Testing & Evaluation Systems of the English Paper at the College of Social Sciences and Law in Kuwait University

Abbas H. Al-Shammari (*)

Abstract

This study aims to analyze the testing and evaluation procedures applied to English language tests at the College of Social Sciences and Law of Kuwait University. The research sheds light on the positive and negative sides of this process. For the purpose of conducting this study, the researcher collected the necessary data from a specimen of 100 exams answer booklets taken from the Exams Department at the College. These materials have been extracted randomly from the final exams of the academic years 2010 through 2012. After scrutinizing the components of the exam booklets of the English language test, the researcher analyzed the results, and classified them into categories according to Bloom's classification. It has been shown that the types of questions applied have many objectives, some of which are completion, summative and performance tests. The evaluation process has also occupied a considerable space of this study. The researcher described evaluation, its purposes and types. The researcher –through this study – confirms that evaluation does not aim at fault finding, but it rather focuses on a systematic measurement of tasks specified for gauging students' competence and progress in the learning process of English language at large.

* Head of the English Unit College of Social Sciences
Law Kuwait University- Kuwait

الهدف من هذه الدراسة هو تحليل نظام الاختبارات والتقييم المتبع في كلية العلوم الاجتماعية والقانون في جامعة الكويت بما يخص مادة اللغة الانكليزية. حيث يلقي الضوء على الجوانب الايجابية والسلبية في هذه العملية. ولغرض القيام بهذه الدراسة، فقد قام الباحث بجمع المعطيات من عينة مؤلفة من مائة كراسة إجابات اختيرت عشوائيا من امتحانات الأعوام الجامعية 2010 وحتى 2012 ضمنا. بعد التمحيص في مكونات كراسة الامتحانات لمادة اللغة الانكليزية قام الباحث بتحليل وتصنيف النتائج إلى فئات وفقا لتصنيف "بلوم". وتبين أن نماذج الأسئلة كانت تشمل امتحانات استيعابية تكميلية لمعلومات ناقصة، امتحانات تحصيلية، وامتحانات مهارات. أيضا نالت مسألة التقييم حيزا هاما في هذه الدراسة. حيث قام الباحث بتفصيلها بدءاً من الغرض منها ، وصولا إلى أنماطها وتحليلها. يؤكد الباحث من خلال هذا البحث بأن وسيلة التقييم ليست لوضع علامات على الأخطاء فحسب، بل إنها عملية قياس منتظمة مكرسة لقياس مقدرة الطلاب و مدى تقدمهم في عملية تعلم اللغة الانكليزية في مجملها.

Introduction

This research is intended to tackle the matter of testing and evaluation systems that are valid at the College of Social Sciences and Law of Kuwait University. The field survey of the study covers the analysis and assessment of the exam results of the English paper. For this purpose, the researcher selected a specimen of 100 exam answer sheets at random. They refer to the academic years 2011 through 2012.

Evaluation is an important factor in the teaching process. It can be viewed as one of the main components of the course—the didactic or behavioral goals, the course content, as well as the techniques and methodology of teaching, (Surry, 2000). The significance and interrelation of these four components of the course and the teaching process are taken into account in this research. The role and effect of these components cannot be identified without the evaluation element. The concept of evaluation should be clarified. Thus, the researcher defined it as the process of decision-making on the value of a task in accordance with systematic and organized information. Also, it can be regarded as the essential criterion in gauging what students have learnt at a teaching institution. Therefore, measuring the competence and progress of students can be the true scale that reflects the level of the students' achievement. In addition, the researcher tackles the concept of measurement, which differs from that of evaluation in being a comprehensive educational concept. Measurement is performed through fixed criteria, or in contrast to another normative reference; the effectiveness of one or more factors in the learning process. On the other hand, the evaluation cares about the value of the educational domain as a whole: the input, the output, and the procedures of the learner's needs. Thus, measurement, as Martorela sees, is a systematic method sued to collect information needed, like the tests that are employed to measure the outcome the students gain in certain subjects. (Martorela, 1998).

Since measurement forms part of the evaluation process, the

researcher used it to measure the outcome the students gained from the academic years 2010 through 2012. This process has been carried out by poking into the analytical and achievement tests, which are tools of measurement, and tools of evaluation at the same time. The researcher relied on the achievement test which is widely used in the assessment of the learning process, in general, and because it is commonly applied in the final examinations of the English paper at the College of Social Sciences and Law in Kuwait University, in particular.

The Problem of the Study

The problem of the study can be defined in providing answers to the following questions:

1. What is the level of the questions the instructors of the English paper care about when designing the final exam questions according to Bloom's classification in the cognitive domain?
2. What type of questions do the instructors include in the exam booklets?
3. What is the purpose behind each type of questions?

The Goal of the Study

This study aims at:

a. Recognizing the nature of the questions the instructors of English paper at the College of Social Sciences are after.
b. Figuring out the strong and weak sides in exam questions.
c. Defining the level of the questions, and its relation to the content of the curriculum.
d. Tracing the reliability of assessment and progress that students have attained at the end of each course.

Procedure of the Study

The researcher depended on:

a. Random samples of 100 students answer sheets of final and midterm exams of the English paper from the academic years 2010 through 2012.
b. Analysis of exam questions prepared by tutors and instructors at the College of Social Sciences and Law.

c. Tables and equations of error analysis.
d. A questionnaire composed of 20 statements administered to students for evaluating the current situation of the exams at the College of Social Sciences and Law.

Foreword to testing and evaluation at the College of Social Sciences and Law

Before going through evaluation and testing processes that are valid at the College of Social Sciences and Law, an introduction to the assigned materials taught at this college is necessary.

For grammar branch, the material contains fundamentals of grammar dealing with the developmental skills specified for graduate students who learn English as a foreign language. The material combines clear and understandable grammar information with a variety of exercises and activities. The approach is eclectic, with texts seeking to balance form-focused language-learning activities with abundant opportunities for engaged and purposeful communicative interaction, (Berrenberg, J.L. 2005). The researcher thinks that the books of grammar include a greatly increased number of "real communication" chances for instructors to use. The texts often include students' real life experience as context, and regularly tackle subjects of interest to stimulate the free expression of ideas in structure as well as in open discussions. The books also back the opinion of many linguists that grammar-based and communicative approaches are not mutually exclusive.

The materials are of reading and vocabulary, as well as four-skill comprehensive ESL/EFL books. They combine communicative activities with skill-building activities to enhance students' academic success. They contain updated content. The pictures and illustrations activate prior knowledge of the topic. The vocabulary preview allows students to anticipate unknown vocabulary items. The pre-reading questions of the comprehension passages help students to do active reading, and they are structured in a style of a magazine article or an

essay. The researcher, too, noticed that the skill development prepares students for standardized tests through reading skills, and critical-thinking activities. The reading texts urge students to evaluate arguments and to do independent research related to the target topics. In addition, the vocabulary and language-learning strategies for synonyms, antonyms, context clues and word families offer students comprehension and self-assessment tools.

For the writing skill, the researcher thinks that the writing is an innovative skill. It can motivate students to write subjectively. In this sense, the wealth of language content that students have acquired through their years of study explores intellectually challenging and contemporary themes which stimulate critical thinking skills while building language acquisition. Themes consist of stimulating topics for reading, writing, listening and speaking. The materials contain creative activities that help students develop language-listening strategies, such as predicting and identifying the main ideas and details.

Process of Evaluation and Testing of English at the College of Social Sciences and Law

English teaching at this college is of atwofold aim. First, to specifically describe and comment on the relationship between curriculum and students evaluation in a certain college; and second, to analyze the kinds of questions which any instructor might pose concerning the harmony between these aims, and the components of ELT.

The staff of the English Language Unit at the College of Social Sciences and Law presided by the researcher has long realized the significance of evaluation and the role it plays in promoting the outcome of learning. A statement of the functions and objectives of evaluation in the English curriculum explicitly states the three factors which contribute towards the effectiveness of language acquiring; namely, the instructor, the student and the teaching materials. However, this research is much concerned with the methods of evaluation and testing of the English paper at the College of Social

Sciences and Law with aview to assessing achievement in the English language learning skills. The researcher found—through checking samples of the English paper exams of different courses starting from the academic years 2010 through 2012—that evaluation is inseparably related to both objectives and goals procedures, and that testing—including quizzes, midterm and final exams—purports to improve the English language teaching-learning process, but it should have a re-evaluation of the current exam system and methods of evaluation as one of its basic goals. Also, this study shows that evaluation has been a sort of hindrance to learning process simply because students study for testing purposes, and not for acquiring the language, the researcher argues. Therefore, the tests and other tools of evaluation determine what the students focus on in their study, and not on what they have achieved. Although the English materials make it clear that evaluation is not to be restricted to testing, the overwhelming majority of the instructors of English at the College of Social Sciences and Law still resort to tests as the sole instrument for the evaluation of achievement. This is a serious defect in the present system of evaluation, the researcher thinks. He suggests that subjective estimates of student work have advantage over tests in term to the extent that the outcome the instructor and the student receive is immediate, constant and informal. Hence, lectures for teaching-learning activities establish situations, which require real language practice and evaluation on the spot. The absence of such evaluation in these situations is bound to encourage students to take their work less seriously.

The Necessity of the Run-on Evaluation

Adrienne defines evaluation as the process of making judgments on the basis of the information collected relative to the learning objectives. While assessment is the process of actual gathering of information to make judgments for evaluation, grading involves assigning a mark as a means of conveying the judgment, (Adrienne, 2004)

The evaluation that the researcher targeted is intended to gauge students' growth, development, and progress toward stated learning

objectives. The assessment instruments have been embedded in the discussion of the instructional strategies to demonstrate the close relationship between instruction and evaluation.   The appropriate instruments that tutors used were constructed to be complementary to the strategies employed, and to the objectives being designated.

The evaluation mechanism, the researcher noticed, was in many aspects sensitive to cultural, linguistic, and community situations as well as to the individual student needs and learning styles. Students were provided with a variety of ways to demonstrate their outcome of learning. So, each responds differently to evaluation. In the end, the overall evaluation, addresses all language strands and balances their orientation, (Ericken, 2002). The assessment and evaluation processes involve multiple perspectives and sources of data, content, process, and product which play an essential role in assessment and evaluation. Thus, students will be able to know what they are required to learn, i.e. (content), 'how' they are expected to learn i.e. (process), and 'what evidence' they will be required to produce as a result of that understanding, i.e. (product). As much as possible, students were introduced to a variety of ways to learn and demonstrate their learning. The content of the language arts program was based on the concepts surrounding language and various oral, literary, and media texts, (Beyer, 2005). The researcher regards the products of the language arts course as the combined results of the content and processes (e.g. an exam paper).

Regarding the types of evaluation which are valid at his College, the researcher confirms that the diagnostic evaluation is done informally and continuously. It is used to assess the strengths and needs of students and to make program modifications. It is used for 'diagnosis' rather than for grading, while some instructors at the College conduct run-on formative evaluation throughout the course. They believe that it improves instruction and learning, and to keep both students and tutors aware of the course objectives. Then they analyze and employ the   results of the formative evaluation to identify the outcome of their work.

Summative evaluation occurs at the end of a unit or a program, (Vaughn, 2005). Other instructors at the College use it with the formative evaluation to determine students' achievement and program effectiveness. They resort to the summative evaluation as a significant part in students' grading. Generally, the researcher recommends and calls upon his staff to follow an appropriate balance of diagnostic, formative, and summative evaluation.

Evaluation Portfolios at the College of Social Sciences and Law

The English language portfolios can be an effective way for students and tutors since they gauge the progress realized over a period of time. Because they are purposeful collections of students' work, portfolios can serve as the basis for evaluation of students' effort, progress and achievement in English language skills. What is applicable at the College of Social Sciences and Law is a term-end portfolio, which assembles a few weeks before a reporting period, and includes selected written products, such as short quizzes, class activity and oral presentation. An end-of-year portfolio also illustrates progress and achievement in a certain course; while the multi-year portfolio acts as a showcase of the students' best results in various English courses of their study years.

The area of testing in language learning has accomplished enormous evolution and modeling. Equally enormous evaluation and modeling are the concern of teachers about the discrepancies between teaching and testing procedures, (Ericksen, S.C. 2002). In this study, for instance, the approach in general, and the syllabus in particular, emphasize oral language skills, but also they prescribe paper and pen tests as evaluation instruments. This in itself is a sufficient reason for the inconveniences and problems that arise during and after the scoring processes, the researcher argues. This realization is fueling a massive and persistent movement calling for redesigning and remodeling the test and evaluation instruments and mechanisms. The researcher purports that it has to be straightforwardly asserted, though self-evident, that instructors have to test students' language skills in order to:

    a. establish what he/she knows (or what he/she can do)

    b. assess how successful they have been in teaching job.

Brown believesThat performance on tests is not necessarily an accurate directory to what might be regarded as students' ability to communicate, but his competence to express and convey a message correctly, certainly, also spontaneously, fluently and appropriately is amore accurate directory. (Brown, 2007) Unfortunately – the researcher confirms- that in our current tests, correctness is the only aspect we test, and not the other features of actual language performance.

One weak side in the tests, the researcher spotted, is that the amount of language produced by the students is actually very little. Their role in the existent system of testing may be confined to selecting alternatives, rather than producing language. Furthermore, the problem with models of tests (the way they are designed currently) is that their value is quantitative but, unfortunately, not qualitative. A random stroke may fall on a correct answer to the item in question. And in the majority of tests the ability to synthesize is missing. In fact, knowledge of the elements of language counts for nothing unless the learner is able to combine these elements in fresh and appropriate ways to meet the linguistic demands of the situations in which he/she wishes to use the language, (Jacobs and chase 1992).

Domains of Evaluation at the College of Social Sciences and Law

Tutors and instructors-made tests, as well as the tests which come from the higher English language pools at the college, are defective in more than one respect, mainly through their failure to correlate with the aims stated in the syllabus, which give emphasis to the four language skills and the communicative aspect of language teaching. In what follows there will be abrief comment on the main points of weakness.

Speaking is entirely ignored in the College of Social Sciences and Law tests. For instance, in all the English paper tests, there is not even one item that could be considered an oral production test. And if some instructors administer some sections orally (e.g. transformation

and substitution)—which indeed they might— this is done through asking students to answer by rote, and do not test their ability to use the language orally in real language situations.

Some may justify this neglect of tests of speaking by arguing that they are highly subjective, time-consuming, and very difficult to prepare. However, regardless of all the objections and shortcomings, the researcher believes that instructors must evaluate the skill of speaking if they are interested in having their students focus on this skill.

While it is understandable why instructors are reluctant to evaluate the students' ability to speak, it is not easy to justify the exclusion of oral comprehension from the English credits of the College of Social Sciences. Oral comprehension tests are easy to prepare and administer, and do not take much time and can be marked objectively. A criticism voiced against oral comprehension tests is that since the questions tend to be in written form, the test becomes in part one reading ability, and the student is sometimes required to use his/her ability of memory and observation, (Madlin. 2003). Moreover, in order to minimize oral comprehension tests, the answer choices are made to be brief and lexically and grammatically simple. It is worth pointing out in this respect that listening comprehension tests have been reported to show the best estimate of the efficiency of the internalized grammar of the non-native speakers.

Poking through a sample of 100 of the answer sheets of the English paper exams from 2010 through 2012 showed that testing of writing presented an entirely different problem, which did not concern the setting of the questions or the administration of the test so much as it concerned the correction of the students' work. If it was not properly corrected, as may will be the case in most credits, the outcome would be invariably minimal.

Another problem arose from the frequent use of free essay or composition. The researcher thinks that the validity of this procedure is certainly questionable, as it is known in advance that a considerable

number of students cannot construct the various types of sentences without committing very elementary errors. One of the weak points in testing writing is that composition or essay writing tests do not give a true account of what the students have learnt, for it is a well-known fact that students can cover up weakness by avoiding structures or other linguistics items which they do not feel confident about, (MaCclain, 2005) And above all, instructors of the English paper should not expect students to write a good composition or essay in English if they cannot do so in their native language.

Structures and vocabulary are quite often based on the exercises in the syllabus. In the college of Social Sciences and Law, the tests of the English language are varied, have wide coverage, and test both receptive and productive skills. A few objections can, however, be raised against the current practice of testing structure and vocabulary in the exams of the English paper at the College of Social Sciences. First, almost all the items are of multiple choice type. Second, grammatical structures and vocabulary are to some extent under-represented (varied number of simple statements with flashing answer among four odd choices). Third, many of the statements are trivial and unrealistic in the sense that they deal with peripheral issues in foreign-language teaching. The researcher suggests that English exam papers should dispense with 'accuracy', and there should be priorities in testing grammar and vocabulary. The choice of items should be based on a careful analysis of the frequency of the occurrence of these items in the language and in the syllabus.

Discrete-Point Tests as a Tool of Evaluation at the College

Another major drawback of instructors-made tests is that they are in most aspects discrete point tests. However, admittedly a few modified forms of cloze tests should be contained in some exams, (Liska, and simson, 2001).

The inclusion of cloze items in the tests represents an improvement. The researcher argues that cloze tests have vivid

advantage over discrete tests in that the latter usually require students to perform highly artificial tasks, and, on the basis of their performance, attempt to infer their level of competence for a different sort of task altogether. Moreover, it has been argued that discrete-point tests often fail to provide students with practice in useful language skills. The cloze tests often fail to provide students with practice in useful language skills. The cloze tests have been used as a measure of the proficiency of non-native speakers in foreign-language situations, and have been found to have a high correlation with TOEFL, (Oller, 2002). Finally, cloze tests are easy to set, and encourage the examinee to behave in a linguistically more creative manner. Nonetheless, the researcher posed criticism against cloze tests, as objectivity in marking cannot be guaranteed, though he found that any contextually acceptable insertion should be permitted.

Errors for Evaluation

The researcher found that language errors made by students at the College of Social Sciences are due to the strategies made by students in language acquisition, and the mutual interference of items within the target language. By treating such errors as a variable that can be measured, it can be claimed that scores on developmental errors can relate to the fluency of the students as contrasted with their accuracy. Therefore, the accuracy-based tests—which are applied to the English paper exams at the College of Social Sciences—should be complemented with fluency-based test items, even if it was in a form of an interview in order to provide a better assessment of the total language competence.

Almost all multiple-choice part of English proficiency tests deals with the accurate responses of the 'testees', (Schwiser, 2003). No attempt is made to measure other deviant, yet acceptable responses, which might have been selected if the students had been given such choices. In addition, making errors is a necessary part of language learning, and in the same way accurate answers indicate language proficiency. Errors will also show a learner's progress and

achievement at the undergraduate stage of language instruction. Much focus is put on the developmental errors and not with errors due to the influence of the mother tongue (interference errors). They do not derive from the transfer from the other language they reflect.

The students' competence at the university, which illustrates some of the general characteristics of language origins, are found within the structure of English itself, and through reference to the strategy by which a second language is acquired and taught, (Sam, 2001).

The researcher thinks that if such errors reflect the students' competence at the College of Social Sciences and Law, then they should be accounted for when assessing students' total competence, (Kuwait University, Exam Dept. 2011).

Given a choice between a correct 'accurate' answer and an incorrect 'deviant' answer of the developmental error type in a number of multiple-choice terms—which is widely applicable to vocabulary, grammar and reading comprehension tests—students' level of competence is, as the researcher believes, more accurately measured by considering both correct and incorrect answers: the core of the developmental errors is related to the students' fluency, which is the maximally effective operation of the language so far acquired.

One way of measuring errors has been recommended by the coordinators as well as the supervisors of the English papers at the English Language Unit (ELU) of the College of Social Sciences and Law. It is done by measuring the frequency of occurrence of various forms. Through controlled elicitation techniques (i.e. free composition, sentence completion structured interviews and story telling) percentages of various errors have been established. When going through the students' answer sheets of the sample, the researcher noticed that they appear in the use of tenses misuse of structures and semantics, due to the influence of the mother tongue, the Arabic language. This research has mainly focused, among other issues, on sections of students from the same linguistic and cultural background. The use of 'introspection techniques' in order to tap students' own intuitions about errors has been much less in evidence.

To the researcher, the chosen errors of this type were not of the type that can impair communication significantly. In all of the errors the emphasis was on the communication of content rather than preoccupation with form. Hence, students should be accorded due credit for choosing these rather than the other distracters which grossly violate the syntactic patterns, and as a result impair communication. By choosing the developmental errors rather than the correct answers, the students indicated 'the maximally effective operation of the language system so far acquired, (Erickson, 2002). And it is in this sense that the scores of DEVELOP can relate to the fluency of the students—their capacities and the demand of the task, rather than a pre-specified type of attainment, and no more linguistic freedom,

The study shows that the regular pattern of errors could be observed in the performance of the students. But, the progress that has been achieved through this pattern can be taken as evidence not of failure but of success and achievement in the learning process.

The researcher's views seem to justify the assumption that, as variables, errors can be quantified and used in the total assessment of students' linguistic competence. This obviously implies that errors will be an inevitable part of the process of second-language development, (William, 2000).

Measurement of Accuracy According to the Researcher

The measurement of errors also gains support from recent observations made by the researcher while checking samples of 450 exams answer sheets of students from the College of Social Sciences and Law regarding the accuracy concept in ELT. It has been found that students' competence in answering the multiple choice questions in vocabulary sections has been the same way when they have had the same type of questions with other subjects.

This implied that not only the correct or the accurate answers should be used for the assessment of competence, but deviant is also acceptable as a language behavior, and it should be taken into consideration. Moreover, if it is assumed that the second language

competence of a language learner at a given time is transitory, then all the language acquired or learned at a certain stage should be the concern of the assessor, not just some of it. This is even more important when language learners are adults using certain 'fossilized' structures: if major concern is with how much is communicated, the fossilized forms should receive due credit as well, (Oller, 2002). The question sheets of the English paper exams between 2010 and 2012 contained repeatedly the same types of questions, especially in grammar section.

Through checking the grammar section of the 1st term English paper final exam of 2012, the researcher found out that error types related to the structure of verbs in English were included in 30 test items, using multiple-choice format; two questions were given for each error type. Only errors in verb structures were chosen because the researcher believes that these errors are more crucial in the process of language learning than errors in prepositions and articles. The grammar sections consisted of 30 statements with subsequent answer. The answer for each question contained four choices: the correct answer (CORRECT), the developmental error (DEVELOP), and two distracters. Students were asked to shade the right answer on the answer sheet. The test was administered to 100 students from the College of Social Sciences and Law. Students' answers to both CORRECT and DEVELOP choices were analyzed for variance.

It was found that the choice of DEVELOP was systematic among the students. However, the researcher's concern was within group variation, i.e. differences between individual students.

One interesting finding of the test was the relationship between the CORRECT and the DEVELOP choices. There was a negative correlation of –0.60 between the two. This means the higher the number of CORRECT is, the lower the number of DEVELOP will be. This is to be expected because of the transitory nature of second language competence. However, the test results show that if a number of students had the same mark for CORRECT, the mark for

DEVELOP was not necessarily the same. The researcher has extracted his findings, listed in tables 1,2, and 3, upon taking the percentage average out of 100 of students' answer sheets of the specimen, for example, 24 students in table 1 received 24 for CORRECT but had a different score for DEVELOP.

Table-1

Student	1	2	3	4	5	6
CORRECT	24	24	24	24	24	24
DEVELOP	3	4	5	6	7	8

This implies that these students had the same language competence if only CORRECT is considered; however, they were at different levels of language development and hence had different competences, if DEVELOP is also taken into consideration. It might be argued that the difference between student 1, 2 and 3, etc. was not really significant, but the one between 1 and 6 definitely was.

Applications and Implications

With an example exam such as the one mentioned here, two scores were established for every student, i.e. CORRECT and DEVELOP. The scores on DEVELOP were used by the researcher to sub-classify students after an initial classification based on the CORRECT results. It is not suggested here that the two scores should be combined as they are. Obviously, the CORRECT score should carry a lot more weight than the DEVELOP score. Nonetheless, it is possible to score for DEVELOP by giving, say, 0.50 to each, and then tallying the results. The final scores –the researcher concludes- listed above would be as in Table 2.

Table 2

Student	1	2	3	4	5	6
Combined Correct & Develop score	25.50	26	26.50	27	27.50	28

The researcher suggests that this is a more accurate profile of these students. Also, consider the partial list in Table3, which is based on CORRECT only, and compare it with the scores achieved by the same students with CORRECT and DEVELOP combined. The researcher found that the rank was different as a result. The figures shown in the table below have been calculated by percentage average out of 100 students' answer sheets which is the specimen of the study.

Table 3

Student	1	2	3	4	5	6
CORRECT	25	26	27	28	29	30
DEVELOP	10	2	8	2	8	2
CORRECT + DEVELOP	30	27	31	29	33	31

It would be interesting to find out the correlation of the combined results with scores on another exam, such as a cloze test that is not item-specific (i.e. where alternative responses are also acceptable).

Thereby, this study has shown that developmental errors were measured and quantified by a multiple-choice test of English language proficiency. If the scores for the correct answers are to be taken as a measure of students' accuracy in second language, the scores can relate to their competence in writing. It logically seems that a

student's total competence may be better measured by a consideration of the two scores rather than of the correct answers alone. This is especially important if it is necessary to make finer distinctions between students who have the same scores for correct answers. Hence, the diagnostic value of developmental errors cannot be ignored. The researcher focused on the structure of verbs in the assigned developmental errors of the grammar section of the final English paper exams. Other grammatical structures can be investigated by the use of similar tests. A consideration of the results in several areas would provide a more accurate assessment of students' acquisition of grammar, and hence their competence in English.

The researcher checked 100 answer sheets of the final exams of the academic years from 2010 through 2012. These papers have been chosen randomly. The researcher classified the questions according to the assigned credits. The researcher analyzed the questions in accordance with Bloom's classification (cognition, comprehension, application, analysis, synthesis and evaluation). The analysis generally showed an average of 30 questions for grammar, 30 questions for vocabulary and 20 questions for reading comprehension. The researcher used repetition and percentage to calculate the frequency of questions. He then, categorized them according to the six levels of Bloom's classification.

The Analysis Validity of the Research

Since this study follows the analytical style of the exams prepared by the tutors and instructors teaching the English paper at the College of Social Sciences and Law, the researcher applied the analytical procedures in defining the questions according to their cognitive level. Then he analyzed them in order to calculate the invariability degree of the study. The researcher used (Holistic Equation, 1969), in calculating the constancy degree as follows:

$$\text{Constancy factor} = \frac{\text{Categories number}}{\text{Total of categories analyzed twice}}$$

The results of the analysis showed high correspondence rate. The general constancy degree of the quizzes and tests questions reached 92%, and 94% in the final exams.

The Outcomes of the Analysis

The findings of the analysis showed that the questions of the exams were of cognition and memorization nature. The memorization questions most often appeared in vocabulary sections.   Students should have memorized the meanings of the vocabulary items given in order to be able to answer the blank-filling, multiple-choice, and the matching-of-definitions questions. The percentage rate of memorization in this part reached 97%.   The application nature questions were of lower rate with 81%. These questions were used in grammar and structure sections. The comprehension questions were employed in the reading part of the exam, at a rate of 93%, and the analysis and synthesis sides manifested in the writing section of the exams, which included essay composition and paragraphing analysis, at a rate of 79%.

These figures reflect the fact that the instructors of the English paper rely partially on Bloom's cognitive classification levels due to the type of teaching followed, on the one hand, and to the nature of the syllabi taught, on the other.

In order to confirm the findings of the analysis, the researcher designed a questionnaire composed of 20 questions, and distributed it among a specimen of 300 students from different units from the College of Social Sciences and Law. The goal of this questionnaire was to recognize the opinions of the students with regard to the mechanism of evaluation and testing at these colleges.   The questionnaire contained the following questions:

Disagree %	Agree Partially %	Agree %	Question	
4	25	71	Exam questions are typically designed.	1
3	13	84	Timing of exam is sufficient.	2
38	15	47	Grammar questions are clearly addressed.	3
9	30	61	Vocabulary items are elaborately chosen	4
40	28	32	Essay system exam shows students' real level.	5
8	76	16	Achievement exams can test more fairly.	6
6	10	84	Level of syllabus matches students' level.	7
10	10	80	Tutors and instructors convey information smoothly.	8.
10	22	68	Reading passages are at students' comprehension.	9
4	8	88	Quizzes help students do well at finals.	10
20	24	56	There is a gap between the material given and the questions of the exams.	11
10	16	74	Midterm exam questions are easier than finals.	12
20	11	69	Some instructors are not fair in evaluation.	13
5	20	75	Students usually obtain the grade they expect.	14
65	20	15	System of exams should be reconsidered.	15
79	21	10	Outcome of evaluation does not reflect students' real level in English paper.	16

27	7	66	Supplementary handouts help students much in understanding the given material.	17
2	9	89	Exam questions cover the syllabus equally	18
2	17	81	Exam system is flexible	19
75	3	22	Students are dissatisfied with their level in English after having exams	20

Analysis of the Questionnaire's Findings

Students' responses to the questions were variable. Five points were allocated for each statement. So, the results were calculated by percentage out of 100. The equation of calculation was like this: total of statement points X 5 divided by 100. The above responses generally reflect their satisfaction with the testing and exams system. Any decline in responses can be referred to either misunderstanding of the statement, or to subjective attitudes, and with some answers to indifference of some students in filling out the questionnaire. For these reasons, fractions in calculation process were approximated to regular figures.

Recommendations for Improving Testing in the College of Social Sciences

To improve testing in this college, the researcher suggests the following guidelines:

1. Oral examination should be included as it has a positive backwash effect on teaching. A skill which is not tested is not taught, and consequently, is not learnt. Tutors and instructors of the English paper at the College of Social Sciences and Law neglect the teaching of the two oral skills (Speaking and listening) because students do not have any oral examinations. When oral examinations are included, this will inevitably lead to a change in the current teaching practices. Students, parents and educational authorities will be interested in oral skills as

much as the written ones. As a result, tutors will have to meet new objectives, and will have to respond to the pressure resulting from the innovation in testing by teaching the oral skills.

2. The shortcomings of the objective discrete-point decontextualized items can be overcome by including natural contextualized items. Using contextualized items with integrative objective techniques, such as improved multiple-choice and multiple-matching, information-transfer and cloze procedure, will be good for testing the language structures, vocabulary, reading and listening comprehension, (Milton, 2004).

3. New types of activities should be included, particularly those which students are expected to encounter in the real world: filling in forms, reading different types of signs and procedures, and writing samples of reports and essays.

4. If the means of scoring the writing test are reliable, the composition will appear a more valid way for sampling the students' writing skill than any other objective tests of grammar. To achieve greater inter-rater reliability (correlation of scores obtained from different markers) of the score given to the subjective items, especially the writing section, the multiple-marking technique is advisable.

5. The fifth leg is the post test work and something almost actually worth the time and effort exerted. This stage constitutes scoring, categorizing and analyzing. Thus, these processes need to be given due care and precision, as the outputs of this work of testing will be the inputs for the period to come.

6. Tutors and instructors have to remodel and modify their techniques to remedy up the shortcomings and weak areas in their teaching.

7. Test items should be based on the content and skills that are most important for students to learn. To keep tract of how well tests reflect the objectives, the researcher advises that tutors may construct a grid listing the course content to tailor their tests accordingly.

8. The test should be valid so that the results will be appropriate

and useful for making decision about an aspect of students' achievement, (Gronlund and Linn, 1990). Technically, validity refers to the appropriateness of the interpretation of the results, and not the test itself. Validity is a matter of degree and considered in relation to specific use or interpretation, (Jacobs and Chase, 1992). For example, the results of writing test may have a high degree of validity for indicating the level of students' writing skill. A practical approach is to focus on content validity, the extent to which the content of the test represents an adequate sampling of the knowledge and skills taught over the semester.

9.   The researcher recommends focus on reliability element in test designing. The test can be regarded as reliable if it accurately and consistently evaluates a student's performance. The purest measure of reliability would entail having a group of students who take the same test twice and get the same scores (assuming that their memories do not hold the test items from the first administration). This is impractical, of course, but there are technical procedures for determining reliability. In general, ambiguous questions, unclear directions and vague scoring criteria threaten reliability. Therefore, the researcher advises that the test should be balanced: to cover most of the main ideas and important concepts in purporting to the emphasis they received in lectures.

10.   The researcher insists that since students vary in their preferences for different formats, so using a variety of methods will help students do their best. Multiple-choice or short answer questions are appropriate for assessing students' mastery of details and specific knowledge, while essay questions assess comprehension, the ability to integrate and synthesize, and the ability to apply information to new situations. A single test can have several formats. So, tutors should avoid introducing a new format on the final exam. If students have all multiple-choice quizzes or midterms, they should not have to write an all-essay final, (Sviniki, 1997).

11.   Diversification of questions to cover the inferior cognitive capacities (memorization, comprehension, application), and the paramount cognitive capacities, (analysis, synthesis, evaluation).

12.   Applying different evaluation means, and not focusing on achievement exams as a sole criterion in evaluating students' answers.

13.   Addressing cognitive questions of various levels in the quizzes, tests, and exams.

14.   Carrying out studies related to The evaluation process through analyzing the other evaluation methods which concern measuring the cognition, affection, and proficiency sides with students.

Conclusion

As an epitome, the questions of the English paper exams at the College of Social Sciences and Law in Kuwait University of the academic years 2010 through 2012 were almost at the memorization and application levels, in accordance with Bloom's classification. The questions rate declined from the second level, in Bloom's classification (comprehension), towards the forth level, the analysis. The organization and evaluation questions were not there. Upon comparing between the level of the questions prepared by the instructors with those included in the syllabi, it was shown that the rate of cognition questions designed by the instructors was lower than that in the assigned material, and the same case was with the comprehension questions.

On the other hand, the positive side in the questions of the exams, in general was their suitability to students' level in English. The English paper exams should be tailored in a way that students come up to a higher level of graduation in acquiring the language skills, and not to be designed to come down to lower level.

References

1- Adrienne, S. (2004), Assessing English Learning Process, London: Penguin Publishing Company,   pp. 116-127.

2- Berrenberg, J.L. (2005), Looking Through Bloom's Classification, San Francisco, Jossey Bass, p. 82, pp 430-458.

3- Beyre, Barry K. (2005), Teaching Thinking in Social Studies. Columbus: Charles E. Merrill Publishing Company, pp. 261-288.

4- Brown, S.M. (1991), To Learn Is to Teaching is to Create the Final Exam, College Teaching, New York: Welly, pp. 150-153.

5- Buchanan, R.W., and Rogers, M. (1990), Innovative Assessment in Large Classes, College Teaching, New York: Welly, pp. 201-208.

6- Chapin, Juner R. and Messick, Rosemary G. (1999), Graduate Social Studies: A Practical Guide. New York: Longman, pp. 122-141.

7- Clark, C., and Imerie, J. (2005), Assessing Students, Appraising Teaching, New York: The Rogers, pp. 330-334.

8- Crooks, T.J. (2001), The Impact of Classroom Evaluation Practices on Students, University of Texas, pp. 438-481.

9- Ericken, S.C. (2002), The Teacher-Made Test, Memo to the Faculty, No. 36, Ann Arbor: Center for Research on Learning and Teaching, University of South Dakota, p. 127.

10- Gregor, T. (1998), Test Partners: A Formula for Success, Published by College of Education, University of Leads, pp. 168-180.

11- Gronlund, N.E., and Linn, R. (1990), Measurement and Evaluation in Teaching, pp. 135-145.

12- Hendrickeon, A.D. (1990), Cooperative Group Test-Taking, Focus, 5 (2), 6, Publication of the Office of Educational Development Programs, University of Minnesota, p. 22.

13- Jacobs, L.C., and Chase, C.I. (1992), Developing and Using Tests Effectively: A Guide for Faculty, San Francisco: Jossy-Bass, pp. 321-348.

14- Kuwait University, English Language (ELU), College of Social Sciences, Statistics Section, Registers since 2000 through 2006.

15- Liska, M. & Simson, H. (2001), Open-Note Exams, Teaching Professor, New York: Macmillan, p. 128, pp. 222-246.

16- MaCclain, O.R. (2005), Content Analysis for the Social Sciences, Addison, Canada, Wesley Publishing House, p. 137, pp. 140-141.

17- Madline, P.S. (2003), The Group Exam, Teaching Professor, 8 (6).

18- Marrlow, D. (2002), Developing and Using Tests Effectively, Cambridge Press, pp, 401-425.

19- Marterrola, John M. (1998), English as a Foreign Language at Higher Studies. New York: Macmillan Publishing Company, pp. 278-299.

20- Maurice J. and Metcalf, Louis A. (2004), Problems in Reflective Thinking and Social Understanding, New York: Macmillan Publishing Company, pp. 318-361.

21- Milton, H., and Eison, J.A. (2004), Making Sense of College Grades, North Carolina: Emerson Publishing House, pp. 223-226.

22- Oller, D.L. (2002), An Experiment with Cloze Tests, McGraw Hill, pp. 54-79.

23- Olmer, U.g. (2005), Testing English as a Second Language, Longman Education, p. 289.

24- Peter, W.Q. (2000), Linguistic Measures of Error Gravity, Publication of University of Yorkshire, pp. 241-277.

25- Perren, S.B. (2004), Discrete-point Test of Integrative Skills, Dublin: Newbury House, pp. 330-345.

26- Sam, H.F. (2001), Effects of Easy-Examination and Multiple-Choice Questions, Journal of Instructional Psychology, 15 (4), 7-11.

27- Schwiser, K.G. (2003), Designing an Authentic Assessment, Educational Leadership, Vol. 43, No. 7, p. 38, pp.49-67.

28- Syurry, T. (2000), Grading & Evaluation, Harvard University Press, pp. 267-288.

29- Vaughn Y.L. (2005), English Language Testing: Some Unsolved Problems, Longman, pp. 228.

30- William, V.R. (2000), A Practical Handbook for College Teachers, Boston: Little Brown, pp. 302-321.