Available online at www.jlls.org

JOURNAL OF LANGUAGE AND LINGUISTIC STUDIES

ISSN: 1305-578X

Journal of Language and Linguistic Studies, 18(Special Issue 1), 359-378; 2022

Evaluating the validity of Saudi English language undergraduate students' results in light of proposed criteria

Sultan Abdullah Almuhaimeed [a 1]

[a] *Department of Curriculum and Instruction, College of Education, Buraydah, Qassim University, Saudi Arabia*

APA Citation:

Almuhaimeed, S. A. (2022). Evaluating the validity of Saudi English language undergraduate students' results in light of proposed criteria. *Journal of Language and Linguistic Studies, 18*(Special Issue 1), 359-378.

Submission Date:15/09/2021

Acceptance Date:08/12/2021

Abstract

Education is the process of acquiring knowledge, skills, values, peers, beliefs, practices and personal development. Mentoring methods include teaching, coaching, storytelling, discussion, and guided research. Education is often led by teachers; however, students can also engage in self-education. As education is a continuous process, it requires rectifying from time to time and that is achieved by conducting regular evaluation and assessment. The role of assessment cannot be denied in any educational process because it works as a course-correction tool. It is indeed one of the pillars of formal education. However, validation of learning results is rarely undertaken in institutions because no comprehensive method for this is available. The current study evaluates international and national best result assessment practices to prepare a balanced model for use in English language Departments in ESL/ EFL situations. The criteria are applied at Mustaqbal University (MU), Saudi Arabia, and results showed that only some of the forty seven criteria are robustly applied, a few partially applied, and some not applied at all. Overall, the study establishes that result assessment needs a careful re-thinking at MU to place it among the most prestigious educational institutions of the world.

Keywords: evaluation; assessment; assessment model; international best practices; language assessment; results assessment

1. Introduction

One of the essential and corner-stone of any teaching is the regular and targeted evaluation of learning outputs. With a great deal of research being conducted in this field, assessment techniques have been undergoing many changes leading to new types of assessment tools becoming available to assess learning outcomes, supporting teachers to obtain the most perfect, targeted and reliable results. Assessment, measurement and evaluation are three basic educational terms and often used interchangeably, yet each term has it's a specific, significant and unique meaning. 'Assessment' has a variety of meanings in its usage and goes as same as the other two terms "evaluation and measurement", yet it is still considered to be used over the other two widely and preferably in the educational arena. Assessment is an essential part of any evaluation process in almost all the educational contexts. Brown (2004) designated assessment as "*any act of interpreting information*

[1] Corresponding author
E-mail address: samhiemied@qu.edu.sa

about student performance, collected through any of a multitude of means or practices." (p. 304). Magno and Ouano (2009) defined it as "*the process of collecting various information needed to come up with overall information that reflects the attainment of goals and purposes.*"(p. 2). The point being driven home here is that assessment can be looked at from different angles: As an act of interpretation; a means of appraising students' achievements or an instance of making a judgment about students' performance. All in all, it relates to estimating the nature, quality, or ability of someone or something.

Measurement is an essential part of assessment, both for learners to know the extent of learning and for the teachers to measure the accomplishment of learning objectives. Magno and Ouano (2009) argued that measurement is characterized by quantification, abstraction, and further analysis. Some assessment results come in the forms of quantitative values that enable the use of further analysis. In terms of measurement, quantification of characteristics or attributes determines the amount of that attribute present.

The assessment results, more or less, are presented in the form of quantitative values that allow for further analysis. In terms of measurement, the dimension of the character or character determines the degree to which traction is present.

When assessment results are used to take decisions and pass judgments, then evaluation takes place (Ellington et al., 1988; Ghaicha, 2016). The term evaluation may be defined differently according to the perspective from which one looks at it. Ghaicha (2016) defined evaluation as "the process of arriving at judgments about abstract entities such as programs, curricula, organizations, institutions and individuals."(p. 213).

Analyzing the results of test statically is a method characteristically used to give full data about student results and verify their validity. However, this analysis gives only results in nutshell, such as frequency and average of the grades. More detailed and comprehensive evaluation of test results still needs to be carefully executed. The issue of student result assessment is highly significant for institutions of higher education to increase the efficiency and effectiveness of the evaluation system. The quality of student assessment affects the quality of university education. Scientific evaluation of student results leads to better performance, together with finding deficiencies that represent unused development opportunities. Result assessment ensures rigor and transparency in evaluation and presents fair opportunities for reform.

Validation is a process in which quality is reviewed. This helps ensure that the assessment system can make informed value judgments (one that confirms a learner holds all the knowledge and skills he/she is certified to have acquired). Validation calls for making sure that the tools used in the assessment have achieved validity, reliability, sufficiency, and authenticity in their evidence. The evidence provides reasonable judgments about whether learning outcomes have been achieved. The validation process ends with making recommendations for future improvements (Australian Skills Quality Authority, 2015).

2. Research Problem

The national development plan of Saudi Arabia as well as Saudi Vision 2030 names education suited to the global job market as one of the primary objectives of the administration. To this effect, education is a heavily funded sector. However, with unemployment reported at 11.6% in 2016, and keeping in view that two-thirds of the population is under the age of 29 years, this figure comes to a staggering number. This inference is corroborated by the reports of a quality audit of 23 Saudi universities by the United Nations Development Programme (UNDP) Regional Bureau for Arab States which state that the assessment process is fraught with weaknesses, primary of these being: i.

Emphasis on recall of content; ii. Poor testing of higher-order cognitive skills; iii. Lack of internal oe external moderation in marking. (UNDP, 2006, p. 5).

In addition, with the use of new assessment tools, changes in technology, manufacturing processes, legislation and graduation necessities which have been named as 'risk indicators' (Australian Skills Quality Authority, 2015), the logical need for validation of student results is also neglected.

Thus, there exists a research gap targeted to be filled by the current study which re-examines traditional validity and reliability concepts which have typically followed a tripartite model comprising:

1. Validity of content: it demonstrates the quality of actions in a specific area of content.

2. Validity of construct: it explains how well the results can be construed as checking about the focused structure of the test.

3. Validity of criteria-related: it is a criterion describing how well the results establish a mutual relation or predict with criteria outside the main assessment.

2.1. Research Objectives

Based upon the existing research gap and pressing need for validation of student results vis-à-vis best national and international practices, the study sets itself the following research aims:

1. Identify the international governing practices used in verifying English language students' results with a detailed overview of current literature and practices,

2. Determine the practices used in validating students' results at Mustaqbal University English language,

3. Develop a model that helps measure the validity of English language students' results.

2.2. Research Questions

The questions of this study are formulated below:

1. What are the international governing practices used in verifying English language students' results?

2. To what extent are the results of English language students at Mustaqbal University validated?

3. What can be a feasible model to measure the validity of English language students' results?

3. Review of Literature

3.1. Theoretical Framework

The quality of student results is assessed by both external, i.e., university graduates, employers of graduates, the Accreditation Commission, and internal environments (Vodák et al., 2013). Valid assessment instruments are produced by first, ascertaining the authenticity and validity of student learning results via suitable approaches, methods and procedures proposed by researchers in the field of evaluation. These include test validly, reliability, difficulty, and discrimination power. Argument-Based Validation is a framework that focuses on validating assessments. Based on this framework, validity pertains to the correctness of the interpretations and utilizing of assessment results, rather than to the assessment instrument itself. Modular-competence is an approach that guides the review of all aspects of the educational process and the inspection and assessment system, including knowledge, skills and practical experience. Additionally, moderation of assessment is an organized procedure that

ensures application of valid assessment material and consistent application of criteria which provides fair academic judgment and reliable outcome in the form of grades essential to monitor students' learning information as an integrated part of the learning process. The following sub-sections enumerate on these aspects.

4. Methods in Evaluating Students Test Results

Having valid data for taking decisions depends on developing a well-planned assessment, the results of which provide information useful for planning improvements. This calls for a review of assessment techniques to ensure that correct procedures for gathering valid data are in place (Australian Skills Quality Authority, 2015; Miller, 2012; Rudolph et al., 1994). As the purpose of the test is to verify that the assessment instruments have achieved the required results, examiners should examine the results in the sample and determine whether they are valid, reliable, relevant and truthful, given that the assessment instruments:

- Complying with the assessment requirements of the relevant outcomes.
- Ensuring that the principles of fairness, flexibility, honesty and reliability are respected.
- Designed to produce results that are legal, relevant and reliable.
- Being suitable for contexts and conditions of assessment.
- Matching the level of difficulty of the tasks to be completed in terms of skills and mental requirements.
- Providing sufficient guidance to clearly explain to the student the tasks to be managed.
- Providing adequate guidance on learning outcomes.
- Identifying reasonable and appropriate adjustments that can be made to achieve the results of the assessment.
- Providing the evaluator with appropriate instructions for gathering evidence, making judgments, and recording the results of the evaluation.
- Being supported by standards of evidence for assessing performance. (Australian Skills Quality Authority, 2015).

Furthermore**,** Miller (2012) proposed the following characteristics of efficient assessment:

1. Validity: the test suitable for the objective testing of the students based on three factors:
   - Content validity: the test able to assess student's knowledge of the subject.
   - Criterion validity: the test is able to measure student's knowledge.
   - Predictive validity: the test is able to predict a student's knowledge during an oral exam, for instance.
2. Reliability: the test is reliable and consistent.
3. Difficulty: the test should be neither too difficult nor too easy.
4. Test discrimination power: the test should show a difference between skilled and unskilled students.

4.1. Argument-Based Validation

It is a framework that focuses only on validating assessments. Nitko (2001) argues that the quality of the assessment is determined using the technical concept of credibility. Legitimacy refers to the

integrity of the interpretation and use of the assessment results, not to the assessment instrument itself used. The argument-based validation framework clearly describes three criteria for assessment. First, assessment results are interpreted; second, reasons are proposed to justify this analysis and how it is relevant to the major uses that lead to the results; third, collected evidence is provided that supports the proposed interpretation in the context of the uses that lead the results.

Evidence in support of the validity argument may be a collection of some kinds: It can be empirical and statistical; it may be established on theoretical and literary research; or it could be the results of a logical analysis. Evidence selection relies on the definition and the specific practice proposed (Nitko, 2001). An English test, as an assumed example, is carried out as part of a computer program for admission to the university. As assuming, the purpose of the test is to assess a candidate's willingness to work at a university with a certain achievement in English (Razak, Krishnasamy & Othman, 2021; Saka, 2020). The argument must support the reasonableness of such interpretation and use of evidence in support of the following suggestions or claims regarding test:

The English content and skills assessed by the examination are in fact necessary for the university level work.

1. Test requirements are based on representative samples of areas relevant to required English language proficiency and skills.

2. The way in which tasks are selected and presented by the computer is consistent with the English frameworks that are the subject of assessment and guidance.

3. Test results are the same and the same for different sets of field assignments.

4. Computer interfaces, testing method, candidates' knowledge of computers, candidates for testing and computer problems do not significantly affect test results.

5. If the test is expected to predict success at the undergraduate level, candidates' performance at the undergraduate level may be well assessed.

6. Students with higher academic performance at university will receive higher scores on computer exams, and vice versa.

Messick (1989), as shown in Table 1, proposed the following validity evidence for assessment.

Table 1. Validity evidence for assessment

Evidence	**Examples of questions needing to be answered**	**Techniques often used to obtain answers**
Correlation between assessment results and the results of other variables (called *external structure evidence*)	a. Are the results of this assessment consistent with those of other similar assessments of these students? b. To what extent does the performance of this assessment method reflect quality or performance as measured by other tests? c. To what extent does performance in this assessment method predict the current or future performance of other valuable actions or measures (criteria)? d. To what extent can the results of the assessment be used to select people for work, school, etc.? What is the size of the error? e. To what extent can the results of the assessment be used to provide students with different recommendations? Is it better to learn when students are distributed this way?	a. Routine operations are identified and analyzed. An assessment of their important characteristics is under development. b. The scores of the assessment compared to the scores are at the expected level. c. Various classification and forecasting errors are analyzed. d. Research shows whether the results of this assessment are consistent with those of other assessments as expected using the proposed definition of student achievement (known as homosexual and concrete evidence).

4.2. Modular-competence approach

It indicates the direction of the review of all aspects of the educational process and the system of verification and assessment (Markova et al., 2014). With this approach, more attention is paid to the assessment of educational achievements, including knowledge, skills and knowledge in general cultural and professional practices and competencies, implemented in the corresponding professional activity (Bobienko, 2012). Particular attention is paid to the development of assessment tools, evidence of the achievement of the declared educational results in the form of competencies. In addition, it is suggested that the choice of assessment instruments is established on several aspects: the validity of the assessment (the methods and results of the assessment should be consistent with the learning objectives); assessment reliability (measuring the accuracy of the assessment system to determine learning outcomes); assessment standard (content is similar to assessment methods, equal time and assessment rules for all subjects). (Robutti et al., 2016).

4.3. Moderation of assessment

It is an organized procedure that ensures using valid assessment material and consistent application of criteria. It provides fair academic judgment and reliable outcome in the form of grades. It ensures appropriate designing and implementation of assessment activities, together with producing valid and reliable results. Using moderation in the assessment system leads to developing academic quality in higher educational institutions (Marg, 2019) as:

1. It tackles any difference in individual judgments of different rater.
2. It confirms that all achievements in grades across courses reflect achievement of same level of standard.
3. It develops a common understanding of the standards.
4. It recognizes performance that demonstrates that standard.

Moderation can be applied to both external and internal techniques of assessment. To implement moderation, following questions need to be answered (Marg, 2019):

1. What are the rubrics used for each of the different types of assessment in the course?
2. Is a standardized rubric used or has the instructor developed his/her own rubric?
3. If the instructor is using a personally framed rubric, or if there is no identified rubric, then how does the assessment measure the learning outcomes?
4. Concerning the difficulty level of the questions, is the difficulty level on the extremes?
5. Concerning, the manner of awarding marks, i.e., has the rating been at the extremes?

A committee should be established, roles assigned, and responsibilities allocated for the moderation process. To ensure neutrality, the moderator should not be the assessor. Staff members should be trained professionally in assessment techniques and moderation procedures. Lastly, all assessment material produced by learner, for example examination sheets, assignments, project reports, and research reports, should be examined.

Moderation is a quality control process designed in accordance with assessment procedures. Monitoring is usually done before the student is summed up because it ensures that the same decisions apply to all assessment results within the same unit of ability. The standards' requirement for judgment does not affect the ability to perform valuation activities or any other process designed to improve the quality of valuation (Australian Skills Quality Authority, 2015). Monitoring students' experience and knowledge is a vital part of the learning process. The term "control" points to the results achieved with

the planned learning objectives. It is necessary to study the students' information in order to obtain awareness about the correct or incorrect ultimate results of the tasks performed. The teacher is aided by correctly organized control over the educational activities of students as that enables him/her to assess students' knowledge, skills, time and supply the necessary support to achieve the educational goals. All this together creates favorable conditions for the development of the cognitive abilities of students and the strengthening of independent work in the classroom. A well-organized discipline allows not only to accurately assess the extent to which students connect the material, but also to learn about their achievements and shortcomings in teaching methods. Therefore, the choice of forms of quality control of the experience gained is very important. (Vaganova et al., 2016).

4.4. Assessment modes and results

Many research attempts have been made to produce valid and accurate student learning results (Abdullateef & Muhammedzein, 2021; Brozova & Rydval, 2014; Cooper, 1994; El-Khawas, 1989; Ferretti et al., 2021; Grainger, 2021; Marshall et al., 2020). These attempts have focused on the evaluation process partially or wholly.

El-Khawas (1989) conducted a survey about how assessment results were used by administrators. Three quarters of the respondents reported that their institutions made use of information gained from assessment activity, but differences are discernible among institutions on the extent of use. Community colleges are in the lead, with 82% reporting some use of assessment compared to 62% doctoral universities. Another way to produce valid student learning results is by comparing student scores on placement exams.

Cooper (1994) stated that one method used by Alabama's Snead Community College to measure student learning outcomes is to make a compression of students' grades on pre-core positional tests, which are two post-graduate assessment tests. Thus, these two results from evaluation tests are taken upon completion of studying the core courses. This data is analyzed to identify similarities among learning outcomes to provide college planners with data to improve student learning.

In addition, Brozova and Ridval (2014) interpreted the results of the applied mathematics and computer science test over a 13-year period. The test consists of two parts: written and oral. Student performance in subjects has been low for some time. The aim of the study was to find out if the lowest scores were a result of test quality or a small number of contact hours. Another reason for low grades can be attributed to the mathematical nature of the subject and the inconsistency of the subject being studied. Due to poor results, students also began to change the grading system.

Grainger (2021) indicated that peer review is regarded as a valid quality assurance method in education. This project investigated the development of assessment literacy, specifically the ability to create quality assessment rubrics, in teaching academics across a range of disciplines. A major strength of the peer review process is that it allows course coordinators, independent assessment experts, tutors and students to work collaboratively to make positive improvements to the assessment task and accompanying rubric as well as strengthening alignment to the teaching program to support students. Likewise Marshall et al., (2020) used comparative judgment to assess student performance as an alternative to traditional grading. Comparative judgment does not require any assessment rules and is based on experts who assess even the relative quality of student work at a high level. The resulting decision data is integrated into a statistical model to give each student a score. There are many benefits to this approach, including improved reliability, validity, and efficiency of estimates. The experts rated the students' responses to the two nationwide assessment tasks proportionally, and the reliability and validity of the results were checked using standard methods. The comparative judgment process is believed to provide reliable and reliable valuation results. Abd al-Latif and Muhammadzin (2021) used

a flexible and humane assessment method: dynamic assessment based on Vygotsky's range of proximal development, seeking mediation through good social training and practice to enhance language learning. They found a statistically significant relationship between dynamic assessment and language learning. Ferretti et al., (2021) addressed the issue of how teachers prepared authentic online assessment, as a key variable catalyzing personal history. They examined teachers' beliefs as part of their personality and rated them as one of the main variables of beliefs. The data showed that teachers did not identify valid assessment methods for online learning during the Covid-19 lockout, largely due to a lack of student oversight. There was a misunderstanding of the definition of continuous assessment with a renewed awareness of the possibilities offered by digital technology in terms of educational personalization.

Thus, teachers are led to consider only summative assessment as a tool to investigate and give feedback on learning. Formative assessment is to be taken into consideration in order to produce accurate student learning results.

From this brief perusal, it is evident that the previous studies concentrated on how to produce valid and accurate student learning results, using various methods and techniques. These methods and techniques included using student assessment results for improvement; comparing student scores on exams; using dynamic assessment; using summative and formative assessment together. While each of these focused on one method, there was no model that could be applied to all situations. However, the current study is an attempt to produce a comprehensive model of producing student-learning results.

4.5. International good test practices

4.5.1. Accreditation Board for Engineering and Technology (ABET)

Regarding international perspectives, assessing student outcomes is an essential issue. In an age of accountability and transparency, assessing outcomes has become an international quality standard. ABET provides a comprehensive definition of assessment student outcomes as, "one or more processes that identify, collect, and prepare data to evaluate the attainment of student outcomes. (ABET, 2020, p. 50).

Concerning ABET Self-Study Questionnaire for academic programs (ABET, 2021), certain criteria mentioned students, student outcomes, and continuous improvement. Concerning students, Criterion 1 speaks about evaluating student performance. It requires summarizing the process by which student performance and progress are evaluated. It entails provided information on how the program ensures and documents that students are meeting prerequisites. Regarding student outcomes, Criterion 3 necessitates establishing and revising student outcomes. As far as continuous improvement is concerned, Criterion 4 speaks about how the results of evaluation processes for the student outcomes have been systematically used as input in the continuous improvement. It requires also describing the results of any changes in those cases where re-assessment of the results has been completed. It refers to providing any significant future program improvement plans based upon evaluations.

Likely, AERA, APA, and NCME jointly developed the Standards for Educational and Psychological Testing (Plake & Wise, 2014). The standards of testing and educational evaluation are divided into three clusters:

1. Design and Development of Educational Assessments: Test results should be clearly described. The impact of tests should be monitored to minimize potential negative consequences. Evidence of validity, reliability, and fairness should be provided. It is important to document design, models, and scoring for tests. Regarding use and interpretation of educational assessments.

2. Using and interpretating educational assessments. Steps should be taken to ensure that test preparation and distribution of student material does not adversely affect the effectiveness of test

decisions. Proof of student motivation must be provided to enable students to learn the content and skills that are being measured on the exam. A decision that has a large impact on the student must take into account not only the results of one exam, but other relevant sources as well. An educational decision based on comparison should consider the degree of overlap between the two constructs and the reliability or standard error of the degree of difference. Statistical staff should professionally interpret student-learning outcomes for administration, registration, and reporting of educational assessments.

3. Administration, Scoring, and Reporting of Educational Assessments: The committee in charge of examination should be proficient in the suitable test administration and scoring measures, adhering to the directions provided by the test developer. Interpretation of results should include the degree of measurement error associated with each score. The process should end with recommendations for instructional intervention.

The Centre for Teaching Excellence at University of Waterloo, Canada, provided detailed information about preparing tests and exams (University of Waterloo, 2021) divided into three parts: pre-assessment, during assessment and post-assessment. This includes reasons for giving an examination to students; guidelines for the instructor on what is to be assessed in terms of learning outcomes; guidelines on deciding what to test and how to test it; qualities of a good exam; providing reliable and valid tests; creating realistic expectations; using multiple question types; offering multiple ways to obtain full marks; keeping tests free of bias; using formative and summative assessment; and using transparent marking criteria; hints on preparing a marking scheme usable by non-experts; and reviewing the marking scheme after the exam. Regarding post-assessment, it recommends reviewing examination results so that the instructor may change how he/she teaches the remainder of the semester; checking for improvement on specific topics or methods over; redesigning the course or the examination for future classes; and assessing teaching practice.

4.6. National Good Test Practices

National Commission for Academic Accreditation and Assessment (NCAAA), in Saudi Arabia issued the Self-Evaluation Scales for Higher Education Programs. These scales contain certain sub-standards indicating the necessity of verifying student results (NCAAA, 2019). In terms of *Governance, leadership, and Management*, the program is committed to applying the institutional regulations to ensure the quality of all aspects of the program, for example courses, teaching, and student achievement standards. The program analyzes the assessment data annually once (for example, performance metrics and measurement data, student achievement, program completion rate, student assessment of the program, courses, alumni and employer feedback). Learning outcomes are used in planning, development and decision-making processes.

From a teaching and learning perspective, assessment strategies and methods are consistent with learning outcomes at both the program and course levels. In addition, teaching and learning strategies and assessment methods vary in nature and level, and the ability to conduct research is developed to enable learners to acquire cognitive skills and a higher level of self-learning. Field learning outcomes are consistent with program learning outcomes. Appropriate learning, assessment and coaching strategies are identified to achieve these learning outcomes. Those responsible for the knowledge of the object are informed about the expected learning outcomes and the nature of the actions assigned to them. Daily practice will be followed based on certain criteria. Instructors adhere to the learning and teaching strategies and assessment procedures listed in the curriculum and course specifications. The institution provides teachers with basic training in the learning and teaching strategies and assessment methods outlined in the curriculum and course specifications, as well as the effective use of modern

technology. At the beginning of each course, students will receive comprehensive information about the course, such as learning outcomes, teaching and learning strategies, assessment methods and timing, and learning outcomes. Courses are usually assessed to ensure the effectiveness of teaching and learning strategies and assessment procedures. Specific methods have been introduced to ensure that assessment methods (e.g., specification, diversity, consistency in the assessment of learning outcomes, distribution of marks, and accuracy of indicators) ensure student achievement.

Evaluation practices of student performance: Al-Baha University, in Saudi Arabia, provides foundations of the design of educational testing and evaluation (Deanship of Quality and Academic Accreditation, 2019). Before initiating an examination, the instructor should identify the characteristics of good testing and the factors affecting validity and reliability. The rules of test preparation and quality should be followed in terms of form and content (e.g., general specifications, rules used in the design of the test paper, rules used for the application of tests). It is vital to follow up with the examination process and progress. Additionally, rating, analysis, and interpretation of results must be monitored.

5. Methodology

5.1. The current practices

This study has been conducted with the results of the English language students at Mustaqbal University. The Department of English, Mustaqbal University, provides many ways that ensure the validity of student results. These are as follows:

5.2. Unification of tests

Male and female sections prepare the same tests for English program in the sense that a male faculty member prepares a test and sends it to his counterpart female faculty member to revise and approve it.

5.3. Independent verification

The Department of English adopts mechanisms for the independent verification of the validity and objectivity of assessment of student achievement. These mechanisms include: i. Tests are evaluated internally by members of English Department according to a set of criteria, ii. Independent verification of tests by a jury from outside the university.

5.4. Report on student results

A detailed report is made on student results in terms of the grades given to students, male and female. It also provides percentages of each grade. The report ends with strengths and weaknesses, together with an action plan for implementing the recommendations.

6. The Rationale for the Proposed Model

Magno and Ouano (2009) indicated that assessment can be applied before, during, and after supervision. Before training, teachers can use the results of the assessment as a basis for goals and recommendations for their plans. These assessment scores are drawn from the previous year's student performance tests, the previous year's student grades, the previous lesson assessment results, or the pre-teaching test results.

6.1. Data analysis and results

With this available background at the University, this study proposed a comprehensive model for evaluating students' results. The suggested model consists of 47 criteria across four stages. The first stage deals with testing in terms of form (test specifications in terms of form) and content (test specifications in terms of content: achieving test validity.) The second stage tackles actual testing: testing follow-up. The third stage has to do with post-testing: rating, analysis, and interpretation of test results. The fourth stage closes the quality loop: implementation before preparing the next test. Table 2 presents the application of these stages of the proposed evaluation scale at Mustaqbal University during the first semester of academic year 2021-2022.

Table 2. Application of the MODEL at the Department of English, Mustaqbal University

Dimension	Assessment Criteria	Evaluation scale			Notes
		Applied	Partially Applied	Not Applied	
First Stage: Pre-Testing in terms of form and content A. Test specifications in terms of form	A. Test specifications of in terms of form: The test should have a set of formal criteria that facilitate the process of dealing with the test for the student and the examinees, the most important of which are:				
	1. Basic data of the university, the college and the department	✔			
	2. Course name, code, and No.	✔			
	3. Target group of students (section and level).	✔			
	4. Test time	✔			
	5. No. of test pages	✔			
	6. Clarity of test instructions.	✔			
	7. Total score of the test.	✔			
	8. Distribution of scores on test questions.	✔			
	9. Content Organization in terms of font type, size, and line spacing.		✔		Some staff are not well qualified in content organization
	10. Diversity of test questions: essay and objective.	✔			
	11. Diversity of objective questions: multiple choice, true and false, matching, and completion.	✔			
	12. Training students in types of test questions during study.		✔		

Dimension	Assessment Criteria	Evaluation scale			Notes
		Applied	Partially Applied	Not Applied	
First Stage: Pre-Testing in terms of form and content B. Test specifications in terms of content (achieving test validity)	B. Test specifications in terms of content (achieving test validity of the test): the test should have a set of criteria related to the test content, ensuring achieving validity of the results obtained from the evaluation process, and the extent to which learning outcomes are achieved. The most important of these criteria are as follows:				
	13. The questions cover the content of the course topics (content validity).		✔		Some tests do not pay attention to covering the course content topics
	14. The wording should be grammatically correct and free from spelling mistakes, ensuring that scientific terms are clearly written in Arabic and English according to the specialization		✔		Some spelling sometimes occur due to not paying attention to revising the test paper after preparing it.
	15. The test questions are related to the intended learning outcomes. (It is preferable to put the intended learning outcome No. as stated in the course description after the question score).			✔	Most tests focus on rememberin g as they use verbs such as mention, state, define, enumerate,
	16. Adequacy of questions to measure all learning outcomes intended by the course based on course description and objectives.			✔	
	17. The questions differentiate between the levels of student achievement.		✔		
	18. Performance verbs are used in formulating questions that measure aspects of knowledge at all levels of thinking.		✔		
	19. Questions are graded from easy to difficult.	✔			Some tests do not observe this criterion.

Dimension	Assessment Criteria	Evaluation scale			Notes
		Applied	Partially Applied	Not Applied	
	20. Questions are clearly and specifically formulated so that they do not confuse students.	✔			Some testes are difficult to understand due to the hasty way of preparing them
	21. The questions are well formulated according to each type of questions.	✔			
	22. Questions are formulated that each question measures one learning outcome.			✔	
	23. Avoiding repeating the same question in different ways of wording.	✔			
Second Stage: During Testing – Testing follow-up	24. Formation of the examination committee.	✔			
	25. Ensuring that the test paper meets quality standards.	✔			
	26. Maintaining discipline in the examination rooms.	✔			
	27. Appropriate setting for the test room in terms of lighting, ventilation and comfort for students.	✔			
	28. Ensuring that students write the required data.	✔			
	29. Students sign attendance sheets.	✔			لإاثغThey sign once when they leave, they should sign twice when coming to exam room for attendance and when leaving.
	30. Collecting answer sheets after test completion.	✔			
	31. Making notes of the examination process.	✔			
	32. Preparing examination reports.	✔			
Third Stage: Post-Testing – rating, analysis, and	33. Receiving the answer sheets from the committee in charge of examination.	✔			

Dimension	Assessment Criteria	Evaluation scale			Notes
		Applied	Partially Applied	Not Applied	
interpretation of test results	34. Specifying criteria of rating each question or part of a question (criterion-referenced - standard-referenced).			✔	
	35. Putting a tick or a cross on each question answer (evidence that the rater read and examined the question answer).		✔		
	36. The score of each question answer is written inside the answer booklet by the rater.		✔		
	37. Ensuring that test scores are added and reviewed.		✔		They are not reviewed.
	38. Revising rating the test paper.			✔	
	39. The total score and grade are signed by the rater.		✔		
	40. Preparing a statistical description of students' performance.			✔	It is for the grades only.
	41. Comparing the academic performance indicators with the course learning outcomes.			✔	
	42. Interpretation of the results by the committee in charge of analysis and evaluation.			✔	
	43. The program quality committee of evaluates the test paper according to the above criteria and prepares a report.	✔			Other criteria are used
	44. Evaluation of the tests by an independent opinion.			✔	
Fourth Stage: closing the quality loop (implementation before preparing the next test)	45. Briefing the course instructor on the quality committee report mentioned in Item No. 43, discussing it with the committee.		✔		
	46. The course instructor takes account of the notes in the report in Item No. 43 and Item No. 44 when preparing a new test.		✔		
	47. Re-evaluating the test continuously to ensure applying all model criteria.			✔	

7. Results

Application of the proposed model clearly demonstrated the extent to which best practices were followed at Mustaqbal University. These results are presented in Table 3 below.

Table 3. Results of applying the Assessment Criteria Model

Stage		Assessment Criteria			Total
		Applied	Partially Applied	Not Applied	
First Stage: Pre-Testing in terms of form and content	A. Test specifications in terms of form	10	2	0	12
	B. Test specifications in terms of content (achieving test validity)	4	4	3	10
	Total	14	6	3	22
Second Stage: During Testing – Testing follow-up		9	0	0	9
Third Stage: Post-Testing – rating, analysis, and interpretation of test results		3	4	6	13
Fourth Stage: closing the quality loop (implementation before preparing the next test)		0	2	1	3
		26	12	10	47
Total percentage		55.3%	25.5%	21.3%	

As shown in Table 3, concerning the first stage of pre-testing in terms of form, 10 criteria are applied, 2 partially applied, and none not applied. What stands out is that most criteria are applied. Regarding the first stage of pre-testing in terms of content, achieving test validity, 4 criteria are applied, none partially applied, and 3 not applied. This weakens the test validity as 7 criteria out of 10 are included in partially applied and not applied criteria. As for assessing the first stage of pre-testing in general, 14 criteria are applied, 6 partially applied, 3 not applied. This indicates that pre-testing stage is not well prepared and needs to be worked upon.

On the other hand, in the second stage, during testing and testing follow-up, Mustaqbal Univeristy applied 9 criteria out of 9. This means that testing and testing follow-up are fully prepared and monitored. As for the third stage related to post-testing: rating, analysis, and interpretation of test results, 3 criteria were applied, 4 partially applied and 6 not applied. This shows a weakness in the post-testing stage. As for the fourth stage: closing the quality loop (implementation before preparing the next test), no criteria were applied, 2 partially applied and one not applied. This also shows weakness in closing the quality loop.

When the criteria are assessed generally, 26 criteria were applied which comes to 55.3% of all proposed criteria, 12 or 25.5% were partially applied, 10 or 21.3% were not applied. This gives the impression that student learning results are not fully validated. Nearly half of the criteria need attention to validate student learning results at Mustaqbal University.

Table 4. Percentage of assessment criteria for each stage of assessing student learning

Stage		Assessment Criteria			Total
		Applied	Partially Applied	Not Applied	
First Stage: Pre-Testing	A. Test specifications in	10	2	0	12

in terms of form and content	terms of form				
	B. Test specifications in terms of content (achieving test validity)	4	4	3	10
	Total	14	6	3	22
	Percentage	63.7%	27.3%	13.6%	
Second Stage: During Testing – Testing follow-up		9	0	0	9
Percentage		100%	0%	0%	
Third Stage: Post-Testing – rating, analysis, and interpretation of test results		3	4	6	13
Percentage		23.1%	30.8%	46.2%	
Fourth Stage: closing the quality loop (implementation before preparing the next test)		0	2	1	3
Percentage		0%	66.7%	33.3%	
Total		26	12	10	47
Total percentage		55.3%	25.5%	21.3%	

Table 4 shows that the second stage achieved the highest percentage of applying the assessment criteria in terms of testing and testing follow-up, at 100%; followed by the first stage of pre-testing in terms of form and content, at 63.7%; and the third stage of post-testing at 23%; and lastly the fourth stage of closing the quality loop at 0%. These results suggest that there is imbalance in applying some assessment criteria at 100% and the others 0%.

The fourth stage achieved the highest percentage of partially applying the assessment criteria in terms of closing the quality loop, at 66.7%; followed by the third stage of post-testing at 30.8%; then the first stage of pre-testing in terms of form and content, at 27.3%; and lastly the second stage of testing and testing follow-up at 0%. Overall, these results indicate that the second stage helps a great deal in the process of producing validated student learning results. On the contrary, the fourth stage weakens this process.

The third stage achieved the highest percentage of not applying the assessment criteria in terms of post-testing, at 46.2%, followed by fourth stage of closing the quality loop at 33.3%; then the first stage of pre-testing in terms of form and content, at 13.6%; and lastly the second stage of testing and testing follow-up, at 0%. This indicates that post-testing in terms of rating, analysis, and interpretation of test results needs considerable effort to improve.

8. Discussion

This research sought to identify the international governing practices used in verifying English language students' results. Certain international bodies were reviewed such as Accreditation Board for Engineering and Technology (ABET); American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME); and Centre for Teaching Excellence, University of Waterloo. Moreover, certain national bodies were reviewed such as National Commission for Academic Accreditation and Assessment

(NCAAA); and Evaluation practices of student performance, Al-Baha University. These international and national good test practices have helped form the model of assessment criteria by providing a basis for formulation. The model was built on the basis that assessment is a process. Therefore, the model consisted of four stages: pre-testing, during testing, post-testing and closing the quality loop. Many studies, such as (Jeltova et al., 2009) and (Council of Chief State School Officers (CCSSO, 2018), confirmed that assessment is considered as a process. Banta et al., (1996) argued that assessment works best when it is ongoing, not episodic.

9. Conclusion

The research attempted to determine international and national best result assessment practices to prepare a balanced model for use in English language Departments in ESL/ EFL situations the practices used in validating Mustaqbal University. The application of the model of assessment criteria to Mustaqbal University English language students' results revealed some weaknesses and strengths. In terms of weaknesses, the third stage of post-testing criteria (rating, analysis, and interpretation of test results) were nor nearly applied. This poor practice weakens the validity of student learning results as rating, analysis, and interpretation of test results are of great importance, as suggested by Magno and Ouano (2009), Marg (2019), and Deanship of Quality and Academic Accreditation (2019). Furthermore, the quality loop was not closed. In this regard, many studies confirm the importance of closing the quality loop, such as Glaskin-Clay (2007), Schoepp and Benson (2016), and Naveed Bin Rais et al., (2021). Overall, nearly half of the assessment criteria were applied, indicating that assessment criteria used in validating Mustaqbal University English language students' results have been reconsidered.

10. Recommendations

Overall, the result assessment at Mustaqbal University is more than satisfactory. There is, however, always a scope for improvement. Thus, it needs a careful re-thinking and on-going progress to place it among the most prestigious educational institutions of the world. Certain recommendations may be set out based on this research. Further work needs to be done to implement assessment criteria concerning validating MU English language students' results. Further research could also be conducted to determine the effectiveness of the suggested model of validating student learning results at other institutions.

References

Abdullateef, S. T., & Muhammedzein, F. (2021). Dynamic assessment: A complementary method to promote EFL learning. *Arab World English Journal (AWEJ), 12*(2), 279-293.

ABET. (2020). *ABET: Accreditation Policy and Procedure Manual.* https://www.abet.org/wp-content/uploads/2021/01/A001-21-22-Accreditation-Policy-and-Procedure-Manual.pdf

ABET. (2021). *Abet self-study questionnaire: Template for a self-study report, 2021-2022 Review Cycle*.https://www.abet.org/accreditation/accreditation-criteria/self-study-templates/

Australian Skills Quality Authority. (2015). Conducting validation. https://www.asqa.gov.au/sites/default/files/202001/FACT_SHEET_Conducting_validation.pdf

Banta, T. W., Lund, J. P., Black K. E., & Oblander F. W. (1996). *Assessment in practice: Putting principles to work on college campuses.* San Francisco: Jossey-Bass.

Bobienko, O. M. (2012). Gauging the materials for the assessment of new educational results. *TISBI Bulletin, 2*(50), 173-183.

Brown, T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education, 11*(3), 305-322.

Brožová, H., & Rydval, J. (2014). Analysis of exam results of the subject' Applied Mathematics for IT'. *Journal on Efficiency and Responsibility in Education and Science, 7*(3-4), 59-65.

Cooper, E. L. (1994). *Using standardized test results to assess student learning.* ERIC Number: ED387154. https://eric.ed.gov/?id=ED387154

Council of Chief State School Officers (CCSSO). (2018). *Revising the definition of formative assessment.* Washington, DC: Author.

Deanship of Quality and Academic Accreditation. (2019). *Designing and assuring quality of exams.* Al-Baha University. https://bu.edu.sa/documents/636739/0/Ref.+3.1.5.2.pdf/3d3d7e54-5de5-9438-2b41-8515aeb17131?t=1593679855218

El-Khawas, E. (1989). *How are assessment results being used? Assessment Update, 1*(4), 1-2.

Ellington, H., Percival, F, & Race, P. (1993). *Handbook of educational technology*. London: Kogan Page.

Ferretti, F., Santi, G. R. P., Del Zozzo, A., Garzetti, M., & Bolondi, G. (2021). Assessment practices and beliefs: Teachers' perspectives on assessment during long distance learning. *Educ. Sci., 11,* 264. https://doi.org/10.3390/educsci11060264

Ghaicha, A. (2016). Theoretical framework for educational assessment: A synoptic review. *Journal of Education and Practice, 7*(24), 212-231.

Glaskin-Clay, B. (2007). Part-time instructors: Closing the quality loop. *College Quarterly, 10*(3), 1-11.

Grainger, P. (2021). Enhancing assessment literacies through development of quality rubrics using a Triad based peer review process. *Journal of University Teaching & Learning Practice, 18*(4), 4-14. https://ro.uow.edu.au/jutlp/vol18/iss4/4

Jeltova, I., Birney, D., Fredine, N., Jarvin, L., Sternberg, R., & Grigorenko, E. (2009). Dynamic assessment as a process-oriented assessment in educational settings. *Advances in Speech Language Pathology, 9*(4). https://doi.org/10.1080/14417040701460390

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535.

Magno, C., & Ouano, J. (2010). *Designing written assessment for student learning*. Philippines: Phoenix Pub.

Marg, B. S. (2019). *Evaluation reforms in higher educational institutions.* New Delhi: University Grants Commission.

Markova, S. M., Gladkova, M. N., Yurtaeva, T.S., Bystrova, N.V., Vaganova, O. I., & Tsyplakova, S. A. (2014). Modular content pedagogical preparation of teachers of vocational training. *Chronicles of the Combined Fund of Electronic Resources Science and Education, 1*(66), 90.

Marshall, N., Shaw, K., Hunter, J., & Jones, I. (2020). Assessment by comparative judgement: An application to secondary statistics and English in New Zealand. *New Zealand Journal of Educational Studies, 55,* 49–71.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York, NY American Council on education and Macmillan.

Miller, I. (2012). *Edukometrie*. https://www.miller.wz.cz/

Naveed Bin Rais, R., Rashid, M., Zakria, M., Hussain, S., Qadir, J., & Imran, M. A. (2021). Employing industrial quality management systems for quality assurance in outcome-based engineering education: A review. *Education Sciences, 11*(45).

NCAA. (2019). Self-Evaluation scales for higher education programs. https://etec.gov.sa/en/productsandservices/NCAAA/AccreditationProgrammatic/Pages/Forms.aspx

Nitko, A. J. (2001). Conceptual frameworks to accommodate the validation of rapidly changing requirements for assessments, In Scott, D. (Ed.), *Curriculum and assessment: International Perspectives on Curriculum Studies,* (Vol. 1). London: Ablex Publishing.

Phan, T. C., Ngo, T. T., & Duong, N. T. (2019). A case study in teaching: The factors determining of assessing the competence of technology-based. *Review of Information Engineering and Applications, 6*(2), 37-45.

Phoenix, AZ: American Council on Education/ The Pryx Press, 13–104.

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME standards for educational and psychological testing?. *Journal of Educational Measurement: Issues and Practice, 33*(4), 4-12. https://doi.org/10.1111/emip.12045

Abdul Razak, A., Krishnasamy, H.N., & Othman, N. (2021). "No trolls left behind!": Using films to address communication apprehension in the English language classroom at Universiti Utara Malaysia. *Journal of Language and Linguistic Studies, 17*(3), 1258-1265. Doi: 10.52462/jlls.89

Robutti, O., Cusi, A., Clark-Wilson, A., Jaworski, B., Chapman, O., Esteley, C., & Joubert, M. (2016). ICME international survey on teachers working and learning through collaboration. *ZDM, 48*(5), 651-690.

Rudolph, L. B., Poje, D. J., & Van Dyke, J. (1994). Using the results of assessment for improvement. *Assessment Update, 6*(5), 4-7. https://doi.org/10.1002/au.3650060504

Saka, F. Ö. (2020). Considerations on the new curriculum of English Language Teaching programmes. *Journal of Language and Linguistic Studies, 16*(3), 1189-1202.

Schoepp, K., & Benson, S. (2016). Meta-Assessment: Assessing the learning outcomes assessment program. *Innovative Higher Education, 41*(4), 287-301.

United Nations Development Programme (UNDP). (2006). *UNDP Annual Report*. https://www.undp.org/publications/undp-annual-report-2006

University of Waterloo. (2021). *Preparing tests and exams*. Centre for Teaching Excellence, University of Waterloo.

Vaganova, O. I., Medvedeva, T. Y., Kirdyanova, E. R., Kazantseva, G. A., & Karpukova, A. A. (2016). Innovative approaches to assessment of results of higher school students training. *International Journal of Environmental & Science Education, 11*(13), 6246-6254.

Vodák, J., Soviar, J., & Lendel, V. (2013). The evaluation system proposal of the businesses preparedness for cooperative management implementation. *Business: theory and practice, 14*(4), 315-322.

AUTHOR BIODATA

Sultan Almuhaimeed is an Associate Professor in the Department of Curriculum and Instruction, College of Education, Buraydah, Qassim University, Saudi Arabia. Dr. Almuhaimeed got his PhD in Curriculum and Instruction with emphasis on TEFL/ TESOL from Kent State University, Ohio, the United States, and MA in Curriculum and Instruction from the same University. He has held many administrative positions so far including, the Dean of Development and Quality, Vice Dean of College of Education for Planning, Development and Quality at Qassim University. Besides being a faculty member at the College of Education, he worked as a lecturer in the Department of Education and Psychology, College of Arabic Language and Social Studies, Qassim University. Dr. Almuhaimeed has a very good record of regional and international conferences and seminars. He is also an active member in many scientific, professional and specialized committees and associations, namely, Head of Standing Committee for Quality, Coordinator of the National Transformation Program Initiatives, Qassim University, among others. Apart from his bright academic and administrative record, he is also a deemed expert of Quality Assurance and Accreditation. His research interests include, but not limited to, learning and teaching strategies, PRESET and INSET, and SLA.