Arab World English Journal
INTERNATIONAL PEER REVIEWED JOURNAL ISSN: 2229-9327
Arab World English Journal (AWEJ) Volume 12. Number2 June 2021 Pp. 49 - 71
DOI: https://dx.doi.org/10.24093/awej/vol12no2.4

Validating English Language Entrance Test at a Saudi University for Health Sciences

Sabria Jawhar
Department of English, College of Science and Health Professions, King Saud bin Abdulaziz University for Health Sciences & King Abdullah International Medical Research Center, Jeddah, Saudi Arabia

Manal Al Makoshi
Department of English, College of Science and Health Professions, King Saud bin Abdulaziz University for Health Sciences & King Abdullah International Medical Research Center, Riyadh, Saudi Arabia.
Correspondent Author: makoshim@ksau-hs.edu.sa

Sajjadllah Alhawsawi
Department of English, College of Science and Health Professions, King Saud bin Abdulaziz University for Health Sciences & King Abdullah International Medical Research Center, Riyadh, Saudi Arabia

Abdulmohsen Alkushi
Department of Pathology, College of Medicine, King Saud bin Abdulaziz University for Health Sciences & King Abdullah International Medical Research Center, Riyadh, Saudi Arabia

Received: 2/23/2021 Accepted: 4/10/2021 Published:6/24/2021

Abstract
This paper aims to validate the English Language Entrance Test for King Saud bin Abdulaziz University for Health Science (hereafter KSAU-ET). It supports the argument regarding using specially designed in-house entrance tests at health universities by showcasing the test's development, administration, and validation process. It presents a new framework for test validation that is informed by various existing frameworks such as Messick (1996), Sireci (1998) and, Weir (2005), with a specific focus on the notions of unitary and practicality. The proposed framework treats validity as a pre-, during, and post-test process that collects evidence from each phase to support the test's overall validity. The data were collected using different tools and through the three stages of the validation process. The test was taken by 474 candidates who applied to join KSAU-HS Stream II medical program. The data confirmed that the test was reliable (alpha > 7) and reasonably meet the university's needs to select the program's top prospective candidates. Nevertheless, the study highlighted the importance of collecting further evidence in future studies and including more selection criteria in the regression model of analysis. Using this framework, the study contributes to the existing body of research that investigate English entrance test validation. It shows that exam validity is a context-sensitive process strongly associated with the purpose for which the exam is used. Finally, the paper discusses pedagogical implications that may help educators at health science universities develop in-house entrance tests in place of standardized tests, which often do not address context, curriculum, or program objectives.

Keywords: English entrance test, health science, language assessment validity, practicality, validation, King Saud bin Abdulaziz University for Health Sciences
Cite as: Jawhar, S., Al Makoshi,M., Alhawsawi, S., & Alkushi, A. (2021). Validating English Language Entrance Test at a Saudi University for Health Sciences. *Arab World English Journal, 12* (2) 49-71.
DOI: https://dx.doi.org/10.24093/awej/vol12no2.4

Introduction

The growing demand for higher education in general and health sciences, in particular, have increased selectivity and pushed universities to use different admission criteria that vary from the standardized test such as the Test of English as a Foreign Language (TOFEL) and the International English Language Testing System (IELTS) to interviews and in-house entrance tests. However, some universities still witness a high dropout rate during the first academic year or students' failure to meet their degree studies' required standards. The internationalization of education is another element that added to universities' pressure to ensure that their newly admitted students have adequate English language proficiency to complete their subject courses in English. Brown (1993), Fulcher (1997), and Wall, Clapham, and Alderson (1994) are a few examples of studies that looked at in-house placement tests in different higher education contexts to ensure a good outcome.

In Saudi Arabia, most universities require students to complete an academic year of a foundation English language program. Students are exposed to academic English reading and writing to help them succeed in their chosen academic programs. The university's selection criteria are mainly international standardized exams that are not developed with the curriculum or program's needs. The mismatch between the programs' requirements and standardized exams often results in students studying English courses that do not support their future needs (Jenkins and Leung 2019; Tomlinson 2020). This paper showcases an in-house entrance test validation process and illustrates how it served its purpose of distinguishing between student applicants based on the program's objectives. This validation process was completed by answering the following questions:

1. To what extent is the KSAU-ET valid in terms of test items and quality?
2. How did the students who took the KSAU-ET perceive the test?
3. Did the test successfully predict the students' GPAs in the first academic semester?

By answering these questions, the researchers demonstrate that the process of test validity and the test results supported the decision not to offer English courses for the Stream II students (see the context section). The test also helped distinguish between students and predict their success versus failure rate in the first academic semester.

In this paper, the researchers tested the validity of the proposed test and its practicality. The study calls for specialized colleges to build up their entrance tests to suit their programs' objectives instead of using ready-made tests limited to testing the candidates' general knowledge of English. Our entrance test was shown to be valid and practical, as reflected in the accepted candidates' results in the academic courses they studied during their first semester at the university.

The following sections discuss the test's goals and the steps to reach the implementation level.

Test's Goals

The test aims to:

a. successfully identify and exclude those at risk of failing their academic degree because of their weak language abilities.
b. predict the probability that accepted applicants will succeed in their studies in the medium of English once they join the university,
c. successfully identify the strength of the applicant's language abilities

Literature Review

Language Test Validity

Test validity in general and language test validity, in particular, have been defined in various ways in the literature. However, most mainstream researchers agree that a valid test should discriminate between test users and provide a meaningful difference by measuring what it is meant to measure (Ginther & Yan 2018; Jenkins and Leung 2019). Brookhart and Nitko (2019) have defined exam validity as the robustness of the interpretation and utilization of assessment results using evidence from different sources. This suggests that the concept of validity applies to how any exam results will be used and not necessarily to its procedures. The purpose of an exam also plays a significant role in its validation. Therefore, instead of asking whether a test is valid, specific questions must be asked about the test scores' uses for a specific purpose (e.g., placing students into specific classes or admitting them into a specific program). This purpose and the situation in which the exam is used determines the degree of its validity. For example, a particular exam may have a high validity score as an admission test in one university and yet score poorly in another. This scoring could be explained by the fact that the exam items match one university's program objectives while not matching the other program. Brookhart' and Nitko's (2019) definition suggests that a conclusion about an exam's validity should not be reached before studying and combining different types of validity evidence.

The process of creating a valid and reliable placement test is a difficult one that involves hard work as the test items must be closely aligned to a curriculum with clear goals and objectives. Once the test is developed, it should be followed by piloting, analyzing, and reviewing the items to ensure that they are reliable and that effective placement decisions can be made (Westrick, 2005). Therefore, when designing an in-house test, one should consider the institute's specific needs and objectives and pilot the test to reflect its curriculum (Dinh, 2019; Inoue, 2006).

However, it is crucial to say that validating tests help provide score-based predictions and theoretical and empirical grounded explanations of these scores (Farley, Yang, Min, and Ma 2020; Xi & Sawaki, 2017). Miller et al., (2013) stress that validity is used to answer two critical questions about tests. The first is related to how appropriate, meaningful, and useful the scores' interpretation is for the results' intended application. The second addresses the effect of the particular uses and interpretations that are made of the results. In this validation process, the test-taking process, strategies, and the consequences of the test should be investigated.

Validity and Practicality

Practicality is as vital an element of any assessment as its validity and reliability. The concept of practicality has saturated the topic of validity in education. The discussion of different types of validity and then different conceptualization of unitary validity is evident in the importance of practicality to the validity. Allen and Yen (2002) distinguish between three major types of validity, i.e., content validity, criterion-related validity, and construct validity. While content validity does not require statistical calculations, criterion-related validity and construct validity are based on statistical measures and correlation testing. The researchers will focus the discussion here specifically on content and construct validity due to its relevance to the current paper.

Content validity refers to the process of rational analysis of the content of the test. It does not require statistical calculation but mainly depends on the individual subjective judgment of the test items. This type of validity is divided into two types, i.e., face validity and logical validity. While face validity discusses the extent to which the test can measure the relevant trait, logical validity is an advanced form of face validity. It measures carefully and logically the extent to which the domain of behaviors is being defined and how the written exam items reflect logically and precisely that domain. Although content validity is often subjective, it is a cornerstone in developing all tests, and test items should be written to meet content validity requirements (Allen & Yen, 2002).

Construct validity focuses on the extent to which a test measures the theoretical construct that it was designed to measure. Unlike content validity, construct validity is an ongoing verification of predictions made about the test scores. These predictions are about how test scores should behave in different situations and are based on current theory regarding the construct or trait being measured. These predictions could be related to group differences, changes in time, age, gender or location, correlation, or how the exam is processed. Construct validity is enhanced if data support the predictions. When there is no data to support the predictions, it could mean that the experiment is flawed or that the theory was not correct, and it should be revised. It could also mean that the test does not measure the trait or the construct it was supposed to measure.

It is critical to mention here that the idea that all tests must rigidly conform to a specific type of validity allows little flexibility, especially when viewed from the practicality framework. Space should be created to negotiate the extent to which the purpose and use of validity could influence the type of evidence needed to validate a test. This brings us to the unitary notion of validity, which stresses that validity should be dealt with as one unit rather than dividing it into different types (Brookhart & Nitko, 2019; Ginther & Yan, 2018; Messick, 1996). Within the unitary notion of validity, different validity types became pieces of evidence to provide interpretations for test scores and support the use of a particular test. The argument here is to bring further evidence that, in nature, represent "the different types of validity" in one place without creating binaries between them. Bachman and Palmer's (1996) work introduced the nation of test usefulness to make Messick's (1996) framework of unitary more accessible to practitioners. Their notion of test usefulness is based on five qualities, i.e., construct validity, reliability, authenticity, inter-activeness, impact, and, most importantly, practicality. Through these different works, the notion of validity in language tests started to focus on score

interpretation for particular test use rather than the test itself, and validation research became more empirically driven (Farlay et al. 2020; Xi & Sawak, 2017).

Weir (2005) proposed a socio-cognitive framework that consists of five types of validity evidence: context validity, theory-based validity, scoring validity, consequential validity, and criterion-related validity. However, according to Weir (2005), those types of validity complement each other and not alternatives. The first two types of validity are linked to the test takers characteristics and state that the tested abilities depend on the test-takers internal mental process. The scoring validity in this framework is placed in the center as it is seen to determine the exam's reliability. This scoring validity includes item analysis, internal consistency, error of measurement and, marker reliability.

The notion of test usefulness paved the way for further development in the validity framework in language testing with the introduction of an argument-based approach to validity by Kane, Crooks and, Cohen (1999) and later represented in the work of Chapelle, Enright and Jamieson (2008), Bachman and, Palmer (2010), and Addey, Maddox, and Zumbo, 2020. The argument-based approach is the most relevant one to this study as it presents "a simple, systematic process for how validation researchers structure validity arguments, linking validity evidence for the development and use of a test" (Im, GH., Shin, D., and Cheng, L, 2019, p. 26). It provides researchers with great flexibility to determine the argument they want to make based on the test's context and purpose. It also gives room for negotiating the types of evidence collected to support any claim regarding the test validity (Addey et al. 2020; Kane, 2013). To sum up, the argument-based approach is grounded on claims backed by data that must be supported by a warrant supported by evidence.

Unlike other proposed frameworks such as Weir (2005), that view the validity items as components that give a sense of binaries, the researchers believe that validity is collecting evidence. It should be implemented throughout the test's different phases, i.e., pre-, during- and post-test. However, the researchers argue that those pieces of evidence should be treated as one unit to substantiate an exam's validity. For us, the validation process's ultimate goal is developing and evaluating evidence for a proposed score interpretation and use based on the context (Farlay et al. 2020; Im et al., 2019; Xi & Sawak, 2017). This paper adds to the body of work supporting the argument-based validity by reflecting on the process through which the entrance test for health sciences has undertaken. It focuses on the dialogue between practicality and validity as the main drive for having a fair and robust entrance test and shows the evidence collected to support the validity argument.

Our framework consists of multiple phases that were informed by Messick (1996), Brookhart and Nitko (2019), Sireci and Faulkner-Bond (2014), Sireci (1998) and Weir, (2005). The framework aims at collecting as much evidence as possible throughout the process of test development, administration, and scoring. It also considers test-takers judgment of the test items using what is typically referred to as face validity (Allen and Yen, 2002). Figure one displays the researchers' understanding of an effective validation framework.

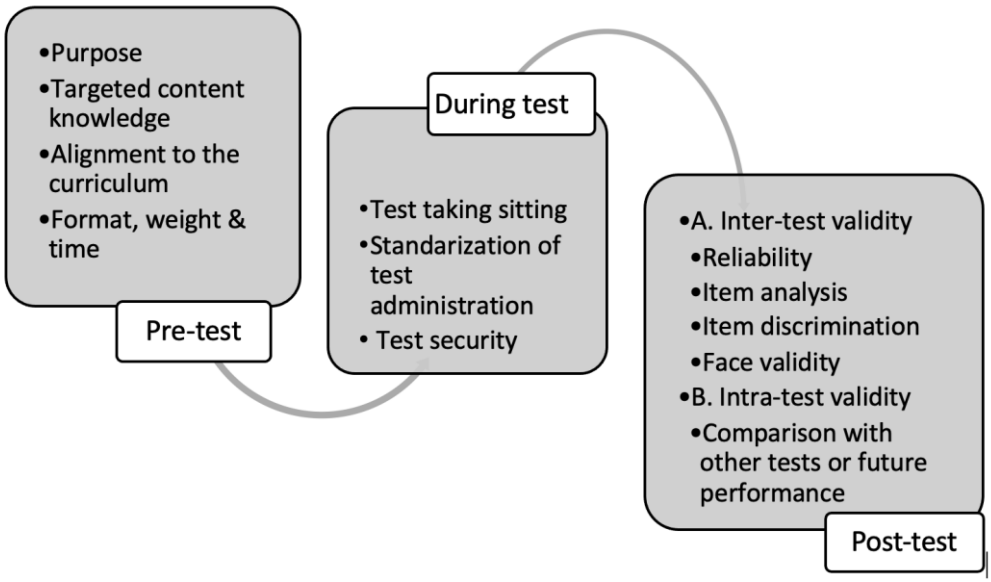


Figure 1. KSAU-HS entrance test validity framework

Methods

As explained in the previous section, the framework governed the research design and implementation from the beginning until the end. This included the instruments development and data collection steps and ended with the data analysis and validation process.

The Context

King Saud bin Abdulaziz University for Health Sciences (hereafter KSAU-HS) is the only Saudi University specializing solely in the health science field. English is the official medium of instruction in KSAU-HS. All students accepted into the university must first, complete three semesters of intensive English language and Basic Science courses in the College of Science and Health Professions (hereafter COSHP). Through COSHP, the university has two entrance points to the College of Medicine (hereafter COM). The first referred to as Stream I, consists of high school graduates admitted based on three main criteria: cumulative percentage of secondary school (natural sciences and no less than 90%), an achievement test, and an aptitude test. Each of these holds a weight of 30% - 40% - 30%, respectively. These Stream I students must complete three semesters in COSHP. At the end of the first academic year, i.e., the preparatory year (two semesters), the students are allocated to different Colleges (i.e., Medicine, Dentistry, Pharmacy, Applied Medical Sciences and, Health Informatics) based on their college choice and cumulative GPA.

The number of seats distributed for student allocation varies annually based on capacity and other mitigating factors. For COM, the number of seats (i.e., spaces) ranges from 75 to 100 for Jeddah and Riyadh female students, respectively, and from 125 to 175 for Jeddah and Riyadh male students, respectively.

The second entry point into COM is through an accelerated program, only offered in the region by KSAU-HS, referred to as Stream II admission. The criteria for entry are; first, a recent bachelor's degree in Basic Science, Applied Medical Science, or Pharmacy. Second, the

cumulative degree GPA must not be lower than "Very Good" (i.e., 3.75/5.00). The candidates must then pass a basic medical science entrance test and personal interview. As the program is highly competitive, a maximum of 25 applicants can be accepted if they successfully qualify.

Previously, Stream II students were required to complete a one-semester intensive English language program to ensure their English language skills were at the same level as their Stream I counterparts in semester three (second-year students). However, it soon became apparent that most students accepted in Stream II were already proficient in English and did not require additional language courses. A change of policy then came into effect in which the intensive language program was replaced with additional science courses. To ensure that new applicants' English language competency for the Stream II program was maintained, an entrance test to assess their English language skills was added to the admissions criteria.

Given that English language skills required to succeed in the Stream II program relate to English as the official medium of instruction (hereafter, EMI) in the health sciences, most standardized English tests would not match the requirements. Therefore, the university introduced the King Saud bin Abdulaziz University Entrance Test (KSAU-ET) to ensure that candidates to the program were qualified to study in English at the level required and keep pace with stream I students.

Participants

This study included four executive committee members from the English language department (two females and two males) and eight Subject Matter Experts (hereafter, SMEs). Four teacher assistants were included for piloting, and oral feedback and, 474 students took part as test-takers.

A separate committee made up of members of the University's Deanship of Admissions and Registration (hereafter DAR), COM, and COSHP reviewed the students' applications and documents. Based on the number of applications, the entry criteria, including GPA, were established, and this committee set a cut-off point. This resulted in approximately 500 students qualifying to take part in the entrance tests. However, on the day of the English test, only 474 applicants attended. The number consisted of 159 male and 315 female participants distributed across two campuses, Riyadh and Jeddah (Deanship of Admissions and Registration, 2019).

Data Collection

The test's validation process necessitated data collection from different sources using different tools based on the requirement of each phase, as illustrated by the framework (see figure one). This process includes the candidates' scores in KSAU-ET, the questionnaires regarding face validity and, their performance (i.e., GPA) in the first semester after joining the university.

Test Instrumentation and Validation

For the development of the test, an executive committee that consisted of faculty members from the English Language departments was formed. The members developed an English Language entrance test to ensure candidates met the criteria for acceptance. To determine the test's scope, the committee prepared a list of the content areas and cognitive abilities that the test is designed to measure in alignment with the curriculum objectives. The test

consists of listening, vocabulary, grammar, and reading comprehension sections (see appendix A). It has a total of 100 multiple-choice questions (hereafter MCQs) that were pitched at an intermediate to upper-intermediate level according to the Common European Framework of Reference for Languages (CEFR). Items were dichotomously scored, with equal weightings for correct answers, and wrong answers were not penalized. The time allocated for the test was three hours, and was administered in a paper-based format. The test was designed to gauge the candidates' English language level and determine whether the successful candidates' scores could be an accurate indicator of their academic success level in their first academic term.

To validate the test's content, the research team used multiple validation tools that were implemented before, during and, after the test was administered. Before the test, the team used a validation framework that entails the four elements of content validity described by Sireci (1998). According to Sireci (1998), content validity covers domain definition, domain representation, domain relevance, and the appropriateness of test construction procedures. As mentioned earlier, the following was done to address these domains.

a. SMEs were invited to evaluate the domain definition involved in the entrance test and assess the test specifications.

b. The SMEs were also asked to evaluate the domain representation and rate all the test items to assess the extent to which the test items are consistent with the curriculum framework (Crocker, Miller, and Franks, 1989; Sireci, 1998)

c. The SMEs were also invited to assess the domain relevance of the test items. They were requested to rate the degree to which the test items were relevant to the test specifications that were initially matched with the course specifications. In other words, they were required to ensure that the test measures all essential aspects of the content domain and that the test did not include irrelevant items (see table one). Following the SMEs rating task, the team used a statistical summary table to show how well each item measures the corresponding objective.

d. The SMEs reviewed the technical accuracy and quality of the test's items (Haladyna and Downing, 1989) as part of the test construction procedures' appropriateness.

e. The SMEs also scrutinized the test for any offensive or inappropriate language that might impact the construct or indirectly disadvantage any test-takers (Ramsey, 1993).

f. Finally, the test was piloted and followed by statistical item analyses to select the most appropriate and delete any inappropriate or problematic items.

Table 1. *Rating task assessing item/objective congruence*

		How well does the item measure its objective? (Circle one)				
Item#	Objective	1 (Not at all)	2	3	4	5 (Very well)
1						
2						
3						
4						
5						

The second phase of the validation process was carried out during the test. It covered the physical conditions under which the test was taken, the uniformity of administration and, test security (Jin 2019; Weir 2005). Evidence related to this phase of the test validity was collected through observations. The aim was to understand whether the test administration's physical conditions were satisfactory and adhere to the general standardized rules and specifications. That included investigating and unifying the test set's actual setting across rooms and campuses, e.g., the lighting, ventilation, tables and chairs, background noise, and quietness of the test halls. The uniformity of test administration was established by ensuring that the test was administered in the same manner across sites, e.g., preparation, timing and, support available during-test. The test-takers security was also addressed by limiting access to the test to only authorized people and not allowing test-takers to make copies of the test or share information during the test session.

The third phase was carried out after the test. In this phase, different tools were utilized to analyze the data and validate the scoring process, including descriptive statistics, reliability (internal consistency reliability), and item analysis (item difficulty and discrimination). Because validity is not a one facet process, as discussed earlier, the candidates' perception regarding the test's suitability was also analyzed through a questionnaire that targeted what is usually referred to as face validity. This phase shows the test's internal property and adds to the evidence collected for the previous two stages' validity, and offers evidence supporting "the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13).

Research Procedures

The data collected through the validation process of this research was analyzed statically and qualitatively. However, the data from observations and initial meetings were analyzed to understate the general feedback related to the test validity. This understanding has led to the development of the final version of the test. Nevertheless, the data collected from the entrance test were statically analyzed, as seen in the results and discussion section. For the analysis of the inter-test properties (construct validity), the researchers used descriptive statistics, reliability, item difficulty, and item discrimination. The analysis was done using SPSS Version 21.0. For the intra-test properties (face validity), with a 5-point Likert scale, a percentage table is used to reflect the degree of the participants' responses.

Results

This section discusses the results of the test validation process based on our proposed framework. During phase one of the test validation process and following the analysis of the SMEs responses to the rating task to assess the congruence between the test items and the course objective, the committee concluded that the test meets the following criteria:

1. The test items correspond to the curriculum objectives.
2. The test specifications cover the domain definition.
3. The test specifications represent and are relevant to the domain.
4. The test construction procedures are appropriate.
5. The statistical analysis of the piloting group showed that the test is reliable, with alpha >0.7. It also showed that the levels of item difficulty were appropriate and that the face validity was sustained. However, the committee was cautious that the teacher assistants'

educational background might have impacted their results as the piloting statistics did not distinguish them.

As reported by the test invigilators and the committee members, the during-test phase observation data revealed that the test adhered to the general rules and specifications as the test's actual setting across venues in each campus. The venues were unified in terms of the lighting, ventilation, tables and chairs, background noise and, quietness of the test halls. The university follows the same standards for building qualities and specifications of equipment and design of lecture halls. The observation reports showed that the test was administered in the same manner regarding preparation, timing and, support available during the test. The test security was monitored by arranging four to eight invigilators based on the students' number and size of the test venue to prevent the test-takers from sharing information. Access to the test was also limited to authorized people only, which ensured validity regarding security.

By securing the validity of the previous two phases, the researchers could then move to the post-test stage of the validation process (Figure one). This phase focused on the inter-and intra-test validation process.

Discussion of the results

In this section, the researchers present a detailed discussion of the validation process based on different phases of the proposed framework (see Figure one).

Inter-test Validity

General Descriptive Statistics

This section addresses the validity of inter-test property, providing evidence for reliability and item analysis, item difficulty, and item discrimination and, face validity from the test-takers perspective.

Test Reliability

The total number of questions used in the listening comprehension section was 20 MCQs. The mean score of the 474 candidates in the listening comprehension was 10.63, and the standard deviation was 3.16. The second section of the test was grammar, and it consisted of 30 MCQs. The mean of candidates' scores in this section was 14.79, while the standard deviation was 5.17. The vocabulary section was made up of 25 MCQs, and the candidates' mean score was 10.06, and the standard deviation was 4.83. The last section, reading comprehension, was made up of 20 MCQs. In this section, the mean score was 9.73, and the standard deviation was 3.85. As illustrated in the table below (Table two), the candidates' mean scores were highest in grammar (14.79) and the lowest in reading comprehension (9.73). On the other hand, while again highest in grammar (5.17), the standard deviation was the lowest in listening comprehension (3.16).

A brief look at the table below (Table two) indicates that the candidates found the grammar section much more accessible than listening comprehension, vocabulary, and reading comprehension, respectively. The Cronbach reliability coefficient was calculated to determine each part of the test's internal reliability or internal consistency. The Cronbach alpha is used to look at how the test items measure the same concept and, hence, is connected to the test's inter-relatedness. On this test, a value higher than 0.7 indicates that the test is reliable.

The test's overall internal consistency, including the four parts, is (>0.7), suggesting a reasonable and acceptable level of reliability. However, when the researchers view each component of the test separately, as can be seen from Table one, the listening comprehension section of the test was the least consistent (Cronbach's α value= 0.594), compared to vocabulary (0.813), grammar (0.719) and reading comprehension (0.719).

Table 2. *Test reliability*

Test Component	N	Mean	SD	Cronbach's α
Listening Comprehension	474	10.65	3.16	.594
Grammar	474	14.79	5.17	.719
Vocabulary	474	10.06	4.83	.813
Reading Comprehension	474	9.73	3.85	.719

Though the reasons for this low score are beyond the scope of this study, and though there is limited research on this topic, some listening comprehension challeges have been documented for Saudi students. Hamouda (2013), for instance, reported that Saudi students are often negatively impacted by spoken accent, pronunciation, audio speed, lack of concentration, anxiety, and lack of vocabulary. In our test, authentic data was used that resembled what students might hear in any classroom setting. The recordings varied in context, speed, and accent. The recordings' quality was judged as very good to excellent by three independent listening comprehension instructors who were consulted as subject matter experts.

After a post-test review of the listening section, the committee decided to keep the section rather than delete it. The rationale for this decision was that candidates would be taught in lecture halls identical to where the test was administered. Thus, the circumstances under which they obtained their scores were more realistic and demonstrated how they would cope in similar classrooms settings. Additionally, the fact that there was no significant difference between the whole group when it came to this part of the test also influenced the decision to keep it. Nevertheless, the listening part questions were subjected to further investigation under the individual test items analysis to assess their difficulty level and were found to be appropriate (see the next section). Aside from the listening section, the researchers can assert that the test has a satisfactory level of internal consistency, i.e., across-items, and can be considered reliable and consequently valid from this perspective (Farlay et al., 2020; Messick, 1996; Xi and Sawak, 2017). By statistically establishing the test's reliability, the researchers moved to a more detailed and in-depth validation, i.e., item analysis, as explained in the framework (Figure one).

Item Analysis (Percent Correct)

For this part of the validation process, the researchers used item analysis under which both item difficulty and item discrimination were calculated. It is essential to mention here that

each question's difficulty is defined as the percentage (calculated across all candidates) of correct answers. The easier the question, the higher the number of candidates who have answered it correctly, and, hence, the higher the difficulty value. If the question was challenging and a lower number of candidates answered it correctly, the difficulty value would be low.

Table 3. *Item difficulty*

Q1	0.63	Q21	0.45	Q41	0.36	Q61	0.26	Q81	0.47
Q2	0.64	Q22	0.63	Q42	0.41	Q62	0.34	Q82	0.59
Q3	0.33	Q23	0.35	Q43	0.35	Q63	0.71	Q83	0.55
Q4	0.49	Q24	0.32	Q44	0.83	Q64	0.52	Q84	0.65
Q5	0.29	Q25	0.39	Q45	0.54	Q65	0.14	Q85	0.47
Q6	0.55	Q26	0.32	Q46	0.51	Q66	0.22	Q86	0.38
Q7	0.69	Q27	0.31	Q47	0.44	Q67	0.16	Q87	0.61
Q8	0.29	Q28	0.63	Q48	0.42	Q68	0.86	Q88	0.33
Q9	0.81	Q29	0.26	Q49	0.33	Q69	0.68	Q89	0.58
Q10	0.68	Q30	0.61	Q50	0.42	Q70	0.33	Q90	0.64
Q11	0.50	Q31	0.32	Q51	0.45	Q71	0.25	Q91	0.38
Q12	0.76	Q32	0.33	Q52	0.46	Q72	0.25	Q92	0.58
Q13	0.86	Q33	0.39	Q53	0.34	Q73	0.31	Q93	0.25
Q14	0.34	Q34	0.41	Q54	0.52	Q74	0.45	Q94	0.37
Q15	0.06	Q35	0.37	Q55	0.54	Q75	0.43	Q95	0.63
Q16	0.36	Q36	0.42	Q56	0.19	Q76	0.60	Q96	0.64
Q17	0.33	Q37	0.67	Q57	0.22	Q77	0.46	Q97	0.33
Q18	0.48	Q38	0.30	Q58	0.21	Q78	0.33	Q98	0.32
Q19	0.34	Q39	0.34	Q59	0.48	Q79	0.42	Q99	0.45
Q20	0.69	Q40	0.31	Q60	0.52	Q80	0.70	Q100	0.51

Table three lists the difficulty of each question. Any test item with a correct answer rate lower than 0.25 or higher than 0.75 was reviewed carefully.

Table three shows that questions 9, 12,13, and 44 received a score of less than 0.25 ($p<0.25$), indicating that these were the most straightforward questions in the test. These questions were included to help reduce anxiety at the start of the test. The difficulty of the items then increases as the test progresses to peak at item number 65.

Table four summarizes the items that scored less than 0.25 or more than 0.75. Questions 56, 57, 58, 65, 66, and 68 are the most difficult questions with a difficulty level of more than 0.75 ($p>0.75$). These items were reviewed, and the committee agreed to include them to allow differentiation between the candidates. Questions with a p-value of less than 0.25 ($p<0.25$) were also reviewed for suitability.

Table 4. *Summary of item difficulty*

P-value	The number of items	Item number
p.<0.75	5	9, 12, 13, 44, 68
0.25>p<0.75	89	1-8, 10-11, 14-44, 59-64, 96-100
p>0.25	6	56, 57, 58, 65, 66, 68
Total number of items	100	

Item Discrimination

The discrimination of each question is defined differently. First, the total score of each candidate is calculated based on the number of correct answers. Based on that score, candidates were divided into three groups of equal size; the high scoring candidates (33.33%), medium scoring candidates (33.33%), and the low scoring candidates (33.33%). The discrimination of each question is defined as follows; the ratio of correct answers of the top-ranked students to those of the bottom-ranked students. Positive values in the discrimination index mean that top students are more likely to answer the question than lower-ranked students, where the reverse holds if the discrimination is negative. Zero values indicate that the question is not discriminating. Research has shown that a test item that receives more than (0.30) is characterized as having high discrimination (Ebel,1966). Table five contains the discrimination values for each question.

Table 5. *Item discrimination*

Q1	0.44	Q21	0.34	Q41	0.13	Q61	0.45	Q81	0.38
Q2	0.36	Q22	0.22	Q42	0.35	Q62	0.57	Q82	0.64
Q3	0.18	Q23	0.18	Q43	0.12	Q63	0.58	Q83	0.34
Q4	0.30	Q24	0.15	Q44	0.23	Q64	0.73	Q84	0.41
Q5	0.16	Q25	0.23	Q45	0.43	Q65	0.18	Q85	0.44
Q6	0.41	Q26	0.12	Q46	0.48	Q66	0.41	Q86	0.20
Q7	0.48	Q27	0.13	Q47	0.58	Q67	0.15	Q87	0.68
Q8	0.12	Q28	0.28	Q48	0.30	Q68	0.34	Q88	0.26
Q9	0.32	Q29	0.26	Q49	0.14	Q69	0.51	Q89	0.48
Q10	0.47	Q30	0.44	Q50	0.22	Q70	0.47	Q90	0.58
Q11	0.08	Q31	0.21	Q51	0.35	Q71	0.25	Q91	0.17
Q12	0.35	Q32	0.20	Q52	0.40	Q72	0.13	Q92	0.41
Q13	0.17	Q33	0.51	Q53	0.32	Q73	0.39	Q93	0.19
Q14	0.26	Q34	0.35	Q54	0.45	Q74	0.56	Q94	0.34
Q15	0.66	Q35	0.31	Q55	0.40	Q75	0.30	Q95	0.42
Q16	0.08	Q36	0.24	Q56	0.26	Q76	0.39	Q96	0.43
Q17	0.12	Q37	0.33	Q57	0.37	Q77	0.11	Q97	0.24
Q18	0.18	Q38	0.25	Q58	0.35	Q78	0.15	Q98	0.39
Q19	0.23	Q39	0.16	Q59	0.64	Q79	0.17	Q99	0.41

Q20	0.33	Q40	0.30	Q60	0.71	Q80	0.46	Q100	0.37

Table five shows that all the test items received positive discrimination index values (PBI), which means that they have discriminated among the applicants. While no item received a zero value, the items varied in their ability to distinguish between the upper and the lower scorers, as shown in Table five. Table six displays the number of highly discriminated items among the participants compared to those that did fairly with a focus on the skills.

Table 6. *Summary of items discriminations index based on tested skills*

Test Component	Total number of items	Items with DI (Negative)	Items with DI (Equal zero)	Items with DI (0.10 -0.30)	Items with DI (> 3)
Listening	20	0	0	10	10
Grammar	35	0	0	18	17
Vocabulary	25	0	0	8	17
Reading	20	0	0	5	15

When interpreting the value of discrimination, the committee compared the value to each item's difficulty index (p-value), particularly those which were intentionally written to be easy or more challenging. For instance, if an item had a discrimination index of < 0.2 but a p-value of >7, the committee considered it relatively easy for most applicants. The last step under the item analysis was an analysis of the distractors. The committee tested the distractors for each item to verify that they were not miskeyed or implausible. This was done by calculating the proportion of the participants' selected answers to the response options. The quality of the items' distractors was found to be satisfactory.

Face Validity

Table seven through Table 12 reveal the candidates' responses from the face validity questionnaire. For instance, Table seven shows how the candidates perceived their proficiency level in the four tested language skills. As illustrated in table seven, the candidates rated their reading skills as the highest, followed by listening, speaking, and grammar. Table seven also highlights the mismatch between the candidates' perceived level of proficiency in listening and their actual performance in this skill, as listening was ranked the lowest in terms of obtained scores. On the other hand, grammar was perceived as the lowest in the candidates' perception of their proficiency level.

Table 7. *Self-rated English language proficiency (Q1-4)*

Level of proficiency	Speaking (%)	Reading (%)	Listening (%)	Grammar (%)
1 (not good)	0.00	0.00	0.00	0.00
2	7.14	4.76	7.14	4.76
3	21.43	14.29	21.43	42.86
4	38.10	40.48	35.71	42.86
5 (very good)	33.33	40.48	35.71	9.52

No response	0.00	0.00	0.00	0.00

When it came to the test's overall difficulty level (Table eight), most candidates classified the test as moderate (73.81%), whereas only 2.38% found it very easy. On the other side of the continuum, 4.76% found the test very difficult.

The candidates were also asked to rate the difficulty level of the test components of the reading passages and the vocabulary. The results show that 38.1% found the reading passages at a moderate level of difficulty, while 11.9% found the reading passages and the vocabulary very difficult. A relatively smaller group rated the reading passages and vocabulary, 7.14% and 2.38 respectively, as easy.

Table 8. *Overall difficulty level and difficulty levels for vocabulary and reading passages (Q5-7)*

Level of difficulty	Overall difficulty level (%)	Vocabulary (%)	Reading passages (%)
Very easy	2.38	2.38	7.14
Easy	2.38	11.90	14.29
Medium level of difficulty	73.81	38.10	33.33
Difficult	16.67	30.95	28.57
Very difficult	4.76	4.76	4.76
No response	0.00	11.90	11.90

The candidates varied in their responses to the appropriateness of the topics used in the test (Table nine). For instance, 26.19 percent judged the topics as adequate. Nineteen percent (19.05) considered the test topics too general, and 11.9% said the topics were not sufficiently focused.

Table 9. *Appropriateness of topics (Q8)*

Topics	Percentage
Topics were too technical	4.76
Topics lacked focus	11.90
Topics were too unbalanced	23.81
Topics were too general	19.05
Topics were adequate	26.19
No response	14.29

More than 60% of candidates responded that the test instructions were either clear or very clear (Table 10). However, a combined total of about 15% of the candidates perceived the test as either very unclear or somewhat unclear. That some candidates found the test instructions unclear is a concern that requires further investigation.

Table 10. *Clarity of test instruction (Q9)*

Clarity level	Percentage
Very unclear	4.76
Somewhat unclear	9.52

Somewhat clear	11.90
Clear	33.33
Very clear	28.57
No response	11.90

Table 11 indicates that the candidate's perception of the entrance test's accuracy to assess their English language proficiency. Most of the candidates perceived the test as accurate (42.86%) or somewhat accurate (30.95%). However, around eight percent of the candidates perceived the test as not accurate or partially inaccurate. Though low, this percentage indicates a need to investigate the reasons behind such perceptions in future research.

Table 11. *Accuracy of the test as perceived by candidates (Q10)*

Accuracy level	Percent
Not accurate at all	2.38
Somewhat inaccurate	4.76
Somewhat accurate	30.95
Accurate	42.86
Very accurate	7.41
No response	11.90

Table 12 shows the candidates' responses to the question regarding the appropriateness of the method of testing. Most candidates (>40%) perceived the test as appropriate (35.71%) or very appropriate (9.52%). However, about 19% of the candidates reported that it was either somewhat inappropriate (16.67%) or very inappropriate (2.38%).

Table 12. *Appropriateness of the test method as perceived by the candidates (Q11)*

Appropriateness level	Percent
Very inappropriate	2.38
Somewhat inappropriate	16.67
Somewhat appropriate	23.81
Appropriate	35.71
Very appropriate	9.52
No response	11.90

While the candidates' negative evaluation regarding the appropriateness of the test's method was not pursued at the time, it is another point worthy of further future investigation.

Intra-test Validity

The KSAU-ET Predictivity of Students' GPA

The students' first academic semester GPA was used as the prediction criterion to determine the predictivity validity of KSAU-ET. The predictor was the students' score in KSAU-ET. According to Zwick (2002), test productivity power can only be determined by its ability to

predict an immediate criterion such as GPA. Following this, efficient regression analysis and the P-value was calculated. The coefficients describe the mathematical relationship between the independent variable (KSAU-ET) and the dependent variable (GPA). To look at the relationship between KSAU-ET and GPA, the researchers used the following estimated linear regression equation [1]:

$$GPA = 2.09 + 0.03\ EPT$$

$$R^2 \text{ is } 10.7\%$$

The number 2.09 is the constant of the simple regression equation, which is interpreted as the value of GPA when the KSAU-ET score is zero. The number 0.03, on the other hand, is the coefficient of KSAU-ET interpreted as the rate of increase in GPA by one unit (score) increase in KSAU-ET. The coefficient of determination, R2, is 10.7%. This means that 10.7% of the GPA variation among students could be explained by variation in students' KSAU-ET scores. Other factors explain the remaining 89.7%.

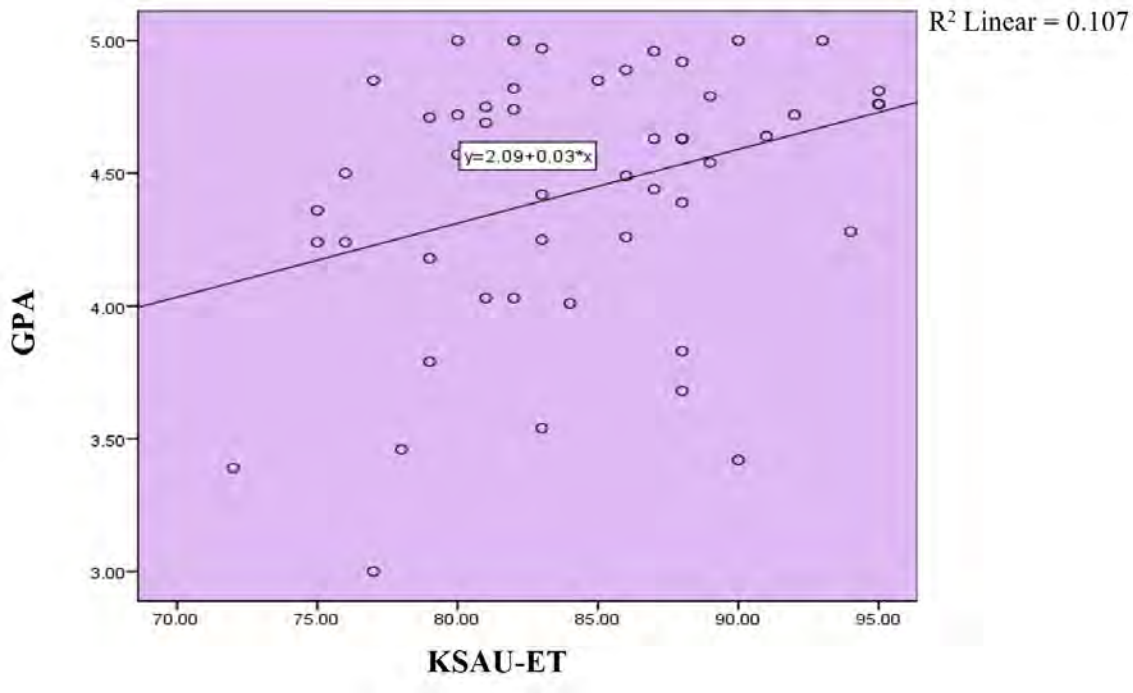


Figure 2. Correlation between students EPT and GPA

It is essential to mention here that the students performed well during the first semester of the academic year at KSAU-HS. The average GPA was 4.2 out of five points. The insignificant predictivity validity result can be attributed to the small number of the students who were accepted and enrolled in the program (50 students in total) in contrast to the number of candidates who took KSAU-ET (474 candidates). This result shows the sample's impact on the coefficients (Ali, 1987; Guan, Alam and Rao, 2019). The result, however, brings to the surface the importance of repeating the predictivity validation process with more significant data and

including other factors such as the candidates' undergraduate specialty, socio-demographic factors, gender and, scores in the different subject matters during the first academic year at the University (Gardner, Liu, and Roberts-Thomson, 2020; Park, 2019; Puddey and Mercer, 2014).

Conclusion

The researchers set out to identify which candidates could successfully complete the intensive Stream II program, a unique bachelor's degree in medicine, from a large applicant pool. This was completed with the development of the KSAU English language entrance test which was used to distinguish candidates and predict their success vs failure in the first academic semester of the program. It is a highly competitive program with limited seats, so the entry criteria, including the English language test (KSAU-ET), are crucial for selecting the best-ranked candidates. The successful candidates are required to complete an intense academic program taught and assessed solely in the medium of English.

As the university specializes in the health sciences field, the researchers needed to develop a curriculum specific entrance test to effectively assess potential students using a highly reliable and valid method. This process may not have been as successful with a commercially available test developed to test general language skills and was not aligned to the program and the curriculum objectives.

Because there is no perfect validation framework, the researchers developed a framework informed by various established frameworks. However, the main driving force behind our framework was the notions of unitary and practicality. The researchers required a framework that suited the context and met the program's needs and objectives. The proposed framework treated validation as a process of pre-, during and, post-test. The process collected evidence for each phase and added it into the test's validity without rendering a particular stage more critical than the others.

Extensive analysis of the KSAU-ET has shown it reliable, based on the data collected from the 474 candidates who completed the test and the face-validity questionnaire. The data confirm that the English entrance test was able to meet the university's needs fairly and reliably. It helped select the top prospective candidates for the Stream II medical program, as demonstrated by the GPAs of those accepted into the program at the end of the first academic semester. Nevertheless, simple regression analysis showed no significant relationship between the KSAU-ET and the students' GPAs, highlighting the importance of collecting further evidence in future studies and including more selection criteria in the regression model of analysis. The research highlighted the importance of in-house English entrance tests for health science universities when a test validation process is carried out systematically through evidence collection in alignment with a program's objectives.

About the authors:

Dr. Sabria Jawhar is an assistant professor of applied and educational linguistics and the English Department Chair at the College of Science and Health Professions (Jeddah) at King Saud bin Abdulaziz University for Health Sciences (KSAU-HS). Dr. Jawhar is a graduate of Newcastle University, UK. She is interested in all aspects of classroom discourse, assessment and use of technology in HE. However, her main focus is on talk-in-interaction. Corpus

linguistics, especially spoken corpora, is another area of her interest. https://orcid.org/0000-0002-1799-8888

Dr. Manal Al Makoshi is an assistant professor of applied linguistics and the English Department Chair at the College of Science and Health Professions (Riyadh) at King Saud bin Abdulaziz University for Health Sciences (KSAU-HS). Dr. Manal holds an MA in TESOL from Portland State University, Portland, OR, USA and a PhD in applied linguistics from the University of Birmingham, UK. Dr. Manal's research interests include language teaching and learning in higher education, corpus linguistics, English for Specific Purposes (ESP) and English for Academic Purposes (EAP)**.** https://orcid.org/0000-0003-2145-1874

Dr. Sajjadllah Alhawsawi is an assistant professor of education and an associate dean in the College of Science and Health Professions, King Saud bin Abdulaziz University for Health Sciences (KSAU-HS). Dr. Alhawsawi holds an MA in TESOL from the University of Exeter, UK, MSc in research methods from the University of Sussex, UK and PhD in education from the Centre of International Education and Development, School of Education and Social work, University of Sussex, UK. Dr. Alhawsawi's research interest includes programme evaluation, teacher education, higher education, instructional design, sociology of education and pedagogical use of ICT university education. https://orcid.org/0000-0002-6175-9892

Dr. Abdulmohsen Alkushi is an associate professor of Pathology & Laboratory Medicine, and Dean of College of Science and Health Professions at King Saud bin Abdulaziz University for Health Sciences (KSAU-HS). He is also a practicing consultant Pathologist at King Abdulaziz Medical City of National Guard. He holds a board from Royal College of Canada in Anatomical Pathology and subspeciality in Gynecological & Breast Pathology. He also holds a MSc in Pathology and Laboratory from University of British Columbia, Vancouver, Canada. He also holds Master of Medical Education (MME) from KSAU-HS. He is a medical graduate from College of Medicine at King Abdulaziz University. Dr. Alkushi's research interest includes health education, curriculum management, language education, breast cancer pathology and women health. https://orcid.org/0000-0001-9329-2097

References

Ali, M. A. (1987). Effect of sample size on the size of the coefficient of determination in simple linear regression. *Journal of Information and Optimization Sciences*, *8*(2), 209-219.

Allen, M.J., & Yen, W.M. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford, UK: Oxford University Press.

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

Brookhart, S., & Nitko, A. (2019). *Educational Assessment of Students plus with MyLab Education with Pearson eText*. London, UK: Pearson.

Brown, J. D. (1993). A comprehensive criterion-referenced language testing project. In D. Douglas, & C. Chapelle (Eds.), *A new decade of language testing research: Selected Papers from the Annual Language Testing Research Colloquium* (pp. 163-184). Washington, DC, USA: TESOL.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.

Chapelle, C. A., Enright, M., & Jamieson, J. (2008). *Building a validity argument for the Test of English as a Foreign Language.* London: Routledge.

Crocker, L.M., Miller, D., & Franks, E.A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2*(2), 179-194. https://doi.org/10.1207/s15324818ame0202_6

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52(4)*, 281-302. https://doi.org/10.1037/h0040957

Deanship of Admissions and Registration (2019). *Admission Requirements*. Retrieved from https://www.ksau-hs.edu.sa/English/Admission/Pages/AdmissionRequirements.aspx

Dinh, M. T. (2019). A review on validating language tests. *VNU Journal of Foreign Studies*, *35*(1),143-154. DOI: https://doi.org/10.25073/2525-2445/vnufs.4343

Ebel, R. L. (1966). *Measuring educational achievement*. India: Prentice Hall.

Fulcher, G. (1997). An English language placement test: issues in reliability and validity. *Language Testing*, *14*(2), 113–139.

Gardner, S., Liu, P., & Roberts-Thomson, K. (2020). Trajectory of performance; the role of selection criteria on student achievement in a Bachelor of Oral Health program. *European Journal of Dental Education, 24*(3), 572-579.

Ginther, A., & Yan, X. (2018). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing*, *35*(2), 271-295.

Guan, T., Alam, M. K., & Rao, M. B. (2019). Sample Size Calculations in Simple Linear Regression: Trials and Tribulations. *arXiv preprint arXiv:1907.*10569, 1-20

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied measurement in education, 2*(1), 37-50.

Hamouda, A. (2013). An Investigation of Listening Comprehension Problems Encountered by Saudi Students in the EL Listening Classroom. *International Journal of Academic Research in Progressive Education and Development, 2*(2), 113-155.

Im, GH., Shin, D., & Cheng, L (2019) Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Lang Test Asia, 9*(14). https://doi.org/10.1186/s40468-019-0089-4

Inoue, N. (2006). What's going on inside the pine tower of babel: Foreign language curriculum reform in a Japanese university. *Languages and Cultures Series*, *16,* 87-115.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, *50*(1), 1-73. DOI: https://doi.org/10.1111/jedm.12000

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational measurement: issues and practice*, *18*(2), 5-17.

Messick, S. (1996). Validity and washback in language testing. *Language testing*, *13*(3), 241-256. DOI: https://doi.org/10.1177/026553229601300302

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*(2), 5-11.

Miller, J. D., et al. (2013). The Five-Factor Narcissism Inventory (FFNI): A test of the convergent, discriminant, and incremental validity of FFNI scores in clinical and community samples. *Psychological Assessment*, 25(3), 748-758. DOI: 10.1037/a0032536

Ozer, I., Fitzgerald, S. M., Sulbaran, E., & Garvey, D. (2014). Reliability and content validity of an English as a foreign language (EFL) grade-level test for Turkish primary grade students. *Procedia-Social and Behavioral Sciences*, *112*, 924-929.

Park, E. (2019). The Effects of Linguistic and Demographic Features of Chinese International Students on Placement Test Levels in Higher Education: Logistic Regression. *Journal of International Students*, *9*(1), 225–241.

Puddey, I. B., & Mercer, A. (2014). Predicting academic outcomes in an Australian graduate entry medical programme. *BMC Medical Education*, *14*(1), 31. DOI: https://doi.org/10.1186/1472-6920-14-31

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. *Differential item functioning,* 38(2), 367-388.

Sireci, S.G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83-117. DOI: 10.1023/A:1006985528729

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100-107. DOI: 10.7334/psicothema2013.256

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing, 11*(3), 321-344.

Weir, C. J. (2005). *Language testing and validation*. Hampshire: Palgrave McMillan.

Westrick, P. (2005). Score Reliability and Placement Testing. *JALT Journal,* 27(1), 71-92.

Xi, X., & Sawaki, Y (2017) Methods of Test Validation. In E. Shohamy, I. G. Or, & S. May, (Eds.), *Language testing and assessment,* (pp. 193–210). United States: Springer International Publishing.

Zwick, R. (2002). Is the SAT a 'wealth test'? *Phi Delta Kappan*, *84*(4), 307-311.

Appendices

Appendix A

King Saud bin Abdulaziz University Entrance Test (KSAU-ET) - MCQ (Multiple Choice Questions) Test Blueprint Summary

The Purpose of the Test

1. To assess the English language proficiency of graduate students applying to enroll in Stream II.
2. To determine the students' skill level in reading, writing, and grammar and ensure that they are linguistically competent to pursue degree-level studies in the medium of English.
3. The English language level required should be at the very least at the exit level of our current semester three (Stream I) students. This means that students should be at the top end of the B2 / low C1 CEFR scale overall (or above).

Test Date: XXX

Test Time: 3 hours

KSAU-ET Assessment Blueprint:

Section	No. of questions	Type of questions - MCQ	Targeted Skill	Weight	CEFR Level
Listening Skills	20	Audio recordings	-Ability to understand straightforward information about general topics -Ability to identify main ideas -Ability to identify specific	20%	B1-C1

			details		
Grammar Skills	35	Fill in the blank: sentences and passages (cloze); and error correction	-Ability to recognize and identify appropriate intermediate and advanced grammatical structures	35%	B1-C1
Vocabulary Skills	25	Fill in the blank, definition and word building	- Ability to deduce word meaning from context -Ability to infer the definition of discipline-specific terms based on context -The ability to identify the right affix based on the meaning	25%	B1-C1
Reading Comprehension	20	Main ideas, supporting ideas, inference, organization and logic, reference and lexical comprehension	- Ability to identify and determine main ideas and supporting details -Ability to utilize direct and implied meaning from reading passages to comprehend the meaning -Ability to skim and scan	20%	B2-C1
	100			100%	

Appendix B

Samples from each test section

Cover page with instructions

King Saud bin Abdulaziz University for Health Sciences

ENGLISH LANGUAGE ENTRANCE EXAM

Time: 1:00 PM to 4:00 PM

Tuesday, November 6th, 2018

Total Questions: 100 – Total Pages: 15

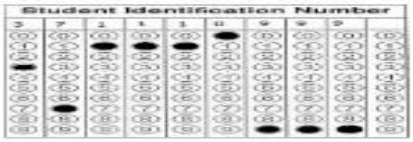
Listening	Grammar	Vocabulary	Reading Comprehension
20 points	35 points	25 points	20 points

Each question carries one mark.

Name: _____ **National ID No.:** _____

Instructions:

- Please ensure that you have no prohibited items in your pockets or on your person in the examination room (i.e. notes, bag, mobile phone, or any other electronic device).
- Write FULL NAME and National ID on the exam booklet and on both answer sheets provided.
- Use **pencil** to WRITE and MARK your National ID on the answer sheets provided.
- Darken the circle corresponding to your choice on the answer sheets provided.
- Ensure that any answers that you have changed are completely erased.
- How to mark *National* ID, sample ⇨



- Mark all exam answers in pencil on **COMPUTER SHEET A** (Listening and Grammar *Questions 1 - 55*) and **COMPUTER SHEET B** (Vocabulary and Reading Comprehension *Questions 1 - 45*) as indicated in the instructions.

Appendix C
Instructions and illustrative examples

Part I. Listening Comprehension

Instructions: This section has four listening prompts. Each prompt will be played twice, and you will have 2 minutes before and after each prompt to answer the questions. Listen carefully to the audios and choose the best possible answer.

Part II. Grammar Skills

A. Instructions Select one (1) correct answer for each question below.

B. Instructions: Complete each blank in the passage below. Select the one (1) correct form of the verb from the choices given.

Three short paragraphs with 3-4 blank spaces each. MCQ options.

C. Instructions: Each sentence has four underlined words or phrases marked A, B, C and D. Choose the letter of the one (1) underlined word or phrase that is NOT CORRECT.

Part III. Vocabulary Skills

A. Instructions: Fill in the blanks with the most suitable word from the list. There are three extra words.

B. Instructions: Choose the correct word from the list that suits the following conditions/explanations. There are two extra words.

C. Instructions: Choose the correct prefix/suffix to complete the underlined word. Disregard any changes in spelling.

Part IV. Reading Comprehension

Instructions: Read the passages below and answer the questions that follow.

One long and one medium reading passage. MCQ questions on main ideas, supporting ideas, inference, organization and logic, reference, and lexical comprehension.