International Journal of Linguistics, Literature and Translation (IJLLT)
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

A Critical Review of the Oxford Q-Skills Placement Test at a Saudi Arabian University

Mohammed Siddique Kadwa[1]* and Ayub Sheik[2]
[1]*Language Specialist, English Language Center, Taibah University, Madinah, Saudi Arabia*
[2]*Professor, Department of English Education, University of KwaZulu Natal, Durban, South Africa*
Corresponding Author: Mohammed Siddique Kadwa, E-mail: dr.m.s.kadwa@gmail.com

ARTICLE INFORMATION

Received: December 08, 2020
Accepted: January 14, 2021
Volume: 4
Issue: 1
DOI: 10.32996/ijllt.2021.4.1.19

KEYWORDS

Cambridge VRIPQ (2013) model, CEFR, English as a Foreign Language (EFL), Preparatory year, Q-Skills Placement Test (QSPT)

ABSTRACT

There is a genuine need to ascertain Saudi Arabian university entrants' English language abilities upon admission. In order to accurately determine the English language levels of students, this study evaluates the Q-Skills Placement Test (QSPT) designed by Oxford University through the most recent evaluative model in English for Speakers of Other Languages (ESOL); the Cambridge VRIPQ (2013) model. The data used to evaluate the efficacy and predictive power of the QSPT is obtained through both quantitative and qualitative approaches. Quantitatively, the QSPT results are statistically analyzed, whilst from a qualitative approach, interviews, and focus group discussions with teachers and students provide depth and insight. The strengths and weaknesses of the placement test are discussed here from a critical perspective with a view towards the improvement of the test. Although the test proved to be valid, it lacked the acknowledgement of the students' context and was not able to discriminate accurately for students who scored less than 30% on the test.

1. Introduction

Although much of second language assessment research during the past 20 years has made tremendous gains by identifying, categorizing, and defining test constructs, the usefulness of a test is often questionable. At the heart of the problem lies the globally recognized CEFR which is used as a reference tool by test designers, publishers, curriculum designers, and language practitioners to describe learners' English language proficiency levels. The levels, however, are problematic as they have been found to be ambiguous at times (Martyniuk, 2010; & Papageorgio, 2009) and restrictive in describing the uppermost C2 level. For instance, Saville (2012) points out that the CEFR does not describe the competence of well-educated native speakers and suggests the possible inclusion of a D level. However, this has not yet been acknowledged in the CEFR scales. Martyniuk (2010) also points to the absence of descriptors below the A1 level. This is of particular concern in Saudi Arabia, where learners' futures depend on their English language scores for local and foreign university entrance purposes, as well as for employment. Therefore, there is a dire need to evaluate the tools that are used for assessment to rationalize and justify their use.

This study investigates the phenomena surrounding language assessment for university purposes from a critical perspective by problematizing and exposing the flaws in the Q-Skills Placement Test. From a philosophical standpoint, the critical paradigm is known for its ability to liberate an individual or group as previous approaches neglected issues surrounding power and social inequalities (Cohen, Manion & Morrison, 2007). In language testing, the use of the critical paradigm is gaining in popularity as testing is "a quintessentially institutional activity, facing increasing scrutiny from this perspective" (McNamara, 2000, p. 76).

The purpose of this paper is to highlight the most pertinent issues raised in a doctoral thesis (Kadwa, 2017) that investigates how students' English language levels are established at a Saudi University. It also seeks to gain a deeper understanding of the reasons for these levels and identifies key areas for improvement of the placement testing system.

2. Background

At Saudi Arabian universities, the teaching of English is a challenging pedagogical process (Alhmadi, 2014 & Al-Nasser, 2015). Firstly, the switch from Arabic as the language of learning and teaching in Saudi Arabian high schools to English at universities

AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT
Your gateway to world-class research

creates numerous literacy and linguistic challenges to Arabic students Abdelgadir & Ramana (2016). To deal with this predicament, a preparatory year of studies was formulated. The primary function of the preparatory year is to prepare school leavers with competent academic skills in order to pursue their studies in English successfully. For example, the fields of medicine, engineering, business, and IT all require students to be proficient in the language of instruction, which is English. Moreover, there are limited places in the prized colleges of medicine and engineering which makes the preparatory year highly competitive.

Secondly, even though English is taught at high schools in Saudi Arabia, the lack of standardized tests of English during the high school phase (Siddiek, 2011) means that students' English-language competencies cannot be accurately established. It could also be one of the causes of students' inability to cope with the level of English in their first year of university studies. This could also be a major contributor to the high drop-out rate by preparatory-year university students in Saudi Arabia (Prokop, 2003).

Thirdly, a lack of cohesion between the secondary schooling phase and higher education organizations with regards to students' English abilities challenges English language teachers at the university level. This is due to the fact that students of mixed linguistic competencies are placed in the same class. This is further convoluted by the competitive nature of the preparatory year, where a limited number of openings are available in undergraduate colleges at the end of the preparatory year with stringent entrance criteria. The prized colleges of Medicine and Engineering select only a small proportion of preparatory-year students every year who are also the highest achievers. It is noteworthy to point out that the entire program uses standardized assessments for all mid-semester and final examinations, with a small component (5-20%) of continuous assessments that are teacher-graded.

Therefore, the challenges faced by students and teachers require two issues to be addressed. Firstly, English language levels need to be determined with precision. Due to the high-stakes nature of the preparatory year program, the English language assessments need to have high levels of validity and reliability. For this, the parameters between high-level and low-level students need to be established and benchmarked against a set of standards like the CEFR. Secondly, students' English language abilities are to be determined against these predetermined levels. Once this is accomplished, the hope is that students will be placed in more homogenous classes and instruction can be modified accordingly. By identifying learners' English language levels and placing students in classes accordingly, the assessment would be of benefit to students and teachers.

3. Literature Review

Second language assessment research is dependent on second language acquisition research. As a result, this study is informed by Krashen's theories in language acquisition and multilingualism. Krashen (1982) had built upon and collated previous research in the field of second-language acquisition theory, which culminated in the proposing of five hypotheses in second-language acquisition. They are (1) the acquisition-learning distinction, (2) the natural-order hypothesis, (3) the monitor hypothesis, (4) the input hypothesis, and (5) the affective-filter hypothesis. Theories in multilingualism are also important as they acknowledge and promote the acquisition and use of multiple languages. This is in stark contrast to previous Eurocentric views of bilingualism and multilingualism, which focused mostly on replacing one language with another. Modern theories in multilingualism are based on more pragmatic, accepting, and pluralistic views of other languages, which contradict previous theories on second language acquisition that viewed languages other than the target language in second-language acquisition as problematic (Garcia, 2009). Theories can influence the outcome of a study and have to be acknowledged for their ability to do so.

3.1. Benchmarking in EFL

The process of benchmarking refers to the setting of standards for policy, curriculum, pedagogy, or test-designing in language studies. It can also be viewed as an agreement designed to promote unity and consistency among policy-makers, curriculum-designers, textbook-publishers, teachers, and test- designers. In testing, benchmarks are very often referred to as criterion-referenced validity, as they set out the standards upon which tests are designed (Hughes, 2002). Once the standards are determined, curriculum designers translate the theoretical constructs into operationally viable course learning outcomes and learning objectives (Bachman & Palmer, 1996; & Gervais, 2016). Benchmarked tests also measure the effectiveness of a language program, establish the progress made by students in a program, and determines the readiness-levels of students before entering a program (Picard, 2006). Therefore, benchmarking is a crucial curriculum design activity and an important validation exercise in high-stakes testing.

In the EFL field, the Common European Framework of Reference (CEFR) is utilized as a referent by many institutions around the world. It has to be noted, however, that it was initially designed to bring together the 14 European languages of the EU by providing "a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc." (CEFR, 2001, p. 1). It is the outcome of years of research by the Council of Europe and was adopted via a resolution by the

European Union Council in 2001 as a means of validation. In simple terms, it has three levels or descriptors used to categorize a language user's proficiencies. They are the basic user, the independent user, and the proficient user. These level descriptors are further subdivided as indicated in Table 1.

Table 1: ***CEFR levels***

BASIC		INDEPENDENT		PROFICIENT	
A1	A2	B1	B2	C1	C2

At each of the six levels, the global scale of the CEFR outlines "Can-Do" statements to describe what a learner is supposed to be able to do. It systematically orders the "Can-Do" statements according to the following categories: Listening-comprehension, reading-comprehension, spoken-production, spoken-interaction, and writing. After the adoption of the CEFR by the Council of Europe, studies (Figueras et *al.,* 2005; Picard 2006) have advocated that language courses and tests be designed based upon the benchmarks and level descriptors found in the CEFR. The CEFR has also been accepted by major English-language testing organizations such as IELTS and TOEFL as the basis upon which tests are designed. The test scores on these high-stakes tests are mapped against the CEFR scales as indicated in the table below.

Table 2: ***Correlation of IELTS and TOEFL with CEFR***

CEFR LEVELS		IELTS	TOEFL iBT
Proficient User	C2	9.0	118
	C1	7.0. 7.5, 8	95
Independent User	B2	5.5, 6.0, 6.5	72
	B1	4.0 -4.5	42
Basic User	A2	-	-
	A1	-	-

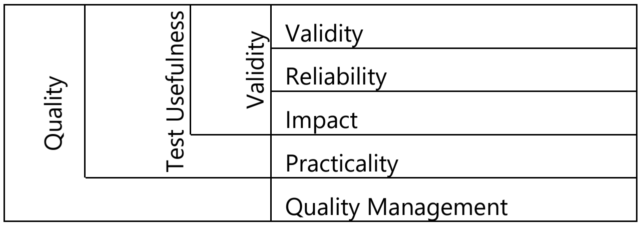
In so doing, test-designers can evaluate their tests by comparing their test scores against the scores of another test designed for a similar purpose as both the tests are weighed against the same criteria. For instance, the IELTS and TOEFL tests can be measured against each other using the CEFR scales as a common point of reference. Similarly, the claims of language curricula and language tests can be evaluated by comparing them against tests that purport to measure concordant constructs. For example, the globally accepted IELTS test describes test-takers' performance by providing CEFR levels for each of the four skills assessed, speaking, writing, listening, and reading. These claims can be ascertained by using a different test, such as the TOEFL or the QSPT to verify the claims. Moreover, the CEFR scales guide publishers in creating learning objectives and appropriate tasks.

3.2. The Cambridge VRIPQ model

The Cambridge VRIPQ (2013) model was chosen for the analysis of the data in this study as it was established to be the most relevant model and allows for more accurate appraisals of high-stakes tests. This was because former theoretical representations in language testing attended solely on validity and reliability discussions. These debates, however, began to shift their focus as Messick (1989) instigated the categorizing and prioritizing of different aspects of construct validity. Messick posits that a unitary approach towards the notion of validity is one that acknowledges and stresses the inter-related nature of the various categories of validity; a perspective shared by another leading scholar, Kane (2006). Additionally, Messick and Kane argue that reliability, as a construct, is merely a part of validity. Later, Bachman and Palmer (1996) built on the construct-validity and reliability concepts by identifying other key notions of 'practicality, interactiveness, authenticity and impact', referring to the usefulness of a test. In more recent times, Weir (2005) classified validity as consisting of five different aspects: Theory-based validity, context validity, scoring validity, construct validity, and consequential validity. Furthermore, Weir (2005) advocates for both holistic and analytical approaches to addressing these constructs.

As such, this study employs qualitative techniques by using interviews to holistically evaluate a language test in the Saudi university context, and analytically by using Pearson's correlation coefficient and a scatter plot. Nonetheless, the contributions made by Kane (2006), Bachman and Palmer (1996), and Weir (2005), along with Cambridge University's quest for a unified inter-related model for evaluating a test, culminated in the production of the Cambridge VRIPQ (2013) model, which can be seen in Table 4. As such, the Cambridge VRIPQ (2013) model accounts for both the social and cognitive dimensions of the test and the test-taker.

Figure 1: ***Illustration of the Cambridge VRIPQ (2013) model***



Such an approach accepts the multi-faceted and interrelated nature of language testing constructs and aims to be as fair as possible. Below, each of these constructs will be discussed separately considering the Cambridge VRIPQ (2013) model.
According to the Cambridge VRIPQ (2013) model, reliability is part of the broader concept of validity and concerns itself with the degree of consistency of the test results and promote freedom from errors in measurement. This requires both criterion-related aspects and scoring-related aspects of the test to be accounted for. In terms of criterion-referencing, all aspects of the test reflect a particular outcome of the course based on the CEFR levels described earlier. Although various forms of statistical evidence can be utilized to measure reliability, this study uses a scatter plot chart to ascertain the reliability of the results of the QSPT.

Impact concerns itself with the consequences of the test and its results on the test- taker, the educational system, as well as society in general. According to the Cambridge VRIPQ (2013, p. 13) model, tests should not be biased or favor certain groups of test-takers. They should also take into account the cultural contexts and cognitive processes involved in the test. Furthermore, the test should avoid having any negative consequences. In order to ascertain the impact of the QSPT, teachers and students were interviewed. This allowed for the phenomena to be understood from more than one viewpoint.

Practicality forms part of the ‘test usefulness’ concept put forth by Bachman and Palmer (1996). It refers to the resources required to design and administer a test in its intended context. The Cambridge VRIPQ (2013, p. 30) model views a practical test as “one that does not place an unreasonable demand on available resources”. Practicality takes into consideration the length of the test, training, and availability of test administrators, test venues, and the costs of the tests. The QSPT is a two-hour long test that does not require much training of administrators. A guideline for proctoring the test was given to each test administrator.

Quality, according to the Cambridge VRIPQ (2013, p. 13) framework, is about the “policies, processes, and procedures that enable an organization” to consistently achieve fitness for purpose. Quality control measures, therefore, would span the entire testing process across the organization using the test, to ensure consistency of the test results and its interpretations. The quality assurance activities in the administration of the QSPT included (1) clearly defined operating procedures when it comes to document control, (2) legally relevant records-management procedures, (3) proper risk management procedures that involved appropriate planning and reflection measures before and after test administration, (4) corrective actions and policies in the effect of errors, and (5) internal audits.

The Cambridge VRIPQ (2013) model was found to be the most recent and complete model when it comes to evaluating an English-language test. Furthermore, the model’s cyclic nature in its quest for persistent enhancement has managed to integrate previously-debated topics surrounding validity, consequence or impact, and quality into a unified model.

4. Research Questions

This study seeks to find answers to the following questions:

1. How are the language proficiency levels of foundation year students determined at a Saudi university?
2. What are the reasons behind Saudi Arabian university entrants’ language proficiency scores?

5. Method

Kadwa’s (2017) study was undertaken through the post-positivist framework of critical realism. The critical-realist paradigm is one that views the ontological nature of existence as interpreted through social conditioning and independent human thought (Cohen Manion & Morrison, 2007; Wahyuni, 2012). Epistemologically, the critical-realist framework recognizes only observable phenomena as credible proof and focuses on exaggerating the phenomena under scrutiny by amplifying and elaborating on the multiple layers of contexts wherein the study takes place. Axiologically, the critical-realist study adopts an external perspective of the phenomena but acknowledges the bias of the researcher. Therefore, a researcher’s bias is the result of socio-cultural interactions, upbringing, and different world or religious views. Since this study is conducted in an Islamic, Arabic setting, the researchers' intimate working knowledge and understandings of the target culture allows them to delve deeper into relevant themes but also has the potential to inhibit their judgments and conclusions by forming generalizations.

To achieve its aims, both qualitative and quantitative data elicitation techniques were employed. Quantitative data is obtained via two placement tests. The tests were conducted at the beginning of the study to establish the language levels of students upon university entrance. The preparatory year program is compulsory for all students in the Medical and Engineering sciences streams in the preparatory year at the university wherein this study takes place, although, in recent years, some Saudi universities are making this optional. The tests were conducted in the first two weeks of the academic year. For placement purposes, two versions of the Oxford University Q-Skills Placement Test (QSPT) were used and were found to be both content-validated and construct-validated as they tested skills and learning outcomes that are addressed in the course books and the Q-Skills series.

Each of the two-hour-long placement tests consists of one hundred- items in multiple-choice-questioning format. In order to verify the data, both versions of the test were used on the same population. Each test has two sections. The first fifty items of each test focus on reading comprehension and writing skills, whilst the remaining fifty items focus on listening comprehension and speaking skills. The listening section begins one-hour into the test whereby the test invigilator plays an audio track for the test-takers. The answers are recorded on a bubble sheet and electronically scanned using a Datawin scanner and associated computer software. In terms of sample size, the placement test was conducted on 1149 students at the same time in different classes at the university. Each class had an English- language invigilator along with an Arabic- speaking instructor from another faculty in the university.

The interpretation of the placement test scores and their correlation with other international tests of English, along with the CEFR scales was supplied to the university by the designers of the test, Oxford University Press. This can be seen in the conversion chart below.

Table 3: *Oxford Q-Skills Placement Test Conversion Chart*

Q: Skills Placement Level and students' grades (%)	TOEFL (Paper)	TOEFL (iBT)	IELTS	CEFR
Level 1 (0-30)	0-393	0-29	0-2.5	A1 (Breakthrough)
Level 2 (31-50)	397-435	30-40	3-3.5	A2 (Waypoint)
Level 3 (51-70)	437-473	41-57	4-4.5	B1 (Threshold)
Level 4 (71-90)	513-547	58-74	5-5.5	B2 (Vantage)
Level 5 (91-100)	550-587	75-90	6-7	C1 (Effective Operational proficiency)

Adapted from: http://www.relod.ru/files/files/tablitsa_urovnei_242.pdf

The reasons explaining students' English language levels were obtained through focus group discussions and interviews with 3 teachers and 6 students at scheduled intervals. This qualitative part of the study involved three teacher participants and six student participants. All the participants were males due to the accessibility of participants and a show of respect for local laws and customs. Although many teachers and students indicated their willingness to participate in the study, the three teacher participants and six student participants were selected based on certain predetermined criteria. For the teacher participants, the researchers had set a minimum of five years of Middle East teaching experience as a requirement. As for the student participants, two students were randomly selected after being categorized according to their averages on both the tests. It is worthwhile to note that the interviews were intended to gain an in-depth understanding of the reasons behind student scores and that a total of nine participants would allow for this. Moreover, the focus group discussions and interviews were held twice over two consecutive semesters to capture any changes in perceptions and attitudes as well as language abilities in the case of the student participants.

For the analyses of the data, two different approaches were taken. Firstly, the quantitative data was analyzed using Microsoft Excel and SPSS for calculating student averages, differences in averages according to CEFR levels, means, and standard deviations. The qualitative data stemming from the interviews and focus groups were transcribed from audio recordings and thematically coded for recurring topics. As such, the qualitative data was meant to substantiate and delve deeper into the phenomena under scrutiny by probing the underlying causes of students' varied test scores.

6. Quantitative Findings

The findings of the first part of this study were quantitative in nature. It involved a comparison of two versions of the same test which were administered on the same student population within a one-week time span. The second placement test used an alternative version of the QSPT which was supplied by Oxford University Press. The second test assesses the same learning outcomes as the first test but with different content. The data was generated through the Datawin bubble-sheet scanning software program which produces the results in Microsoft Excel sheets. The scores were used to classify students in relation to their respective CEFR levels. Thereafter, the data was computed on the SPSS software program to determine the mean and standard deviation of both tests. The same program was used to generate a scatter plot to give a visual depiction of the relationship between the two tests.

6.1. 1st Placement Test CEFR Ranges

The first placement test was taken by a total of 1275 students at the same time, and the results according to the CEFR ranges can be seen in the table below.

Table 4: ***1st Placement Test Results***

CEFR LEVELS	A1	A2	B1	B2	C1	Total
Test Scores	0-30	31-50	51-70	71-90	91-100	
N	217	775	188	89	6	1275

The results of the first placement test indicated that the largest portion of test-takers (61 percent were placed at the A2 (waypoint) level. Of the remaining 39 percent, 17 percent of the students were at the A1 (breakthrough) stage, 15 percent of students were at the B1 (threshold) stage and a further 7 percent were at the B2 (vantage) level. Only six test-takers were at the C1 level which is considered to be the Effective Operational Proficiency level according to the CEFR.

The second placement test, which was administered a week after the first placement test, was selected for the placement of test-takers into homogenous groups. The total number of test-takers increased from 1275 in the first placement test, to 1356 in the second placement test as some test-takers registered after the first placement test. The results of the test according to the CEFR scales can be seen in the table below.

Table 5: ***2nd Placement Test Results***

CEFR LEVELS	A1	A2	B1	B2	C1	Total
Test Scores	0-30	31-50	51-70	71-90	91-100	
N	199	862	211	82	2	1356

The scores of the second placement test show that the majority of test-takers (64 percent) were categorized at the A2 (waypoint) level. The second-largest proportion of test-takers were placed at the B1 (threshold) level and formed about 16 percent of the population. A further 15 percent of the test-takers were at the A1 (breakthrough) level, and 6 percent of students were placed in the B2 (vantage) level. Only two test-takers scored higher and were placed in the C1 (effective operational proficiency) level.

In comparing the data obtained from the first placement test with the data from the second placement test, it is evident that both tests group test-takers in similar-sized populations according to the CEFR scales. This was the first indicator that the tests measure what they purport to measure, meaning that the tests show a high level of validity. So, to verify this, the results of each student who took both the tests were computed using the one-way Anova test to ascertain the differences in means and standard deviations, and a scatter plot to show the correlation of students' scores on both tests in one chart. It is worthwhile to point out that only 1149 test-takers' scores were used in this process as they took both the tests. The placement tests demonstrate a significant level of similarity in both the mean and standard deviation, as indicated in the table below.

Table 6: ***Comparison of Mean and Standard Deviation***

	1st Test	2nd Test
N =	1149	1149
Mean =	42.69	42.15
Standard Deviation =	16.29	15.22

Although the tests provided fairly similar means and standard deviations, the data was further computed on Microsoft Excel to show the difference between the first and second test based on the CEFR levels in Table 3. This was accomplished by using the average scores on both the tests to categorize test takers according to their respective CEFR levels. Thereafter, the difference between the two tests was calculated for each student. Finally, the average difference between the two tests was obtained and can be seen in the third row in Table 7 below.

Table 7: *Table of Differences in Average Scores*

Score	**0-30**	**31 – 70**	**51- 70**	**71 - 90**	**91 - 100**
CEFR Level	**A1**	**A2**	**B1**	**B2**	**C1**
Difference	**14.07**	**6.21**	**6.83**	**4.60**	**3.67**
N	**150**	**736**	**180**	**79**	**3**

The average differences in test-takers' scores indicate that the placement tests are stronger in determining the upper levels of B2 and C1. However, for students who were below 30 on average, the difference between the two tests was significant. In the A2 and B1 categories, the average difference was fairly similar. Therefore, a one-way ANOVA test was conducted using SPSS to verify if similar results would be obtained when comparing the standard deviations from means based on the CEFR levels. This can be seen in the table below.

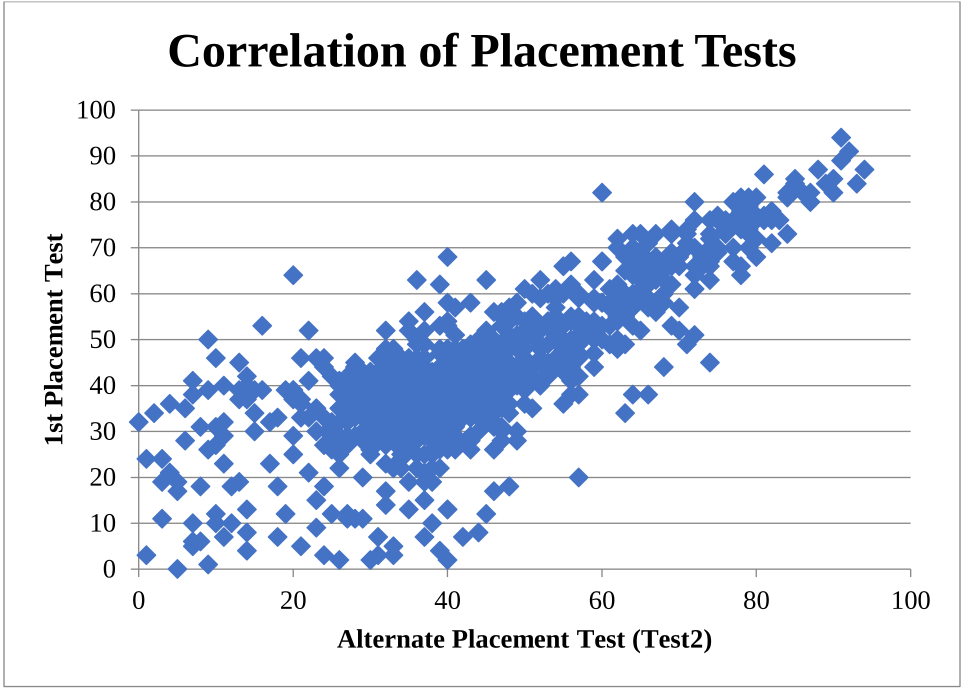
Table 8: *One-way ANOVA Test*

						95% Confidence Interval for Mean			
		N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
1st Placement Test	A1	150	21.7600	10.96528	.89531	19.9909	23.5291	.00	45.00
	A2	736	38.5897	6.48558	.23906	38.1203	39.0590	16.00	63.00
	B1	191	60.6963	7.54369	.54584	59.6196	61.7730	39.00	78.00
	B2	69	79.9565	5.11179	.61539	78.7285	81.1845	69.00	93.00
	C1	3	92.3333	1.52753	.88192	88.5388	96.1279	91.00	94.00
	Total	1149	42.6919	16.29051	.48059	41.7490	43.6348	.00	94.00
2nd Placement Test	A1	150	22.9200	11.67945	.95362	21.0356	24.8044	.00	50.00
	A2	736	38.4063	6.03815	.22257	37.9693	38.8432	17.00	64.00
	B1	191	58.3351	7.54933	.54625	57.2576	59.4126	38.00	82.00
	B2	69	76.9130	4.96677	.59793	75.7199	78.1062	66.00	89.00
	C1	3	90.6667	3.51188	2.02759	81.9427	99.3907	87.00	94.00
	Total	1149	42.1462	15.21957	.44900	41.2653	43.0272	.00	94.00

The one-way ANOVA test confirms the findings from Table 7 by showing higher standard deviations on both the placement tests at the A1 level. The standard deviations then decrease towards the C1 level proving that the Q-Skills Placement Tests are more reliable at the upper levels and decline in reliability as the levels decrease. This trend is noticed on both the placement tests in Table 8 above.

Moreover, the scatter plot chart, which can be seen below, indicates a fairly high level of correlation between the first and second tests. On the chart, the first placement test scores are illustrated on the vertical axis, whereas the second placement test scores are represented on the horizontal axis. The scatter plot confirmed the previously held assumption that a test-taker who scores high on one test will score high on the other. This is graphically evident when the data represented on a scatter plot are linear. In the scatter plot below, the line begins at 0 and continues straight towards 100 diagonally from the bottom left to the top right.

Figure 2: *Correlation of Placement Tests*



Although the two tests indicate fairly similar results for scores above 30, scores between 0 and 30 are randomly scattered. This suggests that scores in this range are not as reliable as the scores which are between the 31 to 100 range. However, it would seem that this is irrelevant because students in the 0 to 30 category will automatically be classified at the lowest level according to the CEFR descriptors, which is the A1 level. Nonetheless, we can conclude that to a large extent, the results are reliable as the two tests give mostly similar results for the majority of test-takers.

6.2. Qualitative Findings

The second part of the study used interviews and focus group discussions with teachers and students to ascertain the reasons behind students' showing varying levels of English on the test. Through the use of coding, recurring themes were highlighted.

6.3. Teacher Interviews

Focus group interviews were conducted with three English language lecturers on a voluntary participation basis. The participants were males from the United Kingdom, the United States, and Egypt respectively. Moreover, each of the participants had been teaching English in Saudi Arabia for between 5 and 10 years. For the purpose of validation, the semi-structured focus group sessions were held on two occasions during the academic year. The first round of interviews was held towards the end of the first semester, whilst the second round was conducted at the end of the second semester. Moreover, the same three teacher participants were used in the one-on-one interviews in both semesters.

Firstly, data from the teachers indicate that the standard of English-language education is very low in Saudi Arabia. After receiving 9 years of tuition in the schooling phase, many students cannot even operate at band A1 on the CEFR scales. Teachers regularly allude to the "quality of teachers" and an inefficient high school system as the major contributors to the. Even at the higher education level, teachers hint at systems that do not address the needs of the students, as well as a "mismatch" between textbooks and assessment practices.

Secondly, the educational fraternity in Saudi Arabia is reluctant to take ownership of the English language. Teachers felt that society had more confidence in internationally designed tests and course materials, as well as expatriate teachers for the delivery of English-language education. Related to this theme is the illusion created by the poor assessment practices at both high schools and the university. The false data originating from high school reports and poorly-designed university-level tests create an impression of high levels of advancement and progress. The teacher participants considered high school tests as not being "credible". This "gimmick", as one teacher described it, confirms that the Saudi Arabian educational fraternity is not prepared to take possession of the English language in Saudi Arabia.

Thirdly, teachers constantly made references to teachers' low levels of motivation, as well as students' low levels of motivation. Teachers were frustrated with various parts of the program. References were also made to money not being spent correctly by state and university institutions. Teachers regularly pointed to students being motivated by extrinsic factors, instead of willing to develop themselves from an intrinsic perspective. Furthermore, learners' low motivation could possibly be symptomatic of Krashen's (1982) I + 1 concept as the course may have been too easy or too difficult for learners.

The fourth recurring theme emanating from the teachers' data is that it is from a Western epistemological worldview. Teachers made regular references to the Islamic belief systems and the Middle Eastern culture not being adequately represented in both the tests and courses of the university. This signifies that the attitudinal and cultural differences between Western and Islamic societies relating to power and identity should not be downplayed, but rather acknowledged and incorporated into the design of both course materials and tests.

The fifth theme that resounds throughout the data coming from teachers concerns the excessive use of quantitative data elicitation techniques in tests. The assessment strategies in Saudi Arabia in general, and specifically at the university where this study was conducted, neglect qualitative data. More emphasis is given to quantitative data through the use of multiple-choice tests. Objective tests are suitable for assessing the receptive skills of reading, listening, vocabulary, and grammar, whilst the qualitatively-obtained subjective data which is used to assess oral and written production is often overlooked or given superficial importance.

The themes, emanating from the teachers' data were mostly of a negative nature. There was very little positive data with regard to their perceptions of the preparatory-year English-language program, as well as Saudi Arabian English-language education in general.

6.4. Student Interviews

One-on-one interviews were conducted with six male students aged between 18 and 20 on a voluntary participation basis. Although many students indicated their willingness to participate in the study, students were selected based on their average scores and CEFR level. Two students were selected from the A1 level (0-30), two students were selected from the A2 and B1 levels (31-70), and two students were selected from the B2 and C1 levels (71-100). For the purpose of validation, the semi-structured interviews were held on two occasions during the academic year. The first round of interviews was held towards the end of the first semester, whilst the second round was conducted at the end of the second semester.

The first theme elicited from the students was that high school level English education is sub-standard. Students frequently made references to feeling "bored", and not developing English-language skills at high schools. Students also confirmed that the assessment practices at high school were not representative of students' English-language abilities.

Secondly, there were two diametrically-opposed views held by the student participants with regards to the use of English as a language of education in Saudi Arabia. Whilst some students felt that there is a need to study English as it is an international language, other students felt that they should only be required to study English for certain fields of study.

The third theme emerging from students concerns their English- language abilities during the interviews. The amount of data and the level of English-language abilities reflected students' scores on the placement test. This confirms the strength of the QSPT in identifying the language levels of test-takers. High-level students and average-level students were able to respond with more qualifying arguments, whilst low-level students could only respond with simple responses to the questions posed to them. Nonetheless, the teacher participants pointed out that the QSPT lacked oral and written production tasks.

As with the data emanating from the teachers, not much of the data that emerged from the student interviews were of a positive nature. Most of the findings and recurring themes were indicative of a sense of boredom and dissatisfaction with English-language education in Saudi Arabia in general, but particularly at the high-school-level.

7. Discussion

This study aimed to establish the English literacy levels of Saudi Arabian students entering university. The language-literacy levels were ascertained using the Oxford QSPT which was found to be fit for purpose. In terms of the Cambridge VRIPQ (2013) test evaluation model described in the literature review section, and Bachman & Palmer's (1996) views of test usefulness, the test was found to be a valid form of assessment. Its validity stems from the fact that the QSPT was criterion-referenced and benchmarked according to the CEFR levels. However, the teacher interviews highlighted the need for the Islamic belief system to be incorporated into the design of the test. The teacher participants also indicated that the test is designed from a Western epistemological perspective with little regard given to the students' culture.

Through quantitative techniques, this study found that upon entrance to a Saudi Arabian university, the majority of students (64 percent) were at the A2 (waypoint) level. This is a reflection of the sub-standard level of English education in the general education sector. The poor standard of English language education confirms the evaluative findings of Alfahadi (2012), Alresheed (2012), and Mahib ur Rahman & AlHalsoni (2013) who call for redress in the English language curricula in Saudi Arabia. This is of particular concern as the government invests billions of Riyals annually into the education system, yet after nine years of English studies in the schooling phase, the majority of students cannot perform at the operational levels of B1 and B2.

Nonetheless, there was a small number of students with extremely high levels of English competencies. The reasons for students' above-average performance in the placement test were attributed to greater exposure to the English language, whilst students who were found to be at a below-average level upon entry to the English-language program turned out to have had very limited contact with the English language. During the student interviews, it became evident that the QSPT was fairly accurate in categorizing students according to the CEFR levels. Higher-level student participants were able to express themselves more than the average level students. Whilst low-level students could only answer in short, simple responses. In addition, the wide range of scores evidenced through the Oxford Q-Skills placement test, when taken in conjunction with the students' school-leaving report of English scores verify teacher observations that the assessment practices at high schools are unreliable.

Still on the topic of validity, the assessment of students' written production and oral production was neglected in the Q-Skills Placement Test. These two areas need to be incorporated into the design of the placement test for it to strengthen its claims of being able to accurately assess test-takers' language abilities. In so doing, the test will counteract the overdependence on the multiple-choice questioning technique and also provide a truer reflection of students' English-language abilities upon entrance to the university preparatory-year program.

In terms of reliability, the Oxford Q-Skills Placement Test was unable to identify students who were below the A1 level. Moreover, the reliability of the scores of test-takers at the A1 level is doubtful. This was established by comparing the average scores on both versions of the test and analyzing the means and standard deviations. In addition, the QSPT could not identify areas of weaknesses, such as reading, writing, grammar, or listening. Such a diagnostic function of the test could have made the assessment more relevant to teachers.

The quantitative data in this study suggest that the CEFR level A1 should not be considered the first level. This became evident with higher levels of deviations from the means on both the placement tests when compared with the other CEFR levels. This finding coincides with Saville's (2012) assertion that the CEFR is in need of more levels. Although Saville (2012) focuses on a possible D level, this study suggests possible levels before the A1 level. For the CEFR to be accepted as a global framework of reference, it has to acknowledge that Arabic learners are more linguistically and culturally distant than European speakers and require CEFR levels to specify such standards.

The qualitative data supports the quantitative data in that those students who scored low and were classified as A1 level students on the QSPT placement tests were unable to hold long meaningful conversations during the interviews. On the other hand, the few student participants who were at the B2 and C1 levels were more able to express themselves. This phenomenon was further corroborated during the teacher interviews where the participants expressed dismay at the wide schism between high and low-level students in the same class. They also complained that teaching multi-level classes is extremely challenging, which is why the use of the QSPT to create homogenized classes was welcomed. Relating the challenges to Krashen's (1982) Input Plus One (I + 1) concept, it is understandable that teachers are pressured, and students feel left out as the pace of the lessons cannot be to every student's liking. The teachers did, however, highlight the fact that the over-reliance on the multiple-choice questioning technique and the lack of oral and written production weakened the credibility of the QSPT.

A limitation of this study, however, was that it only employed male subjects on a single campus in Saudi Arabia. Therefore, the findings should be absorbed from a case study perspective and generalizations should be understood together with complementary studies that were conducted on a wider scale and involve female participants. Another limitation of this study was that the student scores were interpreted entirely upon the recommendation of a conversion chart provided by the publishers of the test. A future study that uses the IELTS or TOEFL to verify or deny the claims made by the publishers would be of great benefit to the institutions that rely on the Oxford QSPT.

8. Implications

The findings of this study have implications for test designers, curriculum designers, and teachers. Firstly, although the QSPT was found to be fairly reliable overall, it was not as convincing for scores that were at the A1 level (0-30). This implies that there is a high possibility that the scores at this level are inflated due to the multiple-choice method of testing. Moreover, stemming from the qualitative data, it is evident that teachers are not convinced with the accuracy of the QSPT's results due to the absence of oral and written production in the test. The findings of this study also imply that the assertions as to the CEFR levels of the test need to be further verified and benchmarked. Lastly, the wide range of student scores suggests that a one-size-fits-all approach should be abolished in the preparatory year English language program.

9. Recommendations

The study suggests that the approach taken by the Ministry of Education in Saudi Arabia concerning the teaching of English in the schooling phase should be revised. The wide range of scores evidenced through the Oxford QSPT, when taken in conjunction with the students' school-leaving report of English scores verify previous assumptions and teacher observations that the assessment practices at high schools are unreliable. The standardization of assessments at certain junctures during the high school phase is a possible solution to this dilemma.

One of the motivating points in selecting the Oxford QSPT for placement purposes was the fact that the Oxford Q-Skills series was used in the program. If all textbook publishers offer institutions a placement test along with their books and materials, a mismatch between the placement test and the course materials would not be prevalent. However, in so doing, publishers need to benchmark their tests against other placement tests from different publishers, or against internationally-recognized tests such as the TOEFL or IELTS.

Another recommendation is that preparatory courses should change their approach to meet the requirements of the student population. Instead of holding a one-size-fits-all outlook, curriculum planners need to create courses that are designed for the students' different language levels. The personalization of courses based upon students' needs would also tie in with the preparatory year program's objectives of developing academic readiness, rather than merely serving as a filter program that only restricts access into prized undergraduate colleges.

The findings of this study also exposed the inability of the Oxford QSPT to accurately discriminate between test-takers who scored below 30 on the test. Similar renowned tests such as the IELTS and TOEFL are also unable to accurately describe low-level performance accurately. This brings into question the effectiveness of the CEFR. One suggestion is to add another level below the A1 level. Another suggestion is for Saudi Arabia to develop its own national curriculum statement for English language education.

Another important recommendation for the publishers of the test is to acknowledge the students' belief systems and cultural practices in the design of the test. One way to do this is to consult with Arabic speaking teachers of English when writing test items for test-takers in Saudi Arabia. Alternatively, bilingual teachers of English could be trained as test writers for a Middle Eastern version of the Q-Skills Placement Test. Since the textbooks are adapted for the Middle East, it would be logical to assume that the placement test can also be adapted to this context.

Lastly, the QSPT should be supplemented by a test of written production as well as a test of oral production. Although this would require assessors to be trained and more time spent on assessing students, the outcome will be a more accurate assessment of students' English language abilities.

10. Conclusion

In probing the assortment of strata related to the pedagogical challenges of teachers and students in the preparatory-year English-language program at a Saudi Arabian university, Kadwa's (2017) study broadens our understanding of how a placement test may be both beneficial and harmful at the same time. The study was undertaken with the express aim of seeking a deeper understanding and insight into the preparatory year by evaluating the QSPT which was held at the beginning of the academic program. This was achieved through the use of the Cambridge VRIPQ (2013) model which allowed for the evaluation of the test from Validity, Reliability, Impact, Practicality, and Quality-related arguments. In light of the findings of this study, there is an urgent need to review the designing and administrative strategies of placement tests for university entrance purposes in Saudi Arabia.

Funding: This research did not receive any funding from any agency in the public, commercial, or not-for-profit sectors.

References

[1] Abdelgadir, E., and Ramana, V.S.V.L. (2016). Challenges of Teaching English to Arabic Students. *International Journal of English Language, Literature, and Humanities* (IJELLH), *4*, 221-227.

[2] Alfahadi, A. (2012). Saudi teachers' views on appropriate cultural models for EFL textbooks: Insights into TESOL teachers' management of global cultural flows and local realities in their teaching worlds [Unpublished doctoral dissertation]. University of Exeter.

[3] Alhmadi, N. (2014). English Speaking Learning Barriers in Saudi Arabia: A Case Study of Tibah University. *Arab World English Journal* (AWEJ), *5*(2), 38-53.

[4] Al-Nasser, A.S. (2015). Problems of English language acquisition in Saudi Arabia: an exploratory-cum-remedial study, *Theory and Practice in Language Studies*, 5(8), 1612-1619.

[5] Alresheed, S. (2012). *Exploring the nature of Saudi English teachers' beliefs and attitudes toward EFL and its effect on their teaching practice*. Paper presented at The Saudi Scientific International Conference 2012. London.

[6] Bachman, L. F.and Palmer, A. S.(1996). *Language testing in practice*. Oxford University Press.

[7] Cambridge VRIPQ (2013). *Principles of good practice – Quality management and validation in language assessment*. Cambridge English Language Assessment.

[8] CEFR. (2001). *Common European Framework of Reference*. Cambridge University Press.

[9] Cohen, L., Manion, L., and Morrison, M. (2007). *Research methods in education.* Routledge.

[10] Figueras, N., North, B., Takala, S., Verhelst, N., and Van Avermaet, P. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing*, *22*(3), 1-19.

[11] Garcia, O. (2009). *Bilingual education in the 21st Century: A global perspective*. Basil/Blackwell.

[12] Gervais, J. (2016). The operational definition of competency-based education. *Journal of Competency-Based Education*, *1*, 98-106.

[13] Hughes, A. (2002). *Testing for language teachers.*: Cambridge University Press.

[14] Kadwa, M.S. (2017). *An Evaluation of an English Language Placement Test within the Saudi Arabian University Preparatory-Year Programme.* [Unpublished Doctoral Thesis]. University of KwaZulu Natal: Durban.

[15] Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4th edition), 17–64. American Council on Education/Macmillan.

[16] Krashen, S. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon Press.

[17] Mahib ur Rahman, M. and AlHalsoni, E. (2013). Teaching English in Saudi Arabia: Prospects and challenges. *Academic Research International, 4*(1), 112-118.

[18] Martyniuk, W. (Ed.) (2010). Aligning tests with the CEFR. Reflections on using the Council of Europe's draft manual. *Studies in Language Testing*, *33*, Cambridge ESOL/Cambridge University Press.

[19] McNamara, T. (2000). *Language Testing*. Oxford University Press.

[20] Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

[21] Papageorgiou, S. (2009). *Setting performance standards in Europe: The judges' contribution to relating language examinations to the Common European Framework of Reference*. Peter Lang.

[22] Picard, M. (2006). *Academic Literacy Right from the Start? A Critical Realist Study of the Way University Literacy is Constructed at a Gulf University.* Rhodes University: Doctoral dissertation at Rhodes University.

[23] Prokop, M. (2003) Saudi Arabia: the politics of education. *International Affairs*, *79* (1), 77-89.

[24] Saville, N. (2012). The CEFR: An evolving framework of reference. In E. Tschirner (Ed.), *Aligning frameworks of reference in language testing: The ACTFL proficiency guidelines and the common European framework of reference for languages*. Stauffenburg Verlag.

[25] Siddiek, A. (2011). Standardization of the Saudi Secondary school Certificate

[26] Examinations and their anticipated impact on Foreign Language Education.

[27] *International Journal of Humanities and Social Science*, *1*(3), 57-64.

[28] Wahyuni, D. (2012). The research design maze: Understanding paradigms, cases, methods, and methodologies. *Journal of Applied Management Accounting Research*, *10*(1), 69–80.

[29] Weir, C. J., Ed. (2005). *Language Testing and Validation: An Evidence-Based Approach. Research and Practice in Applied Linguistics*. Palgrave Macmillan.