استخدام و اتجاهات أعضاء هيئة التدريس بكلية التربية بجامعة الملك فيصل نحو أشكال الاختبارات

د. عبدالله إبراهيم السعادات [١] د. عبدالله عمر النجار [٢]

الملخص : لقد أدى التزايد المطرد في أعداد الطلبة المقبولين بكلية التربية بجامعة الملك فيصل في السنوات الأخيرة إلى نزوع كثير من أعضاء هيئة التدريس إلى ترك استخدام الاختبارات المقالية وذات الإجابة المفتوحة واللجوء إلى استخدام الاختبارات ذات الاختيار من متعدد والأخرى ذات الإجابة القصيرة ، ولكن الملاحظة عن كثب لاختبارات التحصيل (النهائية بالذات) التي تعطى للطلبة في مواد مختلفة بينت بشكل متكرر أن هذا الميل إلى استخدام الأساليب الموضوعية في الاختبارات نتج عنه إعداد اختبارات عاجزة عن تحقيق أهداف المقررات وتغطية محتوى كل مقرر ومادته الدراسية ، وعرض المهارات المكتسبة والتعلم ، وكذلك تطبيق الأساليب الصحيحة في بناء ووضع الاختبارات . وتفترض هذه الدراسة أن أعضاء هيئة التدريس بكلية التربية بجامعة الملك فيصل (وربما غيرها من الجامعات السعوديات أيضاً) بحاجة ماسة إلى :

١- الإلمام بالمستجدات الأخيرة في نظريات وتطبيقات القياس التربوي وتقويم الأداء.

٢- المزيد من التدريب على الممارسات السلمية في تصميم وتطبيق الاختبارات الناجعة.

ولتقرير طبيعة وأفق معادلة الإلمام والتدريب هذه بشكل دقيق ، تم فحص وتحليل عدد جديد من اختبارات النهائية التي تقدم في مواد مختلفة بالكلية ، كما تم استقصاء آراء أعضاء هيئة التدريس بالكلية حول هذا الموضوع ، هذا وقد أجريت هذه الفحوص والتساؤلات في ضوء المستجدات الحديثة في القياس التربوي وتقويم الأداء .

١ - استاذ مشارك – عميد الدراسات العليا بجامعة الملك فيصل – السعودية .
٢ - استاذ مشارك – عميد القبول والتسجيل بجامعة الملك فيصل – السعودية .

Uses and Attitudes of Faculty Members in the College of Education at King Faisal University Towards Test Item Formats

Dr. Abdullah Alsaadat[1] **Dr. Abdullah Alnajjar[2]**

Abstract : -

The ever–increasing number of student enrollment in the college of education at King Faisal University has led many faculty members to abandon essay and open-ended type tests and resort to the quickly scorable multiple choice and short answer techniques. However, close observation of achievement tests, (particularly summative) given to students in different subjects has repeatedly revealed that this tendency towards using objective techniques has resulted in the development of tests that are deficient in their fulfillment of course objectives, coverage of course content and instructional materials, demonstration of acquired skills and /or learning, and application of sound test construction techniques.

To decide on the exact nature and scope of this awareness/training formula, a good number of final achievement tests in different subjects offered in the college is examined and analyzed and the opinions of a representative sample of faculty members are elicited. The aforementioned investigations are conducted in light of recent developments in educational measurement and performance assessment.

This study hypothesizes that faculty members in the College of Education at King Faisal University (and probably other Saudi universities too) are in great need of:

1. Awareness of recent developments in the theory and practice of educational measurement and performance assessment.
2. Further training on sound practices of the design and implementation of effective tests.

1- Associate Professor, University of King Faisal University.

2 - Associate Professor, University of King Faisal University .

Introduction :-

The construction and administration of tests have been explained extensively in the literature as approaches to assess student achievement. However, literature dealing with the effect of test format on student achievement is limited (Melvin, 1987). What is not as clear is whether reviewing different types of tests facilitates learning to different degrees. As (Hills,1981,p.29) maintains "one can hardly decide which kind of test to use based on the arguments and evidence that have been put forward to date".

Two types of test items often used by classroom teachers are: 1) the selection objective type - where the student chooses the correct answer among competing alternatives and 2) the supply type - where the student writes or supplies an answer to a question. Under these two categories, there are different item varieties. Of the various types of objective test items, the multiple-choice format is the most popular for many reasons. First, this type of test item is adaptable to many different situations and can be used to measure almost all the desired outcomes of education. Second, multiple-choice items keep guessing to a minimum degree, especially in comparison with the true and false type. Third, with this type of item format, machine scoring or other forms of answer sheet can always be used (Oosterhof, 1994).

The true and false item is another very common type of objective test items used by the classroom teacher. There are two reasons for its popularity. First, true and false items are easy to make. They do not require the time and thought involved in the construction of multiple-choice items. Another reason is that the true and false item provides for extensive sampling in a given time (Oosterhof, 1994). In such areas as social sciences, it is not unusual for students to respond to 150 to 200 true and false items in a classroom period. One major disadvantage or limitation of true and false items is that they are frequently ambiguous (Stanley & Hopkins, 1981). A good student may see something unintended in the item that the poor student does not. Items that cause such a result lower the usefulness of any test. An additional disadvantage is that guessing is encouraged and favored (Bergman, 1981).

The short-answer item requires students to write an answer in response to the question or problem set by the teacher. The students must supply the answer rather than selecting the answer among alternatives. In other words, this type of item requires a response

composed by the examinee, usually in the form of a word, phrase or sentence. Short answer items are fairly easy to construct, and diminish the likelihood of guessing, but assess mainly factual knowledge (Airasian, 1991).

The essay item represents a very flexible test format (Fraenkel & Wallen, 1990). It can potentially measure any skill that can be assessed with other formats of written tests (Oosterhof, 1994). It allows students to compose their own answer in their own words (Harris, 1969). The answer to an essay item may be short or long, depending on how much the student knows and how full an answer the item requires.

Usually the choice of item format is governed by the nature of the objectives to be measured (Bloom et al., 1981). However, it often seems to be based upon considerations other than student achievement, such as familiarity with a particular format, or attitudes about the superiority of one format over another (Melvin, 1987).

Recently there has been a movement calling for the use of non-multiple-choice test item formats for high stakes testing. This movement emphasizes the use of performance assessments (Myerberg, 1996) .

Two approaches to learning (deep and surface) have been shown to lead to quite different learning quality outcomes. In a deep approach, students have an intention to understand. They focus on what is signified, relate and distinguish new ideas and previous knowledge, relate concepts to everyday experience, relate and distinguish evidence and argument, organize and structure content and adopt an internal emphasis. A surface approach involves an intention to complete task requirements. Students focus on the signs and on discrete elements, memorize information and procedures for assessment, unreflectively associate concepts and facts, fail to distinguish principles from evidence and new information from old, treat the task as an external imposition and adopt an external emphasis (Ramsden, 1988).

Trigwell & Sleet (1990) found that open-ended questions are one way of helping students develop the confidence and the ability to recognize sub-problems at the same time requiring students to complete tasks that lead to deep approaches to learning. Open-ended questions encourage students to think about ways of using data. The students have to decide how data or information can be

used. Memorizing becomes less useful. It is the understanding of the links between concepts that becomes more important.

Garfield (1994) found that traditional forms of assessment of statistical knowledge provide a method for assigning numerical scores to determine letter grades but rarely reveal information about how students actually understand, and can reason with statistical ideas or apply their knowledge to solving statistical problems. Students at the college level need appropriate assessment methods and materials to measure their understanding of probability and statistics and their ability to achieve more relevant goals, such as being able to explore data and to think critically using statistical reasoning.

Performance assessment is an alternative to the traditional methods of testing. It is the direct, systematic observation of an actual student performance and the rating of that performance according to previously established performance criteria. In this type of assessment, students are asked to perform complex performance tasks or create a product. The students are assessed on both the process and the end result of there work. Teachers can use performance assessment to obtain a clear and complete picture of what students know and are able to do (Elliott, 1995). Using performance assessment is considered the best way to discover how students think, or to diagnose where they are having difficulties in learning in a natural context (Ascher, 1990).

Performance assessment can take many forms including: 1) writing essays; 2) conducting experiments; and 3) doing mathematical computations. In language testing, for instance, performance assessment is commonly used for testing writing and speaking. Learners are required to write or present the sample which is evaluated against agreed rating procedures. These samples are elicited in realistic contexts (McNamara, 2000).

According to Ascher (1990) performance assessment includes the following activities:

1. Station Activities: Students proceed through a series of discrete tasks, either individually or in teams, in a given amount of time, much as in a science laboratory.
2. Domain Projects: Students conduct a rich set of exercises designed to explore an idea, concept, or practice central to a particular academic or artistic domain.

3. Portfolios: An extension of domain projects, portfolios consist of several projects completed in a sequence to show students' progress with a subject. Portfolios can include initial plans, drafts, self-evaluations, and feedback from peers and teachers.
4. Videotaping: Although this technology is reliable and inexpensive, its use is still relatively experimental as an assessment technique.

Genesee and Upshut (1999) report that non-conventional performance evaluation activities such as portfolios and conferences, unlike other forms of assessment in which the learner is the object of evaluation, involve the learner as an active collaborator in documenting and monitoring his/her own progress and in identifying learning goals. They stress that classrooms that use only tests or more conventional forms of assessment are often quite different from classrooms where portfolio assessment, for instance, plays a major role. The latter are usually more students centered, collaborative, and holistic. They enhance student involvement in and ownership of their own learning.

Hodges (1995) explored the development of performance events, portfolio assessments, and open-ended questions, with an emphasis on open-ended questions and their scoring. The state-scoring guide placed a 68 percent emphasis on the open-ended questions included in the Kentucky Education Reform Act Assessments (KERA).The open-ended questions of the KERA involved students working individually or together in groups, on simulated real-life problems. The development of the open-ended questions began with central organizers and proceeded through the development of essential questions and the formation of a performance guideline. In his conclusions, Hodges suggests that the rubric developed to score open-ended items must be clear, fair, and reliable. The development of open-ended questions will realign ideas and priorities in the classroom, a change that is the real challenge for the future.

The original impetus for using performance- based assessment was the reaction on the part of educators against the standardized multiple choice tests. Proponents have argued that for high stakes testing, assessments need to involve the direct observation of performance on tasks that are valued in their own right. In other words, the activities need to be authentic. In contrast, the multiple choice test is not authentic since students do not demonstrate their ability to perform particular writing tasks. They

work in isolation and search for one correct answer. Performance assessment is more compatible with curriculum reform and unlike multiple choice tests which focus on factual knowledge and discrete skills, and consider the test to be an indicator of the instructional outcomes, can sustain instruction by focusing on both the process and final products. Thus, performance assessment has been found more appropriate for accountability purposes than multiple choice tests. Performance assessment, therefore, can offer a number of benefits over the use of traditional standardized assessments. The most important benefit is the potential for linking instruction and assessment (McLaughlin & Warren 1995; Linn & Baker, 1996).

Statement of the Problem : -

The ever-increasing number of student enrollment in the College of Education at King Faisal University has led many faculty members to abandon essay and open-ended type tests and resort to the quickly scorable multiple choice and short answer techniques. However, close observation of achievement tests, (particularly summative) given to students in different subjects including English language, Education, Psychology, Sociology, and other social sciences as well as science subjects such as Physics, Math, Chemistry, and Biology has repeatedly revealed that this tendency towards using objective techniques resulted in the development of tests that are deficient in their fulfillment of course objectives, coverage of course content and instructional materials, demonstration of acquired skills and/or learning, and application of sound test construction techniques.

It is the contention of the researchers that faculty members in the College of Education at King Faisal University, (and probably other Saudi universities as well) are in great need of the awareness of recent developments in the theory and practice of educational measurement and performance assessment.

Research Questions :-

The research questions for this study, based upon the statement of the problem, are as follows:

1. What are the attitudes of faculty members in the College of Education at King Faisal University towards types of test item formats?

2. What are the different types of test item formats that faculty member in the College of Education at King Faisal University Use?
3. Do faculty members in the College of Education at King Faisal University use performance assessment techniques in measuring student achievement?

Research Hypotheses:-

The research hypotheses were identified as follows:

1. There are no significant differences between faculty members' attitudes towards test item formats and the employment of these formats in their classes.
2. There is no significant relationship between the use of the test format and the reasons behind that use.

Methodology : -

The Study Population and Sample:

The population for this study consisted of college of education faculty members at King Faisal University (100 faculty members). The sample for this study consisted of (48) faculty members, 19 of them female and 29 male. The sample population includes faculty members from departments of Foreign Languages, Arabic, Education, Educational Administration, Social Studies, Islamic Studies, Physics, Math, Chemistry, and Biology. To obtain information about the type of test item formats used by faculty members in the College of Education at King Faisal University, the researchers collected recent samples of tests written and used by faculty members in the college. These tests comprised 820 questions.

Instrumentation:

The researchers, who took into consideration the characteristics of the target population, developed a questionnaire. This questionnaire includes five parts. Part one was designed to seek demographic information. The second part consists of 17 items designed to evaluate the faculty members' opinions about test item formats. The third part consists of 4 items to examine the reasons behind using different test item formats. The fourth part consists of 3 items to investigate the familiarity of faculty members with different test item formats. The last part contains 3 items to evaluate the faculty members' preferences of different test item formats.

Selection of responses to parts 1, 3, 4 and 5 of the questionnaire were based on a Likert scale format. However, responses to part 2 followed a different arrangement. Respondents were given these selections comprising these test item formats, and were asked to mark the most effective and the least effective test item format among the three. Choices that were selected as most effective item format were given the highest rating (3 points). Those that were considered as least effective item format were given the lowest rating (1 point). Unmarked items were considered as moderately effective test item formats and were given the rating of (2 points).

Reliability testing of the questionnaire resulted in Cronbach Alpha for all parts of the survey (see table 1). These alpha coefficients indicated that the instrument was sufficiently reliable for use in this study.

The validity of the questionnaire was examined by using judges who evaluated each item in the instrument to ensure that these items were measuring faculty members' attitudes toward different test item formats. The researchers concluded that the questionnaire used in this study has good validity.

Table 1 : Item Distribution of Survey Instrument and Their Alpha Reliability Coefficients .

Questionnaire parts	items	subjects	Alpha
MC.	17	48	0.81
SA.	17	48	0.85
ES.	17	48	0.70
Reasons behind using different test item formats.	10	48	0.76
Test familiarity.			
Test preferences.			

Data Collection:

The survey was administered to 100 faculty members in the College of Education at King Faisal University. Forty-eight members (48%) completed the survey. The researchers also collected samples of faculty members' exams comprised 820 questions of recent tests to gather information about the type of test item formats used by faculty members in the College of Education at King Faisal University.

Based on the information presented earlier in this study, exams item formats were divided into two categories. Performance items which comprise free response items such as essay and short answer items, and nonperformance items which includes selection items, namely multiple choice items.

Data Analysis Procedures :

Once the data had been obtained from the questionnaire and the tests, several statistical analyses were performed. Analysis of Variance ANOVA was used to analyze the difference between faculty member attitudes towards each type of item formats and the employment of these formats in the construction of their tests. The researchers also used Pearson Correlation technique to investigate the relationship between the use of the test format and the reasons behind that use.

Results and Discussion : -

First Question:

"What are the attitudes of faculty members in the College of Education at King Faisal University towards types of test item formats?"

To answer this question, the researchers ran the repeated measure test and obtained the Profile Repeated Measure Analysis Withen –Subject Effect which revealed no significant differences in subjects' attitudes towards type of test item format $F(2, 94)=0.237$, $p > 0.05$ (table 2).

Table 2 : Profile Repeated Measure Analysis: Faculty Attitudes (Test Formats) Test of Within-Subject Effect .

Source	SS	DF	MS	F	Sig. of F
Test	15.875	2	7.938	0.237	0.790
Within + Residual	3150.79	94	33.519		

However, calculation of means and standard deviations of all attitude items (table 3) disclosed varying tendencies toward the test item formats concerned.

Table 3
Means and Standard deviations for all Attitude Items .

The Items	MC*		SA*		ES*	
	Mean	SD	Mean	SD	Mean	SD
Assists to measure higher cognitive skills.	1.96	.85	1.90	.63	2.38	1.52
Helps facilitate reliable scoring of answers.	2.46	.77	2.04	.50	1.52	.80
Accurately evaluates students' ability to communicate ideas.	1.40	.61	1.92	.54	2.73	.57
The best measure of students' verbal information skills.	1.25	.53	1.96	.54	2.67	.66
Accurately identifies the learning difficulties of students in the subject.	1.63	.79	1.96	.62	2.42	.79
Helps reduce anxiety of students during the exam.	2.46	.77	1.92	.54	1.56	.80
Tends to cover all of the course materials on the exam.	2.54	.71	2.08	.61	1.33	.63
Provides a clear view of students' understanding of the subject.	1.56	.74	2.04	.62	2.42	.79
Attempts to measure the effective domain of the course objectives.	1.85	.82	2.00	.58	2.15	.90
Attempts to provide consistent results when administered on different groups	2.44	.77	2.06	.63	1.44	.71
Aspires to prepare students to tackle real life problems better.	1.56	.71	1.90	.69	2.42	.79
Help students apply what they have studied better.	1.73	.79	2.04	.68	2.23	.86
A better measure of intellectual information skills.	2.04	.87	2.15	.58	1.96	.92
Help teachers to deal with big numbers of students during test correction.	2.73	.54	2.08	.35	1.21	.58
Easy to construct.	1.52	.68	1.94	.56	2.58	.77
Agrees with the new trends of educational measurement.	2.44	.68	2.10	.59	1.52	.71
The item format that I use most in my test is:	2.00	.80	2.08	.65	2.02	.84

* MC= Multiple Choice SA= Short Answer ES= Essay

As table 3 shows, essay (ES) item format was given highest value in 9 of the 17 statements used to assess the effectiveness of

the test item formats. As the table shows ES item format is considered by respondents as most effective in assisting to measure higher cognitive skills (mean=2.38), and in accurately evaluating students ability to communicate ideas (mean=2.73). Respondents also see ES test item as the best measure of students' verbal information (mean=2.67), and that it accurately identifies the learning difficulties of students in the subjects (mean=2.42). They also consider it to be the most effective in providing a clear view of students understanding of the subject (mean=2.42) and in better preparing students to tackle real life problems (mean=2.42). ES is also seen as the best in helping students to better apply what they have studied (mean=2.23). Finally, respondents agreed that ES is the easiest test item to construct as compared to multiple choice (MC) and short answer (SA) items.

MC test item format, on the other hand, has been given highest value in 6 of the statements of effectiveness of the test formats. MC has been favored by respondents for its ability to help facilitate reliable scoring of the test answers (mean=2.46) as well as its coverage of all course materials in the exam (mean=2.54). It was also found to reduce anxiety during the exam (mean=2.46). Consistency of test results has also been considered as a feature of MC test (mean=2.44), together with its ability to help teachers deal with big numbers of students during test scoring (mean=2.73). One final important remark given by respondents about MC test was its agreement with new trends of educational measurement (mean=2.44).

Although the SA item format was generally the least favored by respondents as highest rating of the test techniques in terms of the 17 statements, it scored highest in two significant accounts. SA was considered the best measure of intellectual information (mean=2.15). It was also considered as the mostly used test format by respondents in their classrooms (mean=2.08).

Such information as is displayed above shows, beyond any doubt, the strong conviction of faculty members in the college of education at King Faisal University of the value and effectiveness of free response essay test items in measuring high order skills and learning that cannot be tested otherwise.

They believe essay tests demonstrate important performative domains and signify learning outcomes better than do the more discrete multiple choice and short-answer techniques. However,

when asked which item format they use most, faculty members elected the short-answer item as the most frequent test type they use in their classrooms. This seemingly contradictory position can, however, be attributed to practical and economical reasons. The use of the essay test, especially with large numbers of students as is the situation in the college of education at King Faisal University would require a great deal of time and effort in evaluating and scoring responses, something that may not be available to the busy classroom teacher. On the other hand, the use of SA technique seems to present an agreeable compromise. By such technique faculty members can guarantee a certain degree of student performance or free response and at the same time test students objectively.

Second Question:

"What are the different types of test item formats that faculty members in the College of Education at King Faisal University Use?"

To answer this question, the researchers analyzed a number of recent tests used by faculty members in the College of Education. Table 4 presents the distribution of these tests among the two major specialties of the college. The Arts tests represent tests prepared at departments such as Foreign Languages, Arabic, Islamic Studies, Education, etc. The Science tests represent tests made at departments such as Physics, Math, Chemistry etc. Both groups of tests were divided into two major categories. Subjective tests, comprising mainly essay test, and Objective tests which include multiple choice, true and false, short answer, and fill-in-blank test items.

Table 4 : Distribution of Faculty Members' Tests Among Specialties and Test Type

	Subjective Tests (Essay)		Objectives tests (MC / TF / SA / FB)		Total	
	F	%	F	%	F	%
Art	321	61.14	204	38.86	525	100
Science	237	80.34	58	19.66	295	100
Total	558	68.05	262	31.95	820	100

MC= Multiple Choice TF= True & False SA= Short Answer FB= Fill in Blank

As table 4 shows, a great majority of tests used by faculty members in the college in both Art and Science specialties (68.05 %) were of the subjective (essay) type, while only 31.95 % of these tests

where of objective type. Preference for subjective type test was more evident in Science subjects (80.34 %) than Arts (61.14%).

This result indicates, the general tendency among faculty members in the Collage of Education at King Faisal University to use subjective (essay) type tests, which reflects the strong belief faculty members have of the effectiveness of these tests in evaluating students achievement as compared to the more structured objective techniques. This result also agrees with the finding discussed in the previous section as to faculty members' attitude towards these test formats.

Third Question:

"Do faculty members in the College of Education at King Faisal University use performance assessment techniques in measuring student achievement?"

To answer this question, faculty members' tests were examined and grouped under two new categories, namely, performance and non-performance tests.

Table 5 : The Analysis of New Categories of Faculty Members Tests.

	Performance tests		Non-Performance tests		Total	
	F	%	F	%	F	%
Art	193	36.76	332	63.24	525	100
Science	119	40.34	176	59.66	295	100
Total	312	38.05	508	61.95	820	100

As shown in table 5, non-performance tests ranked higher in use (61.95 %) in both Art and Science subjects than performance tests (38.05 %), with Arts subjects ranking higher in use of such tests (63.24 %) than Science subjects (59.66 %). On the other hand performance tests were used more frequently in Science subjects (40.34 %) than in Arts (36.76).

This result indicates that preferences shown by faculty members for subjective (essay) tests as displayed in the previous table (table 4) do not necessarily indicate faculty members' awareness of and/or familiarity with performance measures. Although they used a great number of essay tests, the present result shows that most of these tests may have been used in their conventional non-performatve sense. The finding in table 3 supports this conclusion where the majority of respondents find MC items as the technique most in agreement with new trends in educational measurement. The present result may also suggest that many

faculty members in the College of Education at King Faisal University, though convinced of the worth of performance assessment may have considered ES tests a narrow manifestation of performative measures.

First hypothesis:

"There are no significant differences between faculty member attitudes towards test item formats and the employment of these formats in their classes "high, moderate, and low".

To examine this hypothesis, the researchers ran three separate analyses of variance (ANOVA). Results of these analyses are shown in tables 6 through 11 with tables 6, 8, and 10 showing means and standard deviations of the three test formats used, and tables 7, 9, and 11 showing Analysis of Variance of these test item formats.

Table 6 : The Mean and Standard Deviation the Employment of M.C Test Item Format.

Test Format Employment	N	Mean	Std. Deviation
Least	15	22.733	4.131
Moderate	18	29.388	4.692
High	15	30.333	3.903
Total	48	27.604	5.362

Table 7 : ANOVA for the Differences between Faculty member Attitudes Towards MC Test Item Format and the Employment of this Format in their Classes

Source of Variation	SS	df	MS	F	Sig. Of F
Between groups	524.935	2	262.467	14.290	.0001*
Within Groups	826.544	45	18.368		
Total	1351.479	47			

* $p < 0.05$

Table7shows that the ANOVA was significant, F =14.290, $p < 0.05$. This means that there were significant differences between the faculty members attitudes toward MC test item format and their employment of such format in their classes.

To specify to whose favor these differences were, the researchers ran Scheffe test. As depicted in table 8 and figure 1, the Scheffe test results showed that the attitudes of faculty members who indicated moderate and high employment of MC item format differed significantly from those who indicated least employment of such items. Results also showed that there were no significant

differences in attitude towards MC item format between faculty members who employed this format moderately and those with high employment of it.

Table 8 : Scheffe Multiple Comparison Test (MC)

Test Format Empl.		Mean Diffr.	Std. Dev.	Sig.
Least	Moderate	- 6.6556*	1.498	0.000
	High	- 7.6000*	1.565	0.000
Moderate	Least	6.6556*	1.498	0.000
	High	- 0.9444	1.498	0.821
High	Least	7.6000*	1.565	0.000
	Moderate	0.9444	1.498	0.821

*$p < 0.05$

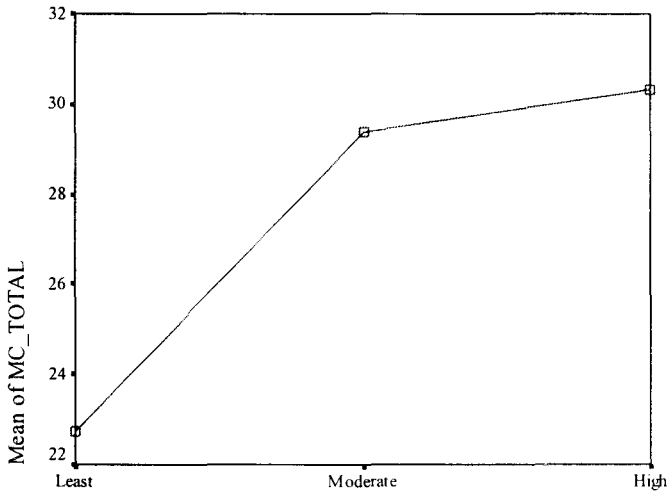

Figure (1)

Table 9 : The Mean and Standard Deviation the Employment of S.A Test Item Format .

Test Format Employment	N	Mean	Std. Deviation
Least	8	23.625	5.475
Moderate	28	27.714	3.598
High	12	31.750	3.137
Total	48	28.041	4.594

Table 10 : ANOVA for the Differences between Faculty member Attitudes Towards SA Test Item Format and the Employment of this Format in their Classes .

Source of Variation	SS	df	MS	F	Sig. Of F
Between groups	324.077	2	162.039	10.918	.0001*
Within Groups	667.839	45	14.841		
Total	991.917	47			

*$p < 0.05$

Table 10 shows that the ANOVA was significant F (2, 45) = 10.918, $p < 0.05$. This means that there were significant differences between the faculty members attitude towards SA test item format and their employment of such format in their classes.

To specify to whose favor the above differences were, the researchers ran Scheffe test (table 11 & figure 2). Data obtained from this test showed that attitudes of faculty members who indicated moderate and high employment of the SA item format differed significantly from those who showed least employment of this format. Also there were significant differences between attitudes of faculty members who indicated highest employment of the SA test format and those with moderate use of it.

Table 11 : Scheffe Multiple Comparison Test (SA) .

Test Format Empl.		Mean Diffr.	Std. Dev.	Sig.
Least	Moderate	- 4.0893*	1.5444	0.038
	High	- 8.1250*	1.7584	0.000
Moderate	Least	4.0893*	1.5444	0.038
	High	- 4.0357*	1.3292	0.015
High	Least	8.1250*	1.7584	0.000
	Moderate	4.0357*	1.3292	0.015

*$p < 0.05$

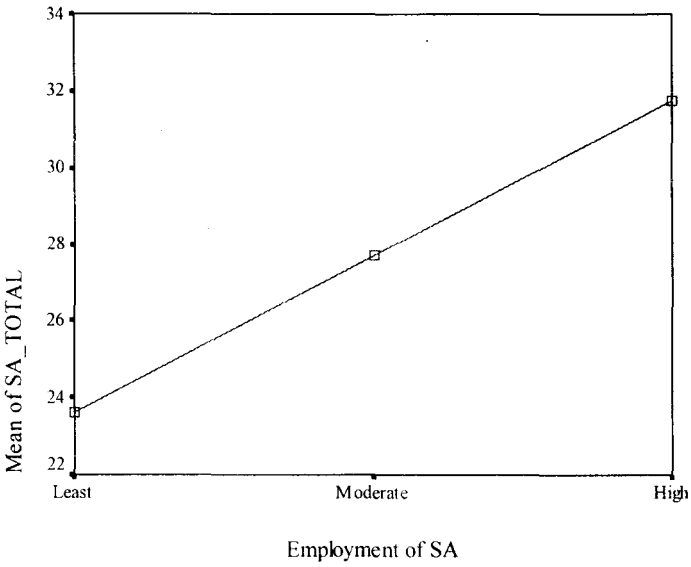

Figure (2)

Table 12 : The Mean and Standard Deviation the Employment of Essay Test Item Format .

Test Format Employment	N	Mean	Std. Deviation
Least	16	26.062	4.373
Moderate	15	28.000	3.982
High	17	31.000	5.612
Total	48	28.416	5.090

Table 13 : ANOVA for the Differences between Faculty member Attitudes Towards ES Test Item Format and the Employment of this Format in their Classes .

Source of Variation	SS	df	MS	F	Sig. Of F
Between groups	204.729	2	102.365	4.548	.016*
Within Groups	1012.938	45	22.510		
Total	1217.667	47			

*p < 0.05

Table 13 shows that the ANOVA was significant F (2, 45) = 4.548, p < 0.05. This result also shows that there were significant differences between the faculty members' attitude towards ES test item format and their use of such format in their classes.

The Scheffe test disclosed a significant difference in attitudes of those with lowest use of ES test item format and those with highest use of it (table 14 & figure 3).

Table 14 : Scheffe Multiple Comparison Test (ES)

Test Format Empl.		Mean Diffr.	Std. Dev.	Sig.
Least	Moderate	- 1.9375	1.7051	0.529
	High	- 4.9375*	1.6526	0.017
Moderate	Least	1.9375	1.7051	0.529
	High	- 3.0000	1.6807	0.215
High	Least	4.9375*	1.6526	0.017
	Moderate	3.0000	1.6807	0.215

*$p < 0.05$

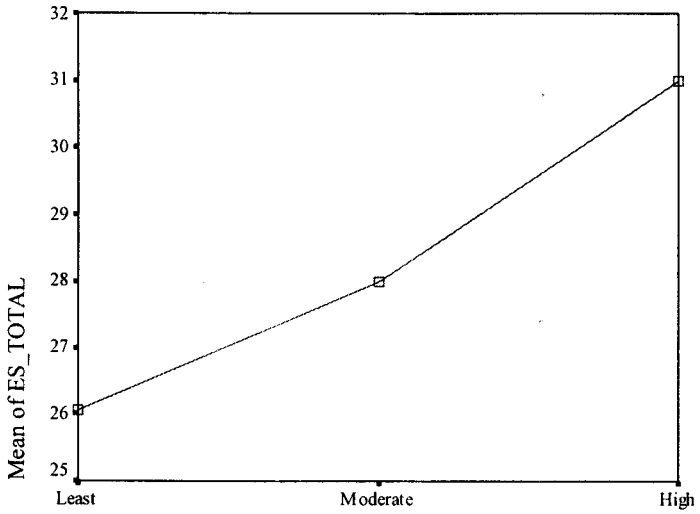

Figure (3)

Second hypothesis:

"There is no significant relationship between the use of the test format and the reasons behind that use"

To examine this hypothesis, the researchers ran pearson Correlation Coefficient.

Table 15 : Correlation Matrix between the Use of the Test Format and the Reasons Behind that Use.

	Use MC	Use SA	Use ES	Reas1	Reas2	Reas3	Reas4
Use MC	1.00						
Use SA	-0.21	1.00					
Use ES	-0.54**	-0.36*	1.00				
Reason1	-0.09	-0.15	0.12	1.00			
Reason2	0.55**	-0.02	-0.58**	0.19	1.00		
Reason3	0.45**	0.03	-0.42**	-0.07	0.63**	1.00	
Reason4	0.48**	-0.04	-0.34*	0.20	0.75**	0.78**	1.00

** Correlation is significant at the 0.01 level.
* Correlation is significant at the 0.05 level.

Table 15 displays information about the relationship between the use of test format and the reasons behind that use. These reasons were identified for respondents as follows:

1. Easy to construct,
2. Easy to correct,
3. Easy to use with large numbers of students,
4. Quick return of test results to students.

As the table shows, there was a significant relationship between the use of MC test item format and the second reason (r=0.55 at the 0.001 level), the third reason (r=0.45 at 0.01 level), and the fourth reason (r=0.48 at 0.01 level). As for the first reason, the table shows no relationship between this reason and the use of MC test item format. This result is reasonable and agrees with the literature on the characteristics of MC test item format. MC test tasks attempt to control in precise ways the particular response required to perform the task. Thus, they are especially useful for assessing particular aspects. They force students to respond to a limited range of alternatives that can be selected carefully to represent the standards of performance of interest to the examiner. Hence, MC tests are quickly scorable as scoring is simply a matter of checking whether the student has chosen the correct alternative. A great deal of care is required for the construction of these items in order to avoid ambiguous or misleading items that are confusing to test taker and produce answers that are meaningless to the examiner. Whether it is worth investing the time and thought needed to devise these kinds of tests depends on how the test results will be used and the importance of the decisions based on those results. Clearly, the

investment of great deal of time and thought is warranted when there is a large number of students to be tested (Genesee and Upshur, 1999).

Moreover, table 15 shows a negative significant relationship between the use of ES test item format and the second, third and fourth reasons mentioned above: r=-0.58 at 0.01 level, r=-0.42 at 0.01 level and r=-0.34 at 0.05 level respectively. Again, this result agrees with the general nature of ES test item format (Harris, 1969). In ES type tests, the range of control on student specific responses is very limited. Such tests permit students not to use tasks that might be of interest to the examiner then the examiner might not be able to assess student performance with respect to certain standards of performance. Students can often find ways of avoiding tasks they do not know or know only poorly. Each student response can be different from other students' responses though not less correct. Hence, a great deal of judgment is called for when scoring ES tests. Consequently, scoring ES tests is much more demanding and requires much more thought than scoring MC tests. Moreover, if ES tests are used to assess proficiency in authentic situations, such as in language use situations, then judgment of appropriateness, effectiveness, and correctness are often called for as these are important standards for assessing language use in situations in which language is normally used (Genesee and Upshur, 1999).

The lack of relationship between this type of format and the first reason above, as shown by table 15, though not in total agreement with what is known about ES item type as reported by the literature (Harris, 1969; Valette, 1977; Hughes, 1993), presents a reasonable conclusion as it may have stemmed again from respondents' belief that ease of construction of the test does not represent a major factor in the selection of item format to use in the test. Also different views that may have appeared in responding to this point may have lead to this conclusion.

Conclusion : -

Our investigation in this study reveals that faculty members in the College of Education at King Faisal University acknowledge and very well appreciate the value of performance assessment in soliciting a clear and complete picture of what learners know and are able to do in contrast to the more structured objective techniques

such as multiple choice items where students only choose answers and have nothing to do to create them.

This belief on the part of faculty members has been manifested in their tendency to use open-ended (Essay) items in their tests, particularly in science departments, which in large part, depend on problem solving to demonstrate student learning.

The insistence on soliciting actual student performance has also been manifested, though minimally, in faculty members' most frequent use of short answer techniques in their tests as scoring and other practical test considerations in large number classrooms do not allow the use of full fledged essays.

The study also shows that faculty members' use of the different test item types is based on their awareness of the features and special advantages of each item type. However, there appears to be some confusion in the minds of faculty members as to which type(s) agrees with recent developments in educational measurement. While, as has been introduced earlier, recent studies emphasize such forms as open-ended items to help learners develop confidence and ability to recognize sub-problems as well as complete tasks that lead to-deep approaches to learning (Trigwel and Sleet, 1990), faculty members in the College of Education at King Faisal University believe that multiple choice test items are most in agreement with new developments in educational testing. This finding suggests that many faculty members in the College of Education may have been overwhelmed with their teaching, research, and administrative duties so long as not to be able to keep abreast of the latest development in educational assessment. Besides, a good number of those teachers are not specialized in education and may have little or no training in educational measurement and testing. Thus, some form of continual program of orientation and feedback on educational measurement and testing practices needs to be organized by the college and made available to its faculty to keep them aware of and in keeping with recent trends and developments in the field.

Surely, our sample of student performance in this study has been limited to conventional paper and pencil classroom tests. Many other forms of performance assessment such as conducting experiments, completing specified projects, and delivering live or audio/video tape presentations, have not been accounted for in the study. Further investigations of the use of such forms by various

departments in the college would disclose more information not only on how much use there is of such performance evaluation measures in the college but also on faculty members' awareness and conviction of performance assessment.

References : -

1. Airasian, P. W. (1991). **Classroom assessment**, New York, NY: McGraw Hill.
2. Ascher, C. (1990). **Can performance-based assessments improve urban schooling?,** New York, NY : Clearinghouse on Urban Education. (ERIC Document Reproduction Service No. ED 327 612)
3. Bergman, J. (1981). **Understanding educational measurement and evaluation**. Boston, MA: Houghton Mifflin Company.
4. Bloom, B. S. , Madaus, G. F. , & Hastings, J. T. (1981). **Evaluation to improve learning**. New York, NY: McGraw Hill Book Company.
5. Elliott, S. N. (1995). **Creating meaningful performance assessments,** Reston, VA: Council for Exceptional Children. (ERIC Document Reproduction Service No. ED 381 985)
6. Fraenkel, J. R. & Wallen, N. E. (1990). **How to design and evaluate research in education,** New York, NY: McGraw-Hill Publishing Company.
7. Garfield, J. B. (1994). Beyond testing and grading: Using assessment to improve student learning. **Journal of Statistics Education,** No. 2,pp.1-11.
8. Genesee, F. & Upshur, J. (1999). **Classroom-based evaluation in second language education,** New Yourk, NY: Cambridge University Press.
9. Harris, D. (1969). **Testing English as second language,** New York, NY: McGrow-Hill Book Company.
10. Hills, J. R. (1981). **Measurement and evaluation in the classroom,** Columbus, OH: Charles E. Merrill.
11. Hodges, R. (1995). Transition to transformation: Open-Ended questions, **ERIC Document Reproduction Service**, No. ED. 390 877.
12. Hughes, A. (1993). **Testing for language teachers**, Cambrindge, GB: Cambridge University Press.
13. Linn, R. & Baker, E. (1996). Can Performance-based students assessments be psychometrically sound ? In G. B. Baron & D. B. Wolf (Eds.), **Performance-based students assessments; challenges and possibilities**. part1. Chicago; Chicago University Press.

14. McLaughlin, M. J. & Warren, S. H. (1995). Using Performance Assessment in Outcomes-Based Accountability System, Council for Exceptional Children ,**ERIC Document Reproduction Service**, No. ED 381 987.
15. McNamara, T. (2000). **Language Testing**. Oxford: Oxford University Press.
16. Melvin, L. R. (1987). **The effects of tests' item format upon the achievement of college level students in an actual classroom setting,** Unpublished doctoral dissertation, The Florida State University, Florida.
17. Myerberg, N. J. (1996, April). **Performance on different test types by racial \ ethnic group and gender**, Paper presented at the annual meeting of the American Educational Research Association, New York, NY.
18. Oosterhof, A. (1994). **Classroom applications of educational measurement,** New York, NY: Macmillan College Publishing Co.
19. Ramsden, P. (1988). **Improving learning, new perspectives,** New York, NY: Nichols Publication.
20. Shepard, L. A. (1995). **Effects of introducing classroom performance assessments on student learning,** Washington, DC. : Office of Educational Research and Improvement. (ERIC Document Reproduction Service No. ED 390 918)
21. Stanley, J. C. & Hopkins, K. D. (1981). **Educational and psychological measurement and evaluation,** Englewood Cliffs, NJ: Prentice-Hall.
22. Trigwell, K. & Sleet, R. (1990). Improving the relationship between assessment results and student understanding. **Assessment and Evaluation in Higher Education**, No.15, pp.190-197.
23. Valette, R. (1977). **Modern Language testing,** New York, NY: Harcourt Brace Jovanovich, Inc.