International Journal of Instruction *October 2021 ● Vol.14, No.4*
e-ISSN: 1308-1470 ● www.e-iji.net *p-ISSN: 1694-609X*
pp. 373-390

Article submission code: 20200913184430
Received: 13/09/2020 Revision: 17/03/2021
Accepted: 09/04/2021 OnlineFirst: 02/08/2021

A UAE Standardized Test and IELTS Vis-À-Vis International English Standards

Maha Al Habbash
PhD Candidate, United Arab Emirates University, UAE, *201370115@uaeu.ac.ae*

Negmeldin Alsheikh
Assoc. Prof., United Arab Emirates University, UAE, *nalsheikh@uaeu.ac.ae*

Xu Liu
Student, Vennesla Municipality, Agder, Norway, *xuliu8619@gmail.com*

Najah Al Mohammedi
PhD Candidate, United Arab Emirates University, UAE, *201370306@uaeu.ac.ae*

Safa Al Othali
PhD Candidate, United Arab Emirates University, UAE, *201790170@uaeu.ac.ae*

Sadiq Abdulwahed Ismail
Dr., Hamdan Bin Rashid AL Maktoum Foundation for Distinguished Academic Performance, United Arab Emirates, *sadiq.ahmed@ha.ae*

This convergent mixed method study aimed at exploring the English context of the widely used Emirates Standardized Test (EmSAT) by juxtaposing it to its sequel, the International English Language Testing System (IELTS). For this purpose, the study used the Common European Framework of Reference (CEFR) international standards which is used as a benchmark for both tests. The study focused on comparing the EmSAT and IELTS test specifications and their alignment with CEFR standards. The test takers reported on the EmSAT in terms of five categories: Test Scoring, Test Presentation and Format, Test Delivery, Test Structure and Preparation Practice. The results of this study revealed that both the EmSAT and IELTS are not aligned rigorously with the CEFR standards. Moreover, the EmSAT aligned mostly with the lower measurement levels of the CEFR while the IELTS aligned with the higher levels of CEFR. The test takers of the EmSAT reported some advantages and disadvantages about the EmSAT. Precisely, the students assigned high agreement with the EmSAT Test Scoring, Test Presentation and Format, Test Delivery respectively and to less degree to Test Structure and Preparation Practice.

Keywords: EmSAT, IELTS, CEFR, test specifications, test alignment, standardized test

Citation: AlHabbash, M., Alsheikh, N., Liu, X., AlMohammedi, N., AlOthali, S., & Ismail, S. A. (2021). A UAE standardized test and IELTS vis-À-vis international English standards. *International Journal of Instruction*, *14*(4), 373-390. https://doi.org/10.29333/iji.2021.14422a

INTRODUCTION

The Emirates Standardized Test (EmSAT) is a national system of standardized computer-based tests applied in the United Arab Emirates which is based on the Emirates national English standards and aligned with the Common European Framework of Reference (CEFR). The EmSAT English standardized test measures grade 12 students' skills and knowledge as they complete their general education and enter higher education institutes. The EmSAT also provides decision makers with data for college admission and placement (The UAE-Ministry of Education [MOE], 2017). Since it was launched in 2016, the EmSAT has gradually replaced the International English Language Testing System (IELTS) as a college entry requirement, and it is mandatory for all grade 12 students (MOE, 2017). The EmSAT English test requires specific English language skills including grammar knowledge, vocabulary knowledge, reading comprehension, and writing skills. Both IELTS and EmSAT are considered standardized tests. While EmSAT focuses on measuring reading, writing, grammar, and vocabulary knowledge, IELTS measures the four main skills: reading, writing, speaking, and listening. The scoring system for both exams are completely different, although both were aligned with the CEFR proficiency descriptors. For example, IELTS' score system is ranged from a band score of 1 to 9, while EmSAT is ranged from 100 to 2000 (MOE, 2017).

Before conducting the study, a preliminary interview was conducted with two instructors in the Foundation program at one of national universities in the UAE. The aim of this interview is to furnish common themes for comparison and probe new ways to examine both tests. The instructors have experience with the two tests and at that time, they were involved in preparing and pr'octoring the EmSAT. General overarching issues emanated from this interview. For example, the two instructors indicated that "they have limited information about the EmSAT test", whereas limited data is available, as they said: "they have only one sample in the EmSAT official website". They also indicated that they build their pedagogical techniques for students' test preparation based on the primary notes that they take while the students are taking the EmSAT. Additionally, they indicated that they neither have sufficient training sessions nor available materials for EmSAT, which made those instructors adapt "a reverse engineering strategy" by which they weave new forms of instructional techniques and content knowledge for directly preparing students for the test. This leaves little room for authentic communication and functional use of the language as they indicated that their ways of delivering instruction for preparing the students are traditional. Furthermore, they specified that the limited timeframe given to prepare students to achieve a score of 1250, which is equal to 5.5 on the IELTS, is very crucial as those instructors were pressed by time to advance students' achievement level in terms of grammar, vocabulary, reading and writing in order for the students to reach a score of 1250. Moreover, they signified that the difficulty levels of the test are unstable as the instructor noted with regard to the writing topics: "one time they gave the students an easy question such as what do you like to eat? Another time was about a solar energy, so there is no gradual transition in questions difficulty. To have a comprehensive picture, the purpose of this article is to explore the EmSAT test through qualitative and quantitative means. So, the study aims at: 1) Comparing the

EmSAT test specifications with the IELTS test specification; 2) Investigating the extent to which the EmSAT and IELTS align with the Common European Framework of Reference (CEFR); and 3) Furnishing a washback mechanism by collecting foundation college students' self-report about their experiences with the EmSAT test; and 4) Checking how these multiple means validate and consolidate the results. Therefore, this study is guided by the following questions:

(1) How are the EmSAT test specifications compared to the IELTS test specifications?
(2) To what degree do the EmSAT and IELTS align with CEFR standards?
(3) What do college students report about their experience with the EMSAT?
(4) How do the converged results from EmSAT and IELTS comparison, alignment with international standards, and students' self-reports inform us about the nature of the EmSAT as a standardized test?

Background

Testing is a part of assessment that is used for different purposes. Educators usually use tests to render information that assists them in making a lot of decisions related to the curriculum, the instruction, or to the learners. Particularly, the purpose of testing is strongly related to the needs of both teachers and learners who benefit from the testing context (Fulcher, 2010). One common type of testing is the standardized test. Standardized tests in language are tests that measure how proficient learners are in using particular language skills. For instance, the four-main international ESOL (English for Speakers of Other Languages) tests including: The Test of English as a Foreign Language (TOEFL), the First Certificate in English (FCE), the International English Language Testing System (IELTS), and the Test of English for International Communication (TOEIC) are constructed by different institutions and based on specific social and cultural contexts, which have distinctive structures, scoring systems, and purposes (Stoynoff, 2011). Within the UAE context, the Emirates Standardized Test (EmSAT) was established to serve the UAE education aims such as entering the global English testing market in the near future (MOE, 2017). Although the EmSAT is currently still locally based, it has aligned with international standards such as the Common European Framework (CEFR), just like those well-established international tests mentioned above. The CEFR was founded in 1949 by an intergovernmental cooperation organization in Strasbourg, France with the aim of providing a common foundation for the expansion of curriculum guidelines, language syllabi and tests (CEFR, 2001). That is, different abilities such as receptive and productive skills are overemphasized in the CEFR as one of the most essential parts that must be mastered by English language learners.

Many studies emphasized the importance and the Value of CEFR (e.g. Alderson, 2002; Figueras; 2012; Garrido & Beaven, 2002; Jones, 2002; Little, Simpson, & O'Conner, 2002; Morrow, 2004; North, 2002; Richterich & Schneider, 1992), while others (e.g. Fulcher, 2004; Goullier, 2007; Hulstijn, 2007; Morrow, 2004) pointed out that changes are necessary when implementing the CEFR guideline into a specific practice due to its "shaky ground" particularly when it used as a base for other tests. For example, Wisniewski (2018) found that there is a low correlation between the CEFR indicators and takers' actual language performance measurements. This result suggests that

interpretation of the test results referring to the CEFR scales should be carefully interpreted. Therefore, test developers should carefully use the CEFR to construct their test results, and to provide test specifications with theoretical evidence to support the validity and reliability of their tests (Davidson & Fulcher, 2007; Fulcher, 2004).

There is no unified definition of test specification, nonetheless, good test specifications contain some common elements such as the measured construct, description of items, timing allocation, scoring criteria, test instructions, and test administration (Brown, 1994; Douglas, 2000; Hughes 2003). For example, Fulcher (2010) and Spaan (2006) suggested some steps for developing a test such as: The explicit statement of the purpose, the measured construct, language skills, language knowledge, scoring system, test process, and other evidences with regard to a test. These aspects must be validated through piloting, revising, and documentation. While, Mislevy, Almond and Luckas's (2003) suggested five specifications that test designers need to acknowledge when designing tests. These five specifications are: 1) Item/Task specifications; 2) Evidence Specifications; 3) Test Assembly Specifications; 4) Presentation Specifications; and 5) Delivery Specifications. The first, Item/Task specifications refer to the given prompts that are designed to elicit inferences about the targeted abilities of the test takers, which is based on Bloom's Taxonomy levels. The second, Evidence Specifications deals with what the test takers are expected to accomplish based on the test instructions. The third, Test Assembly Specifications alluded to the consistency of the number and the range of items included in the test. The fourth, Presentation Specifications, which is based on how well the contents of the test are presented to the test takers. Finally, Delivery Specifications concerned with test administration timing, and test security (Mislevy, et al., 2003; Fulcher, 2010).

In addition to the construct validity achieved by test specifications construction, a standardized test needs to establish a content validity through its alignment with the specified learning outcomes or standards. Alignment is defined as bringing parts or components with proper coordination or coming to a level of agreement (La Marca, Redfield, Winter, & Despriet, 2000). La Marca et al. (2000) pointed out that the assessments must serve in demonstrating students' knowledge and skills with respect to the expectations itemized in the adopted standards and set up in the curriculum frameworks, and thus meaningful interpretations of the students' performance can be made. In particular, test alignment analysis serves to establish content validity by matching three main dimensions, which are: International and national standards, the instructional objectives, and the test items. This kind of matching will nurture the validity evidence to show how the test is an appropriate, representative, and important sample of the content from the international or national content standards. However, matching or alignment is not an easy process, as the standards need to be transformed into measurable objectives thereby allowing items of the test to be aligned and constructed. In fact, alignment analysis should be carried out in advance as a part of test development and construction, so that "Alignment is not a surprise ending or "Eureka! Discovery" (Berk, 2005, p. 20).

The Achieve Method is one of the strategies that should be adopted in the alignment process (Fulcher, 2010). According to Fulcher (2010) the idea of the Achieve Method

depends on looking at each item in the test and trying to identify which learning outcomes and objectives this item is supposed to achieve. This process is called content standard coding, during this process content centrality, performance centrality, source of difficulty, and level of difficulty should be identified by using a measurement scale that was advanced by Fulcher (2010). First, the Content Centrality, which is concerned with the items clarity and the explicitness in measuring the standards which is graded in four measurement scales: a) 0 = Inconsistent; b) 1A = Not Specific Enough whereby standards or objectives are too broad to be assured of item's strong alignment; c) 1B = Somewhat Consistent, where an item assesses only part of a compound objective; and finally, d) 2 = Clearly Consistent. Second, the Performance Centrality concerned with making judgement about whether the cognitive complexity of the items is similar to the cognitive complexity of the objective required in the learning outcomes, and it has the same four measurement scales as of content centrality. Third, Source of Difficulty, is about making a judgment about how difficult the item is with two measurement scales (appropriate = 1, and inappropriate = 0). Fourth, the Level of Difficulty is concerned with making a judgement about whether it is a suitable level for the target learners with (Yes/ No) measurement scale (See Fulcher (2010, p. 287).

Generally, designing a test and constructing its specifications is a recursive process which is executed through iterative steps in a cyclical way such as establishing, revising, drafting, piloting, and getting feedback from both teachers and test takers (Fulcher, 2010). For example, the IELTS was established through a series of revising process in a recursive way. Extensive data were collected through a revision project, then a team of writers was formed to draft the test specifications and they revised the specifications after consulting test takers, teachers, educational administrators, universities experts to ensure the content, context, and the formation are suitable for the targeted test takers. Subsequently, research was conducted to validate the content of the draft test, and finally the IELTS was established (Alderson, 1988; Alderson, Clapham, & Wall, 1995). Another study (Zandi, Kaivanpanah, & Alavi, 2014) which was conducted to investigate the effect of test specifications review on improving the quality of a language test, which found that the quality of the language tests has improved by reviewing the test specifications because the emergent feedback and suggestions rose concerning the usefulness in the actual classroom context.

Test specifications help test designers to align the national curriculum, institutional objectives and cognitive levels of instruction. A similar study was conducted by Abidin and Jamil (2015) in Malaysia that focused on the Malaysian Graduate Admission Test of English (GATE) which is tailored to the Malaysian higher education institutes entry. The GATE was established through three phases which are: Test Development, Test Operation and Test Analysis. In the Test Development phase, a team embarked on reading extensive literature and comparing it to other standardized tests such as TOEFL and IELTS. In terms of Test Operation phase, the team validated the test specifications according to the CEFR Framework, furthermore, the test was piloted through two stages within seven months. Additionally, in Test Analysis phase, the results of the test were analyzed and revised based on the pilot study. Finally, the GATE benchmarked with IELTS in terms of items and difficulty level, and the results were aligned with the

CEFR. The test developers, Abidin and Jamil (2015), highlighted the importance of language knowledge, literature, test specifications, and a validation framework in test construction. This study also emphasizes the importance of reference value of well-established international tests in constructing, validating, and implementing a new test. Given the status of EmSAT as high-stake test in the UAE context, it is crucial to find some mechanisms by which it could be evaluated and analyzed.

METHOD

This study employed a convergent parallel mixed method design by collecting quantitative (questions # 1 and 2) and qualitative (question 3) data sets concurrently. During the research process the two data sets were simultaneously collected. In this study both strands (QUAN= QUAL) are equally prioritized with aim of having complementary data in order to compare, amalgamate, and generalize the findings. The qualitative data set answered the first and second questions. The two sampled data were based on document analysis for both EILTS and EmSAT. The documents analysis for both tests is quantified in terms of their Item/Task specifications based on Bloom's Taxonomy and the samples alignment with the CEFR. In order to that, the Achieve Method Model for test analysis was applied (Fulcher, 2010). Moreover, the quantitative data set answered the third question of the study. A self-reporting Likert survey was used. The survey targeted students in the foundation program (n=194) who had different experiences with EmSAT. The Likert scale ranged from Strongly Disagree (=1) to Strongly Agree (=4) and the survey content items were built using Mislevy et al. (2003) which composed of five categories: 1) Test Scoring; 2) Test Presentation and Format; 3) Test Delivery; 4) Test Structure; and 5) Preparation Practice , which was analyzed by descriptive statistics. Furthermore, the participants were native Arabic speakers and females were the majority participants with 83 percent (n=161) and male participants were 17 percent (n=33). The selection criteria were based on two criteria: The participants should be in the foundation program which required an EmSAT test and their willingness to participate. Therefore, in order to answer the last question, each data set were analyzed independently and strong inferences were made at the results interpretation stage to validate the result.

The validity of the study tools was based on the construct of valid models to analyze the samples documents, and the analysis was done by three raters to establish interrater reliability. The internal reliability among the three raters was consistent, as Cronbach Alpha coefficient was extracted by 0.99. In terms of the quantitative tool, the survey was exposed to experts' consultation to evaluate language and content in which the Lawshe's Content Validity Ratio (CVR) was calculated with a value of 0.89. The Arabic translated survey was validated by back translation techniques, while the reliability of the survey was established by running a Cronbach Alpha coefficient analysis and it was found to be high (0.89).

FINDINGS

Test Specification Analysis: EmSAT vs. IELTS

The first question: How are the EmSAT test specifications compared to the IELTS test specifications? In order to find answer for this question, the Mislevy, Almond and

Luckas's (2003) Test Specifications Model, which was based on Bloom's Taxonomy, was used to compare and analyze the EmSAT item test specifications with IELTS items test specifications. Additionally, the mechanism of calculating items test specifications which was advanced by Bachman and Palmer (2010) and Zimmaro (2016) was used. Following this mechanism for counting and quantifying the EmSAT skills and subskills distribution, the results of counting yielded the following: Reading (*n*= 22), Writing (*n*=1), Vocabulary (*n*= 24), and Grammar (*n*= 25), whereas, the IELTS items distributions yielded the following: Reading (*n*=40), Writing (*n*= 2), Listening (*n*= 40); and the Speaking (*n*=15). For the purpose of comparing between the two tests, we exclusively focused on the reading and the writing skills.

Figure 1 illustrates a comparison between the EmSAT and the IELTS reading items distributions based on Bloom's taxonomy levels. As shown in Figure 1, the results indicated that there is a discrepancy in the distribution of abilities levels in reading between EmSAT test and the IELTS test. As for the EMSAT test most items loaded (52%) on measuring understanding, comprehension and application abilities; whereas the IELT's record showed almost similar results (56%). In terms of measuring the remembering abilities, the EmSAT count for (45%), while IELTS' record showed a slightly lower percentage (37%). In measuring creativity, analysis, and evaluation abilities combined, IELTS record showed a higher count (7%) than the EmSAT which counts for only (3%).

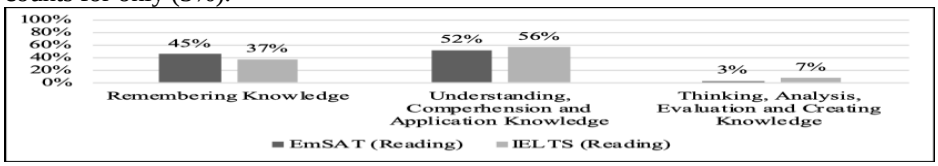

Figure 1
EmSAT Vs. IELTS reading items distribution

In terms of the writing skill, Figure 2 illustrates a comparison between the EmSAT and the IELTS writing items distributions based on Bloom's taxonomy levels. As shown in Figure 2, the results indicated that there is a discrepancy in the distribution of abilities levels in writing between EmSAT writing test and the IELTS writing test. As for the EmSAT writing test most items loaded (50%) on remembering abilities; whereas the IELTS' record showed different results (33.3%). In terms of measuring understanding, comprehension and application abilities, the IELTS count score higher (44.4%), while EmSAT's record showed a slightly lower percentage (40%). In measuring creativity, analysis, and evaluation abilities combined, IELTS record showed a higher count (22.20%) as compared the EmSAT which counts for only (10%).

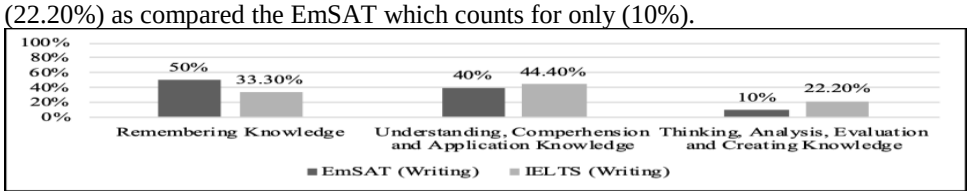

Figure 2
EmSAT vs. IELTS writing item distribution

Test Alignment Analysis: IELTS vs. EmSAT

The second question: To what degree do the EmSAT and IELTS align with CEFR standards? To answer this question, the CEFR alignment analysis framework was used as a backdrop to compare between the IELTS and the EmSAT reading and writing tests because both tests aligned and benchmarked against the CEFR, both of them are used for English language competency and for college admission. Accordingly, the Achieve Method (Fulcher, 2010) was adopted. In this regard, the four criteria of the Achieve Method analysis were used: Content Centrality, Performance Centrality, Source of Difficulty and Level of Difficulty. The CEFR has three main levels of measurement including: Basic Users of the language (A1 and A2), Independent Users of the language (B1and B2), and Proficient Users of the language (C1 and C2).

In general, when the EmSAT reading test was compared to the IELTS reading test in terms of all Achieve Method criteria: Content Centrality, Performance Centrality, Source and Level of Difficulty analysis, the results revealed that the EmSAT aligned with the lower and medium measurement levels of the CEFR standards (A1, A2 and B1); whereas the IELTS aligned with the higher and medium measurement levels of the CEFR standards (B1, B2 and C1). Notably, in three criteria (Content Centrality, Source of difficulty and Level of Difficulty), the EmSAT showed a compatible alignment with the lower and medium levels of CEFR standards (A1, A2 and B1) and it was not specific enough when it comes to the fourth criterion, Performance Centrality. In contrast, the IELTS reading test showed a compatible consistency in terms of Content Centrality, Performance Centrality and the Level of Difficulty with medium and higher measurements of the CEFR standards (B1, B2 & C1). However, Source of Difficulty in the IELTS reading items was not appropriate enough with the CEFR standard.

When the EmSAT writing test compared to the IELTS writing test in terms of all Achieve method criteria, the results revealed that the EmSAT writing test showed a high level of consistency with a medium measurement level of the CEFR standard (B1); Whereas EmSAT appeared to be not specific enough to measure the CEFR measurement levels, B2, C1 and C2. In case of the IELTS writing test alignment analysis, IELTS showed the consistency with higher measurement levels of the CEFR standard (C1 and C2). However, the IELTS writing test consistency with the medium measurement levels of the CEFR standards B1 & B2 showed very little compatibility.

To quantify the data that emanated from the rating of the test alignment of IELTS and EmSAT reading tests in terms of Content Centrality and Performance Centrality (See Table 1), a 4-point Likert scale was used, with scores of (1=not at all, 2=very little, 3= somewhat, and 4=to a great extent), the t-test result revealed that there is a significant difference between the EmSAT and IELTS whereas IELTS scores M=3.88; SD=.21); are higher than the EmSAT (M = 2.90; SD = .57) at ($t = -8.410$, $df = 21$, $p \leq 0.05$). However, there is no significant difference between the EmSAT writing test (M = 2.50; SD = .1.00) and IELTS writing test (M = 2.75; SD = .87), ($t = -.383$, $df = 3$, $p \geq 0.05$).

Table 1
Results of t-test analysis examining differences between EmSAT & IELTS (content centrality and performance centrality aligned with CEFR)

Category	M	SD	t	Df	Sig. (2-tailed)
EmSAT Reading Test – IELTS Reading Test	2.90 3.88	.57 .21	-8.410	21	.000
EmSAT Writing Test – IELTS Writing Test	2.50 2.75	1.00 .87	-.302	3	.783

To quantify the data that emanated from the rating of the test alignment of IELTS and EmSAT reading tests in terms of Source of Difficulty and Level of Difficulty, frequencies anal,ysis was used to compare between the EmSAT and IELTS tests in terms of appropriateness or inappropriateness by using CEFR standards as backdrop. As shown in Figure 3, the results revealed that the EmSAT reading test was rated higher (63.60%) in appropriateness than the IELTS reading test (20%) in Source of Difficulty with the medium measurement level of CEFR standards (B1). In terms of inappropriateness, IELTS reading test was rated a high percentage (80%) in inappropriateness with higher measurement level of CEFR standards (C1 and C2).

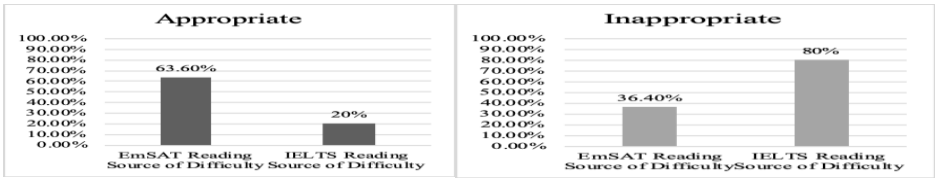

Figure 3
Difference between EmSAT & IELTS reading (Source of Difficulty)

In terms of Level of Difficulty (see Figure 4), the EmSAT reading test was rated high (63.60%) which signified its appropriateness with the medium measurement level of CEFR standards (B1), whereas it was rated (36.40%), which signified its inappropriateness with high measurement levels of CEFR standards (C1 and C2). In contrary, the IELTS reading test rated (62.50%) which signified its appropriateness with higher measurement levels of CEFR standards (C1 and C2). Whereas, it was rated (37.50%), which revealed its inappropriateness with medium measurement levels of CEFR standards (B1 & B2).

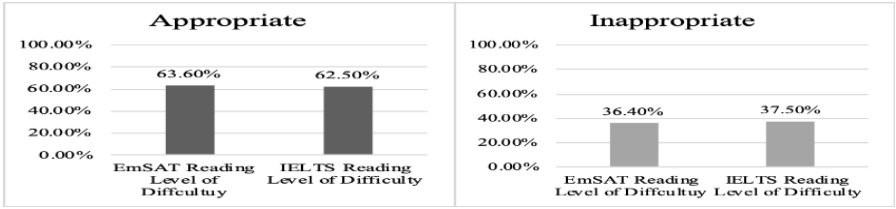

Figure 4
Difference between EmSAT & IELTS reading (Level of Difficulty)

The rating of the test alignment of IELTS and EmSAT writing tests in terms of Source of Difficulty (See Figure 5), which they were based on CEFR Standards, revealed that the level of appropriateness of EmSAT writing test was rated higher (50%) as compared to the IELTS writing test, which was rated only (25%) with the medium measurement level of CEFR standards (B1). In terms of inappropriateness, IELTS writing test was rated high (75%), which indicated that there is a little alignment with the higher measurement levels of CEFR standards (C1 and C2).

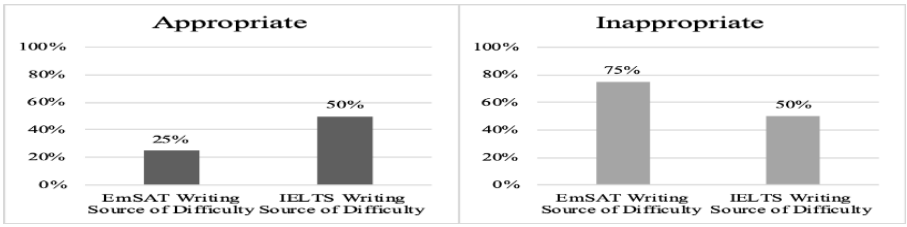

Figure 5
Difference between EmSAT & IELTS writing tests (Source of Difficulty)

In terms of Level of Difficulty (see figure 6), the IELTS writing test was rated at (50%), which indicated its appropriateness with higher measurement levels of CEFR standards (C1 and C2), whereas in terms of inappropriateness, it was rated (50%), which signified that there is a little alignment at the medium measurement levels of CEFR standards (B1 and B2). In contrary, the EmSAT writing test was rated (25%), which indicated its appropriateness with a medium measurement level of CEFR standards (B1). Whereas, it was rated (75%) in its inappropriateness, which pointed out that there is a little compatibility with higher measurement levels of CEFR standards (C1 & C2).

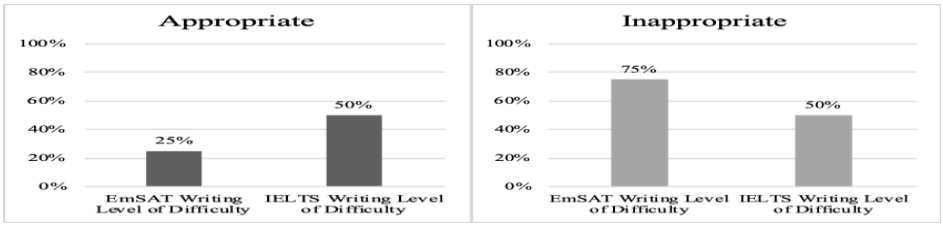

Figure 6
Differences between EmSAT & IELTS writing tests (Level of Difficulty)

Students Report About the Use of EmSAT As A Standardized Test

The third question: What do college students report about their experience with the EMSAT? To answer this question, a survey of 4-point Likert Scale with five main categories including: Test Presentation and Format, Test Delivery, Preparation Practice, Test Structure and Test Scoring was distributed among college students who had previous experiences in taking the EmSAT test. The results displayed in Table 2 showed that the students reported higher in Test Scoring category (M = 3.00; SD = .59) followed by Test Presentation and Format category with mean score of (M = 2.94;

SD=.65) and then Test Delivery category with (*M* = 2.53; *SD* = .54), whereas the Test Structure category had a low mean score with (*M* = 2.46; *SD* = .62) and the Preparation Practice category was reported the least among the five categories (*M* = 2.20; *SD* = .64).

Table 2
Students' self-report on their experiences with EmSAT

Category	M	SD
Test Scoring	3.00	.59
Test Presentation and Format	2.94	.65
Test Delivery	2.53	.54
Test Structure	2.46	.62
Preparation Practice	2.20	.64

These results indicated that in Test Soring category, students' responses revealed that they highly agreed with the Test Scoring criteria for EmSAT test. However, in itemizing this category, the students reported their desire to see the equal distribution of the marks among all the test questions, to know in advance the evaluation rubrics while they partially disagreed that the test score really reflects their English language levels.

In terms of Test Presentation and Format category, students' responses in general revealed that they highly agreed with the EmSAT's Test Presentation and Format. As they reported that they agreed that the instructions of the test were very clear, and they knew what was expected from them during the test. Moreover, they reported they preferred to see the display of text and questions in an intact page rather than scrolling up and down. Additionally, they preferred a "back" and "forth" option, to have a chance for revising their answers. However, they reported that they disagreed on the computer-based test only option, rather, they prefer to have options of either computer-based test or a paper test one.

Generally, in Test Delivery Category students reported that they totally disagreed with the time allotted, which they believe was not enough to answer each question fully. Moreover, students showed partial agreement with the appropriateness of the results announcement time. However, students reported positively about having the opportunity to choose a convenient location of the test centre and having flexibility in changing the test dates when they face with some circumstances. Additionally, they reported some satisfactions with regarding of the ease of website registration.

In Test Structure category, students informed that they totally disagreed that the EmSAT measures their English language skills fairly and their critical thinking skills. Moreover, they reported that the EmSAT test focuses on measuring remembering, comprehension and application abilities. Furthermore, they completely agreed with the idea of evaluating grammar and vocabulary within the reading and writing contexts rather than testing them separately. As with the Presentation Practice category, students collectively reported that they were not provided with examples or an examinee handbook for the EmSAT test, which might contribute in practicing and preparing for the actual test questions. In addition to that, they reported that their teachers were not helping them in preparing for the test.

Additionally, paired sample t-tests were performed on to look for statistically significant differences between the categories. The t-test results are shown in Table 3. Examining the means, it can be seen that there is a significant difference between the Test Format category (*M*=2.94; *SD*=.65) and Test Delivery Category (*M*=2.53; *SD*=.54); (t=8.961, df=193, p≤0.05), Preparation Practice category (*M*=2.20; *SD*=.64); (t=13.936, df=193, p≤0.05), and Test Structure (*M*=2.46; *SD*=.62), (t=9.598, df=193, p≤0.05). However, there is no a significant difference between the Test Format category (*M*=2.94; *SD*=.65) and Test Scoring category (*M*=3; *SD*=.59), (t=-1.244, df=193, p≥0.05). By the same token, significant differences were found between Test Delivery category (*M*=2.53; *SD*=.54) and Preparation Practice category (*M*=2.20; *SD*=.64); (t=7.793, df=193, p≤0.05), and Test Scoring category (*M*=3; *SD*=.59), (t=-10.690, df=193, p≤0.05). However, there is no a significant difference between test Delivery category (*M*=2.53; *SD*=.54) and Test Structure category (*M*=2.46; *SD*=.62), (t= 1.891, df=193, p≥0.05). Finally, significant differences are also found between Preparation Practice category (*M*=2.20; *SD*=.64), Test Structure category (*M*=2.46; *SD*=.62), (t=-5.789, df=193, p≤0.05), and Test Scoring category (*M*=3; *SD*=.59), (t=-15.134, df=193, p≤0.05). Moreover, there is a significant difference between Test Structure category (*M*=2.46; *SD*=.62) and Test Scoring category (*M*=3; *SD*=.59), (t=-11.962, df=193, p≤0.05).

Table 3
Results of t-test analysis examining differences among the five categories

Scale	Comparison	T	df	Sig. (2-tailed)
Pair 1	Test Format – Test Delivery	8.961	193	.000
Pair 2	Test Format – Preparation Practice	13.936	193	.000
Pair 3	Test Format – Test Structure	9.598	193	.000
Pair 4	Test Format – Test Scoring	-1.244	193	.215
Pair 5	Test Delivery – Preparation Practice	7.793	193	.000
Pair 6	Test Delivery – Test Structure	1.891	193	.060
Pair 7	Test Delivery – Test Scoring	-10.690	193	.000
Pair 8	Preparation Practice – Test Structure	-5.789	193	.000
Pair 9	Preparation Practice – Test Scoring	-15.134	193	.000
Pair 10	Test Structure – Test Scoring	-11.962	193	.000

DISCUSSION

The Convergence between the Qualitative and the Quantitative Results

The Fourth question: How do the converged results from EmSAT and IELTS comparison, alignment with international standards, and students' self-reports inform us about the nature of the EmSAT as a standardized test? To answer this question, two strategies were used to merge both sets of data. The first strategy is data transformation that merged data analysis through transforming the qualitative data (document analysis) into quantitative data through quantization (Creswell & Clark, 2011). The second strategy of merging data is a side-by-side comparison for merging data analysis through presenting both the qualitative and quantitative results in a discussion (as illustrated in Figure 7). Based on the merging of the document analysis and the students' self-reports, the EmSAT test specifications do not align with the IELTS test specifications in terms of the targeted English language skills and in in terms of the distributions of levels of

abilities among these language skills. This qualitative result is supported by a quantitative result from the students' self-reports, which confirmed that the items of the EmSAT test do not fairly measure their English language skills and thinking skills. Participants reported that the EmSAT only focuses on measuring understanding and remembering abilities based on Bloom's Taxonomy. The idea generated from these results is that item test specifications should be established carefully to nurture the construct validity of the test itself. It is not only about measuring particular skills and abilities in a random way, but also measuring what the test is supposed to measure in a very comprehensive way. These results are supported by a number of scholars (see Brown, 1994; Davidson & Lynch, 2002; Douglas, 2000; Fulcher, 2003, 2010; Fulcher & Davidson, 2009; Hughes, 2003; Mislevy, Almound, & Lukas, 2003; Mislevy & Riconscente, 2006) who confirmed that test specifications are evidence-driven blueprints which provide an accurate and valid description of the test purpose, the structure of the items, the targeted constructs to be measured, the scoring system, and the consistent content of the test.

In terms of the IELTS and EmSAT tests alignment with CEFR standards, the qualitative analysis revealed that both the EmSAT and the IELTS are not aligned appropriately with the CEFR. To further illustrate, the EmSAT is only aligned with the lowest measurement levels of the CEFR (A1, A2 & B1); whereas, the IELTS is only aligned with the medium and higher levels (B1, B2, C1 & C2). Similarly, the quantitative analysis of students' self-report on the EmSAT revealed that the test is not associated with what the test takers have to learn. It is essential for any alignment to be changed based on what needs to be tested and the context of the test takers as other researchers have highlighted (see Abidin and Jamil, 2015; Alderson, 2002; Figueras, 2012; Fulcer, 2004; Goullier, 2007; Hulstijn, 2007; Morrow, 2004). They also suggested that establishing a test alignment with the CEFR needs to be built on a solid matching that shows the validity and reliability of the results extracted from a test. Additionally, the mixed method results revealed low associations between scales and language performance, which is similar to what Wisniewski (2018) had found. Hence, an interpretation of the results needs to be made cautiously with regard to the CEFR scales.

Quantitative analysis demonstrates other test specifications including: Evidence Specifications, Assembly Specifications, Presentation Specifications, and Delivery specifications. To illustrate that, the students reported that the EmSAT instructions are clear and they knew what was expected from them. Moreover, students highlighted that they had good experiences in terms of the test delivery and assembly. Students mentioned some issues regarding test scoring and presentation; hence they expressed their desire to know the score for each item, and to know in advance the evaluation rubrics and the scoring system. Many of the students agreed that their scores in EmSAT don't reflect their real levels in English language. Moreover, students do not feel comfortable having to scroll up and down option during the exam, instead they prefer a "back" and "forth" option where they can go back and modify and revise their previous answers when they have extra time. They also disagreed with having only a computer-based test option, instead they prefer to have two options where they can choose between a computer-based or a paper-based test. Additionally, students highlighted the

need for an examinee handbook and more accessible samples, as there is currently not one available. Compared these findings about EmSAT to IELTS, it can be indicated that the IELTS scoring system is explained in detail on the official website. For example, each of the 40 reading items has one point, in which a scale of calculating one's score is provided, as the number of the correct items reflects the band score for the test taker. Additionally, rubrics for the writing skills are detailed for the test takers before taking the test (IELTS, 2019). Also, IELTS can be both a paper-based or a computer-based test, and it depends on the test taker's preferences to select between them (IELTS, 2019). Moreover, the instructions of the sample test are very clear where they have the reading followed by its questions, in which test takers can go back to their previous answers when they have the chance to revise and to double check their answers. Furthermore, there are a variety of samples that test takers can use to practice, which are similar to the real test. Additionally, the IELTS has an examinee handbook that explains more about the test's structure, its scoring system, and the measurement criteria. Test takers can access the official website of the IELTS to prepare themselves and understand every single detail about the exam before taking the exam. Generally, all of these specifications affect the structure of the EmSAT test and thus the validity and reliability of the results extracted from the test itself. This idea is supported by Fulcher (2010), Mislevy, et al., (2003) and Zandi, Kaivanpanah, & Alavi (2014) as they stressed the importance of each specification as they scaffold one another and any instability in one of them affects the whole structure of the test.

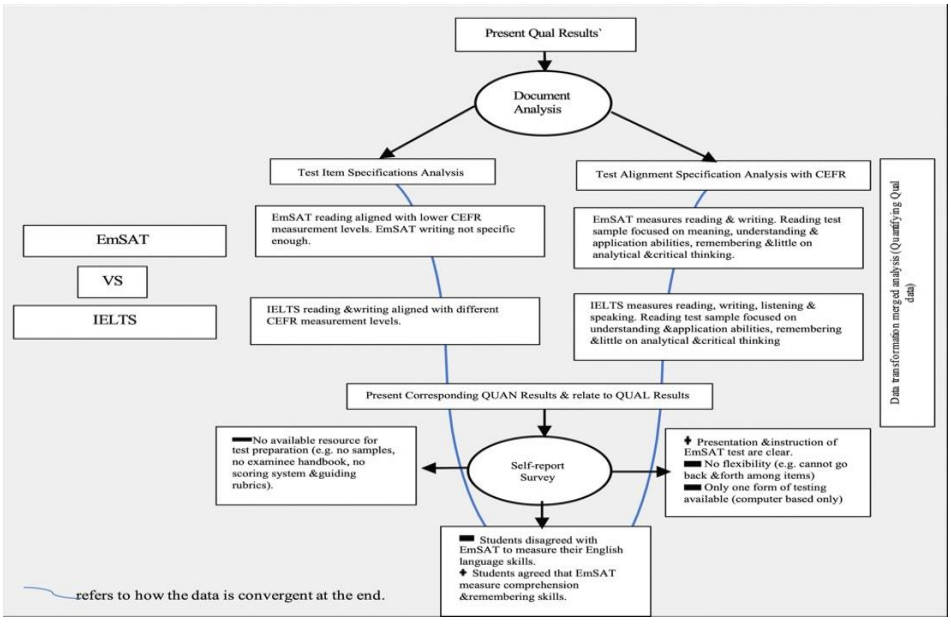

Figure 7
The convergence between qualitative and quantitative data

CONCLUSION AND SUGGESTIONS

Based on the findings of this study some implications and recommendation are sought. First, for establishing a test such as the EmSAT, some considerations should be taken into account. For example, standardized tests must be carefully constructed based on valid test specifications. These types of tests must be piloted, analysed, and reanalysed until the validation is perfectly established. Dealing with the language variable itself is another source of difficulty that must be taken into consideration. More precisely, do they measure content knowledge or language skills? For instance, the EmSAT test analysis indicates that vocabulary, grammar, reading and writing are the only targeted language components to assess test takers' English language levels. The absence of listening and speaking skills leads to a lack of full integration of the four crucial skills. Whereas IELTS covers the four skills in which the target of the test is to measure test takers' English language skills usage instead of focusing purely on measuring the knowledge of the language itself.

Second, it is not only about how the standardized test specifications are established and what kinds of language skills are measured, but also what goes beyond the implementation of a standardized test and how it affects in teaching and learning processes. As La Marca et al. (2000) pointed out that the assessments must serve in demonstrating students' knowledge and skills with respect to the expectations itemized in the adopted standards and set up in the curriculum frameworks, in which meaningful interpretations of the students' performance can be made from the designed assessment. For example, the washback of the EmSAT affected the teaching and learning process, where instructors and students started to deal with the language as a content knowledge instead of a practical and communicative skill. This notion is supported by the two instructors' responses in the preliminary interview conducted at the beginning of this study. They reported that they have limited information and limited data available about the EmSAT. Therefore, they build their pedagogical techniques for students' test preparation based on the primary notes that they take while the students are taking the EmSAT, which made those instructors adapt "a reverse engineering strategy" by which they weave traditional forms of instructional techniques and content knowledge for directly preparing students for the test, which leaves little room for authentic communication and functional use of the language.

Third, the idea of test alignment should be precisely considered in a way that fits the learning context of both learners and teachers. For instance, the EmSAT as a required exam for college admission should not only be aligned with the CEFR standards, but also with the universities' learning outcomes within the Emirati context. Moreover, IELTS, also should be revised to align with CEFR as international standards. As Alderson, 2002; Figueras; 2012; Fulcher, 2004; Goullier, 2007; Hulstijn, 2007; Morrow, 2004) pointed out that changes are necessary when implementing the CEFR guidelines into a specific practice due to its "shaky ground" particularly when it used as a base for other tests, which entail a great concentration on the nature of other tests and their purpose. Additionally, a broader investigation should be considered with other universities in the UAE. This study hopes to contribute to knowledge base and research

base regarding other international English proficiency tests and for test developers to take analytical and critical stances on such international standardized tests. Notably, the current will redound to the benefit of researchers, test makers, assessment policy and curriculum making. For example, researchers, should consider the views of a large number of test trainers in all the universities. As for the test makers, they should take into consideration the intricacies of creating a robust standard test such as the EmSAT which will be used as an entry requirement to all the universities in the UAE, this test should be revised to reflect international standard specifications. The assessment policy makers should consider the different test steps to make the test credible, and valid. As for curriculum planning, the instructional goals should reflect assessment goals and outcomes.

LIMITATIONS AND DELIMITATIONS

One of the limitations of this study is the choice of the purposive sampling. A larger sample of test trainers should be considered, as the limited sample raised valid points about the nature of the text and the preparation stage. Moreover, teachers' views should be taken into consideration, as teachers are responsible for the students' performance. Additionally, one of the delimitations of this study is that the researchers did not make students involve and make time diaries at their different stages of test preparation and test taking. Keeping diaries could track the students' different stages from test preparation to test taking and yield useful results.

REFERENCE

Abidin, S. A., & Jamil, A. (2015). Toward an English proficiency test for postgraduates in Malaysia. *SAGE open*. https://doi.org/10.1177/2158244015597725, 1-10.

Alderson, J. C. (1988). Testing English for specific purposes – how specific can we get? In Hughes, A. (ed.), *Testing English for University Study*. London: Modern English Publications and the British Council, 16–18.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Alderson, J. (Ed.). (2002). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Case Studies*. Strasbourg, France: Council of Europe. Retrieved January 12, 2019, from Council of Europe: http:// www.coe.int/T/DG4/Portfolio/documents/case_ studies_CEF.doc.

Bachman, F. & Palmer, S. (2010) *Language assessment in practice: developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Berk, R. (2005). *Test Alignment and Balancing with State Standards: Every 6 Months or 50 Items, Whichever Comes First*. Retrieved December 5, 2018, from https://images.pearsonassessments.com/images/NES_Publications/2005_03Berk_488_1.pdf.

Brown, H. D. (1994). *Principles of language learning and teaching* (3rd ed.). Englewood Cliffs, N.J: Prentice Hall Regents.

Council of Europe. 2001. *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). London: SAGE.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications.* New Haven: Yale University Press.

Douglas, D. (2000). *Assessing languages for specific purposes*. New York: Cambridge University Press.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal, 66*(4), 477–485. doi:10.1093/elt/ccs037.

Fulcher, G. (2003). Testing second language speaking. London: Longman/Pearson Education.

Fulcher, G. (2004). Deluded by artifacts? The Common European Framework and harmonization. *Language Assessment Quarterly, 1*(4), 253-266

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing, 26*(1), 123-144.

Garrido, C., & Beaven, T. (2002). Using the Common European Framework of Reference in course and materials development: The UK Open University experience. In Council of Europe, Common European Framework of Reference for languages: Learning, teaching, assessment (Case studies) (pp. 25-39). Strasbourg: Council of Europe Publishing.

Goullier, F. (2007). *Council of Europe tools for language teaching*. Common European Framework and Portfolios. Paris: Didier.

Hughes, A. (2003). *Testing for language teacher* (2nd ed.). Cambridge: Cambridge University Press.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*(4), 663–663. https://doi.org/10.1111/j.1540-4781.2007.00627_5.x

IELTS. (2019). About the test. Retrieved from https://www.ielts.org/

Jones, N. (2002). Relating the ALTE framework to the Common European Framework of Reference. In Council of Europe, Common European Framework of Reference for languages: learning, teaching, assessment (Case studies) (pp. 167-181). Strasbourg: Council of Europe Publishing.

La Marca, P. M., Redfield, D., Winter, P. C., & Council of Chief State School Officers, Washington, DC. (2000). *State standards and state assessment systems: A guide to alignment. series on standards and assessments.* Council of Chief State School Officers, Attn: Publications, One Massachusetts Ave.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. M. Downing, & T. M. Haladyna, Handbook of test development (pp. 61-90). New Jersey: Lawrence Erlbaum Associate, Inc.

Mislevy, R. J., Almond, R. G. and Lukas, J. F. (2003). *A Brief introduction to evidence-centred design*. Research Report RR-03–16. Princeton, NJ: Educational Testing Service.

Morrow, K. (2004). *Insights from the Common European Framework*. Oxford: Oxford University Press.

North, B. (2002). *A CEF-based self-assessment tool for university entrance*. In Council of Europe, Common European Framework of Reference for Languages: Learning, teaching, assessment (Case studies) (pp. 146-166). Strasbourg: Council of Europe Publishing.

Spaan, M. (2006). Test and item specifications development. *Language Assessment Quarterly, 3*(1), 71-79.

Stoynoff, S. (2011). Recent development in language assessment and the case of four large-scale tests of ESOL ability. *Cambridge Journals, 42* (1), 1-40.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. New York: SAGE.

Wisniewski, K. (2018). The empirical validity of the Common European Framework of Reference scales. An exemplary study for the vocabulary and fluency scales in a language testing context. *Applied Linguistics, 39*(6), 933-959.

Zandi, H., Kaivanpanah, S., & Alavi, S. M. (2014). The effect of test specifications review on improving the quality of a test. *Iranian Journal of Language Teaching Research, 2(*1), 1-14.