different. Ollennu (2011) noted that Anastasi (1976) contended that different arrangement of items will influence performance. This view is supported by Cacko (1993). Reaserchers in the Research Division, WAEC, Lagos (1993) found that various arrangement of items could influence performance adversely or positively relying upon the subject being referred to. Shepard (1997) declared that small changes in test arrangement can have a vast effect in student performance. Perceiving the significance of proper arrangement test items, Sax and Cromack (1966) and Ahuman and Clock (1971) have argued that tests ought to be constructed in a simple to-hard item difficulty arrangement.

Also Gerow (1980) and Allison (1984) found no difference in performance when items were arranged by a specific order of difficulty or arbitrarily. Soyemi (1980) likewise found no significant contrasts between simple to-hard and hard-to-simple arrangement, simple to-hard and arbitrary order; and hard-to-simple and arbitrary order. The way that there is no agreement among analysts from the literature survey shows that there is a problem and this given the motivation to the investigation.

Then, the objectives of the present study were twofold. First, it tried to probe any significant differentiation between test item arrangements and testee's performance on a test, and second, it explored any significant role of item arrangements and test usefulness criteria. To achieve these goals, the following research questions were formulated:

1. Do test item arrangements have any significant effect on testee's performance on a test?

2. Do test item arrangements have any significant effect on test usefulness criteria (including reliability, validity, authenticity, interactiveness, impact and practicality)?

3. Do test item arrangements have any significant effect on item characteristics (Item Facility, Item Discrimination and Choice distribution /effectiveness)?

3. Method

3.1 Participants

The people who participated in this study consisted of 116 students. They were both males and females between the ages of 18 and 30. All participants were university students in one of the three majors of English including English Teaching, English Translation, and English Literature and students of different language institutes in Iran.

3.2 Instrumentation

Questionnaire, Nelson-Test, Researcher-made Tests (A, B), Reliability of the Instruments:

1. Cronbach's Alpha Reliability Indices of Test Usefulness Criteria (Second Phase= Test A):

Table 1. Reliability Statistics of Test Usefulness Criteria (Second Phase)

	Cronbach's Alpha	N of Items
Reliability	.822	5
Validity	.864	9
Authenticity	.577	2
Interactiveness	.871	9
Impact	.895	10
Total	.955	30

Table 1 displays the Cronbach's alpha reliability indices for the Test Usefulness Criteria Questionnaire during the second phase of the study. The results showed that the questionnaire administered during the second phase of the study enjoyed a reliability of .955. The reliability indices for the five criteria were as follows; reliability (α = .822), validity (α = .864), authenticity (α = .577), interactiveness (α = .871) and impact (α = .895).

2. Cronbach's Alpha Reliability Indices of Test Usefulness Criteria (Third Phase= Test B): Table 2 shows the Cronbach's alpha reliability indices for the Test Usefulness Criteria Questionnaire during the third phase of the study. The results showed that the questionnaire administered during the third phase of the study enjoyed a reliability of .991. The reliability indices for the five criteria were as follows; reliability (α = .954), validity (α = .969), authenticity (α = .892), interactiveness (α = .973) and impact (α = .974).

Table 2. Reliability Statistics of Test Usefulness Criteria (Third Phase)

	Cronbach's Alpha	N of Items
Reliability	.954	5
Validity	.969	9
Authenticity	.892	2
Interactiveness	.973	9
Impact	.974	10
Total	.991	30

3. KR-21 Reliability of Three Tests: Table 3 displays the descriptive statistics and KR-21 reliability indices for the three tests (Test1: Nelson-Test and Test 2 & 3: Researcher-made test with two different item arrangements). The reliability indices were .93, .94 and .83.

Table 3. Descriptive Statistics and KR-21 Reliability of the Three Tests

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	KR-21
Test1	116	4	88	52.82	17.571	308.723	.93
Test2	109	0	74	17.01	14.548	211.639	.94
Test3	80	3	41	17.55	8.896	79.137	.83

3.3 Procedures

1) Having selected the sample, the researchers administered the researchers-made test in two different item arrangements form in four-week intervals. Then, they gave the Test Usefunless Questionnaire already validated Abbasian and Nassirian (2015) based on Bachman and Palmer's (1996) framework. Finally, they went on with the data analysis steps of each test administering processes. The study was run mainly based on Ex Post Facto design justified by Hatch and Farhady (1982), an ex post facto design is regularly utilized when the researcher does not have control over the selection and manipulation of the independent variable. The researcher in such a case looks at the degree of relationship between the two factors instead of at a cause-and-effect relationship.

4. Data Analysis and Results

In addition to the reliability estimation procedures, a paired-samples t-test was run to compare the means of the participants on the second and third tests. Additionally, a repeated measures ANOVA was run to investigate the second research question. However, in order to test the hypothesis, multivariate analysis in the form of repeated measures was run. Finally, as to the third research question, the data related to IF, ID and CD of the various test formats were analyzed descriptively just in terms of percent and mean scores comparisons.

4.1 Testing Normality of Data

The statistical techniques of repeated measures ANOVA and Pearson correlations were run to probe the research questions and test reliability issues raised in this study. These two statistical methods assume normality of the data which was probed by comparing the absolute values of the skewness and kurtosis indices against 1.96. If they were lower than 1.96, the normality of the data was inferred. It should be noted that the ratios of skewness and kurtosis over their standard errors were not computed due to the large sample size of this study as noted by Field (2013, p 229):

Table 4 displays the skewness and kurtosis indices for the test characteristics at the second and third phases of the study. The results show the absolute values of the skewness and kurtosis indices were lower than 1.96.

Table 4. Testing Normality of Test Characteristics at Second and Third Phases

	N	Skewness		Kurtosis	
	Statistic	Statistic	Std. Error	Statistic	Std. Error
Reliability2	108	-.667	.233	1.149	.461
Validity2	108	-.391	.233	.821	.461
Authenticity2	108	-.071	.233	-.641	.461
Interactivenss2	108	-.075	.233	.173	.461
Impact2	108	-.304	.233	.375	.461
Reliability3	74	-.605	.279	-.082	.552
Validity3	74	-.591	.279	.999	.552
Authenticity3	74	-.245	.279	-.204	.552
Interactiveness3	74	-.597	.279	1.130	.552
Impact3	74	-.226	.279	1.343	.552

Table displays the skewness and kurtosis indices for the test characteristics at the second and third phases of the study. The results show the absolute values of the skewness and kurtosis indices were lower than 1.96.

Table 5. Testing Normality of Tests at Three Phases

	N	Skewness		Kurtosis	
	Statistic	Statistic	Std. Error	Statistic	Std. Error
Test1	116	-.239	.225	-.392	.446
Test2	109	1.275	.231	1.115	.459
Test3	80	.998	.269	.606	.532

4.2 Testing the Research Hypotheses

4.2.1 Testing First Null-Hypothesis

The first null-hypothesis postulated that the arrangement of items did not have any significant effect on the examinees' performance on the three tests. A paired-samples t-test was run to compare the means of the participants on the second and third tests. As shown in Table 6, the participants had a higher mean on the second test (M = 19.31) than the third (M = 17.55) one.

Table 6. Descriptive Statistics of Two Tests

		Mean	N	Std. Deviation	Std. Error Mean
Tests	Test2	19.31	80	15.011	1.678
	Test3	17.55	80	8.896	.995

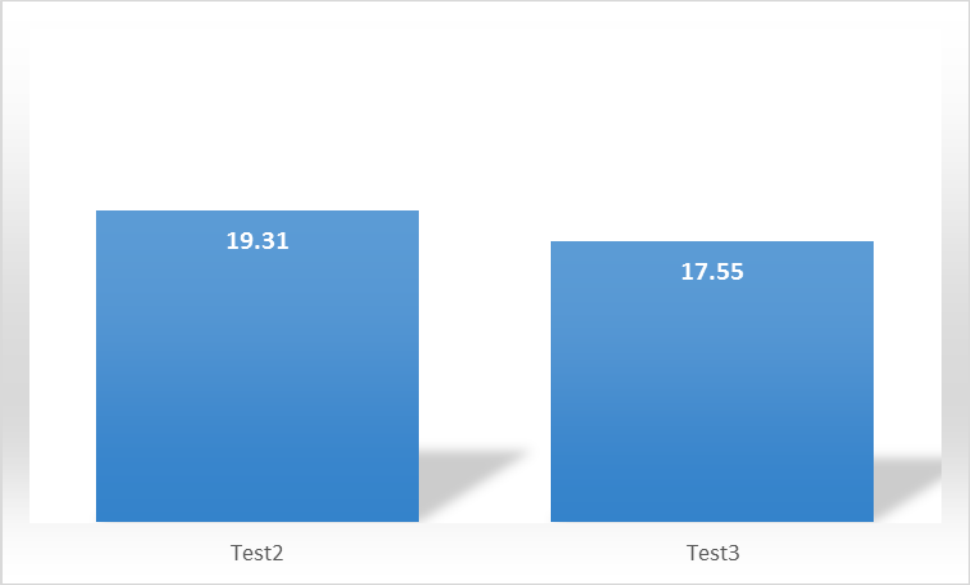

Figure 1. Means on Two Tests

The results of the paired-samples t-test (Table 7) (t (1, 79) = 1.18, p = .239, r = .132 representing a weak effect size) indicated that there was not any significant difference between the performance of the participants' on the two tests. Thus, the first null-hypothesis **was supported**.

Table 7. Paired-Samples t-test; Comparing Second and Third Tests

Paired Differences					t	df	Sig. (2-tailed)
Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
1.762	13.279	1.485	-1.193	4.718	1.187	79	.239

4.2.2 Testing the Second Null-Hypothesis

The second null-hypothesis postulated that the arrangement of items did not have any significant effect on the test usefulness criteria of reliability (i.e., validity, authenticity, interactiveness and impact) at second and third phases. Table 8 shows the description statistics. A repeated measures ANOVA was run to investigate the second research question. However, in order to test the hypothesis, multivariate analysis in the form of repeated measures was run.

Table 8 displays the descriptive statistics for the overall means of the five usefulness criteria. The results showed that authenticity (M = 2.082) had the highest mean. This was followed by the impact (M = 2.08), interactiveness (M = 2.07), validity (M = 1.95) and reliability (M = 1.94).

Table 8. Descriptive Statistics of Overall Usefulness Criteria

Criteria	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Reliability	1.948	.081	1.788	2.108
Validity	1.954	.077	1.802	2.106
Authenticity	2.082	.087	1.910	2.254
Interactiveness	2.073	.083	1.910	2.237
Impact	2.080	.082	1.916	2.243

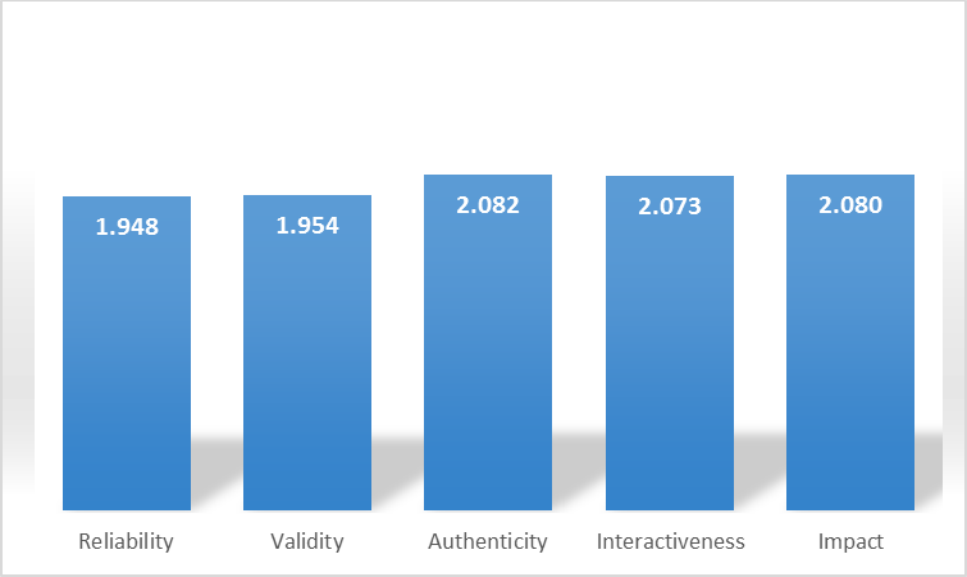

Figure 2. Means on Overall Usefulness Criteria

The present design, as displayed in Table 9 includes two within-subjects factors; tests and usefulness criteria. The repeated measures ANOVA produced three F-values for the effects of the tests, usefulness criteria and their interaction on the performance of the participants on the tests.

Table 9. Multivariate Tests; Usefulness Criteria and Tests

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Criteria	Pillai's Trace	.133	4.284	4	112	.003	.133
	Wilks' Lambda	.867	4.284	4	112	.003	.133
	Hotelling's Trace	.153	4.284	4	112	.003	.133
	Roy's Largest Root	.153	4.284	4	112	.003	.133
Test	Pillai's Trace	.272	42.926	1	115	.000	.272
	Wilks' Lambda	.728	42.926	1	115	.000	.272
	Hotelling's Trace	.373	42.926	1	115	.000	.272
	Roy's Largest Root	.373	42.926	1	115	.000	.272
Criteria * Test	Pillai's Trace	.037	1.063	4	112	.378	.037
	Wilks' Lambda	.963	1.063	4	112	.378	.037
	Hotelling's Trace	.038	1.063	4	112	.378	.037
	Roy's Largest Root	.038	1.063	4	112	.378	.037

Table 9 displays the main results of the analysis. The results (F (4, 112) = 4.28, p = .003, Partial η^2 = .133 representing an almost large effect size) indicated that there were significant differences between the overall means of the five usefulness criteria.

The results displayed in Table 9 (F (1, 115) = 42.92, p = .000, Partial η^2 = .272 representing a large effect size) indicated that the participants had a significantly higher mean on overall second test (M = 2.42) than third test (M = 1.62) (Table 10).

And finally; the results displayed in Table 9 (F (4, 112) = 1.063, p = .378, Partial η^2 = .037 representing a large effect size) indicated that there was not any significant interaction between two tests and usefulness criteria. Additionally, the second and third tests were also compared in term of the five useful criteria. Table 10 shows the total descriptive statistic related to both tests. Clearly, the performance on the second test (2.42) is higher than on the third test (1.62).

Table 10. Descriptive Statistics of Overall Tests

Test	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Second	2.428	.073	2.283	2.572
Third	1.627	.120	1.389	1.865

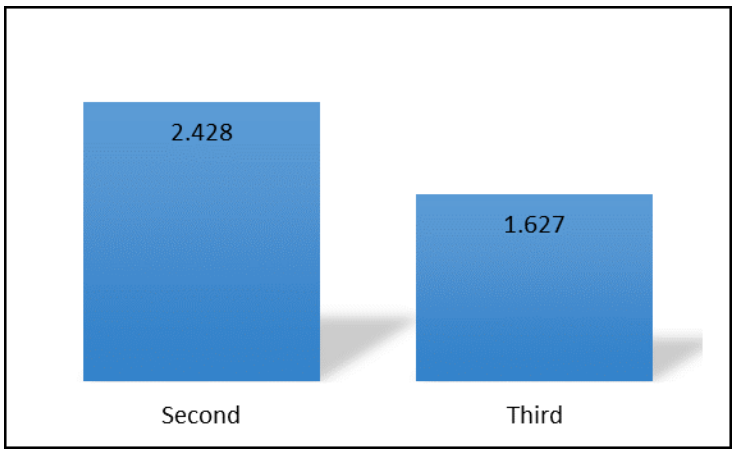

Figure 3. Means on Overall Tests

*Further, Table 11 shows that in all five criteria, the participants performed higher on the second test than on the third test including reliability 2.31, validity 2.33, authenticity 2.48, interactiveness 2.48, and impact 2.51 compared to those on the third test including reliability 1.57, validity 1.57, authenticity 1.67, interactiveness 1.66 and impact 1.64.

Table 11. Descriptive Statistics of Usefulness Criteria and two Tests

Criteria	Test	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Reliability	Second	2.319	.081	2.158	2.479
	Third	1.578	.120	1.340	1.815
Validity	Second	2.333	.075	2.185	2.481
	Third	1.575	.118	1.341	1.808
Authenticity	Second	2.487	.095	2.300	2.674
	Third	1.677	.130	1.420	1.933
Interactiveness	Second	2.482	.080	2.324	2.639
	Third	1.665	.124	1.418	1.911
Impact	Second	2.516	.080	2.359	2.674
	Third	1.643	.122	1.401	1.885

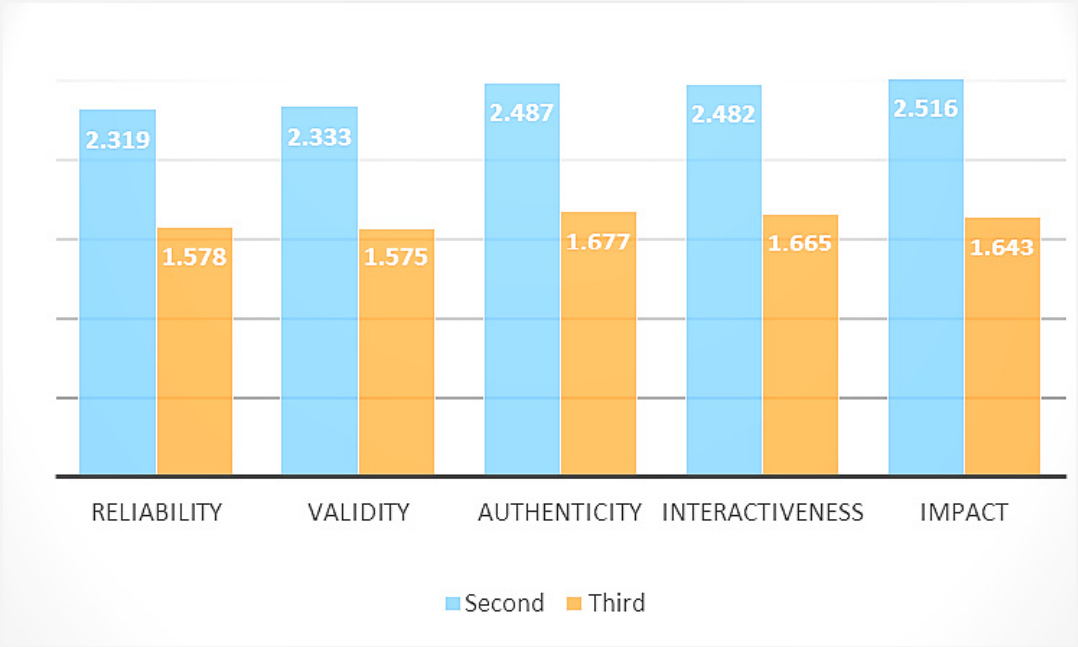

Figure 4. Means on Usefulness Criteria on Two Tests

However, inferentially, pairwise comparisons were run as shown in table 12. Clearly, relationships of authenticity with reliability and validity are 134 and 128, and interactiveness with reliability and validity as 125 and 119, and impact with reliability and validity as 131 and 126 are statistically significant but those of validity and reliability are not significant.

Table 12. Pairwise Comparisons of Overall Usefulness Criteria

(I) Criteria	(J) Criteria	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval for Difference	
					Lower Bound	Upper Bound
Validity	Reliability	.006	.026	.824	-.045	.057
Authenticity	Reliability	.134*	.051	.009	.033	.234
	Validity	.128*	.041	.002	.047	.208
	Interactiveness	.009	.033	.797	-.058	.075
	Impact	.002	.039	.956	-.075	.080
Interactiveness	Reliability	.125*	.042	.004	.041	.209
	Validity	.119*	.031	.000	.057	.181
Impact	Reliability	.131*	.042	.002	.048	.215
	Validity	.126*	.034	.000	.058	.193
	Interactiveness	.006	.026	.806	-.046	.058

*. The mean difference is significant at the .05 level.

The results showed that the following post-hoc comparison tests (Table 12) were significant:

-The overall authenticity criteria (M = 2.082) had a significantly higher mean than the reliability (M = 1.948) (Mean Difference = .134, p = .009).

-The overall authenticity criteria (M = 2.082) had a significantly higher mean than the validity (M = 1.954) (Mean Difference = .128, p = .002).

-The overall interactiveness criteria (M = 2.073) had a significantly higher mean than the reliability (M = 1.948) (Mean Difference = .125, p = .004).

-The overall interactiveness criteria (M = 2.073) had a significantly higher mean than the validity (M = 1.954) (Mean Difference = .119, p = .000).

-The overall impact criteria (M = 2.080) had a significantly higher mean than the reliability (M = 1.948) (Mean Difference = .131, p = .002).

-The overall impact criteria (M = 2.080) had a significantly higher mean than the validity (M = 1.954) (Mean Difference = .126, p = .000).

So, as shown in these tables, the participants had higher means on all usefulness criteria of the second test. Consequently, the second null-hypothesis was rejected.

4.2.3 Testing the Third Null-Hypothesis

In order to test the third hypothesis, first IF, ID and ID percent of each test was calculated item by item as shown in Table 13 and 14.

Table 13. IF, ID, CD Statistics of the Test A

Researcher made test A	Item Nomber	IF	ID	CD
Grammar-Matching	1	0.4	0.68	-
	2	0.4	0.68	-
	3	0.4	0.65	-
	4	0.3	0.48	-
	5	0.4	0.55	-
	6	0.5	0.65	-
	7	0.3	0.65	-
	8	0.5	0.53	-
	9	0.4	0.62	-
Grammar-Multiple Choice	10	0.3	0.37	35
	11	0.2	0.25	21
	12	0.4	0.12	50
	13	0.1	0.15	17
	14	0.2	-0.16	19
	15	0.4	0.37	52
	16	0.4	0.13	48
	17	0.1	0.05	12
Vocabulary-Multiple Choice	18	0.3	0.43	-
	19	0.3	0.39	36
	20	0.1	0	34
	21	0.4	0.6	11
	22	0.3	0.34	50
	23	0.4	0.53	37
	24	0.4	0.46	52
Vocabulary-Dictation	25	0.3	0.56	45
	26	0.3	0.58	-
	27	0.2	0.46	-
	28	0.3	0.53	-
	29	0.3	0.5	-
	30	0.1	0.18	-
Vocabulary-Grammar Completion	31	0.2	0.24	-
	32	0.4	0.5	-
	33	0.3	0.41	-
	34	0.3	0.37	-
	35	0.3	0.29	-

Table 13. IF, ID, CD Statistics of the Test A(continued)

Researcher made test A	Item Nomber	IF	ID	CD
Vocabulary-Make Sentences	36	0.2	0.31	-
	37	0.2	0.37	-
	38	0.1	0.12	-
	39	0.2	0.32	-
	40	0.2	0.32	-
Reading-Question & Answer	41	0.3	0.37	-
	42	0.4	0.5	38
	43	0.4	0.51	43
	44	0.2	0.05	43
Writing-Filling the blanks	45	0.2	0.34	26
	46	0.1	0.18	-
	47	0	0	-
	48	0	0	-
	49	0	0.01	-
	50	0	0.01	-
	51	0.1	0.24	-
	52	0	0.01	-
	53	0	0.01	-
	54	0	0	-
	55	0	0.01	-
	56	0	0.01	-
	57	0	0	-
	58	0	0	-
	59	0.1	0.22	-
	60	0	0	-
Writing-True/False	61	0.3	0.15	-
	62	0.3	0	-
	63	0.4	0.22	-
Speaking-Complete Conversation	64	0.2	0.41	-
	65	0.1	0.29	-
	66	0.1	0.29	-
	67	0.1	0.29	-
	68	0.1	0.2	-
Listening-Complete Conversation	69	0	0	-
	70	0	0	-
	71	0	0	-
	72	0	0	-
	73	0	0	-
	74	0	0	-
	75	0	0	-

Table 14. IF, ID, CD Statistics if the Test B

Researcher made test B	Item Number	IF	ID	CD
Listening-Complete Conversation	1	0	0	-
	2	0	0	-
	3	0	0	-
	4	0	0	-
	5	0	0	-
	6	0	0	-
	7	0	0	-
Speaking-Complete Conversation	8	0.25	0.36	-
	9	0.4	0.5	-
	10	0.25	0.43	-
	11	0.39	0.51	-
	12	0.22	0.37	-
Grammar-Multiple Choice	13	0.33	0.39	36
	14	0.12	0.12	16
	15	0.25	0.46	29
	16	0.27	0.31	33
	17	0.18	0.12	21
	18	0.31	0.44	36
	19	0.29	0.44	34
	20	0.13	0	14
Vocabulary-Multiple Choice	21	0.24	0.41	-
	22	0.16	0.08	27
	23	0.08	0.03	18
	24	0.33	0.56	10
	25	0.38	0.6	40
	26	0.37	0.55	44
	27	0.37	0.67	42
Grammar-Matching	28	0.37	0.67	45
	29	0.28	0.46	-
	30	0.3	0.53	-
	31	0.18	0.32	-
	32	0.25	0.39	-
	33	0.25	0.43	-
	34	0.23	0.36	-
	35	0.21	0.36	-
	36	0.31	0.56	-
Vocabulary-Gap Filling Completion	37	0.18	0.25	-
	38	0.43	0.55	-
	39	0.06	-0.01	-
	40	0.29	0.37	-
	41	0.18	0.29	-
Vocabulary-Dictation	42	0.18	0.29	-
	43	0.2	0.27	-
	44	0.1	0.2	-
	45	0.14	0.22	-
	46	0.14	0.29	-
	47	0	0.01	-

Table 14. IF, ID, CD Statistics if the Test B(continued)

Researcher made test B	Item Number	IF	ID	CD
	48	0.29	0.41	-
	49	0.17	0.24	-
Vocabulary-Make Sentences	50	0.14	0.18	-
	51	0.15	0.27	-
	52	0.12	0.13	-
	53	0.16	0.15	-
	54	0.01	0.03	19
Reading-Question & answer	55	0.08	0.13	2
	56	0.1	0.1	10
	57	0.11	0.18	12
	58	0.12	0.22	-
	59	0.02	0.05	-
	60	0	0	-
	61	0	0	-
	62	0	0.01	-
	63	0.08	0.1	-
	64	0	0	-
Writing-Filling the blanks	65	0.11	0.18	-
	66	0	0	-
	67	0	0	-
	68	0	0	-
	69	0	0	-
	70	0	0	-
	71	0.04	0.08	-
	72	0.04	0.08	-
	73	0.12	0.08	-
Writing-True/False	74	0.17	0.17	-
	75	0.23	0.25	-

Moreover, both tests subsections were compared in terms of IF and ID means scores as shown in Table 15 below. Table 15 shows that in 8 parts from 11 parts of two tests, IF is higher on Test A than on Test B including Gr-Match 0.4, Gr-Multiple Choice 0.26, Vocabulary-Multiple Choice 0.31, Vocabulary-Dictation 0.25, Vic-Filling Completion 0.3, Vocabulary-Making Sentences 0.18, Reading-Question & Answer 0.32 and Writing-True or False 0.33 compared to those on Test B including Gr-Match 0.26, Gr-Multiple Choice 0.23, Vocabulary-Multiple Choice 0.28, Vocabulary-Dictation 0.13, Vocabulary-Filling Completion 0.23, Vocabulary-Making Sentences 0.17, Reading-Question & Answer 0.07 Writing-True or False 0.17. Additionally, Table 15 shows that in 5 parts from 11 parts, ID is higher on Test A than on Test B including Gr-Match 0.61, Vocabulary-Dictation 0.47, Vocabulary-Gap Filling Completion 0.36, Vic-Make Sentences 0.29, Reading-Question & Answer 0.36 compared to those on Test B including Gr-Match 0.45, Vocabulary-Dictation 0.4, Vocabulary-Gap Filling Completion 0.29, Vocabulary-Make Sentences 0.25, and Reading-Question & Answer 0.08. Also, in four other parts ID of Test A is lower than on Test B, and in two parts ID in Test A is equal with in Test B.

Table 15. Mean of IF, ID, CD of the Different Test Formats: A & B

Item Arrangements	Test	Mean	
		IF	ID
Grammar-Matching	Second	0.4	0.61
	Third	0.26	0.45
Grammar-Multiple Choice	Second	0.26	0.16
	Third	0.23	0.29
Vocabulary-Multiple Choice	Second	0.31	0.39
	Third	0.28	0.41
Vocabulary-Dictation	Second	0.25	0.47
	Third	0.13	0.21
Vocabulary-Gap Filling Completion	Second	0.3	0.36
	Third	0.23	0.29
Vocabulary-Make Sentences	Second	0.18	0.29
	Third	0.17	0.25
Reading-Question & Answer	Second	0.32	0.36
	Third	0.07	0.08
Writing-Filling the blanks	Second	0.03	0.06
	Third	0.03	0.06
Writing-True/False	Second	0.33	0.12
	Third	0.17	0.17
Speaking-Complete Conversation	Second	0.12	0.3
	Third	0.3	0.43
Listening-Complete Conversation	Second	0	0
	Third	0	0

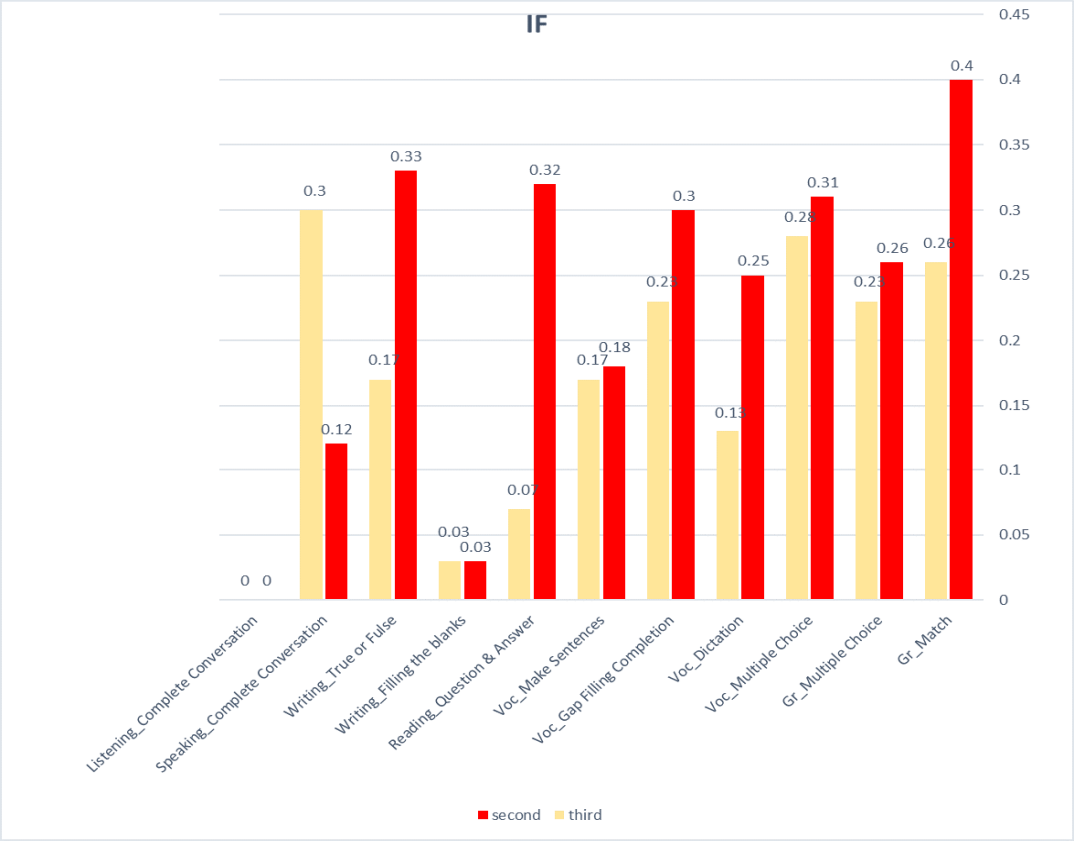

Figure 5. IF Mean for Test A, B

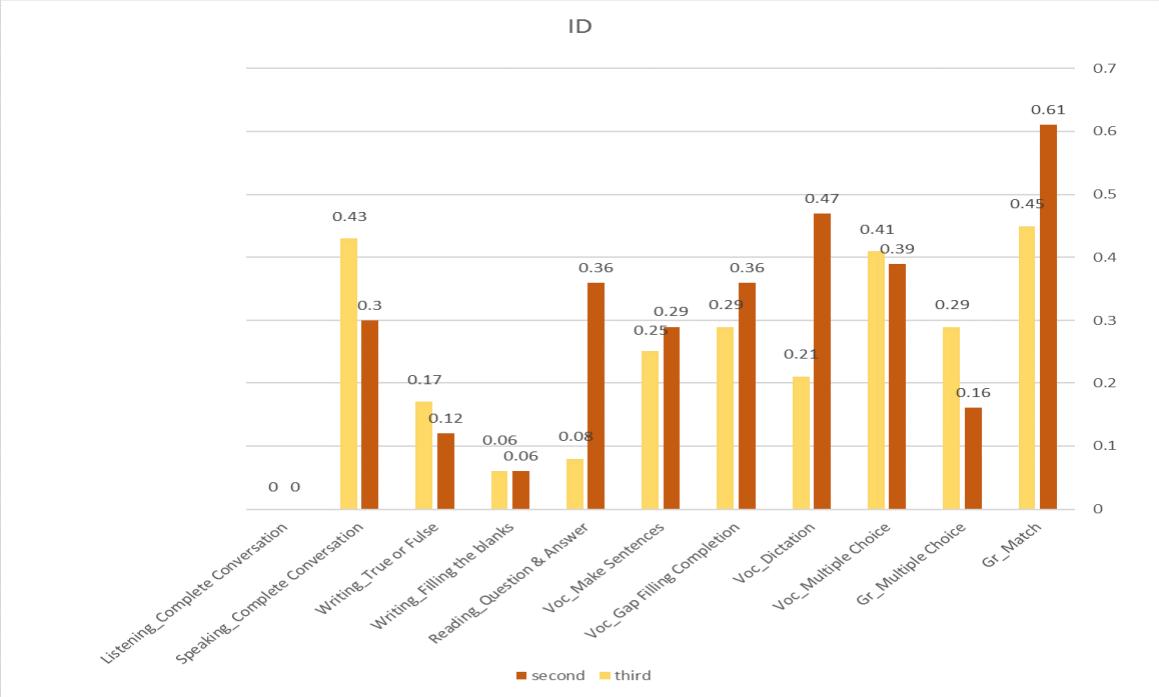

Figure 6. ID Mean Figure of the Tests A & B

Clearly, Table 16 shows that total mean of IF on Test A (0.2) is higher than on Test B (0.15). Also total mean of ID on Test A (0.27) is higher than on Test B (0.24).

Table 16. Total Means of IF, ID in the Tests A & B

	Mean IF	Mean ID
Test A	0.2	0.27
Test B	0.15	0.24

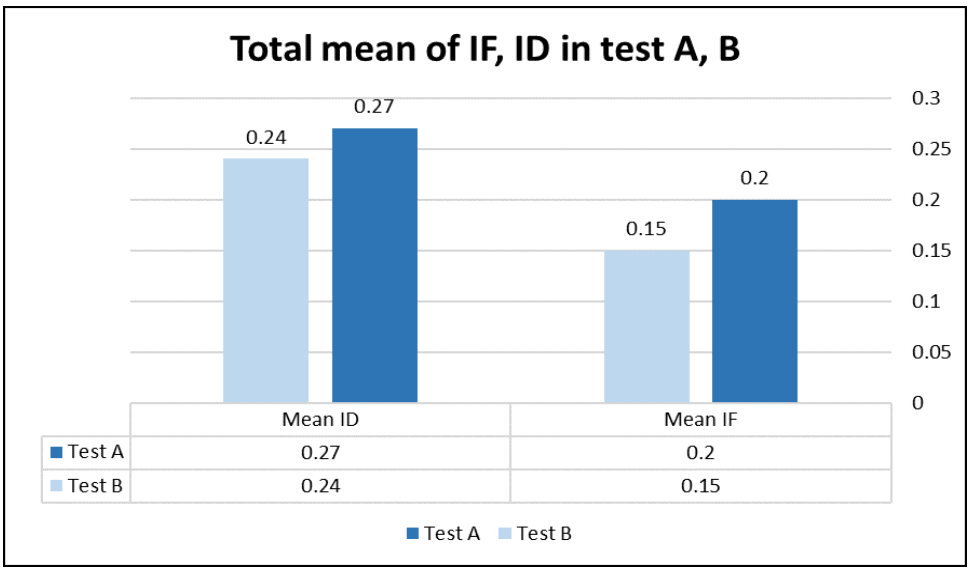

Figure 7. Total Mean of IF, ID in the Tests A & B

Table 17 shows that in Test A, there are 13 questions out of 19 questions that have more students who answered correctly compared to Test B.

Table 17. CD of test A, B

Number of Item		**correct answers**	
Test A	Test B	Test A	Test B
10	13	35	36
11	14	21	16
12	15	50	29
13	16	17	33
14	17	19	21
15	18	52	36
16	19	48	34
17	20	12	14
19	22	36	27
20	23	34	18
21	24	11	10
22	25	50	40
23	26	37	44
24	27	52	42
25	28	45	45
42	54	38	19
43	55	43	2
44	56	43	10
45	57	26	12

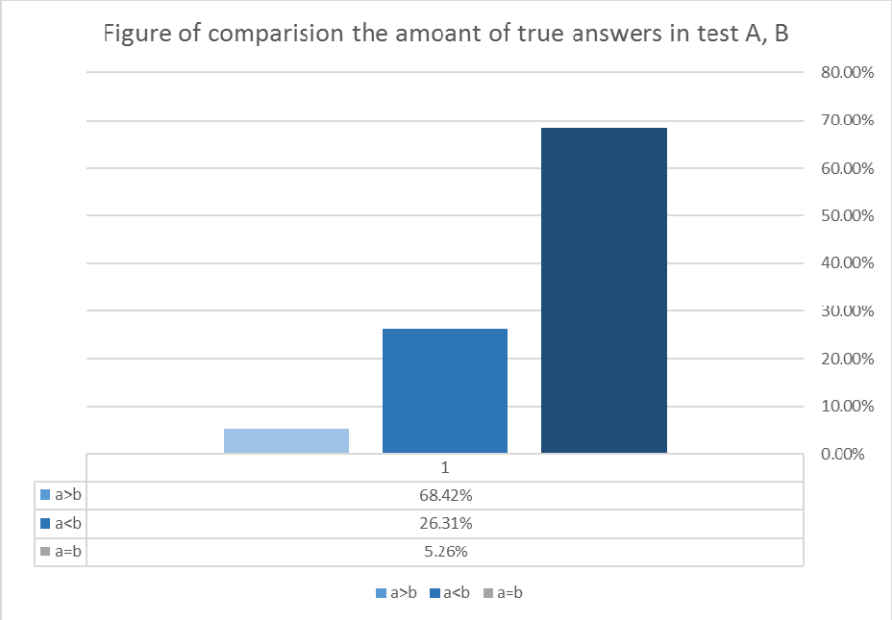

Figure 8. CD of the Test A & B

5. Discussion and Conclusion

To answer the first research question, a paired-samples t-test was run to compare the means of the participants on the second and third tests. With regard to obtained results, the means of second and third phases were not significantly different. It is concluded that the arrangement of items did not have any significant effect on the examinees' performance. This result is inconsistent with finding of the study done by McNeill (1956) who concluded that if

items arranged easily to hard, the performance of testees would be better than the time to arrange items from hard to easy. It is also inconsistent with Hudson (1984) and Huilin Chen (2012) findings who performed various order of questions, easy to hard, hard to easy and random makeup tests, and concluded that if the arrangement of questions is easy to hard, then the testee's motivation is increased and a more reliable test will be produced. Huilin Chen (2012) gained a significant difference in the two tests he administered and observed that if the test starts with difficult questions, the anxiety level of the testers increases and this has a negative effect on their performance. This is in line with Marso' study (1970) in which two tests led to figure out whether any relationship exists between test item arrangements and student performance on power tests. The essential hypotheses were: item arrangements based upon item difficulty, closeness of content, or order of class presentation do not impact test score or obliged testing time. In the first test 122 subjects were randomly deal with three item difficulty arrangements of 139 test items with a 100% difficulty range, and in the second test 156 subjects were randomly assigned to three item content arrangements of 103 items. After effects analyses of variance with test anxiety utilized as a classification factor supported the hypotheses.

To answer the second research question, a repeated measures ANOVA was run to investigate. This design had two within-subject's factors; tests and usefulness criteria. The repeated measures ANOVA produced three F-values for the effects of the tests, usefulness criteria and their interaction on the performance of the participants on the tests and indicated that there were significant differences between the overall means of the five usefulness criteria. Finally, the results indicated that the participants had a significantly higher mean on overall second test (M = 2.42) than third test (M = 1.62) and indicated that there was not any significant interaction between two tests and usefulness criteria. It is concluded that the arrangement of items has significant effect on test usefulness criteria of reliability, validity, authenticity, interactiveness and impact. This is in line with Beckman and Palker (1996) who argue that a good test has six characteristics, which include: the reliability of the test, the validity of the test, the degree of matching and the proximity of the test with real issues (authenticity), the degree of interference the individual characteristics of the testees in the test results (interactiveness), the impact of the test on individuals, the educational system and the community (impact), and the extent of ability to perform the test (practicality).

Statistically, it is still not possible to calculate correlation between two test's item characteristics (IF, ID and CD). According to Hingorjo & Jaleel (2012), Difficulty Index also called Ease Index (IF), describes the percentage of testees who correctly answered the item. It ranges from 0 - 100%. The higher the percentage, the easier the item. So it is clearly that IF of Test A (0.2) is higher than IF of Test B (0.15) and we can say that item arrangements have effect on test characteristics.

According to Escudero and Morales (2000), Item Discrimination (ID), In case the test and an item measure the same ability or competence, we would anticipate that those having a high in general test score would have a high probability of being able to reply the item. We would moreover anticipate the opposite, which is to say that those having low test scores would have a low probability of answering the item correctly. Hence, a good item ought to discriminate between those who score high on the test and those who score low. The higher the discrimination list, the better the item can determine the difference between those with high-test scores and those with low ones. If all the testees in power group reply an item correctly, and all the testees in powerless group answer incorrectly, at that point D = 1. So it is obviously that ID in Test A (0.27) is higher than ID in Test B (0.24) and we can say that item arrangements have effect on test characteristics.

According to Farhady, Jafarpur, Birjandi (1994, p.96), "choice distribution refers to the frequency with which alternatives are selected by the examinees". In the same vein According to Chauhan, P., Chauhan, G. R., Chauhan, B. R., Vaza, J. V., & Rathod, S. P. (2015) frequency distribution according to number of functioning distracter divided in functioning distracters and nonfunctioning distracters groups. As figure 8 shows, Test A has more functioning distractors, so test item arrangements have effect on test characteristics.

This study can inform teachers and students about the factors that affect students' test scores and subsequently the student's destiny, then on the educational system, and ultimately on society. Education administrator and teachers are expected to encourage research in order to come up with the most common test format enhancing their students' performance. Students themselves can take the advantages of self-awareness to see on what format of a test they can perform well. Education system is expected to involve the learners in order to know more on how and on what type of test format they can perform well. Last but not least, education systems and teachers have to appreciate test format and method if they are offer the observation of the most useful test enjoying all the usefulness criteria. It would be benefit for students, educational system and society if test usefulness criteria be in the acceptable range. For instance, if reliability criteria be in its acceptable range, observed scores would be near to true scores. And if validity criteria

are in its acceptable range, it shows that the test measures what is supposed to measure. Also if authenticity criteria be acceptable, it means that the degree of matching and the proximity of the test with real issues is high. Moreover, if interactiveness criteria is in acceptable range, it means that the degree of interference the individual characteristics of the testees in the test results is low. Also if practicality criteria are in true range, it shows that the extent of ability to perform the test is high. And finally if impact criteria are in acceptable range, it means that the impact of the test on individuals, the educational system and the community is good.

References

Abbasian, G. R., & Farhady, H. (2000).Test Method, Level of Language Proficiency, and the Underlying Structure of Language Ability. *Alzahra Journal, 2*(1).

Abbasian, G. R., & Nassirian, H. (2015). *Evaluation of the Iranian State University EFL Entrance Examination Test (UEEET)*.

Ahuman, S. W., & Clock, N. D. (1971). Item difficulty level and sequence effects in multiple-choice achievement tests. *Journal of Educational Measurement, 9*(Summer), 105-11. https://doi.org/10.1111/j.1745-3984.1972.tb00765.x

Allam, S. (2007). *Measurement and Evaluation in teaching process*. Amman: Dar Almasira.

Allison, D. E. (1984). Test anxiety, stress, and intelligence-test performance. *Measurement and Evaluation in Guidance, 16*(4), 211-217. https://doi.org/10.1080/00256307.1984.12022359

Alzude, N., & Elian, H. (2005). *Principles of measurement & Evaluation in Education*. Amman: Dar Alfkr.

Anastasi, A. (1976). *Psychological testing*. New York: Macmillan Press Ltd.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Cacko, I. (1993). *Preparation of good objective test items as a step toward obtaining valid assessment of students' achievement at the SSSCE*. Articles of WAEC Monthly Seminar, Accra, March 1993 ed., 87- 92.

Chauhan, P., Chauhan, G. R., Chauhan, B. R., Vaza, J. V., & Rathod, S. P. (2015). Relationship between difficulty index and distracter effectiveness in single best answer stem type multiple-choice questions. *International Journal of Anatomy and Research, 3*(4), 1607-1610. https://doi.org/10.16965/ijar.2015.299

Chen H. (2012). The Moderating Effects of Item Order Arranged by Difficulty on the Relationship between Test Anxiety and Test Performance. *Creative Education, 3*(3), 328-333. https://doi.org/10.4236/ce.2012.33052

Chidomere, Roland C. (1989). Test Item Arrangement and Student Performance in Principles of Marketing Examination: A Replication Study. *Journal of Marketing Education, 11*(Fall), 36-40. https://doi.org/10.1177/027347538901100307

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Crompton, P. (1996). 12: Evaluation: A practical guide to methods. *Learning Technology Dissemination Initiative, 66*.

Escudero, E. B., Reyna, N. L., & Morales, M. R. (2000). The level of difficulty and discrimination power of the Basic Knowledge and Skills Examination (EXHCOBA). *Revista Electrónica de Investigación Educativa*, *2*(1), 2.

Farhady, H., Jafarpoor A., & Birjandi, P. (1994). Testing language skills: From theory to practice Tehran: SAMT Publications, 1-147.

Gerow, J. R. (1980). Performance on achievement tests as a function of the order of item difficulty. *Teaching of Psychology, 7*, 93-96. https://doi.org/10.1207/s15328023top0702_7

Hatch, E., & Farhady, H. (1982). *Research design and statistics for applied linguistics*. Cambridge, MA: Newbury House.

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *JPMA-Journal of the Pakistan Medical Association, 62*(2), 142.

Hodson D (1984). The Effect of Changes in Item Sequence on Student Performance in a Multiple-choice Chemistry Test. *J. Res. Sci., 21*(5), 489-495. https://doi.org/10.1002/tea.3660210506

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Kufahi, T. (2003). *Measurement & Evaluation in Special Education*. Amman: Dar Almasira.

Leong, L. M., & Ahmadi, S. M. (2017). An analysis of factors influencing learners' English speaking skill. *International Journal of Research in English Education, 2*(1), 34-41. https://doi.org/10.18869/acadpub.ijree.2.1.34

MacNicol, K. (1956). *Effects of varying order of item difficulty in an unspeeded verbal test*. Unpublished manuscript. Educational Testing Service, Princeton, New Jersey.

Madsen, H. S. (1983). *Techniques in Testing*. Oxford University Press, 200 Madison Ave., New York, NY 10016 (ISBN-0-19-434132-1, $5.95).

Margaret, O. I., & Victor, U. I. (2017). Effect of Test Item Arrangement on Performance in Mathematics among Junior Secondary School Students in Obio/Akpor Local Government Area of Rivers State Nigeria.

Marso, R. N. (1970). Test item arrangement, testing time, and performance. *Journal of Educational Measurement, 7*(2), 113-118. https://doi.org/10.1111/j.1745-3984.1970.tb00704.x

Ollennu, S. N. N. (2011). *The impact of item position in multiple-choice test on student performance at the Basic Education Certificate Examination (BECE) Level* (Doctoral dissertation, University of Cape Coast).

Phipps, S., & Brackbill, M. (2009). Relationship between Assessment item Format & Item performance Characteristics. *American journal of Pharmaceutical Education, 73*(8), 146. https://doi.org/10.5688/aj7308146

Plake, B. S. (1980). Item arrangement and knowledge of arrangement on test scores. *The Journal of Experimental Education, 49*(1), 56-58. https://doi.org/10.1080/00220973.1980.11011764

Sax, G., & Cromack, T. A. (1966). The effects of various forms of item arrangements on test performance. *Journal of Educational Measurement, 3*(Winter), 309-11. https://doi.org/10.1111/j.1745-3984.1966.tb00896.x

Sax, Gilbert & Theodore R. Cromack. (1966). The Effects of Various Forms of Item Arrangement on Test Performance. *Journal of Educational Measurement, 4*(Winter), 309-11. https://doi.org/10.1111/j.1745-3984.1966.tb00896.x

Schmitt, John C., & C. James Scheirer. (1977). The Effect of Item Order on Objective Tests. *Teaching Psychology, 4*(October), 144-45.

Shepard, L. A. (1997). The challenges of assessing young children appropriately. In Katheleen M. Cauley (12th ed.), *Educational Psychology*. Shefield: Dubuque Inc.

Soyemi, M. O. (1980). *Effect of item position on performance on multiple-choice tests.* Unpublished M.Ed. dissertation, University of Jos.

Tei-Firstman, R. I. (2011). *Test item arrangement on student test scores*. M. Ed Unpublished thesis, University of Port Harcourt.

Thawabieh, A. M. (2016). A Comparison Between Two Test Item Formats: Multiple-Choice Items and Completion Items. *British Journal of Education*, *4*(8), 32-43.

Thorndike, R., & Hagen, E. (1986). *Measurement & Evaluation in Psychology & Education*. London: MacMillan.

Vander Schee, B. A. (2009). Test Item Order, Academic Achievement and Student Performance on Principles of Marketing Examinations. *Journal for Advancement of Marketing Education*, *14,* 23-29.

WAEC (1993). *The effects of item position on performance in multiple choice tests*. Research Report, Research Division, WAEC, Lagos.

Weiss D. J., & Betz N. E. (1973). *Ability Measurement: Conventional or Adaptive.* Minneapolis, MN: University of Minnesota, Psychometric Methods Program.