Asian Journal of Assessment in Teaching and Learning
Vol 14, Issue 1, 2024 (55-69) /eISSN ***2821-2916***

Do Exam Aims and Content Reflect those of the Curriculum? An Evaluative Study

Abdulhamid Mustafa Onaiba
Department of English, School of Languages, Libyan Academy for Postgraduate Studies, Misurata Branch, Libya

*Corresponding author: a.onaiba@lam.edu.ly

Received: 12 May 2024; **Revised**: 20 June 2024;
Accepted: 28 June 2024; **Published**: 30 June 2024

To link to this article: https://doi.org/10.37134/ajatel.vol14.1.6.2024

Abstract

Language tests, particularly high-stakes language tests, are a powerful tool for evaluating educational outcomes, but their effectiveness hinges on how they are constructed. Failure to construct valid sound tests can negatively impact both teaching and learning. Therefore, continuous research is needed to investigate and evaluate such types of tests. This study examines the congruency between the aims and content of a high-stakes public EFL examination and the prescribed curriculum in Libyan schools. Document analysis was conducted on a sample of the studied test, focusing on its intended objectives and their reflection on the curriculum's goals and content. While the exam offered some practical advantages, the findings revealed a mismatch between the administered exam and the curriculum, and between the stated aims and the actual content of the exam itself. The assessment focused solely on grammar points and reading comprehension based on information cloned verbatim from the prescribed textbook. Writing was assessed indirectly through true-false responses to prompts involving structured or unstructured sentences. Notably, the exam entirely neglected listening and speaking skills. These findings suggest that the current examination system could hinder curriculum implementation and EFL education in Libyan schools. Therefore, the study calls for examination reform to ensure alignment between the tests and the English curriculum, ultimately promoting effective second language learning.

Keywords: *EFL high-stakes tests, Curriculum, Aims, Objectives, Content Validity, Evaluation*

INTRODUCTION

High-stakes tests are a crucial part of language education, impacting what and how teachers teach and what and how students learn (Alderson & Wall, 1993; Dong & Liu, 2022; Desalegn et al., 2023). Due to their importance, teachers may prioritize content covered in the test over broader learning goals (Shohamy, 2020; Gorgodze & Chakhaia, 2021). This can influence students' motivation, learning approaches, anxiety levels, and time spent studying (Dong et al., 2021). There have been attempts to reform these exams to make learning more practical and beneficial for students. Nevertheless, modifying the tests may not guarantee success. Research shows that even reformed tests might not be effective if the tests themselves are not well-designed and have validity problems (Amin, 2021). So, test validity is a key concept here. As mentioned by French et al. (2023), a test valid for one purpose (e.g., comparing students) might not be valid for another (e.g., deciding who passes). The main point is that reforming high-stakes tests requires careful consideration of how they will be used. However, designing and validating language tests is not an easy task. There are several reasons why formal validation of high-stakes tests is demanding and sophisticated, and thus, is rarely done. These reasons include: a) the concept of validity per se is complex and not easily defined; b), there's no single agreed-upon method for validating exams; c), providing a complete picture of an exam's validity requires using

multiple methods, which can be time-consuming and require efforts (Shaw et al., 2012).

When test systems, such as high-stakes tests, carry significant importance for selection, progress, acknowledgment, or accreditation, assurance in the reliability and validity of the test being introduced should be maintained (French et al., 2023). Validity is a core concept in testing and refers to how well a test measures what it claims to measure (Hughes & Hughes, 2020). Different types of validity provide evidence for a test's overall validity. These include: a) face validity which refers to the intuitive holistic stakeholders' judgments about whether an instrument looks like it measures what it was intended to measure (Hughes and Hughes, 2020); b), construct validity, which examines if the assessment measures the target concept); c); content validity, which is the measurement of whether a test contains a representative sample of the relevant language skills and objectives stipulated in a language programme (Fulcher, 2015); d), criterion validity, which assesses whether the assessment scores are associated with the functional behaviors it aims to measure), and; e), consequential validity, which considers if the assessment has the potential for favorable or unfavorable outcomes (Messick, 1984). The current research paper tackles one facet of the type's continuum of validity, namely content validity, the other types are beyond the scope of this research. This paper studies an external high-stakes public examination administered summatively as a school leaving exam in Libyan schools known as the Basic Education Certificate English Examination (henceforth BECEE). To wit, this research investigates documentarily the content of the BECEE and its objectives vis-à-vis the content and objectives of the English curriculum upon which the test is designed.

Moreover, the utilization of curriculum content guides assists in identifying the necessary evidence required to substantiate the assertion that the assessment type, such as BECEE in our case, is comparable with its intended aims and purposes, drawing upon the course objectives as a primary point of reference (Ziebell, 2018). While it is plausible that the assessment might not cover all the proficiencies developed in the courses, it is anticipated to encompass a significant portion of them (Kane & Wools, 2019). Evaluating the course materials' content, i.e., content analysis, would offer insights to ascertain the suitability and representativeness of the language and competencies tested vis-à-vis those assimilated from the prescribed course materials (Ziebell, 2018).

PREVIOUS STUDIES

Investigating the content validity (item analysis) of the third-grade EFL external test administered to primary school students in Turkey, Ozer et al. (2014) used validation—content validity—scales assessed by a group of experts from a public university in the United States. The findings revealed that the test studied has low validity. Similar research that has used expert judgment to validate language tests is Kang and Chang's (2014). The researchers used a critical review of experts to investigate the content validity of the Practical English Language test in Korea PECT. In the same research, construct validity was also investigated by comparing the results of sample students who were asked to take standardized tests after taking the PECT. The results showed that the test is suitable for measuring Korean students' achievement of English learning in public education.

Moreover, other studies have investigated the concept of test validity from teachers' and students' perspectives (such as Wisdom, 2018; Abdulhamid, 2018; Amin, 2021; 2022; Onaiba, 2014; Katiso, 2022; Desalegn, 2023, etc.). For instance, Wisdom (2018) investigated a high school nationwide high-stakes test. The researcher examined the test validity from teachers' and students' perspectives by exploring teachers' perceptions about the influence of the studies exam on Students. The findings indicated that the exam recorded high in terms of its face validity, whilst gaining a low degree of content and construct validity. Furthermore, deploying interviews, Amin (2021) examined students' perception of the high-stakes English language tests administered to junior secondary school students in Bangladesh. Results elicited from students' interviews indicated that there was a discordancy between curriculum content and the content of the investigated exams.

However, in the previously mentioned studies, it was perceived that researchers would not provide in-depth justifications for their claims about their findings, rather, they merely depended on the opinions made by experts and anecdotes raised by some other stakeholders, teachers, and students. Replicating the results of his validation studies, Alderson and Kremmel (2013) commented that "a

sufficient number of studies had already made the problematic nature of 'experts' and their judgments in language testing clear, the field needs to ask itself why it is that 'expert' judgments are still often solely relied upon in these matters" (p. 535). It could be argued here that results from such studies are to some extent doubted and claimed not to be straightforward because teachers and students may hold contrasting perspectives on the validity and fairness of a test, its implementation, and its effects, as noted by Bridglall et al. (2014). Hence, in the current study, the researcher utilized in-depth investigations into content validity utilizing document analysis rather than other stakeholders' opinions and judgments to evaluate the validity of the BECEE.

Literature has documented the adoption of analytical methods to investigate the content of tests. Gashaye and Degwale (2019) have investigated the content validity of high school English Language tests in Ethiopia using textbook analysis, sample paper analysis, and focus group discussions. The study's findings viewed teacher-made English language tests in Ethiopian high schools as insufficiently mirroring the coverage of the textbooks used in the classrooms; thus, it was claimed to be deficient in content validity. Additionally, Katiso (2022) analysed the content and face Validity of the English final examination given to Grade Seven Ethiopian students in the Damboya District. The study used a descriptive survey design and mixed approach to analyze the contents of grade 7 English textbooks and examinations. The results showed that the examinations did not adequately represent the textbook's contents, with grammar and vocabulary dominating. Speaking and writing were weakly represented and listening and reading were ignored. The study concluded that the examinations were ineffective in measuring textbook objectives due to limitations in implementation.

Furthermore, Siddiek (2010) held an analytical investigation of test items where regulations of good examinations were considered while analyzing the test items along with gathering attitudes of language teachers towards the comprehensiveness and content validity of the target tests. The findings were negative towards the test claiming that the test was incomprehensive and lacked content validity. Moreover, Dong et al. (2021), Moritoshi (2002), and Takeno and Moritoshi (2018) have investigated the content of the targeted tests by following Bachman and Palmer's (1996) framework. The researchers reported some deficiencies in the tests they investigated which call for recommendations for some improvements to be made.

Libyan high-stakes exams are no exception; they have experienced validation investigations as well. Regarding the English subject, two washback studies were conducted in the context of the high-stakes public examinations within the Libyan education system. The first was conducted by Onaiba (2014) who investigated the Basic Education Certificate English Exam (BECEE), the studied exam in this research, in terms of its washback effect concerning teachers' perceptions, instructional practices, and curriculum. Teacher questionnaires, interviews with teachers, and inspectors of English and classroom observations were utilized. The researcher contended that the test lacked valid evidence, and it did have washback effect mostly negative. The findings revealed that teachers to some extent hold negative views and opinions towards the studied exam, the BECEE. Onaiba's study also indicated that teachers tailored their classroom teaching practices to fit the exam format and content, pushing teachers to give special priority to focus on the language elements that mostly appear on the final exams ignoring other important elements, which put curriculum content at risk.

Furthermore, in her washback study, Abdulhamid (2019) investigated the relationship between the degree of alignment of components of the Libyan education system and the *washback* of the Secondary Education Certificate Examination of English (SECEE), The researcher studied the exam from washback on teachers and washback on learners, by conducting questionnaires and interviews with teachers and students, accompanied by classroom observation sessions. She reported that the exam lacked validity evidence as well. The study found that there is a limited-to-no degree of alignment between the SECEE and the EFL content standards; SECEE appears to have had negative washback on some teachers and their teaching, but little-to-no negative washback on other teachers; SECEE may have had negative washback on learners and their learning.

Onaiba's (2014) and Abdulhamid's (2019) studies were mainly focusing on the washback effect of the studied tests on teaching and learning in Libyan schools. They dealt with issues related to the kind of washback that can be exerted on teachers' instructional behaviour and students' learning processes due to introducing high-stakes tests, the BECEE, and the SECEE. Both studies concluded that the tests did not attempt to measure adequately the students' achievement as far as the curriculum is concerned. Although the two studies raised serious concerns regarding the validity of high-stakes

tests administered in Libyan schools, they depended largely on stakeholders' anecdotes based on data elicited from questionnaires and interviews. In other words, the two investigations were not conducted to provide in-depth detailed validation insights about the studied test content vis-à-vis its objectives and the objectives and content of the English curriculum prescribed, a gap needs to be filled via this research study.

Moreover, the only study conducted in the context of the current research which has some similarity to this study was carried out by Ghuma (2021). Utilizing content analysis, the researcher investigated the extent to which the SECEE reading comprehension questions of two samples and the reading content of the textbook prescribed are reciprocal. The findings showed that the textbooks cover the pertinent topics necessary to advance reading skills. However, the two SECEE exams focused more on measuring knowledge and content than evaluating reading techniques and skills, consequently the examinations lacked face and content validity. In his study, Ghuma focused only on one aspect of the SECEE, which is reading comprehension, whilst ignoring other aspects such as grammar, vocabulary, pronunciation, etc. Therefore, the current study is an attempt to fill this gap by investigating the BECEE instead of the SECEE, not only in terms of the reading practices in the exam and the textbooks prescribed but also in terms of all language aspects assessed in the exam vis-à-vis its objectives and the objectives and the content of the curriculum.

Notwithstanding, the above review, considering the authors' suggestions for further research, indeed, this research paper is carried out in the Libyan Education System context, as an extension to the existing literature, particularly by critically analysing the BECEE administered to Grade 9 students of the Basic Education Stage, previously known as Preparatory stage/education. The motive is to provide an in-depth investigation rather than superficial comments on the targeted test validity, a gap needs to be filled.

THE CONTEXT OF THE STUDY

This section deals with the place where the study was carried out. It briefly talks about the education and examination system in the study's context. This includes education stages, the prescribed curriculum, objectives, and content. It also highlights how the English exams, particularly the high-stakes ones, are run in Libyan schools.

1. The Education System in Libya

The education system in Libya is centralized, following a top-down policy, i.e., operated under the Authority of Education (i.e. Ministry of Education, as ME). This includes preparing, distributing, and monitoring teaching and examination materials. English is taught as a foreign language (EFL) in four weekly sessions. The teaching of English, which formally starts from the 1st grade of Basic Education at age six, is practiced within a context-restricted environment. The determiners of language learning depend on the classroom teacher's classroom activities based on teaching materials prescribed by the Ministry of Education (henceforth ME) (more illustration on the content and the objectives of the prescribed curriculum of English is provided in the upcoming sections.

Basic Education generally "aims at providing the pupil with necessary principles, behaviour, knowledge expertise, and practical skills" (Otman and Karlberg, 2007: 101), and "learning foreign languages [i.e. English] to communicate with the world" (ME, 2008: 6). In this stage, students are expected to study a range of English language subjects. By the end of this stage, all students sit national public examinations for all subjects they have studied, all of which they must pass with a grade of at least 50% to obtain the Basic Education Certificate (ME, 2008). The Basic Education Certificate Examination in English (BECEE), which is the focus of this study, is one of these exams. Students are eligible for three years of Secondary Education only after gaining this certificate. This requirement forces students to excel in achieving this grade. At the end, students sit another high-stakes public examination to gain the Secondary Education Certificate to register for university education.

2. The Examination System in Libyan Schools

According to their report to the International Conference on Education (2008) in Geneva, the ME (known previously as The General Peoples' Committee of Education) specified general goals of the examination system, that "the examination system is a tool used by which to assess the output of the educational process and to determine whether the student is capable of apprehending the curriculum taught during the school year …" (ME, 2008: 11). The Ministry of Education manages assessment in the Libyan education system by administering curricula that determine the criteria for evaluating students' learning and progress throughout the school year. Examination marks and grades, both formative and summative, represent the most common assessment strategies used in Libyan basic and Secondary Education. English Language examinations in Libyan schools are essential because students' educational futures depend heavily on their scores in these examinations. Two major high-stakes public examinations are administered to schools by the end of each school year as end-of-stage exams. One exam given to Grade Nine basic education students (the Basic Education Certificate English Examination, BECEE), is the focus of this research. The other exam is administered to Grade Three students of secondary Education (i.e. Secondary Education Certificate English Examination, SECEE). Both exams (the BECEE and SECEE) consist of three exam sessions, with the first two being internal exams constructed and administered by schools' English subject teachers. Most of those exams are in paper and pencil formats marked by teachers manually, though they have many similarities with the final exam (the third session exam) in terms of form and content (Elbishti et al., 2022; Ghuma, 2021). The third exam session is the final exam (BECEE), which is an external "high-stakes" examination, usually held in June every year, and is constructed and supervised by a qualified group of inspectors assigned by the Libyan Ministry of Education (Onaiba, 2014).

It is important to highlight here that student marks in the two sessional tests, plus marks gained from homework and participation in classroom activities during the school year, are allocating 40 marks out of 200, hence 20% of the total mark assigned for the English subject in the school year. Whereas the remaining 80% of the total mark is devoted to the third session exam, the final BECEE. Hence, this exam is considered very important for students as it represents the majority of their marks (i.e. 160 out of 200, or 80%) for the school year. Importantly, for a student to be counted successful, he or she must sit the final examination and achieve at least: a) 40% of the required mark of the final examination (BECEE), i.e. 64/160, and; b) 50% of the total required mark in the subject, i.e. 100/200 (Onaiba, 2014, p. 17-18).

This study is essential for several reasons. First and foremost, in the context of the current study, no study seems to have been carried out to critically evaluate the currently administered BECEE and indulge in exploring the kind of relationship that may exist between this exam and the current prescribed English curriculum in terms of their objectives and content. Secondly, hopefully, conclusions drawn from this paper would have important implications for English examination system reform where consistency between both exam content and curriculum content is maintained. Thirdly, this study is significant in terms of its methodology. The study deploys a documentary research method. "This [kind of] method has had little attention compared to other methods … and often marginalized or even used, it only acts as a supplement to the other general social research methods" (Ahmed, 2010, p. 1 - 2). Also, it was contended by Bowen (2009) that "there is some indication that document analysis has not always been used effectively in the research process, even by experienced researchers" (p.27). Therefore, methodologically, this study will add insights to the literature pertinent to research methods used in social and educational research. That is, document analysis can be used as a valuable instrument to collect data to carry out our research studies.

This study aims to explore: a) the extent to which the content of the currently administered examination reflects its aims; and b) the extent to which the aims and content correlate with the current English curriculum. So, the study endeavours to address the following research questions:

1. To what extent does the content of the currently administered examination reflect its aims as envisaged by its constructors?
2. To what extent do the currently administered examination aims and content reflect the objectives and content of the current English curriculum?

METHODS AND PROCEDURES

This research deployed a documentary study research design to address the research questions and meet the research objectives properly. Indeed, documents are used for gathering data in mixed-methods research (Creswell (2015). Nevertheless, as stated by Denscombe (2021), they can be considered a source of data in their own right. Thus, utilising analytical studies to assess language tests has proven to be a valuable method for evaluating the content validity of such assessments. The approaches researchers use to investigate the validity of the targeted tests range from following traditional perspectives and views of validity as isolated methods of collecting validation evidence to adopting an argument-based approach to validation (Chapelle & Voss, 2021). The researchers' choice of the different methodologies is driven by the goals they wish to achieve and the type of questions they seek to answer. Therefore, based on the nature of the current investigation and due to time and space constraints, this study limits its focus to addressing one type of test validity, namely, content validity, via documentary study as the main strategy of the research design to help find accurate answers to the research questions posed.

The first step taken by the researcher was contacting the relevant administrations and offices at the Libyan Ministry of Education to collect the required documents needed for the validation process of the BECEE. Then, a visit was scheduled with both the head of the Inspectorate Office and the head of the Examination Office at the Education and Pedagogy Directory in the city of Misurata where this research was conducted. The researcher obtained some copies of the studied exam, test specifications document, Decree No.6/2007 which includes the objectives of which the studied exam was administered, and a copy of the prescribed EFL textbook, "English for Libya".

As they, i.e. the analysed documents, "do not speak for themselves but require careful analysis and interpretation" (Cohen et al., 2011: 253), a sample of the new exam administered in one of the previous years is evaluated and analysed critically question by question. Mainly, the exam is studied in terms of content. The motive is to see how its objectives, as envisaged by constructors, correlate with the test content and the textbooks (teaching curriculum) prescribed for grade nine students. To this end, it was essential to explore the current textbooks regarding content and objectives, as textbooks produced for teaching in schools are deemed a significant source of research evidence in educational research (Cohen et al., 2017). Other related documents were also subject to review and analysis to further obtain a deeper understanding of the investigation and valid and reliable data. These include the BECEE's test specifications provided by the Educational Inspection Office, and the decree No.6 issued by the Ministry of Education in 2007 (ME, 2007), via which the current BECEE was mandated; besides, some previous studies conducted in this research's context on curriculum reform and implementation, such as those of Orafi and Borg (2009), Onaiba, (2014), Elabbar (2021), Ghuma (2021), and Elbishti et al. (2022).

Textbook content analysis was chosen because it was challenging to contact textbook authors due to the absence of contact information and time limitations. In addition, it is accentuated that content analysis of any materials used, including textbooks, exam samples, and other pertinent documents, would be a valuable source of documentary data (Mayring, 2021; Ozkan, 2023; Klingebiel, 2024). These processes contribute to answering the two research questions about the degree of correlation between the content of the studied exam and the curriculum content. Thus, the researcher decided to use content analysis to pool data from the following documents: a sample of the BECEE and the existing textbooks, besides some archival documents, to elicit information on the technical quality of the test, test specifications, and links to curriculum objectives.

It is important to mention here that my educational experience with the education and examination system in the context of this study, Libya, has helped me shape my ideas and understanding to delve into the analysis of the studied exam and the pertinent reviewed documents. I have been a teacher of English in Libyan preparatory and secondary schools for more than ten years (1995 – 2006). Also, the years of experience (2007 – up to present) at the university level during which I conducted and supervised several research papers and projects have beneficially impacted the embarking and completion of this research study. All this has significantly contributed to the authenticity and trustworthiness of this research.

THE DOCUMENTARY STUDY

This central part of the study explicates the evaluative content analysis of both the BECEE vis-à-vis the prescribed English curriculum to explore the relationship between the content of the studied exam and its set aims on the one hand and the content and objectives of the prescribed curriculum on the other.

1. Content Analysis of the Prescribed Curriculum

This section discusses the English curriculum's content, objectives, and teaching method(s) applied to implement this curriculum.

A perusal of the new curriculum's content shows that it encompasses a series of textbooks called *English for Libya.* The syllabus comprises nine English language levels explicitly written for Libyan students from Grades One to Nine (Grade Nine textbooks were previously known as *Preparatory 3*) for the basic education schools. Other levels of the same series are assigned for the three Grades of Secondary Education. The syllabus was published by Garnet Ltd, Reading U.K., and written by Lucy Frino *et al*., under The Libyan Authority of Education's supervision. For both levels, i.e. basic education and secondary education, the prescribed English syllabus comprises a full-colour course book, a black and white workbook, a teacher's book, and a class cassette.

The objectives of the curriculum as outlined by Orafi (2008, p.15) are:

- For the students to leave school with much better access to the world through the lingua franca that English has become.
- To create an interest in English as a communication tool and help students develop the skills to use this tool effectively.
- To help students use the basic spoken and written forms of the English language.
- To help students learn a series of complex skills: these include reading and listening skills that help get at meaning efficiently, for example, skimming, scanning, and interpreting the message of the text; they also include the speaking and writing skills that help the students organize and communicate meaning effectively.

Additionally, the target language is organized and introduced to students according to topic rather than structure to encourage students to link language and functions to familiar topics and situations (Frino et al., 2021). The course-book sections are "dedicated to reading, vocabulary, and grammar, functional use of language, listening, speaking and writing" (Orafi and Borg, 2009: 245). The curriculum recommends that English be used as much as possible in class to enable students to communicate effectively and fluently (Orafi and Borg, 2009).

Thus, it is evident that Libyan schools' English curriculum, particularly at the Basic Education Stage, is communicatively oriented. The intention is to promote communicative language teaching; as a result, it is highly recommended that teachers of English in Libyan schools adhere to the principles and assumptions of teaching upon which the communicative approach to language teaching is based. This was further evidenced by Ghuma (2011: 35), who affirms that the “Communicative approach is the current teaching method employed in these textbooks. Using communicative language teaching implies using activities that encourage communication and urging students to communicate”. This curriculum is generally welcomed by teachers and inspectors (Shihiba, 2011); however, these “positive views [specifically from teachers] towards communicative activities were not translated into classroom practices” (Orafi and Borg, 2009: 249) (for further elaboration see Ali, 2008; Orafi, 2008; Orafi and Borg, 2009; Shihiba, 2011; Onaiba, 2014; Elabbar, 2021).

2. Content Analysis of the Studied Exam (BECEE)

This section describes the BECEE pattern, aims, and purposes briefly. It also reviews the exam in terms of its format and content. Then, a more detailed content analysis of the exam questions and items concerning its objectives and the current English curriculum's objectives is dealt with in the consequent section.

3. Exam Purposes and Aims

To develop the examination system a decision was issued by the Education Authority (Decree No. 6, 2007) to adopt the *electronic examinations* …. "These exams aim to develop ways and methods of examinations to go in line with modern scientific developments and to introduce computers in examinations to monitor grades, issue certificates and it also enables students to review their results automatically" (ME, 2008: 11).

The exams were first introduced to Grade 9 students of basic Education in May 2009. These new exams were administered as school-leaving exams to two student cohorts: Grade Nine students at the Basic Education Stage and Grade Three students in secondary schools – the former being the main concern in this study.

The aims of the BECEE, i.e. the intended washback, apart from its primary function as a disciplinary tool, as outlined in the decree of the ME No.6 – 2007 are:

- to facilitate how candidates answer the exam questions;
- to cover, comprehensively and equally, all the components and contents of the curriculum;
- to provide students with a gauge of their language learning achievement as far as the material of the prescribed syllabus is concerned;
- to minimize the risk of cheating;
- to score the answer sheets mechanically and disseminate the results quickly, adequately, and as transparently as possible.

4. Exam Design and Format

A perusal of a sample of the current pattern of the BECEE paper shows that it is exclusively based on selected-response items, such as multiple-choice, true-false, and matching items, where candidates are not required to compose their answers. In such a type of test, candidates, grade Nine students in our case, just need to choose the answer they perceive is correct, and transfer them to an answer sheet. The answer sheet is then inserted into computer software to calculate the total score, gaining reliable, quick, and transparent results. This marking based on scanning of answer sheets is deemed a form of automated scoring (Bejar et al., 2016), which has often been cited for its positive impact on marking quality. However, the literature showed widespread evidence for the perception that such auto-sored formats of tests do not yet command public/widespread support as they are deemed to reduce standards and undermine the quality of assessment (Haggie, 2008).

Moreover, according to exam constructors, to facilitate the way of dealing with the current pattern of the BECEE, unlike the old exam pattern, a cover sheet is attached to the new exam paper, which contains information such as student name, city, school, subject matter and the time allotted to answer the exam. More importantly, to make it easily understood, the cover sheet also includes information and instructions for students on how to answer the exam and how they can transfer their answers to the answer sheet. All this information is written in Arabic to ensure students' awareness of how to deal technically with the exam, reducing the risk of misunderstanding on the students' part. This procedure could contribute to the increased ease of exam administration; boosting exam practicality. Moreover, the instruction for each type of question (true/false, multiple-choice, or matching) is translated into Arabic to make it easier for students to know what they are asked to do and to reduce the possibility of students making haphazard responses, a process may improve the face validity of the exam (Katiso, 2022).

Furthermore, each candidate has his/her exam paper with their name on it. Importantly, for test fairness which is deemed an aspect of test validity (Sun, 2022), all exam papers are identical; however, given that whenever there is testing, there is cheating (Fulcher, 2011) and cheating is highly associated with the administration of high-stakes tests (Amrein-Beardsley et al*.,* 2010; Fulcher, 2015; Kim, 2022), the BECEE exam paper for each student within the same region of the exam room is different from his or her neighbour in terms of question order, i.e. sequencing of questions. For example, question 4 for X student may be question 18 for Y student, and so forth. Additionally, better security measures and identification checks are carried out. The purpose of this is two-fold: to prevent exam impersonation and to diminish the risk of cheating, as noticed that cheating was an expected behaviour in the context of this study among students and some teachers with old examination patterns. Students with the current

exam pattern are less likely to cheat because doing so would take too much time to find what question in their paper matches the other student's exam paper. In addition, exam invigilators are usually unaware of English matters, so they cannot assist students. Regardless, if they did so, they would make noise, and it would take time for them to recognize the answers to be leaked to the student(s), a process likely to catch the attention of the exam invigilation and monitoring committee in the venue. Moreover, these committee members will not condone such practices as the examination boards meticulously and selectively appoint them in the region to monitor the examination process. So, all efforts are made to ensure the integrity of test scores by eliminating chances for candidates to attain marks by fraudulent means, which may threaten test validity (Fulcher, 2011).

5. The Critical Analysis of the BECEE

Further to the exam and curriculum review reported above, this section provides a critical analysis of the content (questions and items) of the BECEE, highlighting its pros and cons concerning its objectives and the content of the current English curriculum. Discussion in this section and the curriculum and exam review discussed above lend themselves to be utilized as a documentary study adopted in this research as the primary data collection and analysis tool. It is important to mention that the years of experience in the language classroom as an English instructor in this study's context provided the current researcher with an intimate knowledge of the Libyan Education and examination system, particularly concerning tests and curriculum contents.

To begin with, the exam paper, as shown in Appendix 1, comprises 60 questions/items in total. A perusal of the exam's content reveals that the exam is objective testing techniques based, i.e. discrete-point/selected-response items including multiple-choice with 26 items, true-false with 24 items, and matching with 10 items, representing 47%, 37%, and 16% of the exam questions respectively. It also shows that the exam content is based on the content of the prescribed teaching materials/textbooks. In other words, what is included in the exam stems from what students studied during the school year, making the exam one version/form of other achievement tests.

The exam items appear to be constructed to test students' reading comprehension abilities and grammar competence. However, the emphasis seemed to be on reading comprehension questions whose answers are cloned verbatim from texts stipulated in the student's textbooks. For example, the questions' nature seemed to be virtually based on specific and general information about some events and dates taken from the reading texts in the prescribed course book, as in the case of questions 8, 15, 18, 22, 23, and 46, to name a few. Questions 15 and 46, for instance, were designed based on information from Unit Five, pages 42 and 39, respectively. Thus, one may claim here that this would likely encourage students to memorize these chunks of language rather than understanding or developing the skills they represent (Hughes & Hughes, 2020), a strategy may make the exam harmful rather than useful to the educational process in the context.

Although the exam seems to rate highly in terms of practicality and usability as it can be completed in a reasonable amount of time and presented in an electronic format facilitating its scoreability (Schmitt et al., 2001), the exam sample in Appendix 1 reveals some technical problems. For instance, as shown in Q43, there are spelling mistakes with "furion" instead of "furious". There is also item repetition, for example, Q22 and Q34 test the same thing, and Q9 and Q38 test similar information.

In addition, ambiguity in some items is observed. As indicated in questions 14, 23, and 39, the latter appears to have more than a key. All of this will likely confuse students and thus cause them to misunderstand what they are required to do. Consequently, students will likely lose the marks for these items if the chosen answers are incompatible with the appointed answers inserted into the computer, as the computer programme is designed to accept only one answer for each question/item. Similarly, some exam questions, especially true-false items, tend to be vague regarding the construct it tests – for instance, "Water boil at 100 ºC.", a true or false question (Q7 in the exam sample). Grammatically, this sentence is incorrect regarding subject-verb agreement; but knowledge-wise, based on the students' course-book, it is correct. Conversely, the item stem of Q11, a true or false question, asks students, "Libya is in the south of Africa". In this case, the student might be confused about whether this question tests grammar or knowledge about geography. If it tests grammar, it is true, and if it tests geographical knowledge, it is false, as Libya is located in the north of Africa. Here the student's answer can be wrong

or might be correct, depending on the exam designer's intention and the answer inserted into the computer.

Moreover, the exam items reviewed seem to be thrown into a state of disarray. As a consequence, in some cases, students might become confused (as mentioned above), specifically as the instructions of the questions do not tell students to provide their answers in terms of grammar rules (i.e. exam questions are not split according to the construct(s) measured); instead, it was included among general information questions assigned for testing reading comprehension. So, this suggests that the exam question, in this case, does not test what it was intended to test, a drawback that undermines test validity.

Furthermore, regarding the time allotted for students to answer the exam, which is three hours, it appears that the exam may take much less time to be completed by students. This claim is empirically supported by Beglar and Hunt (1999), who carefully analyzed and validated a revised version of a high-stake national examination in Japan (based on objective-item techniques). The authors found that "less than 35 minutes were required to complete all 72 items, confirming that a large number of this type of item can be completed in a fairly short time" (p. 136). Thus, a three-hour exam of this type in our case would likely harm the practicality of the exam discussed above. It should be noted here that, according to examination board instructions, students/candidates are not allowed to hand over their answer sheets and leave the examination room unless half of the exam time has elapsed, i.e. 90 minutes. This may encourage some students to look around the examination room with the intention of cheating, an activity that contradicts one of the main purposes of the exam outlined earlier in this paper, 'to minimize the risk of cheating'.

Further, skimming the exam sample brings about a feeling that the candidate in such exams seems to act as if they are solving crossword puzzles rather than taking a language test (See Q41, and Q44 where all options seem distractors, with no key). In many cases with a purely objective-item exam, especially when students do not know the key answer, they have a 50% chance of being right in true-false questions (Hughes & Hughes, 2020); by analogy, hence, students will have a 25% chance of being right in four-option question items, as is the case in the studied exam. Consequently, students' scores in such exams might not reflect the students' actual levels in English, leading to “construct-irrelevant contamination in score interpretation” (Messick, 1984), which was later coined by Haladyna et al. (1991, p. 4) "test score pollution". Indeed, students may gain high scores but cannot express themselves in real-life situations or even write sentences in English correctly.

Moreover, a review of the exam sample attached to this paper and some other samples of the same exam, particularly vis-à-vis the textbooks prescribed, revealed that the exam suffers from some defects. First, the writing skills are untested in this testing syllabus; students are not asked to produce written English in any form. Second, the communicative language testing items, mainly speaking items, are entirely ignored. These two findings clash with one of the curriculum objectives, i.e. “to help students use the basic spoken and written forms of the English language” (Orafi 2007, p. 15). So, as the prescribed teaching syllabus is based on the theory of a communicative approach to language teaching, this suggests that the exam lacks construct validity, as the less consistent a test is with the theory that underlies the course of study, the less a test has construct validity (Fulcher, 2015).

The third issue with the studied exam is that it lacks content validity. This is supported by Henning (1987) who affirms that the content validity of a test is ultimately “concerned with whether or not the content of the test is sufficiently *representative* and *comprehensive* for the test to be a valid measure of what it is supposed to measure” (p, 94; original italics). The primary concern here is whether the exam questions/items can adequately represent the English curriculum content. In other words, the question is how a 60-item test, mostly testing reading and grammar objectively, can adequately represent the content of the whole or at least the most portion of the prescribed syllabus, which is described as a good representation of a positive language curriculum programme (Orafi and Borg 2009). It is crystal clear that the BECEE has a tremendous effect on the content of the current curriculum of English, narrowing the curriculum. This particular finding contradicts one of the main aims assigned by test constructors, i.e. to cover, comprehensively and equally, all the components and contents of the curriculum.

This review indicates that the current BECEE will likely bring about a negative washback, rather than a positive one, as it is evident that it does not accord with the objectives of the current English curriculum. This is because a test is considered to have beneficial washback when preparation

for it does not dominate teaching and learning activities, narrowing the curriculum Hoque (2016). A test that aligns with the course objectives is likely to have positive washback; a test that does not align with the objectives is likely to exert negative washback.

Although the current curriculum of English is communicatively based and covers a wide range of skills, it appears that the currently administered exam (the BECEE) has not shifted the focus from testing students' grammar knowledge towards communicative competencies. Thus, the content of both the BECEE and curriculum reviewed indicate that congruence between the two is tenuous, in the sense that some aspects, which are deemed pivotal, writing, speaking, and listening, are untested. This violates the objectives of the curriculum prescribed and contravenes one further crucial objective of the BECEE pattern: "to cover comprehensively and equally all the components and contents of the curriculum", outlined above. For better or worse, teaching to the test is a common reality, and a test that includes topics that are not in the curriculum may consequently lead to a change in the curriculum so that it is more in line with what is being assessed.

Discordance between the content of high-stakes achievement tests and the content of the prescribed curriculum was also documented in the literature as reviewed in section 1.2 (Siddiek, 2010; Gashaye and Degwale, 2019; Katish, 2022; Amin, 2021; Katiso, 2022). For instance, Katiso (2022)'s study found that the investigated exams did not sufficiently represent the textbook contents. "The tests primarily focused on grammar and vocabulary, and speaking, writing, listening, and reading skills were all underrepresented" (p. 73). The grammar and vocabulary dominated, speaking and writing were weakly represented, and listening and reading were ignored from the examinations. Similarly, Amin (2021) asserted that the conclusion drawn from students' responses suggested a wide range of mismatches between the curriculum and assessment practices.

FUTURE DIRECTION

Given the above discussion, the researcher might contend that the currently implemented English exams can be evaluated as tests of some successes and more failures. Regarding the exam content, in the first instance, heterogeneity is observed between the exam content and its aims and between the exam content and the content of the current curriculum of English. This is because the question papers hardly represent the entire curriculum, a drawback that undermines the exam content validity. Consequently, the currently implemented exams would not help achieve the desired aims, the intended washback.

In addition, the exam tends to divide language into discrete points such as grammar and reading comprehension items, and tests them separately, in multiple-choice and true-false questions. Hence, the current exam mission seems less authentic, which may lead to inaccurate deductions about students' language abilities according to test scores and thus produce low levels of construct validity, consequently leading to test-score pollution.

Furthermore, it seems that the main reason for the overflow of selected-response items in the current BECEE paper is their convenience for machine grading. This convenience may sacrifice the test validity, which would mislead EFL teaching and learning in Libyan schools, and it may contradict other essential objectives of the test. Thus, one may argue here that this kind of examination pattern suffers from some defects. They appear to be not authentic in measuring students' linguistic abilities, as they did not include integrative/construct-response items, ignoring the inclusion of some important aspects of language, listening, and speaking, leading to curriculum misalignment. Moreover, although the literature has shown the increasingly pervasive use of constructed response testing items in the most recent educational reforms (Bejar et al., 2016), the BECEE seems to limit the constructs to be measured because it depends solely on discrete-response testing with deliberate inclusions of selected-response items. This by default, may have far-reaching repercussions for language education in the context.

As a consequence of the above-highlighted shortcomings of the status quo of the BECEE, one may underscore that EFL teaching and learning in the context may be directed from syllabus-oriented teaching to test-oriented teaching, test alignment instead of curriculum alignment. The BECEE has strayed from its purpose and aims. It's no longer a tool to improve education, but a source of fear that prioritizes the test itself as an end over quality teaching and learning as a goal. Further, if the status quo of the examination system continues to divorce instruction from content, teachers may simply continue

to accommodate it to their current and future modes of teaching, a situation that will undoubtedly have serious repercussions for Libyan EFL classrooms. Therefore, the researcher suggests that the EFL testing system in Libyan EFL classrooms be revisited due to the crucial role that sound language examinations have in determining the what and the how of teaching and learning.

This study's findings may draw the Ministry of Education personnel's attention, recommending they take serious reformatory steps to improve the current BECEE. That is, having that what is tested is taught and learned, our findings suggest incorporating testing the elements of speaking, listening, and writing into the current BECEE to yield significant improvements in student's language proficiency, a step that may have great benefits to develop and improve English teaching and learning and raising standards and quality of education in Libyan schools. The proposed reform aims to improve consistency in language teaching by implementing significant changes to classroom assessment and instruction. In other words, exam questions and items should not be restricted to testing rote learning and memory or recognition of knowledge. Additionally, the system should measure higher-order outcomes based on comprehension and meaningful learning in order to assess students' application of information in both educational and real-world contexts.

However, any attempt to meaningfully reform the examination system must be accompanied by a comprehensive reform of the whole educational system, particularly providing well-established teacher programmes including teacher preservice preparation and in-service training and supervision as well as giving schools proper infrastructures and state-of-the-art equipment. To this end, there will be much to gain and little to lose in taking such steps forward in the direction of test reform. In essence, the performance of such examinations should be evaluated and traced by expert personnel to feed back into the process of language education and assessment in schools and contribute to bridging the existing gap between the examination and the English curriculum.

Moreover, this study suggests some directions for further research. One direction can be a study that investigates other forms of test validity other than the one studied in this research paper, content validity, which were highlighted in the first section of this paper. That is, to investigate the face validity of the BECEE, by utilizing qualitative and quantitative research design via questionnaires and interviews. The purpose is to probe into stakeholders' (teachers and students) beliefs and attitudes towards the current BECEE. This will add insights into the findings of this study contributing to the degree of effectiveness, practicality, and credibility of the studies exam, the BECEE.

One more direction for future research could be a validation study of the new examination in terms of its construct validity and reliability. It is to design a performance-based test paper based on a BECE paper, for the convenience of comparison. The overarching aim is to find any significant differences in the construct validity between the BECE paper and the performance test paper concerning exam takers. Such studies can produce practical implications: surveying stakeholders to acquire quantitative and qualitative data for the BECEE is worthwhile because their responses afford insights into the educational and pedagogical consequences of the exam under investigation, providing rich and important information about the broad validity of the exam and its impact at the micro and macro levels.

CONCLUSION

In conclusion, it is hoped that this study's implications, suggestions, and recommendations will benefit the education system of the country in general, and EFL education in Libyan schools in particular. In other words, it may aid the relevant policymakers in developing appropriate policies and procedures for enhancing the education and examination system and figure out how to better support teachers' performance and talents across all education stages. This will therefore probably increase students' motivation to learn the language efficiently. I would recommend taking a bottom-up rather than a top-down policy, giving more weight to teachers' and inspectors' suggestions, by involving them in the implementation of the teaching and assessment practices, for better future outcomes, as today is the best guide to tomorrow.

ACKNOWLEDGEMENT

The authors would like to thank the Libyan Academy for Postgraduate Studies for providing the support and facilities.

FUNDING

The authors declare that no financial support was received for the research, authorship, and publication of this article.

DATA AVAILABILITY STATEMENT

Data will be made available on request.

CONFLICT OF INTEREST

The authors declare no conflicts of interest regarding the publication of this paper.

REFERENCES

Abdulhamid, N. (2019). *What is the relationship between alignment and washback? A mixed-methods study of the Libyan EFL context* [Doctoral dissertation, Carleton University].https://doi.org/10.22215/etd/2018-13456

Ahmed, J. U. (2010). Documentary research method: New dimensions. *Indus Journal of Management & Social Sciences*, *4*(1), 1-14.

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im) possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, *30*(4), 535-556. https://doi.org/10.1177/026553221348956

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied linguistics*, *14*(2), 115-129. https://doi.org/10.1093/applin/14.2.115

Ali, M. A. A. (2008). *The oral error correction techniques used by Libyan secondary school teachers of english* [Doctoral dissertation, University of Sunderland].

Amin, R. (2021). *Students' perception on high-stakes English language testing in Bangladesh: expectations vs reality* [Doctoral dissertation, Brac University]. http://hdl.handle.net/10361/15665

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000-word level and university word-level vocabulary tests. *Language testing*, *16*(2), 131-162. https://doi.org/10.1177/026553229901600202

Bejar, I. I., Mislevy, R. J., & Zhang, M. (2016). Automated scoring with validity in mind. *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*, 226-246. https://doi.org/10.1002/9781118956588.ch10

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, *9*(2), 27-40. https://doi.org/10.3316/QRJ0902027

Bridglall, B. L., Caines, J., & Chatterji, M. (2014). Understanding validity issues in test-based models of school and teacher evaluation. *Quality Assurance in Education*, *22*(1), 19-30. https://doi.org/10.1108/QAE-12-2013-0053

Chapelle, C. A., & Voss, E. (Eds.). (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge Applied Linguistics. Cambridge University Press; 2021:i-i.. https://doi.org/10.1017/9781108669849

Cohen, L. Manion, L. and Morrison, K. (2011) *Research Methods in Education.* (7th ed.). Routledge.

Cohen, L. Manion, L. and Morrison, K. (2017) *Research Methods in Education.* (8th ed.). Routledge. https://doi.org/10.4324/9781315456539

Creswell, J. W. (2015). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Pearson India.

Denscombe, M. (2021). The good research guide: *Research methods for small-scale social research projects*. (7th ed.). McGraw-Hill Education.

Desalegn, G., Disassa, R., & Kitila, T. (2023). The influence of high-stakes English examinations on students' out-of-classroom English learning practices: A comparative study. *Education Research International*, *2023*(1), 1108951. https://doi.org/10.1155/2023/1108951

Dong, M., & Liu, X. (2022). Impact of learners' perceptions of a high-stakes test on their learning motivation and learning time allotment: A study on the washback mechanism. *Heliyon*, *8*(12), e11910. https://doi.org/10.1016/j.heliyon.2022.e11910

Dong, M., Fan, J., & Xu, J. (2021). Differential washback effects of a high-stakes test on students' English learning process: Evidence from a large-scale stratified survey in China. *Asia Pacific Journal of Education*, *43*(1), 252-269. https://doi.org/10.1080/02188791.2021.1918057

Elabbar, A. A. (2021). Reforming the Libyan education system: Seven articulated years via a strategic planning pyramid. *London Journal of Research in Humanities and Social Sciences*, *22*(14), 11-26.

French, S., Dickerson, A., & Mulder, R. A. (2024). A review of the benefits and drawbacks of high-stakes final examinations in higher education. *Higher Education*, 88, 893-918. https://doi.org/10.1007/s10734-023-01148-z

Frino, L. Mhochain, R.N. O'Neil, H. and McGarry, F. (2021) *English for Libya, Preparatory 3: Teacher's Book,* Reading-UK: Garnet Publishing Ltd.

Fulcher, G. (2011) 'Cheating gives lies to our test dependence: Policymakers are using language tests to carry a larger social burden than they can reasonably bear'. *The Guardian Weekly, Tuesday 11th October 2011,* Learning English (4), https://languagetesting.info/articles/store/cheating.pdf. Online accessed on 12 November 2023.

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. Routledge. https://doi.org/10.4324/9781315695518

Gandini, E. A., & Horák, T. (2020). Promoting positive washforward through personalised test feedback and other benefits: Piloting a computer-based testing system. *Language Learning in Higher Education*, *10*(1), 235-244. https://doi.org/10.1515/cercles-2020-2012

Gashaye, S., & Degwale, Y. (2019). The content validity of high school english language teacher made tests: the case of Debre Work preparatory school, East Gojjam, Ethiopia. *International Journal of Research in Engineering, IT and Social Sciences*, *9*(11), 41-50.

Ghuma, M. A. (2021). Content validity of national exams of English reading in Libya. *Libya Bulletin for Studies 7th issue. Dar Azzawyah Lelketab*, (7). 59-78.

Ghuma, M.A. (2011) *The transferability of reading strategies between L1 (Arabic) and L2 (English),* unpublished PhD dissertation. University of Durham.

Gorgodze, S., & Chakhaia, L. (2021). The uses and misuses of centralised high stakes examinations-Assessment Policy and Practice in Georgia. *Assessment in Education: Principles, Policy & Practice*, *28*(3), 322-342. https://doi.org/10.1080/0969594X.2021.1900775

Green, A. (2013). Washback in language assessment. *International Journal of English Studies*, *13*(2), 39-51.

Haggie, D. (2008). The strategic use of marking technologies to support innovation and diversity in assessment. *International Association of Educational Assessment, September, in Singapore. http://www. iaea2008. Cambridge assessment. org. uk/ca/digital Assets/180436_Haggie. pdf*.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher*, *20*(5), 2-7. https://doi.org/10.2307/1176395

Han, B., Dai, M., & Yang, L. (2004). On the Problems of CET. *Foreign Languages and Their Teaching*, 2, 17-23.

Henning, G. (1987) *A guide to language testing: development, evaluation, research.* Cambridge: Newbury House.

Hoque, M. E. (2016). Teaching to the EFL curriculum or teaching to the test: An investigation. *The EDRC Journal of Learning and Teaching*, *1*(1), 1-25.

Hughes, A. & Hughes, J. (2020). *Testing for language teachers*. (3rd ed.). Cambridge University Press.

Kane, M. T., & Wools, S. (2019). Perspectives on the validity of classroom assessments. *Classroom assessment and educational measurement* (pp. 11-26). Routledge.

Kang, M. K., & Chang, H. J. (2014). Ensuring validity of practical English certification test of local office of education in Korea. *Pan-Pacific Association of Applied Linguistics*, *18*(2), 111-122.

Katiso, A. (2022). *Analyzing content and face validity of english final examination prepared by Damboya district education office: Grade seven in focus* (MA Dissertation, Haramaya University).

Klingebiel, S., Hartmann, F. L., Madani, E., Paintner, J., Rohe, R. A., Trebs, L., & Wolk, T. (2024). Methods for data collection and analysis. *Exploring the effectiveness of international knowledge cooperation: an analysis of selected development knowledge actors* (pp. 43-50). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-55704-0_4

Mayring, P. (2021). *Qualitative Content Analysis: A Step-by-Step Guide.* United Kingdom: SAGE Publications.

ME (Ministry of Education) (2008) *The Development of education, the national report of Libya presented to the international conference on education, session (48), Geneva, November, 2008*, Libya Authority of Education 25 - 28.

Messick, S. (1984). The psychology of educational measurement. *Journal of Education Measurement*, *21*(3), 215-237. https://doi.org/10.1111/j.1745-3984.1984.tb01030.x

Moritoshi, T. P. (2002). *Validation of the test of English conversation proficiency*. (Master's dissertation). University of Birmingham, UK. *(Available online: http://www. bhamlive1. bham. ac. uk/Documents/collegeartslaw/cels/essays/matefltesldissertations/MoritoshiDiss. pdf)*.

Onaiba, A. M. E. M (2014). *Investigating the washback effect of a revised EFL public examination on teachers' instructional practices, materials, and curriculum* (Doctoral dissertation). University of Leicester. https://hdl.handle.net/2381/28561

Orafi, S. M. S. (2008). *Investigating teachers' practices and beliefs in relation to curriculum innovation in English language teaching in Libya* (Doctoral dissertation). University of Leeds.

Orafi, S. M. S., & Borg, S. (2009). Intentions and realities in implementing communicative curriculum reform. *System*, *37*(2), 243-253.

Otman, W., & Karlberg, E. (2007). *The Libyan economy: Economic diversification and international repositioning*. Springer Science & Business Media.

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.

Özkan, U. B. (2023). Validity and reliability in document analysis method: A theoretical review in the context of educational science research. *The Journal of Buca Faculty of Education*, (56), 823-848. https://doi.org/10.53444/deubefd.1258867

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language testing*, *18*(1), 55-88. https://doi.org/10.1177/026553220101800103

Shaw, S., Crisp, V., & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy & Practice*, *19*(2), 159-176. https://doi.org/10.1080/0969594X.2011.563356

Shihiba, S. E. S. (2011). *An investigation of Libyan EFL teachers' conceptions of the communicative learner-centered approach in relation to their implementation of an English language curriculum innovation in secondary schools* (Doctoral dissertation). Durham University.

Shohamy, E. (2020). *The power of tests: A critical perspective on the uses of language tests*. Routledge.

Siddiek, A. G. (2010). The impact of translation on language acquisition and knowledge transfer in the Arab world. *European Journal of Social Sciences*, *16*(4), 556-567.

Sun, M. (2022). Validity and fairness of TOEFL iBT reading test. *Learning & Education*, *10*(8), 141-142.

Takeno, J., & Moritoshi, P. (2018). Re-examining the English Proficiency Level of Japanese EFL Learners. *Chugokugakuen Journal*, *17*, 35-39. https://doi.org/10.18282/l-e.v10i8.3094

Wisdom, S. C. (2018). *Teachers' perceptions about the influence of high-stakes testing on students* (Doctoral dissertation). Walden University.

Ziebell, N. (2018). Curriculum alignment: Performance types in the intended, enacted, and assessed curriculum in primary mathematics and science classrooms. *Studia paedagogica*, *23*(2), 175-203. https://doi.org/10.5817/SP2018-2-10