Arab World English Journal
مجلة اللغة الانكليزية في العالم العربي

Volume 2, Number 1 **pp.47-71**

Predicting Foundation Year Students' Performance on International Proficiency Tests

Ahmed Shaker Al Kilabi, PhD
Department of English Language
and Literature,
College of Applied Sciences, Nizwa
Laila Zahir Al-Salmi
University of Texas, El Paso.U.S.A

Abstract

The present study has two aims. First, it attempts to validate secondary school scores by using data collected from 347 foundation year students' performance on six different tests in addition to their secondary school scores. The hypothesis underlying this validation was that the secondary school scores were inflated and they did not reflect the real level of students' proficiency. The Secondary School Scores were found to have moderately high correlation co-efficients with all other six sets of scores. These scores, if proved valid, could be used as indicators of students' level of performance on any international general proficiency test such as the IELTS. This study also endeavors to predict students' level of performance by using any one of the seven sets of scores for the sample of the study. A pilot study was conducted involving 35 students who took the IELTS in May 2006. The results of the pilot study showed that the only good predictors of the level of performance on IELTS were a computer-based proficiency test and a May 2006 Achievement test.

Key Words: assessment, Achievement test, proficiency test, Foundation Programs, IELTS

Predicting Foundation Year Students' Performance on International Proficiency Tests

Test construction has become part of everyday life. It is no longer restricted to the work of psychologists and educators, but extended to cover a wide spectrum of specializations and interests. People are being tested by various methods for several purposes including determining eligibility for university enrollment.

The validation of a language test has witnessed considerable developments and the accurate validation of language tests has become one of the major concerns preoccupying foreign/second language teachers and researchers alike. At a practical level, interpretations, or inferences, are considered at two levels of generality: on the basis of the total score for the test and at the level of individual items. Second, it is emphasized that validation is not a one-off exercise but an ongoing process. The third principle is that validation involves the accumulation of various kinds of evidence to support the anticipated interpretations of the test scores under the general rubric of construct validity.

The present study endeavors to deal with the possibility of replacing an international test with a locally designed one since the purpose is to evaluate students' proficiency in English in order to join an academic program in the Colleges of Applied Sciences in Oman. This study aims to investigate and validate secondary school scores and to establish the reliability of the criterion variables used in the study. It also aims to predict students' level of performance by using any one of the seven sets of scores obtained for the sample of the study.

Rationale and Significance

The study was undertaken in response to the dire need for a valid and reliable test that could be used as a yardstick in measuring the proficiency in English of Omani secondary school leavers. In order

to attain that goal the study draws on the analysis of results obtained from two semesters, representing seven sets of scores. However, it is worth mentioning that the ideal assessment of Omani students' level of proficiency in English, for those embarking on university courses of applied sciences, may lie in taking the IELTS or any other internationally recognized test such as TOEFL. The biggest hurdle in the way of attaining this is the large number of students who need to take the test every year, and the amount of money required not only for the test fees but also for covering students' various expenses when they come from distant areas of the Sultanate.

The present study represents an attempt to explore the possibility of building up a test or a battery of tests locally - or using one of several tests developed locally. These tests can be validated so they can be used as a standard measure to evaluate students' level of proficiency in English in a reliable and valid manner.

Literature Review

A considerable number of studies were conducted in the field of test validation and this in itself reflects the paramount importance of validating the tests for the purpose of using them in educational institutions. The following survey focuses on previous studies that are relevant to the purposes and aims of the present study.

Ito (2005) conducted a validation study on the English language test of the Japanese nationwide university entrance examination; the Joint First Stage Achievement Test (JFSAT). The study investigated the reliability and validity of the most widely taken English language test in Japan. Two studies were reported. The first examined the reliability and concurrent validity of the JFSAT-English test. The reliability was reported to be acceptable. Criterion validity was estimated by correlating the JFSAT-English test and the English Language Ability Measure (a carefully constructed cloze test) and was found

to be satisfactory. The second study reported on a construct validation study on the test through an internal correlation study. The JFSAT-English test was divided into five subtests. Examination of the correlation matrix indicated that the paper-pencil pronunciation test had low validity with almost no significant contribution to the total test score. . Ito (2005) argued that the JFSAT-English test could be used as a reliable and to a certain extent valid measure of English language ability; one of its 5 subtests, the paper-pencil pronunciation test, reported low validity compared to the other 4 subtests. Yet the paper-pencil pronunciation test should be eliminated and a listening comprehension test might be included as one of the subtests in the JFSAT-English test.

Eckes et al. (2005) reported on contributions from seven European countries that pinpoint major projects, problems, and prospects of reforming public language assessment procedures. Each country has faced unique problems in the reform process, yet there have also been several common themes emerging, such as a focus on multilingualism, communicative skills, standardization, references to the Common European Framework of Reference for Languages (CEFR) and certification. Eckes et al suggested that future work needs to develop these themes further and to study impact and support issues as well as standardization and validation. Accordingly, tests should be developed in terms of clearly defined specifications. Language assessment requires measuring instruments constructed on the basis of sound psychometric criteria and appropriately chosen test methods. Reforming language assessment practices needs to be embedded in continuing efforts at establishing the highest quality possible, encompassing all relevant aspects ranging from the objective measurement of examinee proficiency and item difficulty to precise definitions of test administration conditions and scoring systems.

Jin and Yang (2006) reviewed the College English Test (CET), the world's largest language test administered in China nationwide, in terms of examinee population and its development, score

interpretation and test validation made it well received and established as a large scale standardized EFL test inside and outside China. The study adopted an in-depth analysis of the test content, test format, and candidates' performances on each component in order to present an overview of the proficiency of the CET test-takers. The authors discussed the issue of reform of the CET as a response to the pressing social need for college and university graduates with a stronger communicative competence in English. However, in spite of the challenges facing the CET in relation to measuring the students' speaking and listening abilities, the test is still recognized for its consistent marking, rigorous administration and comparable scores. The CET has now become well-established as a large scale standardized EFL test. It is held in high esteem by language testers and teachers inside and outside China and is well received by the public. The study suggests how to find a way forward for the CET to meet the ever-changing needs of society.

AsSarmi and AlHajri (2006) examined the results of the final achievement test that was administered in the five newly transformed colleges of applied sciences in Oman in May 2006. The test was administered to 1650 male and female students in these colleges. The study attempted to determine the cut-off point that was equivalent to the Level 4.0 on the IELTS, to establish the relationship between students' scores in their final exam of the general secondary certificate and their scores on the achievement test and to determine the predictive power of certain variables such as the diagnostic tests to predict students' scores on the Foundation Year Program achievement test. It was found that the cut-off point was 58% on the achievement test which was equivalent to Level 4.0 on the IELTS. The researchers reported a positive and moderately high relationship between the diagnostic test and the final achievement test ($r = .68$). This means that students who performed well on the diagnostic test obtained higher marks on the achievement test. It was reported in the study that students' scores on the final exam

of the general secondary school certificate were positively correlated to the final achievement test **scores** (r = .50).

The Present Study

Subjects

The sample of the study consists of 347 male and female students who studied English in the Foundation Year 2005/2006 at Nizwa College of Applied Sciences. Apart from gender and level of proficiency in English, the students were homogenous in terms of age, learning experience (all of them spent 8 years of studying English at schools) , and mother tongue (i.e. Arabic).

Hypotheses

This study tested two competing hypotheses about the relationship between test scores and students' proficiency in English. It was postulated that:

1. Secondary School Scores are highly inflated and they do not reflect the real level of students' proficiency in English.
2. These scores, if proved valid, could be used as indicators of students' level of performance on any international general proficiency test.

Description of the Tests

The scores obtained from a set of eight tests were used for the purpose of the present study. They are Nizwa College Teachers' Evaluation scores, Students' Secondary School Scores, Pencil & Paper Test, Computer-based Test, October 2005 Diagnostic Test, January 2006 Achievement Test, May 2006 Achievement Test and IELTS. However, the IELTS test scores were obtained from the sample students who sat for the test for the purpose of deciding on the bench mark that would determine who could move up to the Colleges of Applied Sciences Academic Program. The sample consisted of 160 students from 5 Colleges. However, only the 35 sample of student who study at the College of Applied Sciences at Nizwa

was used in the pilot study. The October 2005 Diagnostic test, the January 2006 Achievement test and the (Final) May 2006 Achievement test were set by the Ministry of Higher Education whereas the other sets of scores, not counting the IELTS, were designed by the college for continuous assessment purposes.

The October 2005 Diagnostic Test was conducted in October 2005. This test consisted of three parts: Listening, Grammar and Writing. The listening component was divided into two sections which contained 19 questions. The first section contained 9 questions in which the students were asked to listen to sentences and choose one picture (from a set of four pictures) that portrays the sentence. For the second section, the students listened to words and marked the right word. The total score for this component was 25. The reading component consisted of comprehension questions as well as grammar questions. It was divided into four sections (total number of questions 40 MCQ, total score 20). The writing part was in two sections. In the first section, the students were asked to write 5 sentences using the correct given tenses. In the second section, the students were asked to write two paragraphs on two given topics. The total score for this component was 30. The test was conducted in one day and lasted for two hours and a half.

The January 2006 Achievement Test was conducted in January 2006 and it consisted of 4 components: Listening, Reading, Grammar and Writing. The test was conducted in two days having listening and reading on the first day and grammar and writing on the second. The listening component contained 20 questions divided into 2 parts. The first part had two tasks. The first task contained 5 sentences; the students had to listen to an interview and then decided whether the sentences were true or false. For the second task, the students listened again to the same interview and accordingly decided if the listed five activities were useful for teachers or students. The second section had two tasks as well. In the first task the students listened to an extract from a lesson and then they were asked to answer 5 multiple choice questions. In the second task, the students listened again to the same extract and answered five

short answer questions. The total score for this component was 20. The reading component consisted of two texts. Text one included two tasks; the first was multiple choice questions and the second was locating information in paragraphs. The second text also included two tasks. The first was true or false statements and the second was short answer questions. The total mark for this section was 20. The grammar test included 10 parts with a total mark of 40. In the writing section the students were given pictures and asked to write a narrative about what happened to the character in the pictures. The total score for this component was 15. The test lasted for four hours.

The Final Achievement Test was conducted in May 2006. The test consisted of three parts: a listening test, a reading test and a writing test. The listening test included two sections; in each section there were two tasks. The questions varied between true and false questions, multiple choice questions and completion questions. The total mark for this component was 25. The reading test also consisted of two sections and in each section there were two tasks. The first task in the first section asked the students to put sentences in the order they appeared in the reading text and the second task included short answer questions. On the other hand, the first task in the second section was a matching question and the second task included true or false questions. The total for this component was 25. The writing test included two writing tasks. The first task asked the students to describe their country (Oman) to a visitor and the second task was about the students' opinion about TV providing supporting reasons for their opinion. The total for this component was 25. This test was conducted in one day and lasted for 3 hours.

Validity

Bachman (1990) contended that the most important quality of test interpretation or use is validity, or the extent to which the inferences or decisions we make on the basis of test scores are meaningful, appropriate, and useful.

In the same vein of thought, Messick (1989) defined validity as a unitary but multi-faceted concept. Accordingly, testing may result in inferences being made about test-takers' abilities, knowledge, or performance and in decisions being made about admitting test-takers to a course or hiring them for a job. The consequential aspects of validity include wash-back and social responsibility. Consequences refer to the value implications of the interpretations made from test scores and the social consequences of test use. More recently, Messick (1996) suggested that "Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores".

Test Validation

The process of validation starts with the inferences that are drawn and the uses that are made of scores. These uses and inferences dictate the kinds of evidence and logical arguments that are required to support judgments regarding validity (Kunnan, 1998).

According to McNamara (2000), test validation is the process of investigating the defensibility of the inferences about candidates that have been made on the basis of test performance, and it is the main focus of testing research. Furthermore, McNamara (1996, 2000) suggested trialing the test materials and procedures prior to their use under operational conditions. This stage involves careful design of data collection to see how well the test is working. A trial population will have to be found, that is, a group of people who resemble in all relevant respects (age, learning background, general proficiency level, etc.) the target test population. (AlKilabi, 2004)

Kunnan (1999) argued that the purpose of test validation in language testing is to ensure the defensibility and fairness of interpretations based on test performance. It also seeks to establish a solid basis according to which we propose that individuals are admitted or denied access to the criterion setting being sought, and to ensure that it is a sufficient and fair basis. In this respect, McNamara (2000) stated:

> *Test validation involves thinking about the logic of the test, particularly its design and its intentions, and also involves looking at empirical evidence – the hard facts – emerging from data from test trials or operational administrations. If no validation procedures are available there is potential for unfairness and injustice. This potential is significant in proportion to what is at stake. (p. 32)*

Reliability

The reliability of the Pencil and Paper and Computer-based tests was determined by using the split half method. Co-efficient values of 0.7731 and 0.7534 were obtained for the first and second respectively, representing the reliability co-efficients of these two tests (1).

The item-total score reliability was used to determine item reliability for each one of the three tests (October 2005 Diagnostic Test; January 2006 Achievement Test; May 2006 Achievement Test). This was achieved by using the product moment correlation between marks for an item and the corresponding total marks gained by the subjects on all other items in the test (Bachman, 1990). In the light of the correlation coefficients, the three tests can be considered reliable tests.

Table 1: Internal Consistency of the October 2005 Diagnostic Test

		LISTNING	GRAMMAR	TENSES	WRITING1	WRITING2	TOTAL
LISTNING	Pearson Correlation	1	.427	.436	.479	.518	.792
	Sig. (2-tailed)	.	.000	.000	.000	.000	.000
	N	337	337	337	337	337	337
GRAMMAR	Pearson Correlation	.427	1	.394	.354	.371	.609
	Sig. (2-tailed)	.000	.	.000	.000	.000	.000
	N	337	337	337	337	337	337
TENSES	Pearson Correlation	.436	.394	1	.400	.381	.660
	Sig. (2-tailed)	.000	.000	.	.000	.000	.000
	N	337	337	337	337	337	337
WRITING1	Pearson Correlation	.479	.354	.400	1	.798	.815

	Sig. (2-tailed)	.000	.000	.000	.	.000	.000
	N	337	337	337	337	337	337
WRITING2	Pearson Correlation	.518	.371	.381	.798	1	.828
	Sig. (2-tailed)	.000	.000	.000	.000	.	.000
	N	337	337	337	337	337	337
TOTAL	Pearson Correlation	.792	.609	.660	.815	.828	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.
	N	337	337	337	337	337	337

** Correlation is significant at the 0.01 level (2-tailed).

Table 2: Internal Consistency of the January 2006 Achievement Test

		SPEAKING	LISTENING	READING	GRAMMAR	WRITING	TOTAL
SPEAKING	Pearson Correlation	1	.406	.368	.401	.412	.627
	Sig. (2-tailed)	.	.000	.000	.000	.000	.000
	N	346	346	346	345	346	346
LISTENING	Pearson Correlation	.406	1	.514	.386	.379	.720
	Sig. (2-tailed)	.000	.	.000	.000	.000	.000
	N	346	346	346	345	346	346
READING	Pearson Correlation	.368	.514	1	.512	.420	.728
	Sig. (2-tailed)	.000	.000	.	.000	.000	.000
	N	346	346	346	345	346	346
GRAMMAR	Pearson Correlation	.401	.386	.512	1	.576	.808
	Sig. (2-tailed)	.000	.000	.000	.	.000	.000
	N	345	345	345	345	345	345
WRITING	Pearson Correlation	.412	.379	.420	.576	1	.736
	Sig. (2-tailed)	.000	.000	.000	.000	.	.000
	N	346	346	346	345	346	346
TOTAL	Pearson Correlation	.627	.720	.728	.808	.736	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.
	N	346	346	346	345	346	346

** Correlation is significant at the 0.01 level (2-tailed).

Table 3: Internal Consistency of the May 2006 Achievement Test

		SPEAKING	LISTENING	READING	WRITING	TOTAL
SPEAKING	Pearson Correlation	1	.291	.408	.386	.672
	Sig. (2-tailed)	.	.000	.000	.000	.000
	N	347	347	347	347	347
LISTENING	Pearson Correlation	.291	1	.298	.255	.659
	Sig. (2-tailed)	.000	.	.000	.000	.000
	N	347	347	347	347	347
READING	Pearson Correlation	.408	.298	1	.310	.672
	Sig. (2-tailed)	.000	.000	.	.000	.000
	N	347	347	347	347	347
WRITING	Pearson Correlation	.386	.255	.310	1	.746
	Sig. (2-tailed)	.000	.000	.000	.	.000
	N	347	347	347	347	347

TOTAL	**Pearson Correlation**	**.672**	**.659**	**.672**	**.746**	**1**
	Sig. (2-tailed)	**.000**	**.000**	**.000**	**.000**	**.**
	N	**347**	**347**	**347**	**347**	**347**

** Correlation is significant at the 0.01 level (2-tailed).

Statistical Procedures

Two statistical procedures were used to achieve the objectives of the present study. The Pearson product moment correlation co-efficient was used to establish the relationship among the seven tests that were given to the students. Tables 1, 2 and 3 above show the matrix of correlations among the items of some of these tests, whereas Tables 4 and 7 give the matrix of correlations among the different sets of scores used in the Pilot and Main Studies respectively.

The multiple regression analysis is usually used in discovering the best predictors of the students' performance on the criterion variable (Bachman, 2004). The step-wise regression program was used to identify the predictor variables of students' performance on the May 2006 Achievement Test.

The Pilot Study

A pilot study was conducted involving 35 students who took the IELTS in May 2006. These students were randomly selected from the population of the main study and used as a sample in the pilot study. The purpose of this study was to establish the relationship between the IELTS scores and the scores obtained by students on the remaining tests.

Table 4 below shows the values of correlation among the variables used in the pilot study and it can be seen that almost all the variables have strong positive relationship with the IELTS, except for the variable "Secondary School Scores" which has a weak relationship; use it attained a non-significant value of correlation (=.313). On the other hand, the May 2006 Achievement Test, the Computer-based Test and Teachers' Evaluation are all strongly correlated with the IELTS score (r=.606), (r=.580) and (r=.559) respectively.

It is worth mentioning that the secondary school scores obtained reasonably high correlation co-efficients with most of the sets of test scores used in the Pilot Study, e.g., Teachers' Evaluation (r = .480), Diagnostic Test (r = .512), Achievement Test Jan (r =.521) and Computer-based Test (r = .548). `

Table 4. Correlation Co-efficients among the Eight Variables of the Pilot Study

		Teacher's Evaluation	Secondary School Scores	Oct 2006 Diagnostic Test	Jan 2006 Achievement Test	Pencil & Paper Test	Computer-Based Test	May 2006 Achievement Test	IELTS
Teacher's Evaluation	Pearson Correlation	**1**	**.480***	**.422**	**.767****	**.417**	**.635****	**.573****	**.559****
	Sig. (2-tailed)	.	.004	.013	.000	.024	.000	.000	.000
	N	35	34	34	34	29	30	35	35
Secondary School Scores	Pearson Correlation	**.480***	**1**	**.512***	**.521***	**.219**	**.548***	**.348**	**.313**
	Sig. (2-tailed)	.004	.	.002	.002	.263	.002	.043	.072
	N	34	34	34	34	28	30	34	34
Oct 2006 Diagnostic Test	Pearson Correlation	**.422***	**.512***	**1**	**.631****	**.358**	**.446**	**.389**	**.518***
	Sig. (2-tailed)	.013	.002	.	.000	.062	.014	.023	.002
	N	34	34	34	34	28	30	34	34
Jan 2006 Achievement Test	Pearson Correlation	**.767****	**.521***	**.631****	**1**	**.489***	**.517***	**.621****	**.536****
	Sig. (2-tailed)	.000	.002	.000	.	.008	.003	.000	.001
	N	34	34	34	34	28	30	34	34
Pencil & Paper Test	Pearson Correlation	**.417***	**.219**	**.358**	**.489***	**1**	**.359**	**.386***	**.411***
	Sig. (2-tailed)	.024	.263	.062	.008	.	.072	.039	.027
	N	29	28	28	28	29	26	29	29
Computer-Based Test	Pearson Correlation	**.635****	**.548***	**.446***	**.517***	**.359**	**1**	**.504***	**.580****
	Sig. (2-tailed)	.000	.002	.014	.003	.072	.	.005	.001
	N	30	30	30	30	26	30	30	30
May 2006 Achievement Test	Pearson Correlation	**.573****	**.348***	**.389***	**.621**	**.386***	**.504***	**1**	**.606****
	Sig. (2-tailed)	.000	.043	.023	.000	.039	.005	.	.000
	N	35	34	34	34	29	30	35	35
IELTS	Pearson Correlation	**.559****	**.313**	**.518***	**.536**	**.411***	**.580****	**.606****	**1**
	Sig. (2-tailed)	.000	.072	.002	.001	.027	.001	.000	.
	N	35	34	34	34	29	30	35	35

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

It was also intended to investigate the predictive power of other variables in relation to the IELTS scores. Seven variables were used in the equation and the IELTS Scores were used as the dependent variable.

The results of the pilot study also showed that the only predictors of the level of performance on IELTS were the computer-based Test and the May 2006 Achievement Test. The remaining five variables were excluded from the equation and hence did not contribute to the prediction of the dependent variable.

Table 5. Summary Table of Multiple Regression Analysis (Pilot Study)

Step No.	Variables Entering	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
						R Square Change	F Change	df1	df2	Level of Significance
1	CBT	.589	.347	.327	4.39590	.347	17.010	1	32	.000
2	MAT	.670	.449	.413	4.10316	.102	5.729	1	31	.023

To further corroborate the results of the pilot study, the researchers used the scores of May 2006 Achievement Test and IELTS collected for 160 male and female students who took the IELTS on May 20, 2006. These students were from the five newly transformed colleges.

Table 6. Correlations of Students' Scores on the May 2006 Achievement Test and IELTS

		May 2006 Achievement Test	IELTS
IELTS	Pearson Correlation	1	.595
	Sig. (2-tailed)	.	.000
	N	160	160
May 2006 Achievement Test	Pearson Correlation	.595	1
	Sig. (2-tailed)	.000	.
	N	160	160

** Correlation is significant at the 0.01 level (2-tailed).

The results showed a moderately high correlation value of .59 between May 2006 Achievement Test and the IELTS.

Results and Discussion

In order to test the first hypothesis, which stated that "Secondary School Scores are highly inflated and they do not reflect the real level of student's proficiency in English," a comparison was drawn between secondary school scores for the sample of the study (347 male and female students) and their scores on six college tests, i.e., October Diagnostic Test, Teachers' Evaluation, January 2006 Achievement Test, May 2006 Achievement Test, Pencil-and-paper Test, and Computer-based Test.

As can be seen from Table 7 below, the mean score of Secondary School Scores was found to be the highest compared to the mean of any set of scores given in the table. It may be noted that the mean score of the secondary school scores is 79.76 whereas it is only 50.64 on the October Diagnostic Test, the first test the students took in the college when they were first admitted to the foundation year program. The mean score of the January 2006 Achievement Test is 61.82. Students obtained even lower mean scores on other tests that were administered in the college such as the pencil-and paper and the computer-based test (46.45 and 46.42 respectively).

Table 7. A comparison of Secondary School Scores with College Tests

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
Secscore	347	48.50	50.00	98.50	79.7622	10.58575	112.058	-.639	.131
Octtest	347	66.00	18.50	85.50	50.6464	10.50386	110.331	-.074	.131
Jantest	347	66.00	28.00	94.00	61.8271	11.48430	131.889	-.251	.131
Maytest	347	66.30	32.50	98.80	71.9784	9.43375	88.996	-.620	.131
Teaevalu	347	70.06	25.88	99.94	65.6919	2.32280	5.395	-.192	.131
Papencil	347	32.00	12.00	76.00	46.4582	5.57247	31.052	-.117	.131
CBT	347	32.00	14.00	78.00	46.4236	5.47214	29.944	.071	.131

In the light of the above result, the first hypothesis was accepted. This means that there is sufficient evidence to support the postulation that secondary school scores were highly inflated

and they did not reflect the real level of students' proficiency in English. In other words, students usually tend to obtain lower scores on other English tests than the general secondary school examination.

In order to test the second hypothesis, which stated that "These scores, if proved valid, could be used as indicators of students' level of performance on any international general proficiency test," Secondary School Scores were validated using the Concurrent Criterion Relatedness method (concurrent validity). The other tests were used as external criteria to establish the validity of the secondary school scores (Weir 1990). Table 7 shows the criterion-related evidence that demonstrates a relationship between test scores and some criterion which is believed to be also an indicator of the ability tested. According to McNamara (1996), concurrent validity is a kind of criterion-related validity which is obtained through concurrent administration of a newly developed test with another well-known standardized test of which the validity is already established.

However, a cursory look at the correlation matrix shown in Table 8 below would reveal that the secondary school scores have consistently obtained high and positive correlation values with most of the other test scores, e.g., Teachers' Evaluation (r = .723), Diagnostic Test (r = .709), Achievement Test Jan (r = .735), and Achievement Test May (r = .703). The secondary school scores also obtained moderately high correlation co-efficients with the remaining two sets of test scores, i.e., Pencil and Paper Test (r=.452) and Computer-based Test (r = .468). Therefore, in the light of the moderately high correlation co-efficients, it can be stated that the secondary school scores have a good degree of association with the sets of scores used in the study. Hence, the second hypothesis was accepted.

This result is corroborated by the findings that were reported by AsSarmi and AlHajri (2006) where they emphasized the positive and high correlation between the secondary school scores and the Achievement Test May (r = .50)

It must be pointed out that these scores should not be discarded as useless altogether. Actually, they can be used as an indicator of students' level of proficiency in English if the problem of scores' inflation is solved. The common observation of the researchers is that students who enter the Foundation Programme with higher scores on the final exam of the general secondary certificate tend to have a greater chance of improving their level of English and obtaining better results on the tests administered in the college. Therefore, these scores usually give higher values than the real levels of the students are.

Table 8. Correlation Co-efficients among the Seven Variables of the Main Study

		Teacher's Evaluation	Secondary School Score	Diagnostic Test	Achievement Test Jan	Pencil & Paper Test	Computer-Based Test	Achievement Test May
Teacher's Evaluation	Pearson Correlation	**1**	**.723****	**.695****	**.738****	**.524****	**.482****	**.693****
	Sig. (2-tailed)	.	.000	.000	.000	.000	.000	.000
	N	347	347	347	347	347	347	347
Secondary School Score	Pearson Correlation	**.723****	**1**	**.709****	**.735****	**.452****	**.468****	**.703****
	Sig. (2-tailed)	.000	.	.000	.000	.000	.000	.000
	N	347	347	347	347	347	347	347
Diagnostic Test	Pearson Correlation	**.695****	**.709****	**1**	**.716****	**.420****	**.454****	**.646****
	Sig. (2-tailed)	.000	.000	.	.000	.000	.000	.000
	N	347	347	347	347	347	347	347
Achievement Test Jan	Pearson Correlation	**.738****	**.735****	**.716****	**1**	**.488****	**.525****	**.715****
	Sig. (2-tailed)	.000	.000	.000	.	.000	.000	.000
	N	347	347	347	347	347	347	347
Pencil & Paper Test	Pearson Correlation	**.524****	**.452****	**.420****	**.488****	**1**	**.647****	**.452****
	Sig. (2-tailed)	.000	.000	.000	.000	.	.000	.000
	N	347	347	347	347	347	347	347
Computer-Based Test	Pearson Correlation	**.482****	**.468****	**.454****	**.525****	**.647****	**1**	**.494****
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.	.000
	N	347	347	347	347	347	347	347
Achievement Test May	Pearson Correlation	**.693****	**.703****	**.646****	**.715****	**.452****	**.494**	**1**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.
	N	347	347	347	347	347	347	347

However, in the Pilot Study these scores were found to have obtained low correlation co-efficients with the Achievement Test May (r = .348) significant at the 0.05 level. This may be explained by the fact that the sample used in the Pilot Study was relatively small (i.e. 35 students), whereas correlational studies usually require larger samples. However, the researchers believe that further studies with larger samples may reveal a connection between the students' achievement in these tests which could have been possible if the sample who sat of the IELTS exam was larger.

In an attempt to achieve the second major objective of the study, i.e., to predict students' level of performance by using any one of the six sets of scores for the sample of the study, the Stepwise multiple regression was used for that purpose.

In the main study, the May 2006 Achievement Test was used in the regression equation as the dependent variable, whereas the other variables such as the October 2005 Diagnostic Test, the January 2006 Achievement Test, Secondary School Scores, Teachers' Evaluation, Paper and Pencil Test, and Computer-based Test were used as predictors. The IELTS results were not used in the main study because only thirty five students from Nizwa College of Applied Sciences took that test, which is rather a limited number. The researchers were ambitious to get the IELTS scores for the 160 students from the five colleges who took the international test, but their efforts to get these results from the concerned colleges came to no avail. That may explain why they had to content themselves with the data available to them and exclude the IELTS scores from the main study.

Table 9 below gives the result of the regression equation. As can be seen from that table, only four variables were entered into the equation. These variables were the January 2006

Achievement Test (ATJ), Secondary School Scores (SSS), Teachers' Evaluation (TE) and Computer-based Test (CBT). The remaining two tests, i.e., October 2005 Diagnostic Test and Pencil and Paper Test were not entered into the equation and these two variables were excluded. The results shown in the table below indicate that 50.9% of the variance in students' performance on May 2006 Achievement Test can be explained by their performance on the January 2006 Achievement Test. It may also be noted that Secondary School Scores contributed to 6.6% of the variance in students' performance on the final test.

Table 9. Summary Table of Multiple Regression Analysis (Main Study)

Step No.	Variables Entering	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
						R Square Change	F Change	df1	df2	Level of Significance
1	ATJ	.715	.511	.509	6.60841	.511	360.102	1	345	.000
2	SSS	.761	.579	.577	6.13598	.069	56.170	1	344	.000
3	TE	.777	.603	.600	5.96687	.024	20.775	1	343	.000
4	CBT	.781	.611	.606	5.92033	.007	6.415	1	342	.012

The major finding of the study is that the January 2006 Achievement Test is a good predictor of students' level of performance on the final test. The Secondary School Scores come second in predicting students' performance on the same test.

The other two predictors, namely teachers' evaluation and the computer-based test, contributed slightly to the prediction of students' performance on the final test.

Conclusion

The secondary school scores were found to be unrealistic measures of students' level of proficiency because they were highly inflated. In other words, these scores tend to carry bigger values than the actual levels of the student are and, therefore, they are not indicative of the actual level of proficiency the students have attained after leaving the secondary school.

The findings of the study indicate that some of the criterion variables were found to be good predictors of students' level of performance, e.g., the Computer-based test and May 2006 Achievement Test in the pilot study and January 2006 Achievement Test and the secondary school scores in the main study. Both January 2006 and May 2006 Achievement tests were used in the common and simultaneous assessment of students' level of performance.

The secondary final examination scores in English can be made more realistic and reliable by improving the examination paper and the marking procedures followed by secondary school teachers. By the same token, this could lead to the more arduous task, i.e., secondary school examinations could be improved by initiating a validation process in order to make the scores obtained from these exams reliable predictors and real indicators of students' level of proficiency in English.

The researchers believe that the validation process, which needs to be applied to the secondary school examination in English, is worth the time and effort that would be invested in it. This would surely take into account the interest of the students and the benefits that they may gain from using realistic and reliable scores in their final evaluation. Once this process is in effect, it is expected that the streaming of the students in their foundation year can be carried out using the validated scores.

End Notes:

(1) According to this method, the subjects take the test in the usual way, but each subject is given two scores. One score is for one half of the test, the other second score is for the other half. The two sets of scores are then used to obtain the reliability coefficient as if the whole test had been taken twice. In order for this method to work, it is necessary for the test to be split into two halves which are really equivalent, through the careful matching of items. In fact, where items in the test have been ordered in terms of difficulty, a split into odd-numbered items and even-numbered items may be adequate. (Grunlund and Linn, 1990).

References

AlKilabi, A. (2004). Testing the tests: Validating two placement tests. *Jordan Journal of Applied Science, 7*, (2), 1-19.

AsSarmi, A. M. & AlHajri A. A. (2006). *A Study of the results of the Foundation Year Programme 2005/2006 in the specialised colleges.* Unpublished manuscript. Ministry of Higher Education, Muscat, Oman. (In Arabic).

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Eckes, T., Ellis, M., Kalnberzina, V., Pi˘zorn, K., Springer C., Szollás, K. & Tsagari, C. (2005). Progress and problems in reforming public language examinations in Europe: cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France and Germany. *Language Testing, 22,* (3), 355–377

Gronlund N. E. and Linn R. L. (1990). *Measurement and evaluation in teaching.* 6th Ed., New York: Macmillan Publishing Company,77.

Ito, A. (2005). A validation study on the English language test in a Japanese Nationwide University entrance examination. *The Asian EFL Journal Quarterly, 7,* (2), 90-116.

Jin Y. & Yang, H. (2006). The English proficiency of college and university students in China: As reflected in the CET. *Language, Culture and Curriculum, 19,* (1), 21-36.

Kunnan, A. J. (1998). Approaches to validation in language assessment. In A J Kunnan (ed.) *Validation in Language Assessment: Selected papers from the 17th Language Testing Research Colloquium*, Long Beach. Mahwah, NJ.: Lawrence Erlbaum Associates.

Kunnan, A. J. (1999). Recent developments in language testing. In W. Grabe et al. (eds.) Annual Review of Applied Linguistics, *A Survey of Applied Linguistics. 19,* 235-253.

McNamara, T. (1996) *Measuring second language performance*. London: Longman.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press. 32-48.

Messick, S. (1989). Validity. In R. L Linn (ed.). *Educational measurement (3rd Ed.).* New York: Macmillan. 13-103.

Weir, C. J. (2005). *Language testing and validation: an evidence-based approach*. Basingstoke: Palgrave Macmillan.