EFL Proficiency Testing at the University of Dar es Salaam: Performance of Candidates

Erasmus Akiley Msuya

Department of Foreign Languages and Linguistics, University of Dar es Salaam,P.O. Box 35040, Dar es Salaam, Tanzania

Abstract

This study sought to measure the University of Dar es Salaam EFL students' proficiency so as to get deeper and comprehensive insights of the candidates' variability in their linguistic ability across test areas, namely; comprehensive reading, writing, listening, and grammar and vocabulary. The variability was in terms of sex and level of education. The study adopted a framework of measurement of proficient levels developed by American Council on the Teaching of Foreign Languages (ACTFL) (2012) in which learners are rated at four levels: superior, advanced, intermediate and novice and in which the last three were further subdivided into high, mid and low. 136 Udsm EFL participants were involved in the study. These were test takers who sat for the University of Dar es Salaam proficiency test at different times between 2009 and 2013. These were of different education backgrounds and were picked randomly using their test scripts.The findings showed that, in the whole, the students' performance was good since all groups of candidates performance ranged from 'intermediate' to 'intermediate proficient levels with Udsm alumni taking the lead with a mean of 85.5% while ELT short term students were the last with a mean score of 56.3%. In terms of gender, males outperformed females in the four out of five groups, even though the difference was only marginal. As for the content areas that were tested, the candidates' performance was the highest in the area of vocabulary where their overall mean score was 83 while listening was the most underperformed content area with the mean score of 28.

Keywords: Test, EFL, Language Proficiency, Score Analysis

1. Introduction

Languages have been taught and their learners assessed in various modes depending not only on the theory of language teaching the individual institutions or nations subscribe to but also defined by the predominance of the theory in particular era. For example, Bachman and Palmer (1996) observe that in 1970s language testing practice was informed essentially by a theoretical view of language ability as consisting of skills (listening, speaking, reading, and writing) and components (e.g., grammar, vocabulary, pronunciation). In 1980s, however, a new wave of applied linguists, including Widdowson (1983), Savignon (1983), Canale and Swain (1980) viewed language use as the creation of discourse, or the negotiation of meaning, and of language ability as multi-componential and dynamic. This line of thought, observes Bachman (1999), forced language testers to take into consideration the discourse and sociolinguistic aspects of language use, as well as the context in which it takes place.

In all these different theoretical perspectives, tests are used as the measurement instruments designed to elicit specific behavior, directly or indirectly (Shohamy, 2001). Tests serve different purposes, but mainly five as given by Spolsky (2001), namely; i) being a competitive selection device, ii) to provide information on the quality of the "product" to those who are paying for an education system, iii) to process and certify that an individual has achieved a specific level of technical or professional skill, iv) for prediction or prognosis of the probable results of training, and v) as an integral part of all good teaching.

As for proficiency tests, the major purpose is to pinpoint strengths and weaknesses in the language abilities of the students and thus make decision as to who should be allowed to participate in a particular program of instruction (Henning, 1987). Bachman (1990:16) defines language proficiency as "knowledge, competence, or ability in the use of a language, irrespective of how, where, or under what conditions it has been acquired." For The Council of Chief State School Officers (CCSSO), a fully English proficient student is ideally able to use English to ask questions, understand teachers and reading materials, test ideas, and challenge what is being asked in the classroom. The testing linguistic areas for most proficient tests, according to CCSSO (1992), are a) *Reading* – where the learner's ability to comprehend and interpret text at the age and grade appropriate tested, b) *Listening* focusing on the learner's ability to understand the language of the teacher and instruction, comprehend and extract information, and follow the instructional discourse through which teachers provide information, b) *Writing* the aim of which is the learner's ability to produce written text with content at the age and grade-appropriate level, and d) *Speaking* focusing on the ability to use oral language appropriately and effectively in learning activities (such as peer tutoring, collaborative learning activities, and question/answer sessions) within the classroom and in social interactions within the school. Oller and Damico (1991) had earlier called this compartmentalization of test areas the discrete point approach, which, he contends, consist of separable components of phonology, morphology, lexicon, syntax, and so on, each of which could be

further divided into distinct inventories of elements (e.g. sounds, classes of sounds or phonemes, syllables, morphemes, words, idioms, phrase structures, etc). By this model, the ideal assessment would involve the evaluation of each of the domains of structure and of the skills of interest. Then, all the results could be combined to form a total picture of language proficiency. However, Valdés and Figueroa (1994) caution that that language proficiency testing should as much as possible involve contextualized language processing.

Whatever the testing perspective, proficiency tests may be at different levels. The first and the most popular are international category which tests the candidate's proficiency in English who come to join English as native language countries like USA, Great Britain and Australia. The tests of this category are mainly five. The first one is TOEFL (Test of English as a Foreign Language), which is designed to measure the English proficiency of non-English speaking people and is documented in more than 2400 American colleges and universities which require the *TOEFL* test scores from non-English speaking students in order to admit them to a program[1]. The second is IELTS (The International English Language Testing System) and is the world's most popular high stakes English-language test for study and is designed to reflect real life use of English — at study, at work, and at play[2]. The third is ELPT (The English Language Proficiency Test) which can be sat for instead of or in addition to TOEFL for college entrance depending upon requirements of the schools in which the student was planning to apply[3]. The fourth one is ITEP (The International Test of English Proficiency), which is a language assessment tool that measures the English skills of non-native English speakers and is supported by more than 600 institutions, and is available in more than 40 countries. It is also used by businesses, and governments such as Saudi Arabia, Colombia, and Mexico for large-scale initiatives[4]. The last is TOEIC (The Test of English for International Communication), which is an English language test designed specifically to measure the everyday English skills of people working in an international environment (http://www.examenglish.com/TOEIC/).

The second level of proficiency tests are language tests offered by individual universities, examples of which are:

1. New York University-School of Continuing and Professional Studies (NYU-SCPS), which measure the candidate's proficiency in more than 50 languages and their scores can be used by universities to grant academic credit or to grant advanced placement in the candidates' language of study (*http://www.scps.nyu.edu/about.html*).
2. Utah State University (UTS) Language testing which is offered for UTS students in the Languages, Philosophy and Communication Studies Department in the following languages: Arabic, Mandarin Chinese, French, German, Japanese, Russian, and Spanish (*http://lpsc.usu.edu/testing5.aspx*).
3. Purdue University Language Placement Testing, offered by school of Languages and cultures and tests in measuring proficiencies in French, German, Japanese, Latin, Russian, Spanish, and Spanish for Heritage Speakers (*http://www.cla.purdue.edu/ Language_Placement_Testing.html*).

In Africa, we have an example of the Test of the proficiency for translators or interpreters, and which also screens prospective candidates for a job, offered by Wits Language School in Johannesburg, South Africa. The University also has the English for Academic Development and Testing Unit which conducts proficiency testing for individuals or corporations needing general English assessments or Wits admissions[5]

In Tanzania, a number of universities also offer proficiency tests; examples are: a) Sokoine University of Agriculture (SUA) which conducts English Language Proficiency Tests (ELPTs) and aptitude tests to its prospective first year for screening purposes on assumption that proficiency in English language has a significant relationship with the student's academic achievement (Wilson and Komba, 2012), b) Dare es Salaam University College of Education, which offers English language proficiency test for students requesting the service for various purposes (*http://www.duce.ac.tz/principal-message.html*) and c) Open university of Tanzania does the same although it is less strict, focusing mainly on oral proficiency.

As for The University of Dar es Salaam, proficiency testing and certification began over ten years ago with some holistic scoring approach based upon the candidate's ability to comprehend and respond to the examiner's oral prompts and thus create a string of conversation with the examiner. So the focus was on aural-oral aspects. Then about four years later the test was formalized to have components and formats resembling those of TOEFL, only not quite like it in vigilance and comprehensiveness. Its mode of scoring also changed from holistic and impressionistic judgment of the examiner into numeral scores characterized with unequal distribution of testing areas in reading comprehension, grammar, vocabulary and listening comprehension. However, the scores were only summed, then converted to percentage and submitted thus to the candidate (or any requesting organization) without aggregating. Further, for six years now, tests have been administered and

---

[1] https://www.encomium.com/webmentor/abouttoefl.html

[2] *https://www.ielts.org/test_takers_information/what_is_ielts/ielts.aspx*

[3] *http://en.wikipedia.org/wiki/English_Language_Proficiency_Test)*

[4] *http://en.wikipedia.org/wiki/International Test of English proficiency.*

[5] http://www.witslanguageschool.com/Services/ProficiencyTestingandScreening.aspx.

scoring done in the same manner without a deeper and comprehensive analysis of the candidates' performance comparing their performance generically as well as across test areas, on the one hand, and comparing performance of test testers categories according to their educational status, on the other. This study was carried out this aim.

2. Literature Review and Theoretical Framework

2.1 Literature Review

Review of related literature has been organized according to discrete testing areas as follows:

a) *Vocabulary*: in the area of vocabulary, a number of studies have been conducted but three are of interest to us here. The first is Qian's (2008) study on the predictive power of discrete and contextualized vocabulary items on assessing the reading performance, it was found that, in assessing reading performance, discrete-point vocabulary items and fully contextualized vocabulary items provide a similar amount of prediction. The second is Laufer and Goldstein's (2004) testing of the size and strength of vocabulary of 435 English as Second Language learners with the trial of bilingual computerized test. Their notion of size was delimited to the number of words the learners know, and the strength as a combination of four aspects of knowledge of meaning that are assumed to constituted a hierarchy of difficulty: passive recognition, active recognition, passive recall and active recall. They noted that that the hierarchy was present at all word frequency levels and passive recall was the best predictor of classroom language performance.

A more recent study was by Öztürk (2012) who investigated the effect of context on the performances of students in achievement vocabulary tests on 123 elementary students at Afyon Kocatepe University English preparatory program, using two different tests, discrete and contextualized with the same target vocabulary items. The results revealed that students performed better in the contextualized test, and that there was a significant difference between the performances of students.

b) *Listening:* Listening, a comparably least focused language skill in the realm of testing has equally had few empirical studies. Among such studies is Smit's (2006) quasi-experiment that sought to determine whether the recognition and interpretation of discourse markers will enhance students' listening comprehension in academic lectures in the University of Namibia (UNAM). The findings show that there was generally low success rate that could not be attributed solely to the students' disadvantaged past as most of the students who took part in this experiment came from urban areas where teaching was of better resourced and they seemed to know English well in its written form as they appeared to be proficient in reading and writing English texts. However, this knowledge did not seem to assist them in the listening process.

Another study is the one by Chiang and Dunkel (1992) who tested the listening comprehension of 388 high- and low- intermediate listening proficiency Chinese EFL students and found that attempts to comprehend and retain English lecture information was probably thwarted by a number of cognitive and linguistic factors as well as by academic and cultural issues such as inability to anticipate discourse markers and logical relationships in the English lecture and inability to detect the main points of the lecture or to grasp the usual goals of particular genres of discourse situations of which the discourse is a part.

Earlier on, Young (1994), cited in Smit (2006), sought to identify some of the more prominent microfeatures that contribute to the macro-structure of the university lecture using seven two-hour lectures from third and fourth year courses. She noted that an acquaintance with the correct schematic patterning of lectures will greatly assisted students.

c) *Reading*: A notable study on reading in our context is by Chang (2012) who investigated the effect of timed reading (TR) and repeated oral reading (RR) on 35 adult students of English as a foreign language. The overall findings indicated that increasing the reading amount for the TR group improved reading rates and comprehension. However, increasing the reading rate for the RR group did not have a negative impact on reading comprehension. Other studies by Cushing-Weigle and Jensen (1996) and Chang (2010), showed that, despite students registering reading rate gains, the readers did not show significantly better comprehension. This could be attributed to the fact that the readers' rates might have still not reached the optimal level that could promote comprehension.

d) *Writing:* Writing skill has been widely researched in the area of language testing both at interpersonal communication and formal or genre-specific levels. For example, Wang and Bakken's (2004) academic writing needs assessment of English as a second language clinical investigators revealed that mistakes such as inappropriate format, limited vocabularies and simple sentence patterns, lack of organization and coherence and use of flowery speech without conciseness were common amongst English as Additional Language (EAL) students and less common amongst English as Foreign Language (EFL) students.

Stephen, Welman and Jordaan (2004) investigated the impact of English language proficiency on academic success of first-year black and Indian students in human resources management at a tertiary institution. The findings strongly confirmed the hypothesis that English language proficiency has a significant impact on black student success rates.

Maher (2011) studied aimed how much influence academic writing ability has on academic performance in tertiary students in South Africa and found that academic writing ability is not a major predictor of and contributor towards academic performance.

e) *Grammar*: This is yet another area of language testing that has several empirical studies. We will take only a few examples here. Angolf and Sharon (1971) and Johnson (1977) employed university students at a Western State College and at Tennessee University, respectively, and used the original five section version of TOEFL and found that native speakers of English, on average, performed reasonably better relative to ESL testees. Clark (1977) made score analysis of college-bound high school students in New Jersey and found that the mean number of correct items (n=150) on two test forms was about 135 (90%) for native speakers and about 89 (59%) for ESL speakers even though the Structure and Written expression was comparatively more difficult to both groups.

f) Studies in Tanzania

At the local scene, Criper and Dodd (1984) conducted a study to assess the language proficiency of Tanzanian students at all levels observing whether the level they have would facilitate learning in the medium of English. They found that the level of ELP among most Tanzanian students was so low that it hindered learning at an alarming rate.

Mwinshehe's (2001) experimental studies to argue for use of Kiswahili as Medium of Instruction (MOI) rather than English revealed that the experimental group (which used Kiswahili for teaching and testing) did better in proficient-based examination than the control group, which used English.

In 2002, Dooey and Oliver assessed the predictive validity of the IELTS test as an indicator of future academic success. To do this, a small scale quantitative study was carried out amongst first year undergraduate students from diverse non- English speaking background and who were admitted on the basis of their IELTS scores. Their findings showed little evidence for the validity of IELTS of language proficiency as a predictor for academic success.

Wilson and Komba (2012) studied the relationship between English Language Proficiency (ELP) and academic performance in Tanzanian secondary schools by administering an ELP test and making a review of students' reports. They noted that there is a positive relationship between ELP and students' academic achievement.

Elisifa (2013) assessed the level of English language proficiency of the Open University of Tanzania undergraduate students. She found out, inter alia, students had more difficulties in presenting their subject matter clearly and precisely. Moreover, over half of them had problems with regard to producing linguistic outputs whose propositional content were congruent to the expected subject matter.

2.2 Theoretical Framework

The current study adopted measurement of proficient levels developed by American Council on the Teaching of Foreign Languages (ACTFL) (2012) in which learners are rated at four levels: superior, advanced, intermediate and novice. The approach to ACTFL is based on what a testee can do with language in terms of speaking, writing, listening and reading in a spontaneous and non-rehearsed context. For each skill ACTL rates a candidate in any of the five major levels of proficiency: Distinguished, Superior, Advanced, Intermediate, and Novice. The major levels, Advanced, Intermediate, and Novice are further subdivided into High, Mid, and Low Levels. These levels are operationalized so as to show the ranges that are descriptive of what an individual can and cannot do with language at each level irrespective of the time, the duration and the place the target language was acquired. These rating scales form a hierarchy the higher level of which subsumes all levels below it.

The rationale for the adoption of this mode of scoring is its focus on evaluating an individual's *functional language ability* (emphasis ours). In the current study we also rated grammar and vocabulary using the same scales, since the tasks were given in context. We also did not include the speaking skill.

3. Methods and Materials

The study involved 136 Udsm EFL test takers who sat for the test at different times between 2009 and 2013 from different education backgrounds as summarized in Table 1 below. Participants were picked randomly using their test scripts.

Table 1: Sample Strata

Sample Strata			
	Male	Female	Total
Udsm Alumni	23	19	42
ELT Short term Students	22	11	33
Sec. Level	1	0	1
Non Udsm Candidates	19	9	28
PG Students	6	4	10
UG Students	15	7	22
Total	86	50	136

From Table 1 above, one can note that a total of 136 test scripts were collected and analysed in this study, 86 (63%) of which were males and the remaining 50 (37%) females. The substrata for each sex category were four: i) 42 (31%) Udsm alumni of whom 23 (55%) were males, ii) 33 (24%) ELT short tem students[1], 11 (33%) among whom were females and the rest males, iii) 28 (21%) Non-Udsm candidates of whom 19 (68%) were males, and iv) The bona fide students who made a rather rich texture as it had 22 (16%) Udsm undergraduate students, 10 (7.4%) Udsm postgraduate students and 1ordinary level secondary school pupil.

The general compositional texture of respondents shows that there were fewer females as compared to males. Furthermore, the respondents belonging to, or with some affiliation with, Udsm formed the largest group since it engulfed all the groups (Udsm students-undergraduate and postgraduate, Alumni and ELT short term students) except one, the non Udsm candidates. Among Udsm respondents, the largest group was the alumni who were 107 in total, which is 39% of the whole Udsm group, followed by ELT short term students who are 33 (31%) while the smallest group was that of undergraduate students, who are 10 (9%). Most of Udsm Alumni who came for the test were those who had applied for postgraduate studies and were thus asked for evidence of their English language proficiency.

4 The Findings
4.1 Overall Performance
4.1.1 General Proficiency Level Analysis
We began our analysis with the overall performance of the testees was according to their proficiency levels as illustrated in table 2 below.

Table 2: The Overall Performance of Testees

	Frequency	Percent
Advanced Proficient	16	11.9
Proficient	36	26.7
Intermediate High	34	25.2
Intermediate	28	20.7
Intermediate Low	16	11.9
Novice	3	2.2
Novice Low	2	1.5
Total	135	100.0

The data in table 2 above show that the majority of the candidates (36 out of 135, which is 26.6%) were proficient, closely followed by the intermediate high proficiency level with 34 (25.2), suggesting that more than 50% of the sampled candidates were either proficient or intermediate high. Only a few (11.9%), however, were at the advanced level of proficiency. Nonetheless, it is clear that the majority of the candidates (a total of 86, which is 63.8% of all candidates) performed at levels above intermediate.

All things being as per the ideals of language testing, viz. reliability and validity, these 63.8% of the sampled candidates, could do an array of communicative tasks including, inter alia, i) understanding the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization, ii) interacting with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party, iii) using English language flexibly and effectively for social, academic and professional purposes' iv) producing clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices, and v) expressing him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in the most complex situations (Council of Europe, 2011).

Only a few (3 (2.2%) were at novice and novice low (2 (1.5%) candidates) levels signaling their

---

[1] *These were Mozambican candidates for higher education in Tanzanian universities under Tanzania-Mozambique exchange programme who came for 8 months intensive EFL teaching before starting theirs classes.*

ability in English language being limited to, inter alia, ii) using basic interpersonal phrases such as self introduction and language use related to where one lives, people one knows and things he/she has, ii) communicating using sentences and expressions related to areas of most immediate relevance and communicating in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters (European Council, 2011). Since the grand majority of the candidates sat for the test as a requirement for meeting the perquisites for acceptance for advanced degrees, their being at such lower levels suggest their need for attending courses not only to boost their Basic Interpersonal Communication Skills (BICS) but also their Cognitive Academic Language Proficiency (CALP) (Cummins, 1979) so that they can go beyond Oller' s (1979) claim that all individual differences in language proficiency could be accounted for by just one underlying factor, which he termed 'global language proficiency'. CALP, argues Cummins (2000), measures the extent to which an individual has access to and command of the oral and written academic registers of schooling.

4.1.2 Proficiency Levels per Test Testees' Groups

Another analysis was made of testees according to their clusters of educational status, as summarized in Figure 1 below.
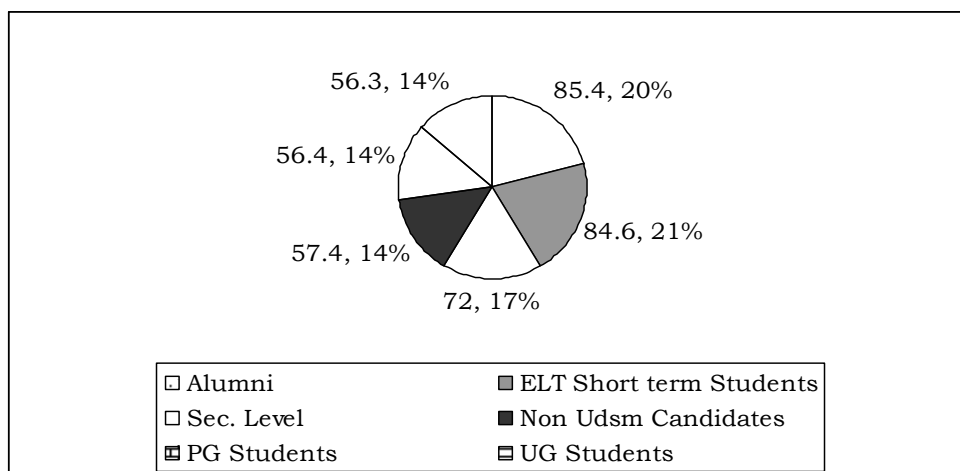
**Figure 1: Overall Performance of Testees**

The general impression from figure 1 above is that the performance is good as all groups of candidates ranged from 55% and 86% (which translates from 'good' or 'intermediate' to 'excellent' or 'advanced proficient').

However, there is also intergroup variability in the performance. The groups that have 'excellent' scoring and are thus interpreted as being 'high proficient' are Udsm alumni (who are graduates from the University from different degree problems in different academic years). These got a mean score of 85.5% and consisted of 20% of all candidates. The second group is Non-Udsm candidates, the grand majority of whom were also alumni of other universities and they were the largest single group of all testees. Their mean score was 84.6%. One unique case for this high proficient is a secondary school pupil with a 72% score!

None of the group scored B+ (meaning 'good' or 'proficient') but the remaining three groups were at 'B' stage, which means they were 'good' or 'intermediate' in their proficiency of English; These were Udsm Undergraduate students, Udsm Postgraduate students and ELT short term students with mean scores of 57.4%, 56.4% and 56.3%, respectively.

4.1.3 Gender-based Comparative Performance

One important key factor for variability in language proficiency among the candidates is that of sex. It has been widely acknowledged that women are better than men in general language proficiency. So we were interested in finding sex based variability in performance among the test takers and the results are as summarized in Figure 2 below.
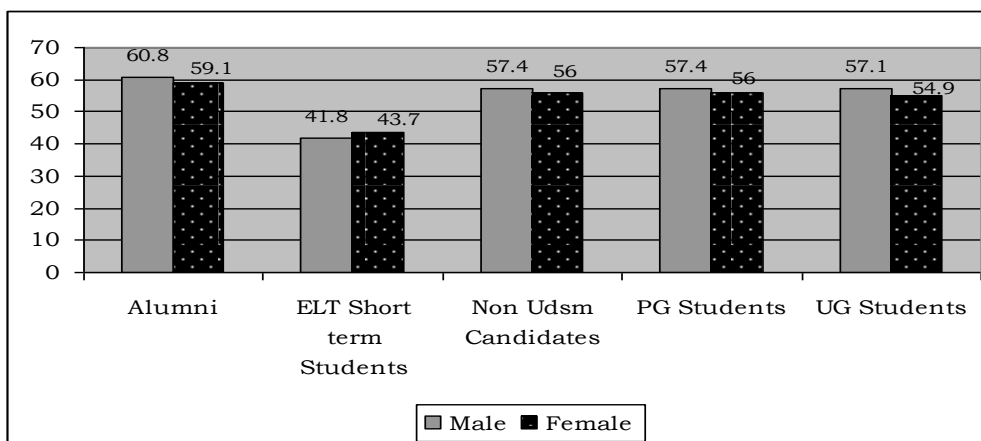
Figure 2: Comparative Gender-wise Performance

As revealed in figure 2 above, males have generally outperformed females in the four out of five groups. However, the difference is only marginal since it ranges from 2.2 (for UG students) and 1.4 (for PG students). Marked comparatively unique is ELT short term group in which females did comparatively better than males, though again marginally at 1.1%.

Generally, comparatively significant aggregate difference in at the alumni where males had an average of 'B+' as contrasted to females' average of 'B'. All the others are such that both sex-based groups were at the same aggregate of 'B' except ELT short term group both groups of which had an average of 'C'.

4.2 Proficiency According to the Candidates' Educational Status
As explained earlier on, there were five main categories of test takers each of which was explained in 3.1. When each stratum was given its respective test takers across the columns representing the proficiency scales, the data are as summarized in table 2 below.
Table 2: Performance per Educational Status

	A		B+		B		C		D		E		Total
	N	%	N	%	N	%	N	%	N	%	N	%	
Alumni	9	23	11	28	11	28	6	15	3	8	0	0	40
ELT Short term Students	0	0	3	9	4	12	11	32	14	41	2	6	34
Non Udsm Candidates	2	7	12	43	7	25	4	14	3	11	0	0	28
PG Students	0	0	5	50	2	20	3	30	0	0	0	0	10
UG Students	4	18	4	18	8	36	4	18	2	9	0	0	22
Total	15	48	35	148	32	121	28	109	22	69	2	6	134

Key: A =Advanced Proficient
    B+ = Proficient
    B = Intermediate High
    C = Intermediate
    D = Intermediate Low
    E = Novice

The data in table 2 above show that the alumni group is comparably of higher level of proficiency than the rest as 31 out of 40 (which is 79%) of its members at proficiency levels above intermediate. Out of these, 20 (50%) are either proficient or advanced proficient. Non Udsm candidates, which are also graduates of other universities, ranked the second with 14 (50%) at points above proficiency and another 25% at intermediate level, which adds to 75% of all its members above intermediate. This is very telling in terms of empirical findings that the more the learners are exposed to a foreign language via input enhancement the more likely they will have their proficiency increased, as was the case for the studies by Lee and Huang (2008) on effects of pedagogical interventions with visual input enhancement on grammar learning and by Jabbarpoor and Tajeddin (2013) on The Effect of Input Enhancement, Individual Output, and Collaborative Output on Foreign Language Learning focusing on English Inversion Structures in Japan.

That fact is validated by the ELT short term students who were Mozambican pre-entry university learners who took the test in their 6th or 7th week of intensive English language training and to most of them, the weeks were their very first time they have encountered serious proficiency-based English language training. This group was the last of the five groups with only 3 (9%) and 4 (12%) at proficient and intermediate high levels, respectively, while none was at advanced proficiency stage.

However, an adverse situation to the argument for duration of exposure to English as a foreign

language is presented by contrastive analysis of PG students and UG students where the former had none in the advanced proficiency level while the latter had 4 (18%). Nonetheless, the situation is mitigated by 50% of PG students at proficient stage and none below intermediate as contrasted to UG students' 18% at proficient level and 9% at intermediate low.

4.3 Overall Performance in Test Areas

Debate has been raging over testing language proficiency generically, following Oller's (1991) argument on the integrative or holistic approach, on the one hand, and measuring plurality in the same learner through what scholars like Cummins (2001) refer to as *discrete language skills* which involve the learning of rule-governed aspects of language (including phonology, grammar, and spelling) where acquisition of the general case permits generalization to other instances governed by that particular rule, on the other hand. It is in the light of the above arguments that the examiners divided the test into five sections, namely; i) reading comprehension with three passages, the first with 12 test items, the second with 10 items and the thirst with 10 items; ii) a section on grammar with 15 items (which we deemed 'Grammar A') and the second with 40 items which we called 'Grammar B'); iii) vocabulary and meaning with two sections, one with 25 items (which we referred to as 'Vocabulary A') and another with 10 items (which we called 'Vocabulary B'), and iv) Listening comprehension, with a total of 6 items. Thus our analysis hereunder will be under seven subheadings: i) The overall comparative analysis of candidates in the test areas, ii) Reading Comprehension, iii) Grammar A, iv) Grammar B, v) Vocabulary A, vi) Vocabulary B, and vii) Listening comprehension.

4.3.1 The Overall Comparative Performance

The initial task with regard to test areas was computation of means for each of the six areas and such means were put in a single excel column and comparative summary was created as illustrated in Figure 3 below.
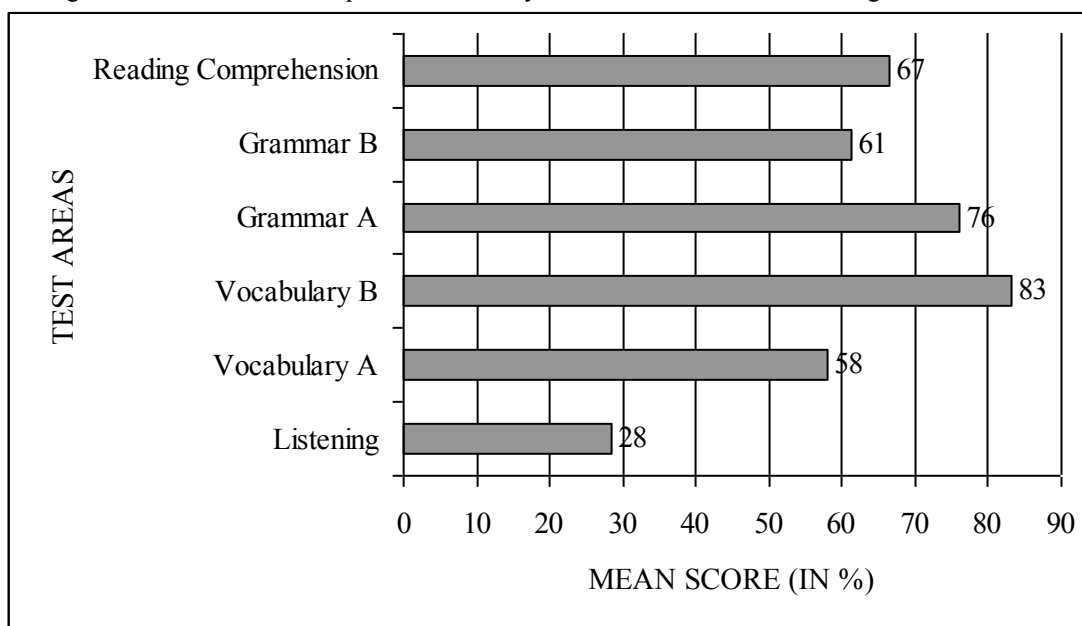
Figure 3: Comparative Mean Performance per Test Areas

The data in Figure 3 shown that the candidates' performance was the highest in the area of vocabulary B where their overall mean score was 83 (which is an A aggregate, indicating their being of high proficiency in this vocabulary aspect. In this content area, the candidates were given ten sentences each of which had one word in italics and one gap. The candidates were required to choose from the list of words lettered A to E the word that is most nearly opposite in meaning to the word in italics which will correctly fill the gap in the sentence. In other words, the candidates, in their majority, were able to correctly identify the word which is an antonym to the word italics word. According to Ellis and Sinclair (1989), knowing vocabulary involves understanding the word when it is spoken or written, recalling it when needed, using it with the correct meaning and in a grammatically correct way and in the right situation, pronouncing and spelling it correctly, knowing which other words can and cannot be used with it, knowing if it has positive or negative connotations and when or when not to use it.

This was in contrast to vocabulary A where the mean score was 58 (the aggregate of C, which translate as intermediate). This latter test area consisted of 25 items, which were numbered gaps in the passages given and below the passages were lists each with five choices offered in columns lettered A to E. The candidate were to choose the word that was the most suitable to fill the numbered gap. The whole aim in vocabulary A was to test what Öztürk (2012) termed the candidates' knowledge of contextual relevance of the vocabulary as defined by

the linguistic context occurring in the sentences. Comparing the two vocabulary tests areas one ca see the students were far better in antonyms than in the contextual vocabulary area, contrary to findings by what Öztürk (2012) in which 123 elementary students at Afyon Kocatepe University English preparatory program performed better in the contextualized test than in a discrete vocabulary test area. Whatever the case, the major focus is on how well   the students have mastered a vocabulary skill that they have been taught (or learnt by themselves) (Read, 2000).

In the grammar area, Grammar A, which consisted of 15 sentences each with an underlined words or phrases marked A to and the candidate was to choose the one item which makes the sentence ungrammatical, was yet another where the candidates performed very well, with a mean score of 76 (which is of A aggregate, and thus translating to advanced proficient.

Conversely, Grammar B, which was of *Cloze procedures*, type in which every nth word in a text omission of (somewhere between every fifth or tenth word) is omitted (Žlábková, 2007). To this category the candidates' mean score was at B+ (that is, 61 translating to 'advanced'). This means students were more conversant were more conversant with grammaticality judgment items defined by Rimmer (2006) as grammaticality judgment test, where subjects make an intuitive pronouncement on the accuracy of form and structure in individual decontextualized sentences. Such findings are in line with a study by Rahimy and Moradkhani (2012) on the effect of GJ (Grammaticality Judgement) tasks as a classroom activity on Iranian EFL learners' grammatical patterns where they found that the learners in the experimental group received higher scores in grammatical patterns after being treated with 10 sessions of GJ tasks.

Ranking third is reading comprehension, in which the candidates were given a three passages, one on Alzheimer's disease after which 12 multiple choice questions were provided, the second on tea to which 10 questions were given, also of multiple choice type, and a third on eyes to which ten questions were also given. The major focus of questions were contextual meanings of the words as used in the passage while a few others were retrieval of factual information from the passage. The overall candidates' performance was 67 which translate to a B+ aggregate, similar to 'very good' or 'proficient'. This good command in this receptive skill implies the learners being able to productively engaging in consumption of information   (academic but also social) using written sources. It also means the candidates, particularly those who were applicants for advanced degree, proved they could engage in high level of discourse community where intensive and extensive interaction with the text is a norm.

The most underperformed test area is listening for comprehension, where the students were given six multiple choice questions and instructed to take some minutes reading the questions after which a passage was read to them twice. Then the candidates were to answer the questions depending what they remember from the passage read to them. In this test area, the mean score is at 28, which is of D aggregate, translating poor of novice low. What this means is that listening proved a real problem to the extent that the learners either could not have comprehended what was being read or were unable to recall what they heard.

4.3.2 Specific Test Areas

a) Reading for Comprehension

Testing reading comprehension, posits Hughes (2000), the examiner seeks from the candidates, among others, the ability to identify examples presented in support of an argument, identifying referents of pronouns, using context to guess meaning of unfamiliar words, and understanding relations between parts of text by recognizing indicators of discourse. The reading compression passages given were mainly in the light of the above ideals. Additionally, the questions were of multiple choice nature which Hughes (ibid.) require the candidate to provide "evidence of successful reading by making a mark against one out of a number of alternatives" (p.120) and to which Ashworth (1982) hails as extending over a wide range of knowledge in one test and thus encouraging the testees to develop a wide background of facts and abilities[1]. This section had two reading tasks; the first was a one paragraph text on Ahlzeimer's disease with 12 items while the second was a five paragraph passage on tea plant after which the candidates were to answer ten questions. Thus the total number of items for this section is 22. The candidates were clustered into their respective proficiency level in the reading comprehension passage and the results are summarized in Table 3 below.

---

[1] *Ashworth (1982:30) also emphasizes the point that these item types 'are also able to measure high level of thinking" and thus dismisses the critique of multiple choice items that they measure lower levels of cognitive abilities.*

**Table 3: Candidate's Proficiency Levels in Reading Comprehension**

Proficiency Level	Frequency	Percent
advanced proficient	46	34.1
proficient	19	14.1
intermediate high	13	9.6
intermediate	23	17.0
intermediate low	16	11.9
novice	9	6.7
novice low	9	6.7
Total	135	100.0

The data in table 3 above show that slightly over 50% had proficiency levels above intermediate while another 38.5 % were at intermediate proficiency level (if we combine intermediate high, intermediate and intermediate low levels). In other words, only about 13.2% were at a novice and novice low levels. So, generally the reading section was done very well suggesting that the learners could, in their majority retrieve information of the passage as well as infer and contextualize meaning as given by the examiners.

b) Grammar A

The rationale for testing grammar as a separate item in language testing is as attested by Hughes (2000) that grammatical ability, or rather lack of it, sets limit to what cab be achieved in the way of skills performance. He adds that 'the successful writing of academic assignments, for example, must depend to some extent on command of more tan the most elementary grammatical structures' (p.142). Testing grammar for a second language/foreign language, attests Harris (1969), will have the testing of control of the basic grammatical patterns of the spoken language, which would not pose challenge for native speakers.

That was the very rationale the examiners might have in bringing in the grammar section in the examination, the total items of which were 15 which were of the types referred to as *error recognition multiple – choice items* (Dastoshadesh, Brijandi and Jalilzedeh, 2003) in which the candidate is given sentences each with four highlighted words or phrases, marked A, B, C and D, out of which there is one the presence of which would render the sentence ungrammatical. The candidate is instructed to identify that item which must be changed for the sentence to be grammatical. The candidates' performance in this test is is summarized in Figure 3 below.
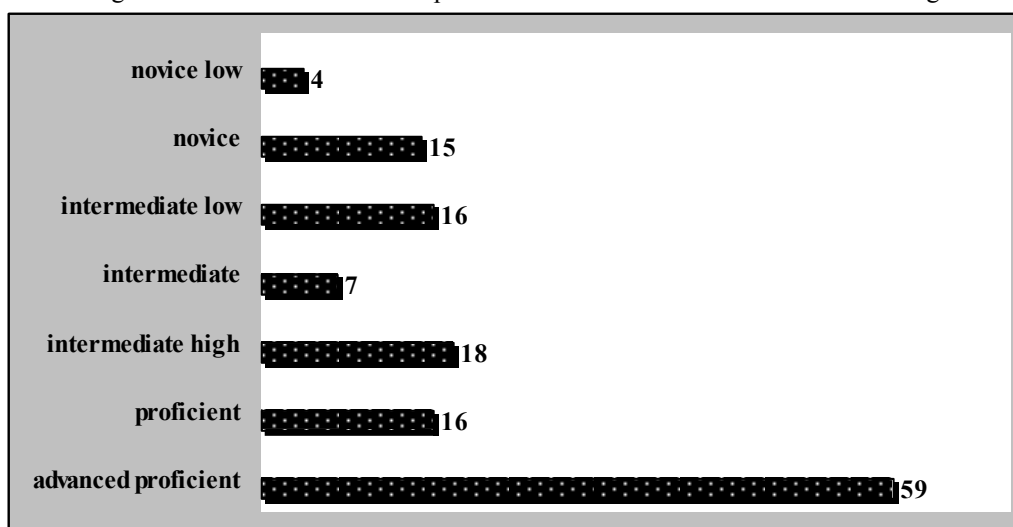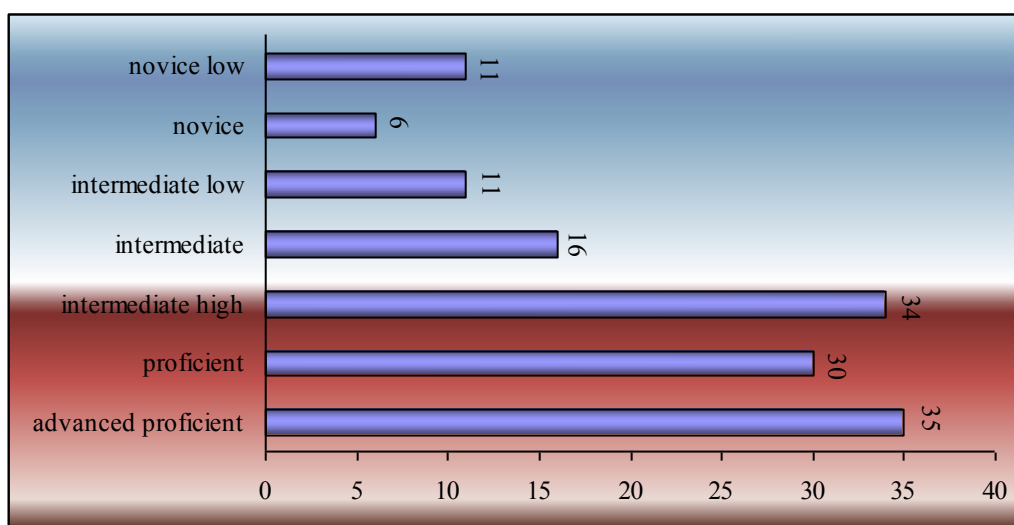

Figure 3: Candidates' Performance in Grammar

The data from figure 3 shows that the majority of the respondents (59 in total, which is 44%) are of advanced proficient level, while 16 (which is 12%) and 18 (which is 13%) were of proficient and intermediate high levels, respectively. Conversely, only 4 (3%) and 15 (11%), were at novice low and novice levels, respectively.

A study which is more or less similar to the afore going findings was by Dastoshadesh, Brijandi and Jalilzedeh (2003) investigated the relationship between Iranian EFL students' performance on an error recognition test and their writing ability. The findings showed that there is no statistically significant relationship between test takers' performance on the error recognition test and their writing ability.

c) Grammar B

Grammar B was unlike Grammar A by its diversity of grammar areas ranging from tag questions (items 1-12), subordinating conjunctions (items 13-16), prepositions (items 17-22), verb forms (23-26) and use of conjuncts and comparative adjectives, among others. The performance in this area is as summarized in Figure 4 below.

Figure 4: Comparative Performance in Grammar B per Proficiency levels.

The data in figure 4 above show that learner variability across proficiency levels between advanced proficiency and intermediate high was not as high as was the case of Grammar A (ref. Figure 3). This is because the learners who proved advanced proficient were 35 (which is 26%) closely followed by those at intermediate high proficiency (34, which is 25% of all testees) while the proficient level learners in this test area were 30 (22%). The same also goes for low achievers in this areas in which the number is equal (11 (8%) each for intermediate low and novice low while the learners at novice level were only 6 (4%). Thus, while the distribution of candidates across proficiency levels showed more robustness in this test area than in Grammar A, the rift between high achievers and low achievers can still be noticed due to the fact that there were 34 intermediate high candidates while the intermediate were only 16, with a range of 14 between the two scales.

d) Vocabulary A

In addition to control of grammatical forms and syntactic patterns, observes Weir (1990), one has to be proven proficient in knowledge and effective use of lexical resources. According to Lado (1964), producing a vocabulary means that a unit can be recalled almost instantaneously with its proper structural position in accordance with the context, the lack of mastery of which, observes Chastain (1988), is the most common cause of students' inability to say what they want to say during communication activities. Thus, posits Read (1997), if vocabulary knowledge is accepted as a fundamental component of second language proficiency, it is natural to accept that one of the primary goals of language testing will be to assess whether learners know the meanings of the words they need to communicate in a successful way in L2. It is from such framework of argumentation that the examiners had two separate vocabulary sections.

As noted earlier on, vocabulary A focused on learner's knowledge of register-specific vocabulary item in which the candidates were given passages with 25 numbered gaps which indicated the missing words. The candidates were required to choose from the list of words lettered A to E given in five columns the word that is most suitable to fill the numbered in the passage. The candidates' performance in terms of proficiency levels are presented in Figure 5 below.
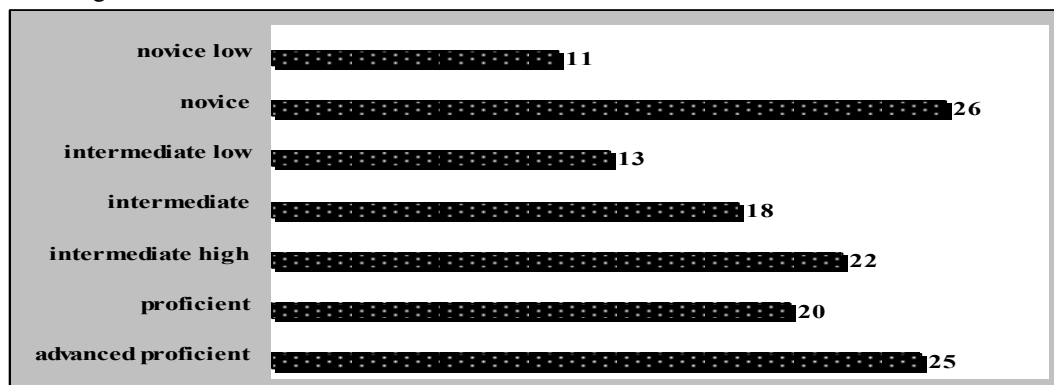
Figure 5: Performance in Vocabulary A

As can be seen in figure 5 above, a total of as (39%) were proficient (which included 25(19%) at advanced stage of proficiency and 20 (15.5%) at proficient stage). These, if we take Anderson's (1985) checklist

of linguistics abilities of FFL learners, are able o use vocabulary and idioms rarely distinguishable from ha of educated native speakers and though they can occasionally make use of inappropriate terms or rely on circumlocution.

A fair number were at intermediate stage (18, which is 14% at A intermediate high and 18, which is 10%, at intermediate), pointing to their frequency use of correct but not quite appropriate terms and a limited ability to expand a machines of a given word beyond its denotation. However, quite a relatively big number, slightly above hose at advanced proficiency level were at novice stage who according to Anderson (1987), have so limited vocabulary and so frequently misused that I they are to produce their own (unguided) composition, he readers would be forced to rely on their own interpretation. The learners at the (novice stage), as observed by the British council Criteria (in Hughes 2000: 6) tend to display gross inadequacies of vocabulary, accompanied with misspelling and punctuation.

In short, the performance is such that 83 (64%) candidates were at stages from intermediate and above, pointing to the fact that at least over half have a fair mastery of vocabulary and a fairly wide range of knowledge of contextual and technical (specialized) means of terms beyond the common care or denotative meaning.

e) Vocabulary B

Vocabulary B consisted of 10 items each of which has four choices one of which is correct. The candidate's task was to identify the word from the choices, the meaning of which is nearly opposite to the italicized word in a sentence given but which will also be correct one to fill the gap in the sentence. Hughes (1989) call this type of vocabulary 'tests recognition testing' in which the candidate recognizes either using co-text clues or and the prior knowledge of the italicized word meaning to make a choice of near antonym of the word list. The findings of the candidates' performance in this test area are summarized in figure 6 below.
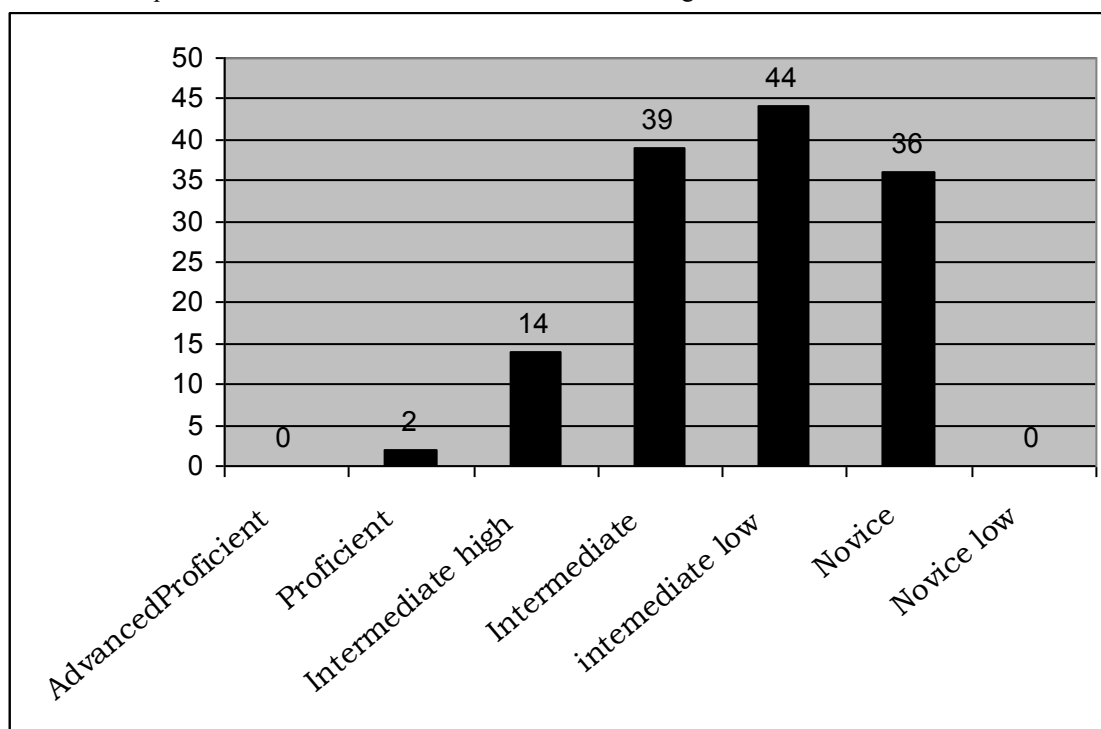
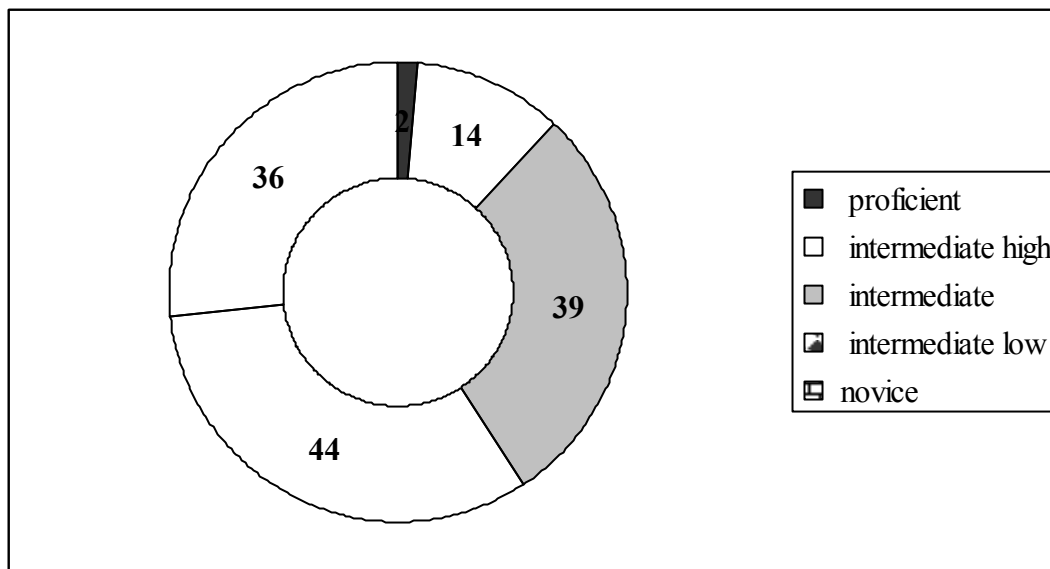**Figure 6: Candidates' Proficiency in Antonyms**

In this test area, unlike in vocabulary A, the majority, a total of 33 (25%) and 19 (15%), were at novice low and novice stage respectively. This is a contrasted with 16(12%) which were at proficient and advanced proficient stages. The most difficult items for the majority of the candidates were the multi word items like tangled up the wrong response chosen by the candidates of which was unpicking, and well informed; the majority of whom chose 'clueless' for a response. Hence he need for instructing learners on the use and meaning of multi-word verbs.

Being in at novice stage as already indicated for the case of vocabulary A, implies their quite limited ability to retrieve meaning of a given vocabulary word sometimes even in complex circumlocution, nor can they co-textually, predict the vocabulary (through, in this context, opposite of the focal word) in a given string of word (Cohen, 1996). Wilcox and Mediana (2013) contend in their study of effects of semantic and phonological clustering on L_2 vocabulary acquisition of novice learners, that vocabulary is critical in second language acquisition as I enrich semantic clustering. Their study involved investigation of four semantic categories and found that semantically clustered words (not phonologically similar) were more difficult to learn.

Yet, Jones (2002), reiteration the importance of a mastery of autonomy in language use, posits that as many as one in every fifty sentences might include autonym pairs serving one or more of several discourse functions. Further, studies that fit well this section on autonyms include Justeson and Kartz (1991), Fellbaum (1995) and Jones (2002) – all based in corporate of adult directed written language. These studies have shown that antonyms frequently co-occur in sentences, often in recognizably constrictive constructions.

f) Listening

Omailley, Chamot and Kupper (1988) reiterate that listening comprehension is on active process in which individuals focus on selected aspects of aural input to construct meaning from passages and relate what they hear to existing knowledge. This part of the test consisted of 6 items to which testees were provided with the questions before hand and were given time to skim them before the Passage was read to them twice, first at a slower rate and the second time in a normal tempo. The testees performance is summarized in Figure 7 below.

Figure 7: The Candidates' Proficiency in Listening Skill

Here, too, like in vocabulary B, the performance is largely at intermediate low (44 candidates, which is 34%) intermediate (39 (30%), and novice 36 (28%). In other words, only 16 (a mere 12%) were at intermediate high and above; out of these 16, only 2 (less than 2%) were related proficient. In other words, using ACTFL proficiency guidelines 2012 for listening, the majority of the learners (over 85%) are able to understand information from sentence – length strong of speeds, are utterance at a line, and since he majority were unlikely to recognize the speech they cannot anticipate, they failed items 3, 4, and 6 which required then o get an overall meaning of the passage before they could answer the question correctly.

Kurits (2012) sympathetic to PFL plight with listening comprehension difficulties admits that listening is the most difficult (though also most neglected) skill to learn, even though it is at the heart of language learning, without which learners cannot understand second language (or even L_1) speakers nor comprehend a variety of L_2 multimedia such as DVDs and he Internet. Rost (2001) adds rather emphatically that a key difference between more successful and less successful acquires related in large part to their ability to use listening skill as a means of acquisition (p.94). In Noro's (2006) study that examined the psychological reality of the construct of listening stress' concluded that "the difficulties were same sort of 'stress or to their" (p.64).

Thus, L2 learners (like the ones in the current study) who have limited processing ability with less linguistics knowledge argues Kurita (2012), will rely or their ability to make use of the available resources o interpret what they hereby top-down processes. Also, Stahr's (2009) study found that the vocabulary size and depth of vocabulary knowledge were both significantly correlated with listening comprehension. In his earlier work in 1993, Eastman produced evidence that L2 learners face an importer obstacle in distinguishing content words and function words when their L1 does not resemble English rhythmically.

5. Conclusion

From the above findings a number of issues are worth bringing forth. First, the findings have confirmed the long held belief that the longer one gets exposed to the target language the more proficient that person is likely to be. This has been the case where the Alumni did far better than the continuing students and worst performers were occasional ELT students. Secondly, there are situations where the BICS level is at its best when the CALP is at its minimal, as was attested by an outstanding performance of a secondary school candidate in comparison to continuing students, both at undergraduate and postgraduate levels. Third, it is not always the case that women

excel in language proficiency when compared to men, given the findings of the current study. Lastly, listening skill, being the most underperformed, is a sends a wash back signal to EFL teaching where experience and official language curricula and syllabi at secondary school level and even at university level, highly stigmatize it or it is not rightly taught.

References

ACTFL (2012). Proficiency Guidelines 2012 © Actfl, Inc., 2012. [Online]Available: *www.actfl.org* (27th August, 2014).

Anderson, J. R. (1985). Cognitive Psychology and Its Implications (2nd ed.). New York: W. H. Freeman.

Anglolf, W.H. and Sharon, A.T. (1971). A Comparison of Scores earned on the Test of English as a Foreign Language by Native American College Students and Foreign Applicants to U.S. Colleges. TESOL Quarterly, 5,129-136.

Ashworth, A.E. (1982). Testing for Continuous Assessment. A Handbook for in Schools and Colleges. London: Evans Brothers Limited.

Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

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

Canale, M., & Swain, M. (1980). Theoretical bases of Communicative Approaches to Second Language Teaching and Testing. Applied Linguistics, 1, 1-47.

Council of Chief State School Officers (1992). Recommendations for Improving the Assessment and Monitoring of Students with Limited English Proficiency. Alexandria, VA: Council of Chief State School Officers, Weber Design.

Chang, C-S. (2010). The Effect of a Timed Reading Activity on EFL Learners: Speed,

Chang, C-S (2012). Improving Reading rate Activities for EFL Students: Timed Reading and Repeated Oral Reading. Reading in a Foreign Language. 24,1,56-83.

Chastain, K. (1988). Developing Second Language Skills: theory and practice. (3rd Ed.). New York: Harcourt Brace Javanovich, Inc

Chiang, C.S. and Dunkel, P. (1992). The Effect of Speech Modification, Prior Knowledge and Listening Proficiency on EFL Lecture Learning. In TESOL Quarterly. 26, 2, 345 – 373.

Clark, J.L. (1977). The Performance of Native Speakers of English on the Test of English as a Foreign Language (TOEFL Report 1). Princeton: Educational Testing Service.

Cohen, D.C. (1996). Second Language Learning and Use Strategies: Clarifying the Issues. A revised version of a paper originally prepared for presentation at the Symposium on Strategies of Language Learning and Use, Seville, Spain, December 13-16, 1994.

Cushing-Weigle, S., and Jensen, L. (1996). Reading Rate Improvement in University ESL Classes. CATESOL Journal. 9, 55–71.

Dastoshadesh, A., Birjandi, P. and Jalilzedeh, K. (2011). Error Recognition Tests as a Predictor of EFL Learners' Writing Ability. International Journal of English Linguistics.1, 2,252-257.

Elisifa, Z. S. (2013).Assessment of Levels of Written Language Proficient at the Open University of Tanzania. Unpublished Doctoral Thesis. Open University of Tanzania.

Fellbaum, C. (1995). Co-occurrence and Antonymy. International Journal of Lexicography. 8, 281–303.

Harris, D.P. (1969). English Testing as a Second Language. Toronto: McGraw-Hill Book Company.

Henning, G. (1987). A Guide to Language Testing: Development, Evaluation, Research. Rowley, MA: Newbury House.

Hughes, A. (2000). Testing for Language Teachers. Cambridge Handbooks for Language Teachers series. Cambridge: CUP

Jabbarpoor, S. and Tajeddin, Z. (2013). The Effect Of Input Enhancement, Individual Output, and Collaborative Output on Foreign Language Learning: The Case of English Inversion Structures. RESLA. 26, 267-288.

Jones, S. (2002) Antonymy: A Corpus-based Approach. London: Routledge.

Justeson, J. S., and S. M. Katz (1991) Co-occurrences of antonymous adjectives and their contexts. Computational Linguistics. 17, 1–19.

Kurita, T. (2012). Issues in Second Language Listening Comprehension and the Pedagogical Implications.

Lado, R. (1964) Language Testing; The Construction and Use of Foreign Language Tests: A Teacher's Book. New York: McGraw-Hill.

Laufer, B. & Goldstein, Z. (2004) Testing Vocabulary Knowledge: Size, Strength and Computer Adaptiveness. Language Learning. 53, 4,399-436.

Lee, S. and Hung-Tzu Huang, H. (2008).Visual Input Enhancement and Grammar Learning: A Meta-Analytic Review, SSLA. 30, 307–331.

Maher, C. (2013). Academic Writing Ability and Performance of First Year University Students in South Africa.

Unpublished Masters of Education Dissertation, University of the Witwatersrand.

Mwinsheikhe, H. (2001). Science and the Language Barrier: Using Kiswahili as a Medium of Instruction in Tanzania Secondary Schools as a Strategy of Improving Students' Participation in Performance in Science. M.Phil. Thesis, University of Oslo.

Noro, T. (2006). Developing a construct model of "listening stress": A qualitative study of the affective domain of the listening process. Annual Review of English Language Education in Japan, 17, 61-70.

Oller, J.W. Jr. and Damico, J.S. (1991). 'Theoretical Considerations in the Assessment of LEP students'. In E. Hamayan & J.S. Damico (Eds.), Limiting bias in the Assessment of Bilingual Students. Austin: Pro-ed publications.

O'malley, M. Chalmot, A. & Kupper, L. (1988). Listening Comprehension Strategies in Second Language Acquisition. Applied Linguistics. 10, 4, 418-437

Öztürk, G. (2012). Effect of Context in Achievement Vocabulary Tests. Journal of Educational and Instructional Studies in the World. 2, 4,126-134.

Qian, David D. (2008) From Single Words to Passages: Contextual Effects on Predictive Power of Vocabulary Measures for Assessing Reading Performance. Language Assessment Quarterly. 5, 1,1-19.

Rahimy, R. and Moradkhani, N. (2012). The Effect of Using Grammaticality Judgment Tasks on Iranian Efl Learners' Knowledge of Grammatical Patterns.

Read, J. (1997) Assessing vocabulary in a second language. In Clapham, C. & Corson, D. (eds.) Encyclopedia of Language and Education. Language Testing and Assessment. 7, 99-107.

Read, J. (2000). Assessing Vocabulary. London: Cambridge University Press.

Rimmer, W. (2006). Grammaticality Judgment Tests: Trial by Error. Journal of Language and Linguistics. 5, 2, 246-261.

Rost, M. (2001). Teaching and researching listening. London: Longman.

Savignon, S. J. (1983). Communicative competence: Theory and Classroom Practice. Reading, MA: Addison-Wesley.

Schmitt, N. (2000) Vocabulary in Language Teaching. New York: Cambridge University Press.

Shohamy, E. (2001). The Power of Tests. London: Longman.

Smit, T. C. (2006). Listening Comprehension in Academic Lectures: A Focus on the Role of Discourse Markers. Unpublished Masters Dissertation, University of South Africa.

Spolsky, B. (2001). Assessment and Testing: Notes for the third Millennium. Testing into the third Millennium.[Online]Available: *http://www.russnet.org/fipse/* March, 2015

Stahr, L. S. (2009). Vocabulary Knowledge and Advanced Listening Comprehension in English as a foreign language. Studies in Second Language Acquisition, 31,577-607. doi: 10.1017/S0272263109990039.

Valdés, G. and Figueroa, R. (1994). Bilingualism and Testing a Special Case of bias. Norwood, NJ: Ablex Publishing Corporation.

Wang, M., and Bakken, L. L. (2004). An academic writing needs assessment of English-as-a-second-language clinical investigators. The Journal of Continuing Education in the Health Professions, 24, 181-189.

Weir, C.J. (1990). Communicative Language Testing. Hemel Hempstead: Prentice-Hall.

Widdowson, H. G. (1983). Learning Purpose and Language Use. Oxford: OUP.

Wilcox, A. and Medina, A. (2013).Effects of semantic and Phonological Clustering on L2 Vocabulary Acquisition among Novice Learners. System, 41,4, 1056-1069.

Wilson, J. & Komba,S.C. (2012). The Link between English Language Proficiency and Academic Performance: A Pedagogical Perspective in Tanzanian Secondary Schools. World Journal of English Language. 2, 4,1-10.

Young, L. (1994). "University lectures − Macro-structure and Micro-features". In Flowerdew, J. (ed.) Academic Listening. Cambridge: Cambridge University Press. Pp. 159 −176.

Žlábková, L. (2007). Testing Grammar: Using Multiple Choice Tests versus Translation. Bachelor Thesis, Faculty of Education, Masaryk University Brno.